Metagenomic and metatranscriptomic analysis of microbial communities in the oxygen minimum zone off Peru

Dissertation zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

Vorgelegt von
Harald Schunck
Kiel, Juni 2012

1. Referent: Prof. Dr. Julie LaRoche, GEOMAR | Helmholtz-Zentrum für Ozeanforschung Kiel
2. Referent: Prof. Dr. Philip Rosenstiel, Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität zu Kiel

Tag der mündlichen Prüfung: 28. Juni 2012
Zum Druck genehmigt: 28. Juni 2012

Content

Summary

The sequencing of nucleic acids is a powerful tool to investigate multi-cellular organisms, single-cell microorganisms, as well as whole microbial communities. Historically, sequencing was limited to known organisms or genes, but recent technological advances in sequencing technologies with high-throughput have led to a much broader application, including also unknown targets. High-throughput sequence data can reveal both the structure (phylogeny) and the functional role of the organisms of interest. Thereby, microbial community DNA describes the genetic potential and community RNA the genetic activity. High-throughput sequencing can also be used to investigate the functional role of highly diverse and mostly unknown microbial communities and place them into a broader, biogeochemical context. The target of analysis in this study was the microbial community within the oxygen minimum zone (OMZ) off Peru, an ecosystem of high importance for nutrient cycling processes, which has not yet been described in detail with molecular techniques. Furthermore, this area represents one of the most productive fishing grounds in the world and therefore is of crucial importance for human feeding.

During a cruise with the research vessel Meteor in January 2009, water samples were collected within the OMZ off Peru. In a first experiment the influence of the sampling time on the decay of fragile RNA sequences was evaluated. With a delay of only 20 minutes, the structure and function of the microbial community already changed selectively. Sequences affiliated to the largest phylogenetic group in the initial time point, the β-proteobacteria, decreased from a relative abundance of 30 to 5% of the whole microbial community. This corresponded to a loss of about 85% of all β-proteobacterial sequences. After five hours, β-proteobacterial sequences decreased by almost 97%.

A major effort of this work further comprised the establishment of a bioinformatic analysis pipeline for metatranscriptomic and metagenomic datasets. This pipeline allowed the identification of unknown sequences with different methodological approaches, starting with the Basic Local Alignment Search Tool (BLAST), followed by profile hidden Markov model scans of the ModEnzA Enzyme Commission groups and of the Pfam protein families and finally the recruitment of sequences onto reference genomes. All obtained data and metadata were stored and organized in a suitable and user-friendly database system (MySQL database) making it available for analysis via a common web browser using a phpMyAdmin application. Additionally, a java-based program (FROMP) was developed, allowing the metabolic profiling of microbial communities using the Enzyme Commission groups.

With these tools in hand, a combined metagenomic and metatranscriptomic analysis was carried out, targeting a microbial community obtained from sulfidic (anoxic and hydrogen sulfide containing) waters within the Peruvian OMZ. Until now, the sulfidic plume detected was the largest one ever reported for ocean waters, covering an area of ~8000 km^2 between Lima and Pisco and containing an estimated amount of ~3.5 x 10^4 tons of highly toxic hydrogen sulfide.

The high-throughput sequence data from the microbial community of the sulfidic plume was complemented with water column profiles, flux calculations of nutrients and data from satellite remote sensing. Furthermore microbial (group-specific) cell counts and rate measurements of carbon dioxide fixation and nitrogen transformation processes were included in the analysis. The combined analysis revealed that the microbial community of the sulfidic waters was dominated by several sulfur-oxidizing γ- and ε-proteobacteria, showing highest similarity to the SUP05 cluster bacterium, *Candidatus* Ruthia magnifica str. Cm, *Candidatus* Vesicomyosocius okutanii HA and *Sulfurovum* sp. NBC37- 1. Most likely, the microbial community was exploiting a wide range of oxidants (oxygen, nitrate, nitrite, nitric oxide and nitrous oxide) to detoxify the waters. The sequence data also suggested that organisms related to e.g. the δ-proteobacterium *Desulfobacterium autotrophicum* HRM2 were carrying out sulfate reduction or sulfur disproportionation, and thus forming hydrogen sulfide within the water column.

Additionally, dark carbon dioxide fixation in the aphotic zone accounted for one quarter of the total inorganic carbon fixation. The produced biomass might fuel further sulfate reduction and thereby form hydrogen sulfide, which could stabilize the sulfidic waters. Due to anthropogenic activities like ocean eutrophication and global warming, which will intensify oxygen and nitrogen depletion, sulfidic OMZ waters might become more frequent and widespread in the future. This could have harsh consequences for fish stocks and the living conditions in densely populated coastal areas.

Zusammenfassung

Das Sequenzieren von Nukleinsäuren eröffnet weitreichende Möglichkeiten, um vielzellige Organismen, einzellige Mikroorganismen oder auch ganze mikrobielle Gemeinschaften zu untersuchen. Ursprünglich ließen die Sequenzierungstechnologien nur eine Analyse von bekannten Organismen oder Genen zu, neueste Sequenzierungstechnologien mit hohem Durchsatz erlauben aber mittlerweile erheblich breiter gefächerte Anwendungen, wie etwa die Untersuchung von unbekannten Zielen. Die Daten aus solchen Sequenzierungen lassen sowohl die Struktur (Phylogenie) als auch die Funktion der untersuchten Organismen erkennen. Dabei beschreibt die DNA einer mikrobiellen Gemeinschaft das genetische Potential und ihre RNA die genetische Aktivität. Modernste Sequenzierungstechnologien können inzwischen auch genutzt werden, um hochkomplexe und weitgehend unbekannte mikrobielle Gemeinschaften zu untersuchen und diese in einen größeren, biogeochemischen Kontext zu setzen. Das Ziel dieser Studie war die Untersuchung der mikrobiellen Gemeinschaft in der Sauerstoffminimumzone (SMZ) vor der Küste Perus, eines für Nährstoffkreisläufe sehr wichtigen Ökosystems, das aber auf molekularer Ebene noch nicht im Detail untersucht worden ist. Dieses Gebiet stellt außerdem einen der reichsten Fischgründe weltweit dar und ist somit äußerst wichtig für die Nahrungsversorgung des Menschen.

Im Rahmen einer Ausfahrt mit dem Forschungsschiff Meteor im Januar 2009 wurden Wasserproben aus der SMZ vor Peru genommen. In einem ersten Experiment wurde untersucht, wie sich die Dauer der Probennahme auf den Abbau von fragilen RNA Sequenzen auswirkt. Bereits nach einer Verzögerung von nur 20 Minuten änderten sich die Struktur und die Funktion der mikrobiellen Gemeinschaft selektiv. Sequenzen, die Ähnlichkeit mit der im ersten Zeitpunkt größten phylogenetische Gruppe hatten, den β-Proteobakterien, nahmen in ihrer relativen Häufigkeit von 30 auf 5% ab. Dies entspricht einer Abnahme der β-proteobakteriellen Sequenzen um etwa 85%. Nach fünf Stunden waren fast 97% der β-proteobakteriellen Sequenzen nicht mehr nachweisbar.

Ein großer Teil dieser Arbeit umfasste außerdem die Etablierung eines bioinformatischen Analysesystems für metagenomische und metatranskriptomische Datensätze. Dieses System erlaubt die Identifizierung von unbekannten Sequenzen mit verschiedenen Methoden: zunächst das Basic Local Alignment Search Tool (BLAST), dann Profil Hidden Markov Model Scans der ModEnzA Enzyme Commission Groups und der Pfam Protein Families und schließlich die Rekrutierung von Sequenzen auf Referenzgenome. Alle Daten und Metadaten wurden in einer geeigneten und benutzerfreundlichen Datenbank (MySQL Database) gespeichert und verwaltet, um sie für Untersuchungen zugänglich zu machen. Diese können mit einem gewöhnlichen Internetbrowsers und einer phpMyAdmin Applikation durchgeführt werden können. Zusätzlich wurde ein Java-basiertes Programm entwickelt (FROMP), das die Kategorisierung der Stoffwechseleigenschaften von mikrobiellen Gemeinschaften mithilfe der Enzyme Commission Groups erlaubt.

Mit diesen bioinformatischen Techniken war es möglich, einen gekoppelten Metagenom-Metatranskriptomanalyseansatz einer mikrobiellen Gemeinschaft in sulfidischen (anoxischen und schwefelwasserstoffhaltigen) Gewässern innerhalb der SMZ vor Peru durchzuführen. Es handelte sich dabei um die größte Ausdehnung sulfidischen Meerwassers, die je beschrieben worden ist. Insgesamt war zwischen Lima und Pisco eine Fläche von ~8000 km^2 betroffen, die geschätzte ~3.5 x 10^4 Tonnen hochgiftigen Schwefelwasserstoff enthielt.

Die Sequenzierungsdaten der mikrobiellen Gemeinschaft aus dem sulfidischen Wasser wurden mit Profilen der Wassersäule, Flux-Berechnungen von Nährstoffen und Fernerkundungsdaten von Satelliten ergänzt. Zusätzlich wurden mikrobielle (gruppenspezifische) Zellzählungen, Messungen der Fixierungsrate von Kohlenstoffdioxid und der Rate von Stickstoffumwandlungsprozessen in die Analyse mit einbezogen. Die Gesamtheit der Ergebnisse lässt erkennen, dass das die mikrobielle Gemeinschaft des sulfidischen Wassers von mehreren schwefeloxidierenden γ- und ε-Proteobakterien dominiert wurde, die große Ähnlichkeit zu dem SUP05 cluster Bakterium, *Candidatus* Ruthia magnifica str. Cm, *Candidatus* Vesicomyosocius okutanii HA und *Sulfurovum* sp. NBC37- 1 aufwiesen. Sehr wahrscheinlich nutzte die mikrobielle Gemeinschaft eine Reihe verschiedener Oxidationsmittel (Sauerstoff, Nitrat, Nitrit, Stickstoffmonoxid und Distickstoffmonoxid (Lachgas)), um das Wasser zu entgiften. Weiterhin deuten die Sequenzdaten darauf hin, dass Organismen mit Ähnlichkeit z.B. zu dem δ-Proteobacterium *Desulfobacterium autotrophicum* HRM2 Sulfatreduktion oder Schwefeldisproportionierung ausübten und damit zur Bildung von Schwefelwasserstoff innerhalb der Wassersäule beitrugen.

Außerdem betrugen im Dunkeln gemessene Kohlenstoffdioxidfixierungsraten der aphotischen Zone ein Viertel der gesamten anorganischen Kohlenstoffdioxidaufnahme. Die so produzierte Biomasse könnte die Sulfatreduktion weiter antreiben und damit Schwefelwasserstoff bilden, was die sulfidischen Gewässer stabilisieren könnte. Aufgrund anthropogener Aktivitäten wie der Eutrophierung der Ozeane und der Klimaerwärmung, welche die Sauerstoff- und Stickstofflimitierung intensivieren werden, könnten auch sulfidische Gewässer in Zukunft häufiger und weitverbreiteter auftreten. Dieses könnte fatale Folgen für Fischbestände, wie auch für die Lebensbedingungen an dicht besiedelten Küstengebieten haben.

Introduction

1. Introduction

1.1. Genetic research

It was a matter of debate for a long time, whether certain characteristics of living organisms are inherited from ancestors or are formed by the surrounding environment. First experiments carried out by Gregor Mendel led him to the conclusion that phenotypic traits of plants are indeed inherited (Mendel, 1866; Mendel, 1870). Although Mendel's experiments were unappreciated for many years, at the turn of the 20th century his monumental work and ideas were rediscovered and pursued (e.g. Correns, 1899; von Tschermack, 1900) and today Mendel is considered to be the founder of the science of genetics.

From the early 20th century on, scientists tried to pin down the substance in living cells that actually contained the genetic information. It took until 1944, when Oswald Avery and co-workers found first proof that desoxyribonucleic acid (DNA) could be the actual carrier of the genetic information (Avery et al., 1944). Ten years later in 1953, James Watson and Francis Crick, with a crucial contribution of Rosalind Franklin, solved the three-dimensional structure of DNA, and proposed the double helix (Watson and Crick, 1953). The specific pairing of the desoxyribose-bases purine (adenine and guanine) and pyrimidine (cytosine and thymine) suggested that the order (namely the sequence) of these bases constitute the genetic information and thus the phenotypic traits of living organisms.

Nucleic acid sequencing technologies

The first nucleotide sequence that was determined was a gene encoding for a structural protein of a virus (Min Jou et al., 1972). At that time, this was a major scientific achievement. In the following years, different techniques for the determination of sequence information were developed, e.g. by Maxam and Gilbert (Maxam and Gilbert, 1977). However, it was not until late 1977, when Frederick Sanger and colleagues described a new technique for sequencing using chain-terminating inhibitors, which became the golden standard in genetic research for several decades (Sanger et al., 1977). The ability to determine unknown DNA sequences *de novo* was a milestone on the way of relating phenotypic traits and characteristics to a specific part of the DNA-sequence (namely a gene). Identifying these genes and their functions in a cell increased the understanding of life and had tremendous impacts not only on life sciences, but on sciences in general and on society.

The application of Sanger-Sequencing focussed not only on the analysis of selected genes, but also on whole chromosomes and genomes, which was later termed genomics. The bacterial

Haemophilus influenzae genome was the first to be completely sequenced, having 1.83 million base pairs (Fleischmann et al., 1995). The human genome, more than 1000 times larger (2.91 billion base pairs), was published in 2001 (Lander et al., 2001; Venter et al., 2001).

In 1998, a new sequencing technique was developed, which is based on real-time monitoring of DNA synthesis, normally referred to as pyrosequencing (Ronaghi et al., 1998). This technique enabled the first high-throughput sequencing run, which delivered several million base pairs of sequence information, far exceeding the capacity of Sanger-Sequencing-technology (Margulies et al., 2005). Pyrosequencing caused a revolution in genetic research and as a consequence, sequencing efforts became not only more productive, but also faster and cheaper (Gilbert and Dupont, 2011). Rapid technological improvements also led to a variety of other sequencing techniques, which are all termed high-throughput sequencing:

- FLX pyrosequencing (Roche): 300-600 base pairs per sequence, 400-800 million base pairs per run
- Illumina (Solexa) sequencing: 100–300 base pairs per sequence, ca. 200 billion base pairs per run
- SOLiD sequencing (Applied Biosystems): 50–75 base pairs per sequence, 100-300 billion base pairs per run
- DNA nanoball sequencing, Helioscope single molecule sequencing, Single Molecule SMRT sequencing, Single Molecule real time (RNAP) sequencing: variable in length and number of sequences

Until now, more than 3000 genomes of all domains of life (eukaryota, bacteria and archaea) have been successfully sequenced and many genome projects are currently ongoing. High-throughput sequencing opened the common use of sequencing technology also for other research areas beyond molecular biology, e.g. environmental microbiology, and is now commonly applied throughout many scientific disciplines. The targets of sequencing efforts can be higher multi-cellular organisms, both cultured and uncultured single-cell organisms, viruses and mixed environmental samples of unknown composition (metagenomics) (Gilbert and Dupont, 2011).

Metagenomics

Metagenomics is an experimental approach, which is turning away from single genes and organisms and is trying to unravel taxonomic diversity and to connect genetic information to environmental functions. The concept, first formulated by Handelsman in 1998, revolves around the sequencing of whole microbial communities of unknown composition, thus bypassing the need for isolation and cultivation of individual species (Handelsman et al., 1998; Streit and Schmitz, 2004; Chen and Pachter, 2005). The randomly acquired sequence information represents the most abundant genes and intergenic regions to be found in the microbial community of that particular environment. The metagenomic approach tries to correlate the abundance and variability of detected genes to biogeochemical and ecological patterns and processes, namely to the function of the whole environment (DeLong, 2009). This experimental setup is of major importance, since up to 99% of all microorganisms are not readily culturable, and in fact, most of the microbial species have never been described (Amann et al., 1995; Pace, 1997; Streit and Schmitz, 2004; Glockner and Joint, 2010). One of the first large-scale metagenomic sampling efforts was carried out by Craig Venter and colleagues in the Sargasso Sea (Figure 1) (Venter et al., 2004).

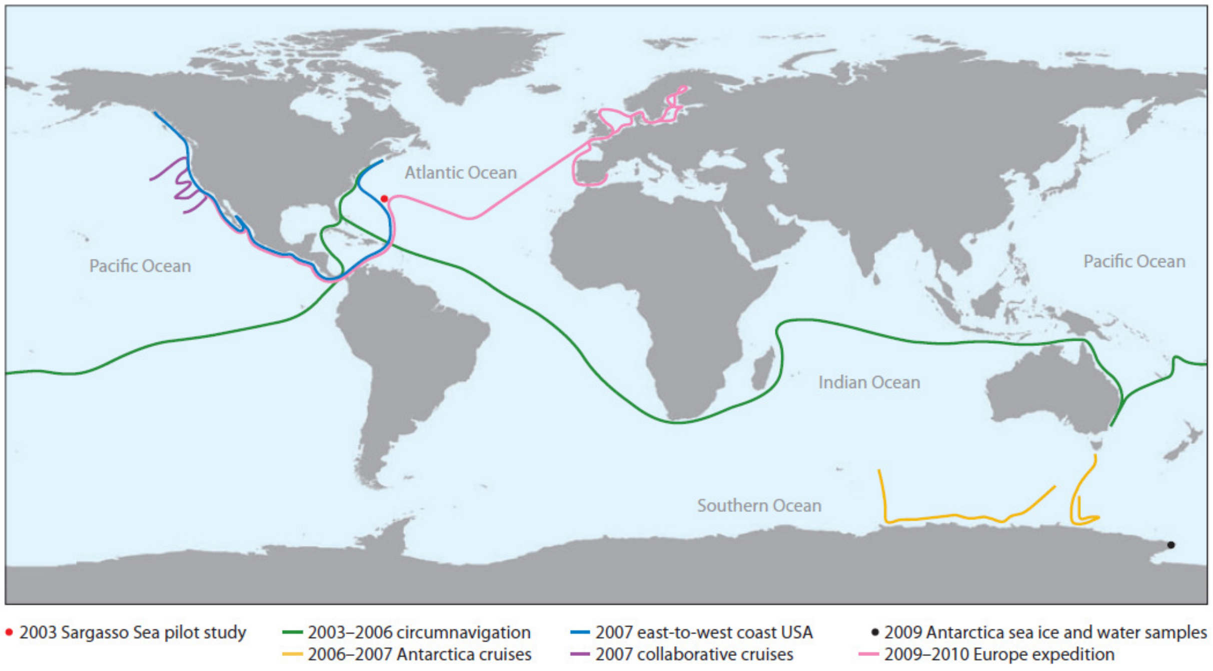

Atlantic Ocean
Pacific Ocean
Pacific Ocean
Indian Ocean
Southern Ocean
● 2003 Sargasso Sea pilot study
2003–2006 circumnavigation
2006–2007 Antarctica cruises
2007 east-to-west coast USA
2007 collaborative cruises
● 2009 Antarctica sea ice and water samples
2009–2010 Europe expedition

Figure 1: Map displaying the sampling site of the Sargasso Sea Project and the cruise tracks of the Global Ocean Sampling expedition (from Gilbert and Dupont, 2011).

The output of their sequencing effort, although still using the Sanger-based capillary sequencing technology, was more than one billion base pairs of information, the discovery of

148 new bacterial phylotypes and 1.2 million previously unknown genes (Venter et al., 2004). The Global Ocean Sampling expedition, a world-round circumnavigation with regular sampling stations carried out a few years later, delivered 6.3 billion base pairs from 7.7 million reads from Sanger-based capillary sequencing; six times more data than the Sargasso Sea Project (Rusch et al., 2007). The size and complexity of these datasets were so immense that several scientists around the world are still analyzing the gathered sequence information with specific questions and hypotheses (e.g. Zhang and Gladyshev, 2008; Gianoulis et al., 2009; Haque et al., 2009; Barz et al., 2010; Raes et al., 2011; Toulza et al., 2012).

However, even these large-scale metagenomic studies are only scratching the surface of extremely complex and diverse microbial ecosystems. Assuming that in every millilitre of seawater one can find approximately 1 million microorganisms with an average genome size of 2 million base pairs each, the Global Ocean Sampling expedition sequenced only the DNA that is present in ~0.003 millilitres of seawater (Gilbert and Dupont, 2011). Although large parts of the sequence information were probably repetitive and thus redundant, the Global Ocean Sampling expedition was far from identifying all rare organisms and genes and consequently did not capture the full diversity of the ecosystems sampled.

Nevertheless, metagenomic approaches applied to microbial communities of low complexity indeed hold the potential to fully characterize an ecosystem. A metagenomic study on a low diverse biofilm has shown to lead to the reconstruction of several genomes of so far unknown bacteria (Tyson et al., 2004).

Metatranscriptomics

Although sequencing techniques were first limited to DNA, ribonucleic acid (RNA, in the form of complementary DNA = cDNA) soon became an additional target for sequence analysis. In analogy to metagenomics, metatranscriptomics represents the sequencing of the expressed RNA in an environment. Of interest for metatranscriptomics can either be ribosomal RNAs (rRNA), the most commonly used phylogenetic marker, small RNAs (sRNA) or messenger RNAs (mRNA), which hold the active functional information of living organisms. Since the vast majority of RNAs in living cells is actually rRNA (Sorek and Cossart, 2010), studies targeting total RNA will recover an overwhelming amount of ribosomal sequences. In order to target mRNA, most metatranscriptomic studies apply special molecular biological methods, developed to enrich for mRNAs. The first study explicitly and exclusively targeting mRNA isolated from a natural environment was carried out by in 2005 (Poretsky et al., 2005). Although the sequencing was carried out with traditional Sanger-technology, a modified version of the protocol was also used in this thesis (see methods section).

Two high-throughput metatranscriptomics sequencing approaches followed in the next years (Leininger et al., 2006; Bailly et al., 2007) both focussing on microbial soil samples. In 2008, the first high-throughput-sequenced metatranscriptome from a marine environment was published (Frias-Lopez et al., 2008), pioneering this techniques for standard use in marine microbiology. A few other studies were published in 2008 as well, which are also considered to be pioneers in metatranscriptomics (Gilbert et al., 2008; Urich et al., 2008).

To fully understand diversity, complexity and function of natural microbial ecosystems, the coupling of metagenomic (genetic potential) with metatranscriptomic (genetic activity) techniques is desirable. Especially for poorly characterized ecosystems, like sewage treatment plants, hot springs, hydrothermal vents or oxygen minimum zones the combination of both techniques opens new possibilities. Only few studies so far have followed this approach for marine habitats (Frias-Lopez et al., 2008; Gilbert et al., 2008; Shi et al., 2009; McCarren et al., 2010; Stewart et al., 2011). For future analysis, the description of common metabolic characteristics of natural environments is worthwhile, enabling comparability between different samples from different projects.

1.2. Analysis of nucleic acid sequence data

High-throughput metagenomic and metatranscriptomic datasets are among the largest in marine ecology and oceanography (Gilbert and Dupont, 2011). The fact that merely four different characters are produced (namely a, g, c and t) should not mislead to an underestimation of the complexity of such datasets. The need for high-throughput computational approaches to cope with the sheer magnitude of the produced data is immense. Rapid advances of the high-throughput sequencing technologies have so far been followed by only slower improvements in suitable analysis tools and capacities to store the data. Computational resources and strategies, databases and applications are at the moment operating on the borderline of feasibility (DeLong, 2009) and until now, large parts of the data still remain difficult to interpret due to this bottleneck in bioinformatic analysis (Gilbert and Dupont, 2011). In order to make the sequence information and the metadata (relevant environmental parameters) available to other researchers for comparative studies, they should be reported and managed according to clear standards (Field et al., 2008). This was formulated in the Minimum Information about a Genome Sequence (MIGS) approach, which aims at standardizing and organizing the publication of genomic and metagenomic data, and in the improvement of its exchange (Field et al., 2008).

Identification of unknown sequences

The identification of unknown sequences *de novo*, also referred to as annotation, is based on their homology to sequences deposited in (publicly available) databases. The amount of sequences in such repositories doubles approximately every 18 months and is expected to increase even faster if high-throughput sequencing is used more routinely (Gupta, 2008; Glockner and Joint, 2010). The deposited sequences are derived from previous sequencing efforts, but must not necessarily be linked to any supplemental information on structure, function or origin (Gilbert and Dupont, 2011). Hence, some general considerations must be addressed:

- The majority of sequences in public databases has no experimental proof for the structure or function they encode and is based solely on similarity and the presence of conserved domains. Thus, a basically unknown sequence will often be used to identify another unknown sequence.

- High sequence homology does not necessarily imply functional homology. Very similar sequences do not automatically encode for a similar function, but sequences with as low as 20% identity can indeed lead to proteins with highly similar three-dimensional structures and identical functions (Bourne et al., 2010).
- Even if a sequence is correctly annotated in the database, the metabolic context as derived by physical, chemical and biological parameters might lead to a different function. Thus, there is no absolute proof on the functional role of the enzyme a sequence encodes, and ecological interpretations could be biased.

However, the information gained from bioinformatic profiling of the microbial community structure and function for large datasets is still much better than that for very small datasets or for single sequences (Gilbert and Dupont, 2011), and the current bioinformatic techniques are the only ones available for metagenomic and metatranscriptomic datasets. Although many of them still rely on principles developed for manual analysis of only few single sequences or for the reconstruction of genomes from single organisms, more and more bioinformatic tools are indeed designed specifically for metagenomic and metatranscriptomic datasets (see below). The major challenge lies here in the streamlining and automating of techniques and tools – the construction of bioinformatic analysis pipelines, which allow gathering and management of the obtained data as well as integrative analysis and interpretation.

Sequence-sequence alignments

The most commonly used tool for identifying unknown sequences is the Basic Local Alignment Search Tool (BLAST) from the National Centre for Biotechnology Information (NCBI) (Altschul et al., 1990). It was developed in 1990 and has undergone several modifications and improvements since then. The principle of BLAST is the usage of heuristic local similarity algorithms. The program searches for highly conserved stretches within two sequences (sequence-sequence alignment), allowing larger parts of unmatched or gapped stretches and sets up a matrix of similarity scores for all possible pairs of residues (Altschul et al., 1990). The maximum value of similarity is computed by enlarging or shortening the local sequence alignment and the values for all available local alignments within one sequence pair are summed up. The local alignment of homologous sequence stretches allows the BLAST algorithm to initially exclude large amounts of sequences from the database before calculating the final, overall sequence similarity. Hence, the great advantage of BLAST is the speed of the search and the relatively low computational resources required. On the other hand, a problem associated with BLAST is that it is mostly used with the NCBI nucleotide and protein databases or the Swiss-Prot database (from the Swiss Institute of Bioinformatics), which are primary databases and not curated. Furthermore, though two sequences which are significantly similar over the entire length are likely to be homologous, as much as 50% similarity over a short sequence can occur just by chance.

BLAST can be used for both nucleotide (DNA and RNA) and amino acid (protein) sequence analysis. In the latter case, a translation of the nucleotide into the amino acid sequence, including all six open reading frames, can optionally be carried out by BLAST automatically. The different BLAST-searches that can be performed are:

- BLASTn: nucleotide query vs. nucleotide sequence database
- BLASTp: amino acid query vs. amino acid sequence database
- BLASTx: translated nucleotide query vs. amino acid sequence database
- tBLASTn: amino acid query vs. translated nucleotide sequence database
- tBLASTx: translated nucleotide query vs. translated nucleotide sequence database

Earlier developed heuristic algorithms next to BLAST include FASTA (Lipman and Pearson, 1985; Pearson and Lipman, 1988), which also compare sequences locally and pairwise. FASTA algorithms are also used for both phylogenetic and functional identification of sequences, of course dependent and limited upon the information available in the selected databases, too.

Profile-sequence alignments

Predicting a function of an unknown sequence can also be achieved by using profile-sequence alignments. This approach does not compare sequences pairwise, but rather uses a multiple alignment or a multiple alignment’s profile hidden Markov model (profile HMM) of known genes to compare it to the unknown sequence. A profile HMM contains much more information than a single sequence. Thus, the advantage is an improved detection of conserved patterns and motifs (domains) in unknown sequences.

The success of a profile-sequence alignment is dependent upon the quality of the profile HMMs used for the alignment. Profile HMMs are probabilistic models of sequences belonging to the same protein family and are calculated from a multiple alignment of sequences from that family. The underlying sequences are deposited in primary databases, e.g. the Swiss-Prot database. Profile HMMs themselves are stored in a secondary database like the Pfam protein families database (Bateman et al., 2004; Finn et al., 2010). Especially the Pfam-B database entries are of high quality, because they are manually curated, while Pfam-A is not manually curated).

The profile-sequence alignment can be carried out by different programs; the two most widely used are Position-Specific Iterated BLAST (PSI-BLAST) searches and implementations of profile hidden Markov Models, e.g. HMMER (Finn et al., 2011).

Profile-sequence alignments are primarily used for assigning functions to an unknown sequence and offer only a limited possibility of determining phylogenetic information of an unknown sequence. The latter can be achieved through the construction of phylogenetic trees, e.g. with the ARB software package (Ludwig et al., 2004), but this approach has quite some limitations, because it is a labour-intensive manual process that requires significant effort.

Reconstruction of biological functions

Once unknown sequences have been phylogenetically and functionally identified, the reconstruction of metabolic pathways, or more precisely, the interpretation of the composition of identified sequences can be used to characterize the sampled environment.

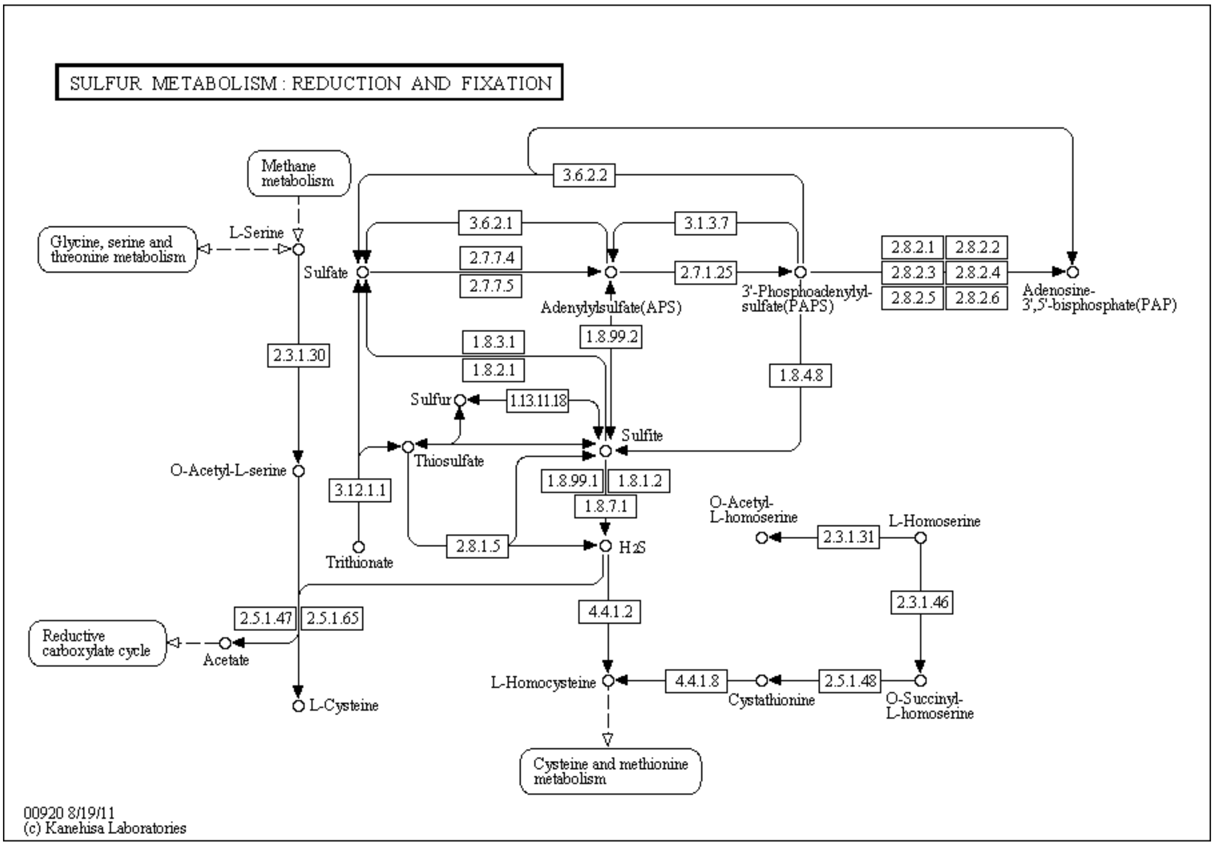


Figure 2: The sulfur metabolism pathway with all involved EC numbers and links to other KEGG-pathways as presented in KEGG (source: http://www.genome.jp/kegg/).

One approach is to map the sequence information on metabolic pathways (Figure 2).
KEGG (Kyoto Encyclopedia of Genes and Genomes) offers a collection of metabolic pathways on its homepage (Ogata et al., 1998). The metabolic pathways in KEGG, which are compiled by the Japanese Biochemical Society, display information from a collection of scientific publications and books and are based on the famous wall chart 'Biochemical Pathways' from the company Boehringer-Mannheim (Ogata et al., 1998). Regularly updated and verified, KEGG now includes 153 pathway maps. KEGG uses the concept of Enzyme Commission numbers (EC numbers), which categorize biological functions independent of structure similarities. The collection thus includes structurally different enzymes, which catalyze a similar or identical function, but it excludes structurally similar enzymes which carry out a different biological function. EC numbers consist of four digits separated by periods, uniquely classifying every known biological function, and can be directly transferred onto the metabolic pathway maps of KEGG (Ogata et al., 1998).
Functional gene analysis can also be performed by using hierarchical clustering of functional categories. These functional categories can be quite arbitrary groups, e.g. a set of manually selected genes, pathways or whole metabolisms, but can also be comprised of the predefined KEGG-categories (Xie et al., 2010; Eloe et al., 2011).

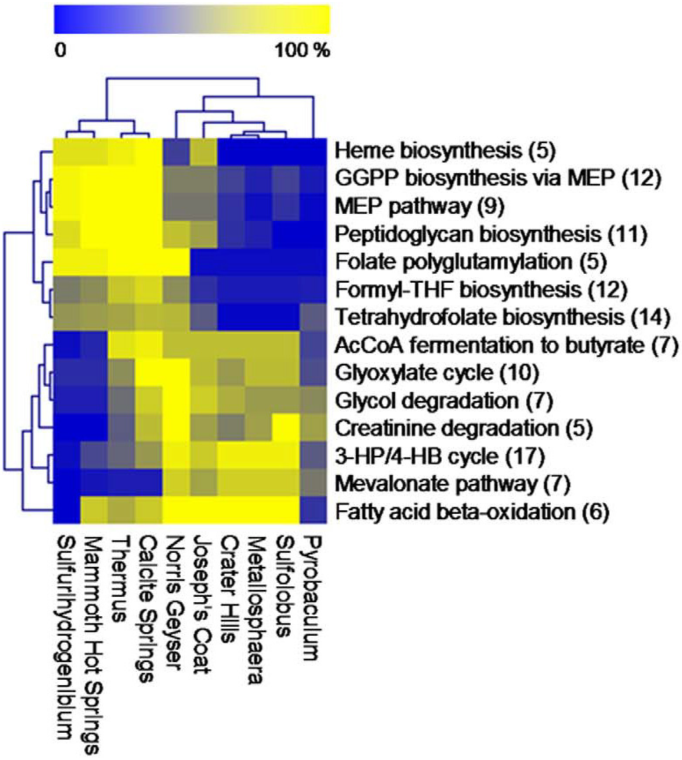


Figure 3: Two-way clustering of 10 metagenomic samples and 14 biochemical pathways. Heatmap scaling indicates pathway completeness in percent (from Inskeep et al., 2010).

Hierarchical clustering is often carried out as a two-way clustering approach (biclustering), which includes the comparison of different functional categories and different samples. Figure 3 shows one of the first metagenomic biclustering approaches that was published, displaying 14 functional categories and 10 different samples with a heatmap scaling (Inskeep et al., 2010). The displayed scaling has to be considered qualitative and not quantitative, since only relative comparisons are carried out.

Stand-alone analysis tools

As high-throughput sequencing is used by more and more scientists with no formal knowledge in bioinformatics, several tools were developed to provide a complete analysis of nucleic acid datasets without the need to write specific scripts or programs. In 2007, the first stand-alone metagenome analysis tools (MEGAN) was published (Huson et al., 2007). In the same year, the web-based metagenomic rapid annotations using subsystems technology (MG-RAST) annotation platform was made available, offering a whole variety of analysis tools. MG-RAST also serves as a repository for publicly available metagenomes and metatranscriptomes (Meyer et al., 2008). However, many of the analysis done by MEGAN and MG-RAST can be considered ‘black box’ metagenomic analysis – a user unfamiliar with bioinformatic principles might not choose adequate threshold cut offs for certain parameters. Also custom-designed analyses, to answer specific scientific questions are difficult to achieve with the web-based programs alone.

Furthermore, a recent study has shown that the use of the Lowest Common Ancestor (LCA) approach, which MEGAN employs, can result in an increased number of insignificant hits being assigned to relatively higher taxonomic levels (root, cellular organisms, bacteria, etc.) and thus reducing the specificity of the analysis. This is in parts due to the fact that MEGAN was originally intended to analyze genomic DNA from the woolly mammoth and not metagenomes of mostly new and unknown microorganism (Poinar et al., 2006). The algorithms which are now used in MEGAN partly solve this problem by employing a bit-score cut off parameter in order to filter out insignificant hits, thereby improving the specificity of assignments (Huson et al., 2007; Haque et al., 2009).

1.3. Study area

Eastern Boundary Upwelling Systems are areas in the oceans of high biomass production which support some of the most productive fisheries worldwide (Carr, 2002; Chavez and Messie, 2009). They are found in the Eastern Tropical North and South Pacific (ETNP – USA and Mexico and ETSP – Ecuador, Peru and Chile) and in the Eastern Tropical North and South Atlantic (ETNA – Mauritanian upwelling and ETSA – Benguela current upwelling) (Stramma et al., 2008) and are responsible for about 50-58% of the global fish catch, despite the fact that they comprise only 0.1-0.2% of the total ocean area (Ryther, 1969; Pauly and Christensen, 1995). The Peruvian coastal waters alone supports a catch of 10^7 tons of Anchovies (Montecino and Lange, 2009).

Marine oxygen minimum zones

The high primary production in surface waters in Eastern Boundary Upwelling Systems is driven by the upwelling of nutrient-rich waters (Friederich and Codispoti, 1987; Helly and Levin, 2004; Karstensen et al., 2008; Ulloa and Pantoja, 2009). Upwelling is an upward water movement process caused by a combination of trade winds, the Earth's rotation (Coriolis force) and an offshore movement of surface waters, referred to as Ekman transport (Ekman, 1905). Upwelling transports cold, dense and nutrient-rich deep water towards the ocean surface, replacing the warm and usually nutrient-depleted surface water. This input of nutrients into photic surface waters enables high photosynthetic biomass production by the phytoplankton community. A significant proportion of biomass sinks out of the surface layer and is remineralized via microbial respiration at intermediate water depths (usually between 100-1000 m), leading to severe O_2 depletion (Wyrtki, 1962; Dugdale, 1972; Helly and Levin, 2004; Lam and Kuypers, 2011). These oxygen-depleted waters are referred to as oxygen minimum zones (OMZs). In addition to OMZs in Eastern Boundary Upwelling Systems, oxygen-depleted waters are also present in enclosed coastal basins like the Baltic Sea (Brettar and Rheinheimer, 1991; Brettar et al., 2006; Glaubitz et al., 2009) and the Black Sea (Jorgensen et al., 1991; Luther et al., 1991; Sorokin et al., 1995) as well as in the northern Indian Ocean (Helly and Levin, 2004; Paulmier and Ruiz-Pino, 2009) (Figure 4). Nevertheless, OMZs are often not totally oxygen-free waters, but are rather featuring low O_2 concentrations, often below the detection limit of the sensors used, and immediate O_2 consumption if new O_2 is introduced into the system (Stevens and Ulloa, 2008; Ulloa and Pantoja, 2009; Finster and Kjeldsen, 2010). Despite their restricted spatial extent, they are of great interest to humans from an ecological as well as from an economical perspective, since low oxygen concentrations are lethal for most multi-cellular organisms, including fish (Danovaro et al., 2010; Seibel, 2011).

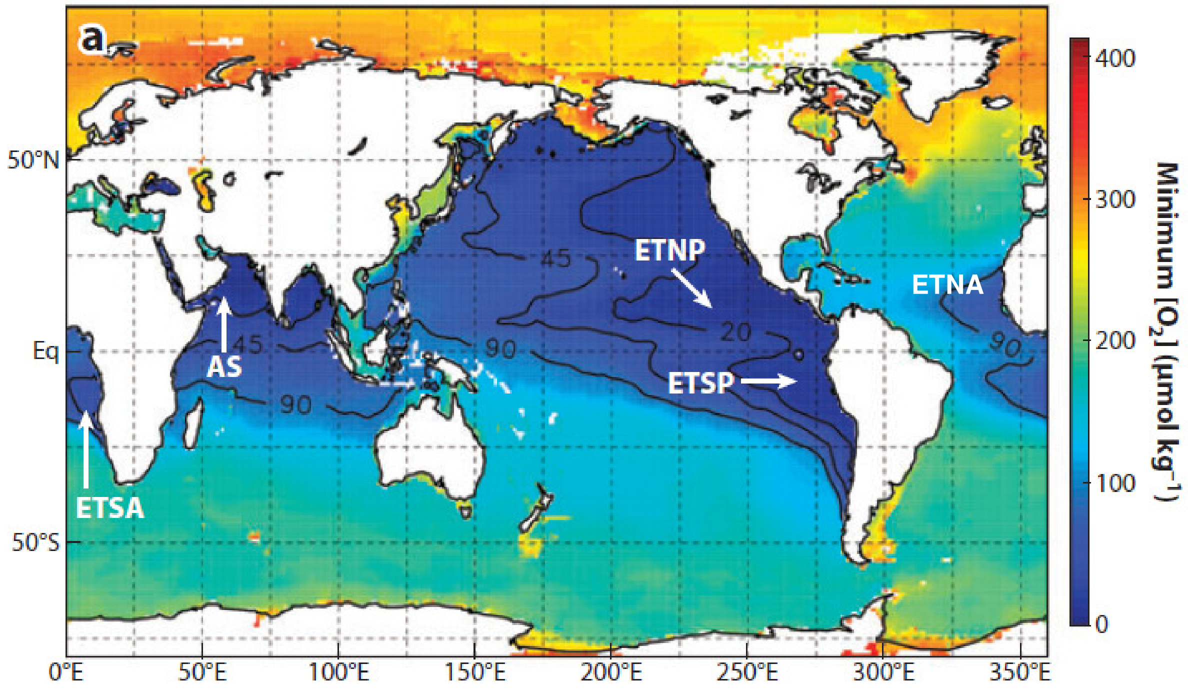


Figure 4: Oxygen minimum zones (ESTA - Eastern Tropical South Atlantic, AS - Arabian Sea, ETNP - Eastern Tropical North Pacific, ETSP - Eastern Tropical South Pacific and ETNA - Eastern Tropical North Atlantic) in the world's oceans. Shown are the minimum oxygen concentrations (µmol kg^{-1}) in vertical water column (modified after Lam and Kuypers, 2011).

Oxygen-depleted waters are predicted to increase in both frequency and size (Stramma et al., 2008). One major reason is enhanced nutrient input due to anthropogenic activity (e.g. use of artificial fertilizers in farming), which results in the eutrophication of coastal waters, and thus enhances surface water productivity (Naqvi et al., 2000; Beman et al., 2005). Secondly, ocean warming in the course of global climate change will decrease the solubility of O_2. In addition to this effect, the predicted increase in surface water temperatures will also intensify the stratification of the ocean, which leads to a reduced gas exchange of surface with subsurface waters, eventually reducing the transfer of O_2 to deeper waters (Sarmiento et al., 1998; Grantham et al., 2004). Thus, the future oceans might potentially suffer a decrease in fish populations and other commercially important aquatic species, which could lead to undersupply of seafood in coastal areas.

Respiration processes and nitrogen-loss

Oxygen-producing photosynthetic organisms dominate marine surface waters and consequently, oxygen-dependent heterotrophic respiration is also by far the predominant type of respiration in surface waters. But also in deeper oceanic waters and close to the seafloor O_2 is usually present and preferentially used (Orcutt et al., 2011).

In contrast to most multi-cellular organisms, many microorganisms can still carry out oxic respiration at concentrations that would be lethal for many higher organisms. The limit for those microorganisms can be even below the detection limit (~50 nM) of state of the art oxygen-sensors (Revsbech et al., 2009; Kalvelage et al., 2011). Nevertheless, the community in OMZ waters is diverse and complex, and distinctly different to oxygenated open-ocean and deep-sea communities (Lam and Kuypers, 2011).

Generally, microorganisms living in OMZs are not energy-limited. A great supply of biomass is generated by the photosynthetic phytoplankton community in surface waters. These reduced carbon compounds sinking to deeper waters represent a freely available source of energy for heterotrophic microorganisms. Since O_2, the most widely used terminal electron acceptor during the respiration of organic matter is scarce, a pronounced OMZ should rather be considered as being limited in this terminal electron-acceptor. The scarcity or absence of O_2 gives rise to different physical and chemical properties for the microbial community. Instead of O_2, other oxidized compounds are utilized for respiration. Some of the most important compounds are listed in Table 1.

Table 1: Common electron acceptors used for respiration of organic matter, corresponding redox partners and concentrations in the world's oceans, ordered according to the potential energy that they theoretically can provide (modified after Cypionka, 2010 and Lam and Kuypers, 2011).

Electron acceptor	Redox couple	Concentration
Oxygen	O_2 / H_2O	Variable (partly limiting)
Nitrate	NO_3^- / N_2	30 µM or less (partly limiting)
Manganese dioxide	MnO_2 / Mn^{2+}	nM or less (limiting)
Nitrate	NO_3^- / NH_4^+	30 µM or less (partly limiting)
Iodate	IO_3^- / I^-	0.2-0.5 µM (not limiting)
Ferric oxide	Fe_2O_3 / Fe^{2+}	nM or less (limiting)
Sulfate	SO_4^{2-} / H_2S	28 mM (not limiting)
Carbon dioxide	CO_2 / CH_4	variable (not limiting)

Due to the relatively high concentrations of nitrate (NO_3^-) and iodate (IO_3^-) in seawater, these compounds play a greater role in respiration processes than manganese dioxide (MnO_2) and ferric oxide (Fe_2O_3), which occur typically only in nanomolar concentrations and are considered to be limiting (Lam and Kuypers, 2011). Especially NO_3^- and related oxidized nitrogen species like nitrite (NO_2^-), nitric oxide (NO) and nitrous oxide (N_2O) are the second preferred terminal electron-acceptors after O_2 (Cypionka, 2010). A special characteristic of oxidized nitrogen species, in addition to their function as electron acceptors is that they also serve as a major and essential nutrient for growth and the building-up of biomass. Thus, a lack in nitrogen will also limit biomass production.

The reduction of oxidized nitrogen species during respiration of organic matter (heterotrophic denitrification) is a stepwise process ($NO_3^- \rightarrow NO_2^- \rightarrow NO \rightarrow N_2O \rightarrow N_2$) and will result in the formation of dinitrogen gas (N_2). N_2 is relatively inert and not available as a nutrient for most living organisms, except for those few organisms (diazotrophs), which are capable of reducing N_2 gas and form ammonia (NH_4^+). Eventually, through heterotrophic denitrification processes, the oceans will loose fixed nitrogen to the atmosphere (Figure 5) (Emery et al., 1955; Codispoti et al., 2001; Gruber, 2004).

Next to the heterotrophic denitrification, the anaerobic ammonia oxidation (anammox), coupling the reduction of NO_2^- and the oxidation of NH_4^+ also leads to the formation of N_2 (Kuypers et al., 2005; Hamersley et al., 2007; Lam et al., 2009; Jensen et al., 2011). Both processes occur in such magnitude in OMZs that they account for approximately 30-50% of the fixed nitrogen loss in the global oceans, despite the relative small size of OMZs (Codispoti et al., 2001; Castro-Gonzalez and Farias, 2004; Galloway et al., 2004; Codispoti, 2007; Lam et al., 2009; Jensen et al., 2011; Lam and Kuypers, 2011). As an ultimate result of the nitrogen-loss, a deficit of inorganic nitrogen relative to inorganic phosphorous (N^*) will develop, which is considered to be a common characteristic of oxygen-depleted marine systems (Deutsch et al., 2007; Ulloa and Pantoja, 2009). A low N^* value can potentially favor those few diazotrophic microorganisms, who are capable of fixing (reducing) N_2 and form NH_4^+.

The close proximity of reduction and oxidation of distinct chemical species like nitrogen within OMZs and adjacent areas with increasing O_2 concentrations leads to an enhanced cycling of elements in these ocean regions (Figure 5) (Ward et al., 1989; Lam and Kuypers, 2011). Since the use of a distinct redox couple for respiration determines the energy that can be obtained by the reaction, a clear spatial separation of reactions and chemical zones should occur. Microorganisms that support the redox reactions are thought to use the energetically most efficient one if they are able to switch between different modes of respiration. An "oxygen-switch" has been hypothesized and is widely applied throughout different oceanic models (Strous et al., 1997; Paulmier et al., 2009).

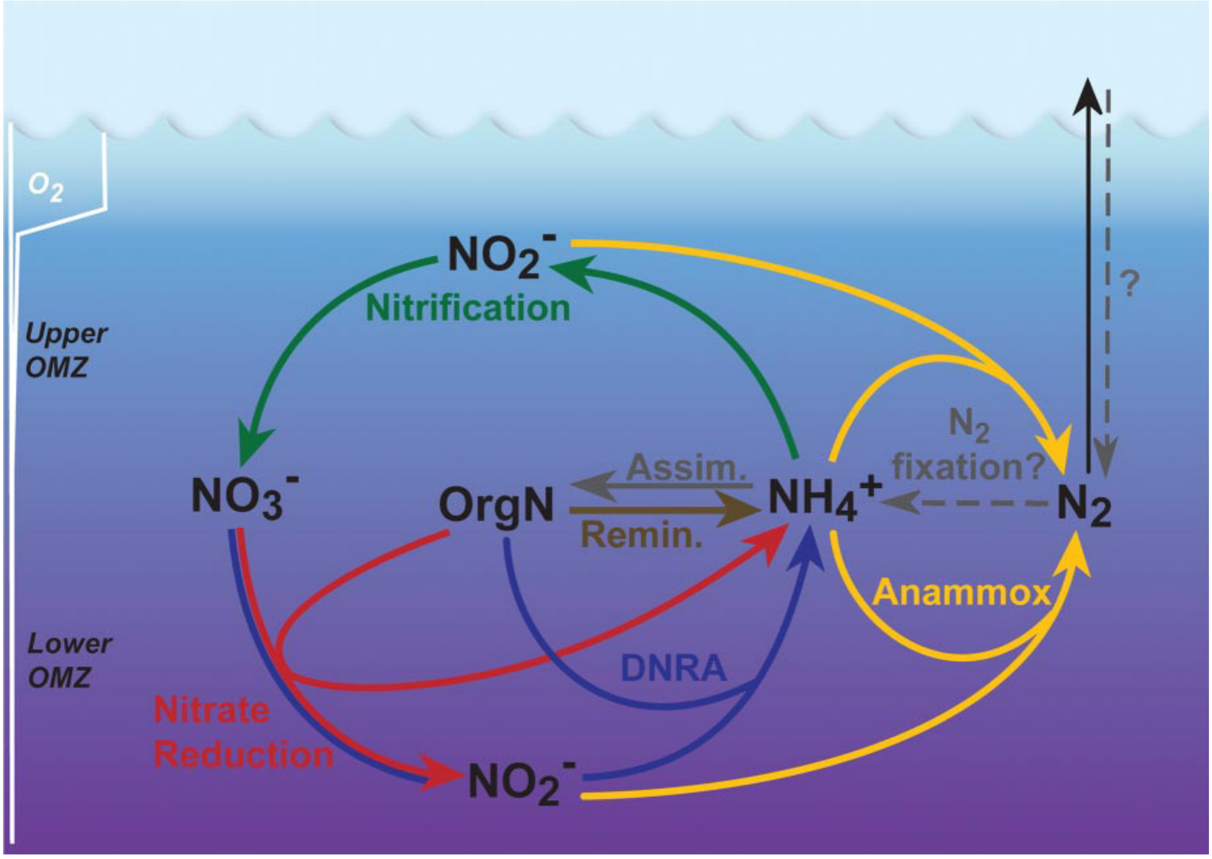


Figure 5: Schematic diagram of the nitrogen cycle in the Peruvian OMZ. Anammox (yellow), nitrification (green), dissimilatory nitrate reduction to ammonia (DNRA, blue), nitrate reduction (red), remineralization of organic matter (brown), ammonia assimilation (grey), nitrogen fixation (dashed grey) and oxygen concentration (white). Microaerobic conditions are suggested for the upper part of the OMZ (from Lam et al., 2009).

If a certain threshold of minimal O_2 concentration is reached, the second preferred electron acceptor after O_2 is used instead and the set of needed genes are transcribed exclusively then (Lam and Kuypers, 2011). However, recent findings suggest that respiration profiles actually do overlap and are occurring continuously (Figure 6). This might be in part due to the fact that some microorganisms are limited to the use of only one redox couple for respiration. They will make use of this electron acceptor, even at very low concentrations or otherwise are forced to shut down their metabolism completely (Canfield and Thamdrup, 2009).

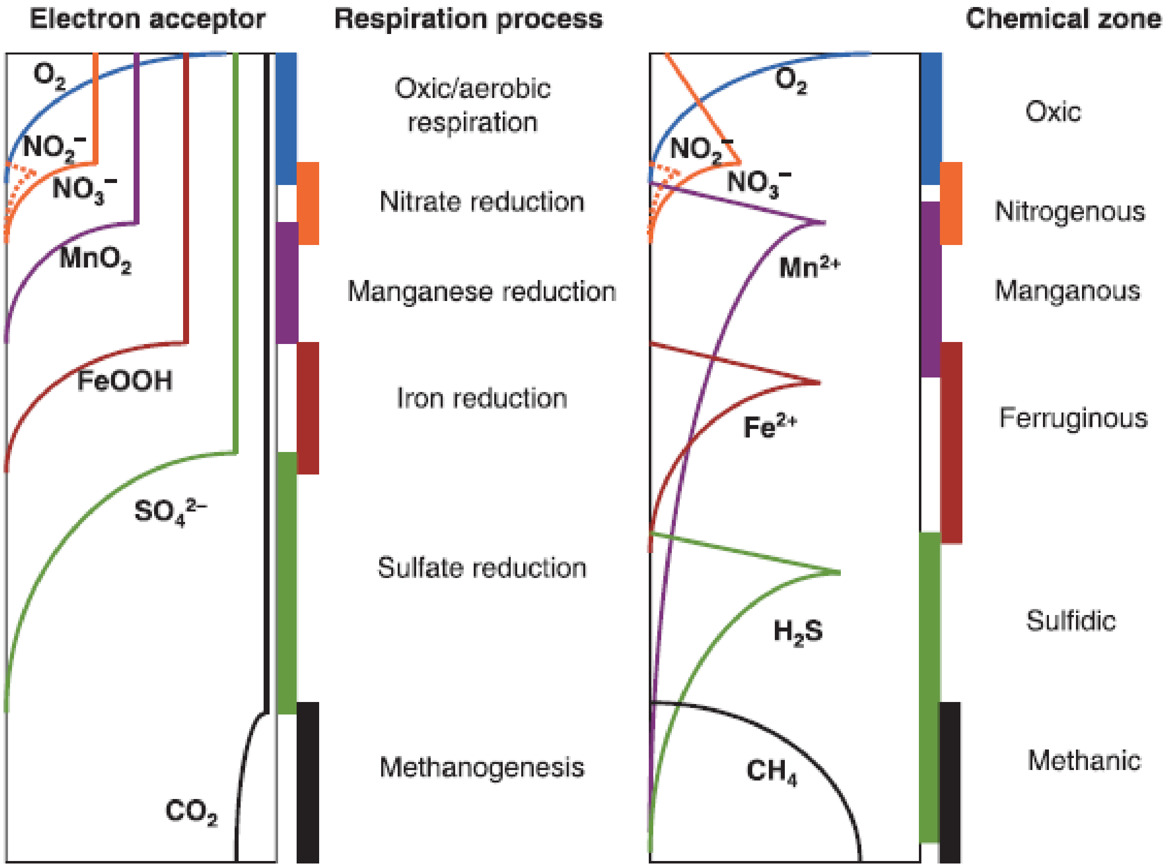


Figure 6: Schematic cartoon displaying the depth distribution of common electron acceptors and the names used to represent the respiration processes and the chemical zones (from Canfield and Thamdrup, 2009).

Sulfidic ocean waters

On the lower end of the energetically favourable redox-reactions, the reduction of sulfate (SO_4^{2-}) and carbon dioxide (CO_2) are found (Table 1 and Figure 6). While the reduction of CO_2, concurrently with the oxidation of hydrogen gas (H_2) can produce methane (CH_4), SO_4^{2-} reduction leads to the formation of hydrogen sulfide (H_2S). H_2S is a toxin that is more deadly to multi-cellular life than cyanide. It has been invoked as a cause for massive fish mortality and anomalously low fish catch in certain upwelling regions (Copenhagen, 1954; Hart and Currie, 1960; Hamukuaya et al., 1998).

Due to the high concentration of SO_4^{2-} in seawater (~28mM), SO_4^{2-} reduction is commonly occurring in anoxic sediments, coupled to the degradation of biomass (Jorgensen, 1982). Sulfidic waters develop, when O_2, NO_3^- and NO_2^- are depleted within the water column and H_2S is released from underlying sediments. This outgasing can either happen in eruptive bursts or continuously (Bruchert et al., 2003; Lavik et al., 2009; Shao et al., 2010; Orcutt et al., 2011). So far, it is not known how long H_2S can remain in a water column or if it can accumulate without the initiation of sulfidic sediments through large-scale SO_4^{2-} reduction within the water column. There have been speculations about SO_4^{2-} reduction even in the water column of moderate OMZ waters (Canfield et al., 2010), however, it is assumed that the produced H_2S will immediately be reoxidized (Hannig et al., 2007; Lavik et al., 2009), such that an accumulation event of H_2S in the water column without a sediment-initiation seems improbable.

Apart from sulfidic chemocline and bottom waters of enclosed basins like the Baltic Sea (Brettar and Rheinheimer, 1991; Brettar et al., 2006; Glaubitz et al., 2009) and the Black Sea (Jorgensen et al., 1991; Luther et al., 1991; Sorokin et al., 1995; Glaubitz et al., 2010), the Cariaco basin off Venezuela (Zhang and Millero, 1993; Taylor et al., 2001; Hayes et al., 2006) and the fjord Saanich Inlet in Canada (Tebo and Emerson, 1986; Walsh et al., 2009), sulfidic waters have been reported only very infrequently for open-ocean environments (Dugdale et al., 1977; Naqvi et al., 2000; Lavik et al., 2009) and can be regarded as extreme specifications of OMZs. The main detoxification process of H_2S is the reverse reaction of its formation, the oxidation of H_2S back to SO_4^{2-} (Figure 7).

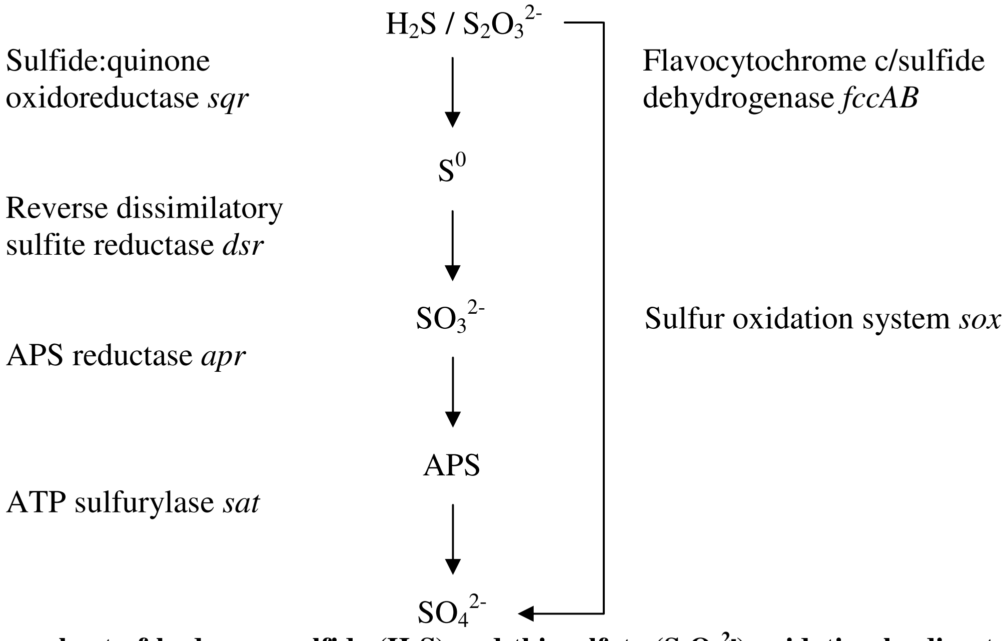

H_2S / $S_2O_3^{2-}$
Sulfide:quinone oxidoreductase *sqr*
Flavocytochrome c/sulfide dehydrogenase *fccAB*
S^0
Reverse dissimilatory sulfite reductase *dsr*
SO_3^{2-}
Sulfur oxidation system *sox*
APS reductase *apr*
APS
ATP sulfurylase *sat*
SO_4^{2-}

Figure 7: Flow chart of hydrogen sulfide (H_2S) and thiosulfate ($S_2O_3^{2-}$) oxidation leading to the formation of SO_4^{2-}. Shown are the genes encoding for enzymes responsible for the transformations. The set of genes allows no conclusion whether the oxidant (redox partner) is O_2 or NO_3^- (modified after Walsh et al., 2009).

In this process, the microorganisms do not use reduced carbon compounds as the energy source, but H_2S. H_2S is oxidized and either elemental sulfur (S^0), sulfite (SO_3^{2-}) or SO_4^{2-} is formed. The oxidation process coupled to the reduction of NO_3^- is termed sulfur-driven autotrophic denitrification or anaerobic sulfur oxidation. Under nitrate-limiting conditions (limitation in electron acceptors) the H_2S oxidation is hypothesized to proceed only up to the formation of S^0, which requires only the transfer of two and not eight electrons, as needed for SO_4^{2-} formation (Lavik et al., 2009; Walsh et al., 2009).

The speed of H_2S detoxification by microorganisms and its connections to other processes still remain largely uncertain. Especially, the question whether and how sulfidic waters in oceanic OMZs can persist and stabilize over longer periods of time, as it is observed in the Baltic and the Black Sea, remains unclear. Furthermore, the impact of man-made environmental changes on oxygen-depleted environments is not known; sulfidic systems might increase in frequency in the future, aggravating the uncertainties of the development of fish stocks and human feeding.

Microbial diversity

The inhabiting organisms of OMZs are largely unknown. Due to the toxicity of low O_2 concentrations for most multi-cellular organisms, OMZs mainly support the growth of microorganisms (Danovaro et al., 2010). Most of those are still uncultured and unknown. Although recent studies targeted the analysis of microbial diversity and functional activity in oxygen-depleted waters (Castro-Gonzalez et al., 2005; Molina et al., 2007; Stevens and Ulloa, 2008; Woebken et al., 2008; Molina et al., 2010), the methodological approaches were mainly Sanger-sequencing-based and focused only on few single genes or ribosomal sequences. So far, only a handful of studies have applied high-throughput sequencing on open-ocean OMZ-waters (Canfield et al., 2010) and only one study analyzed coupled metagenomic and metatranscriptomic datasets (Stewart et al., 2011).

Some of the abundant microorganisms detected there were archaeal ammonia-oxidizer similar to *Nitrosopumilus maritimus*, autotrophic anammox-planctomycete related to *Candidatus Kuenenia stuttgartiensis* and ubiquitous α-proteobacterial *Candidatus Pelagibacter* species. Even less is known about microorganisms living in sulfidic OMZ waters. Most information about sulfur-metabolizing microorganisms are either derived from studies targeting sulfidic deep-sea hydrothermal vents (Sunamura et al., 2004; Huber et al., 2007; Nakagawa and Takai, 2008; Huber et al., 2010; Xie et al., 2010) or terrestrial sulfidic springs and caves (Engel et al., 2004; Engel, 2007; Macalady et al., 2008; Porter and Engel, 2008; Chaudhary et al., 2009; Chen et al., 2009; Niederberger et al., 2009; Porter et al., 2009; Inskeep et al., 2010). This is mainly due to the fact that these sulfidic conditions are mostly permanent and stable and therefore easier to sample.

These studies have shown that main players in sulfidic habitats are often proteobacteria. Commonly affiliated to either SO_4^{2-} reduction or H_2S oxidation are groups of α-, γ-, δ- and ε-proteobacteria (Suzuki et al., 2005; Urakawa et al., 2005; Lavik et al., 2009; Walsh et al., 2009; Yamamoto et al., 2010). Especially γ-proteobacteria related to chemolithoautotrophic gill symbionts of deep-sea hydrothermal-vent clams have been detected (Lavik et al., 2009; Walsh et al., 2009; Zaikova et al., 2010). The genomes of *Candidatus* Vesicomyosocius okutanii HA (Kuwahara et al., 2007), *Candidatus* Ruthia magnifica str. Calyptogena magnifica (Newton et al., 2007) and the metagenome of SUP05 cluster bacterium (Walsh et al., 2009) have been published. Other inhabitants, discrete populations of α- and ε-proteobacteria, probably related to *Arcobacter spp.* and *Roseobacter spp.*, seems to play an important role, too (Lavik et al., 2009). But further taxonomic assignments are missing and the microbial community in sulfidic ocean waters, as well as their functional mechanisms for H_2S detoxification remains largely unknown until now.

1.4. Aims of this thesis

The aim of this thesis was the coupled metagenomic and metatranscriptomic characterization of microbial communities from the oxygen minimum zone off Peru. While most samples were taken in waters displaying ‘classical’ oxygen minimum zone features, one set of samples was collected during the occurrence of a sulfidic event, the massive accumulation of toxic hydrogen sulfide in the water column. Sulfidic ocean waters display an extreme specification of oxygen minimum zone waters and so far have been detected very rarely. The underlying biogeochemical processes are still largely unknown.

Thus, the objectives of this thesis were the description of microbial communities within sulfidic waters and the identification of differences when compared to communities from ‘classical’ oxygen minimum zone waters. The aim was to analyze the phylogenetic diversity, the metabolic potential and the metabolic activity in detail using high-throughput metagenomic and metatranscriptomic sequencing techniques.

In order to achieve these goals of this thesis, the following steps were necessary:

- Modifying the most up-to-date method of nucleic acid sample preparation for high-throughput metagenomic and metatranscriptomic sequencing.
- Assessing the influence of the sampling time on the detected microbial community structure and function from samples obtained from oxygen-depleted waters.
- Establishing a bioinformatic analysis pipeline for high-throughput sequence data, allowing the identification of unknown sequences with three different methodological approaches. Next to the commonly used BLAST-searches, profile hidden Markov model scans and the recruitment of sequences onto reference genomes was aspired.
- Storing and organizing the obtained data and metadata in a suitable and user-friendly database system. This database system should be assessable by users with only little background in bioinformatics.

In order to gain a holistic view of the microbial processes in oxygen-depleted and sulfidic waters, the high-throughput sequence data had to be combined with water column profiles, rate measurements, flux calculations, data from satellite remote sensing and microbial cell counts, obtained through collaborations with the Institute for General Microbiology in Kiel and the Max-Planck-Institute for Marine Microbiology in Bremen.

Material and Methods

2. Material and methods

2.1. Water collection and handling

All samples used for this thesis were collected during research vessel Meteor cruise M77/3 on the Peruvian shelf between December 27th, 2008 and January 24th, 2009. Water was pumped from depth directly on board using a pump-conductivity-temperature-depth water sampler (pump-CTD) and filled in 4.5 litre polycarbonate bottles (Nalgene). For each sample 1,5-2 litres of water were prefiltered through 5 or 10 µm pore size filters (Millipore/Durapore Membrane filters) and then collected upon 0.2 µm pore size filters (Millipore/Durapore Membrane filters) using a vacuum pump (Sartorius eJet). From the time point the water was pumped on board, less than 18 minutes elapsed until the filters were put in microcentrifuge reaction tubes and flash frozen in liquid nitrogen.

2.2. Sample preparation and sequencing

DNA and RNA were extracted using the DNA/RNA-Allprep kit (Qiagen) with minor modifications in the protocol for the lyses step: The frozen filters were crushed with a disposable pestle and incubated with 200 µl lysozyme (10 µg/µl) and 1mM EDTA at ambient temperature for 5 minutes. Subsequently, 40 µl of Proteinase K (10 µg/µl) was added and samples were incubated for another 5 minutes. After adding 500 µl buffer RLT-Plus (containing 10 µl/ml β-mercaptoethanol) the manufacturer's instructions were followed. Total RNA was eluted in 50 µl nuclease-free water and a subsequent step of DNA digestion with the Turbo DNA-free kit (Ambion) was carried out. The rRNA was removed with Epicentre mRNA only prokaryotic and Microb*Express* kits (Ambion). Cleaned mRNA was subjected to an *in vitro*-amplification step using Message*Amp* (Ambion). Finally, cDNA was created by using superscript III cDNA synthesis kit (Invitrogen) with random hexameric primers (Qiagen). Leftover reactants and reagents were removed using the PCR Mini Elute Kit (Qiagen). The cDNA was immediately stored at -80°C until pyrosequencing. Throughout the whole sample preparation DNA and RNA-samples were subsequently quantified using nano-litre spectrophotometry (NanoDrop) and checked for degradation with BioRad Experion (RNA Standard & High Sense, see Figure 8). Furthermore, nuclease-free plastic consumables and nuclease-free water and reagents were used to hinder any possible degradation or contamination of RNA or cDNA. 50 µl of each DNA/cDNA-sample were sequenced with a GS-FLX pyrosequencer (Roche). Each sample was loaded on one quarter of a PicoTiter plate.

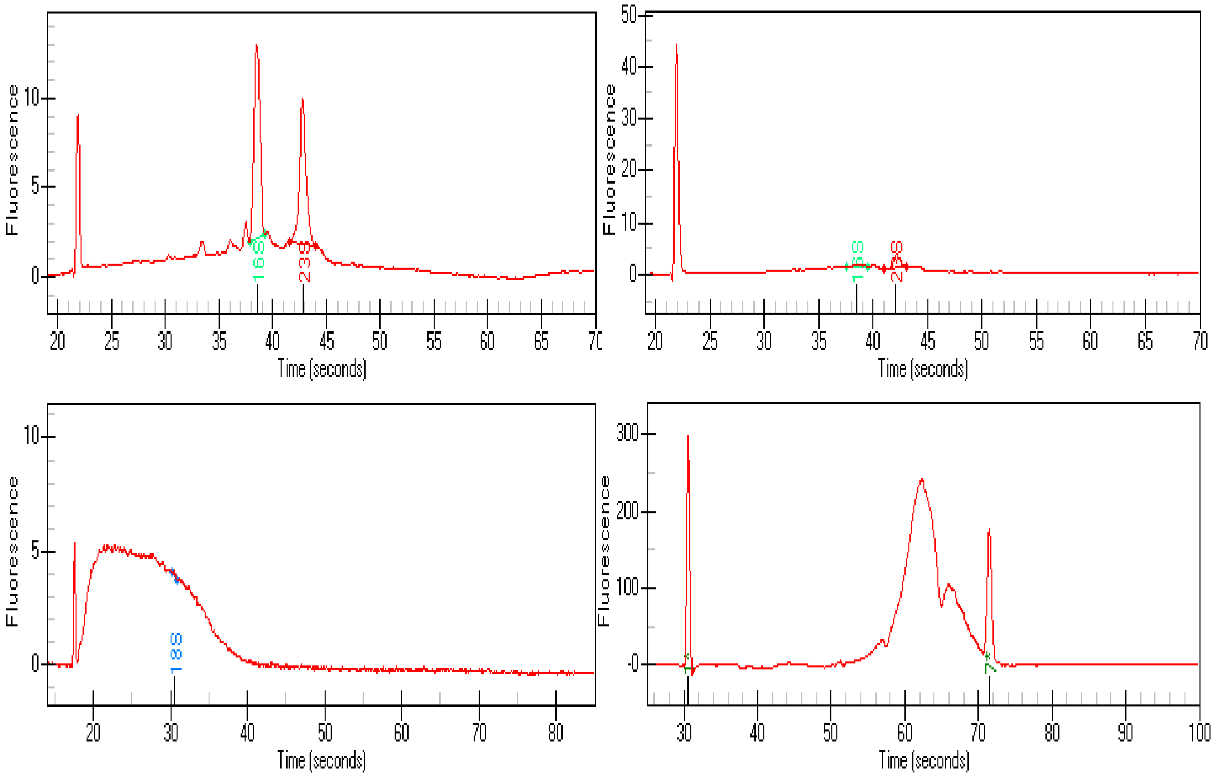


Figure 8: Electropherograms as obtained with the Experion Automated Electrophoresis System (Bio-Rad). The x-axis depicts time (relative size of DNA/cDNA and RNA) and the y-axis fluorescence (relative concentration of DNA/cDNA and RNA). First and last peaks (the last peak is only visible in the figure on the bottom right) are markers for time (relative size) estimation. The positions of predicted 16S or 18S ribosomal RNAs are depicted for all RNA samples. Top left: Total RNA. Top right: rRNA-depleted total RNA. Bottom left: amplified rRNA-depleted total RNA (aRNA). Bottom right: cDNA.

2.3. Data organization and analysis

An overview of bioinformatic analysis pipeline is shown in Figure 9. The raw sequences were clustered using Cd-hit (Li and Godzik, 2006) with a sequence identity threshold of 98% and a word length of 8 to identify cluster representative sequences. The rRNA sequences in these cluster representatives were identified by a BLASTn-search (Altschul et al., 1990) against the SILVA database (Pruesse et al., 2007) including both prokaryotic and eukaryotic small and large ribosomal subunit sequences with a bit score cut off of 86. Applying the same cut off, MEGAN (Huson et al., 2007) was used to make taxonomic assignments for the ribosomal cluster representative sequences and to split these from the non-ribosomal cluster representative sequence pool. Those cluster representatives without a hit in the SILVA database were subsequently compared against the non-redundant database of NCBI using BLASTx with a bit score cut off of 35. For both BLAST-searches, the top hit was used to characterize the cluster representative sequences. The non-rRNA cluster representative sequences were further scanned with profile hidden Markov models of the Pfam protein families (Finn et al., 2010) and the ModEnzA Enzyme Commission groups (Desai et al., 2011). Subsequent results were mapped to the KEGG reference pathways using the FROMP program (see manuscript B).

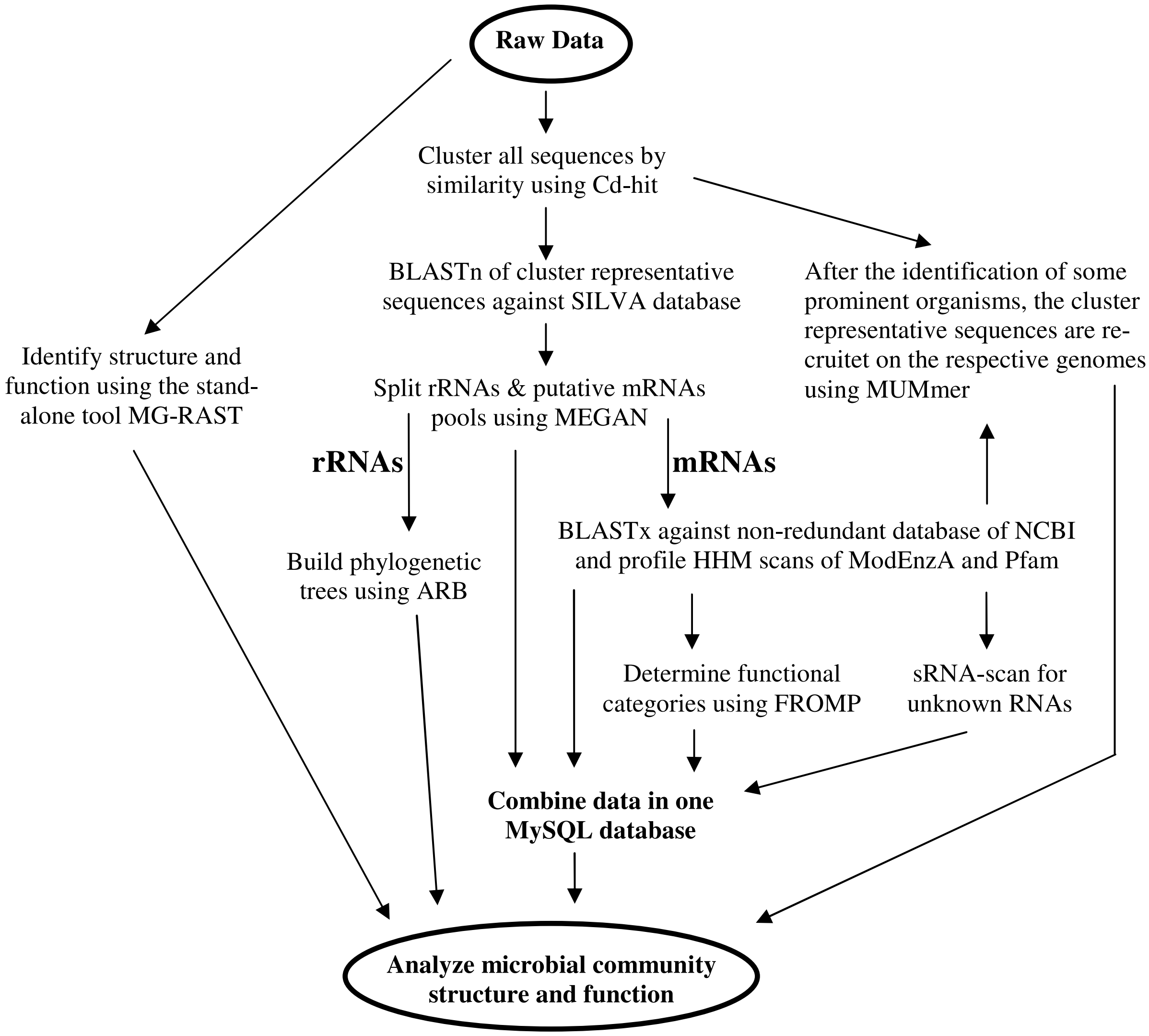

Raw Data
Cluster all sequences by similarity using Cd-hit
BLASTn of cluster representative sequences against SILVA database
After the identification of some prominent organisms, the cluster representative sequences are recruitet on the respective genomes using MUMmer
Identify structure and function using the stand-alone tool MG-RAST
Split rRNAs & putative mRNAs pools using MEGAN
rRNAs
mRNAs
Build phylogenetic trees using ARB
BLASTx against non-redundant database of NCBI and profile HHM scans of ModEnzA and Pfam
Determine functional categories using FROMP
sRNA-scan for unknown RNAs
Combine data in one MySQL database
Analyze microbial community structure and function

Figure 9: Flowchart of the bioinformatic analysis pipeline. A detailed description of the methods is also found in manuscript B and in the methods section of manuscript C.

The sequences, clustering information from Cd-hit, results from the BLAST-searches, the Pfam and EC-scans and the taxonomic assignment from MEGAN for each cluster representative sequence were added to a MySQL database for further analysis with phpMyAdmin via a common web browser. The EC numbers and Pfam assignments were imported to FROMP, which subsequently translated the Pfam assignments to EC numbers. All EC numbers were then exported from FROMP, along with an EC activity matrix of the different samples, and used as an input for hierarchical clustering (Eisen et al., 1998) with the

MultiExperiment Viewer program (Saeed et al., 2003). The EC activity matrix (with sample sizes equalized to the smallest sample, randomly selecting the same amount from the other samples) and the EC counts for each sample were also used to calculate the inverse Simpson's index (1/D) where D = Σ P_i^2 and P_i representing the proportional abundance of species i, and the Evenness E = (1/D)/S with S being the number of unique species. Similar calculations were also performed for the taxonomic assignments at the phylum level from the BLASTx searches normalized to total number of sequences having a BLASTx hit.

For the phylogenetic analysis of the selected genes, the representative rRNA sequences or sequences mapped to a specific EC number in FROMP were collected. To this, sequences with a matching BLAST-hit were added, collected from the SwissProt (Boeckmann et al., 2003) database. All sequences were assembled with the Celera Assembler (Huson et al., 2001) using parameters to build contigs with 12% error overlaps, and allowing 14% error globally, or with the TGICL assembly program (Huang and Madan, 1999) using the default parameters. Unique contigs and singletons were imported to ARB and aligned to a reference database and a maximum likelihood or a neighbour joining tree was constructed with PhyML (Guindon and Gascuel, 2003) in the ARB program (Ludwig et al., 2004).

The sequence data was further recruited on the (meta-) genomes of some prominent organisms using the MUMmer program (Kurtz et al., 2004). The fragment recruitment plots were created with the 'R' software package (www.R-project.org). The recruited reads were re-assessed using a BLAST search against the reference genomes and fragments hitting more than one genome with a bit score difference of less than 5% between the first and the second hit were discarded, giving rise to a non-overlapping set of reads for each genome. These reads were then BLASTed again to genome to calculate the average coverage for each base over non-overlapping windows of 300 base pairs from the reference genomes. The coverage of the reads for each metatranscriptomic sample in each reference genome window was normalized by the total number of BLASTx hits for that sample and divided by coverage of the same window from the corresponding metagenome reads (which had also been similarly normalized). This value, the expression ratio (odds ratio), was then corrected for the differences in sizes of the metatranscriptome and metagenome and plotted for selected regions of the reference genomes using customized R and PERL scripts (see manuscript B).

Results

3. Results

This thesis is based on the following manuscripts:

- **Manuscript A – RNA sampling in oxygen-depleted waters**
  Schunck, H.§, Desai, D.K.§, Großkopf, T., Schilhabel, M., Rosenstiel, P., LaRoche, J. (2012) Influence of the sampling time on the selective decay of abundant β-proteobacterial RNA-sequences obtained from oxygen-depleted waters off Peru. In preparation for submission.

  Contribution: H.S. designed the research, collected the samples, carried out the experiments and measurements, analysed the data and wrote the manuscript.

- **Manuscript B – Bioinformatic analysis of high-throughput sequence data**
  Desai, D.K.§, **Schunck, H.**§, Löser, J.W., Lommer, M., LaRoche, J. (2012) Fragment Recruitment on Metabolic Pathways (FROMP): A tool for comparative metabolic profiling of metagenomes and metatranscriptomes. In review in *Bioinformatics*.

  Contribution: H.S. designed the research and assisted in programming and scripting. H.S. designed the tool and tested its functionality. H.S. assisted in preparation of the manuscript.

- **Manuscript C – Microbial communities in sulfidic ocean waters**
  Schunck, H.§, Lavik, G.§, Desai, D.K., Großkopf, T., Kalvelage, T., Löscher, C.R., Paulmier, A., Mußman, M., Holtappels, M., Contreras, S., Siegel, H., Rosenstiel, P., Schilhabel, M.B., Graco, M., Schmitz, R.A., Kuypers, M.M.M., LaRoche, J. (2012) Giant hydrogen sulfide plume in the oxygen minimum zone off Peru stimulates high chemoautotrophic carbon dioxide fixation. Submitted to *PLoS ONE*.

  Contribution: H.S. designed the research, collected the samples, carried out the experiments and measurements, analysed the data and wrote the manuscript.

§Authors contributed equally to the work

Further contributions to manuscripts that do not form part of this thesis:

- **Manuscript D**

  Großkopf, T., Mohr, W., Baustian, T., **Schunck, H.**, Gill, D., Kuypers, M.M.M., Lavik, G., Schmitz, R.A., Wallace, D.W.R., LaRoche, J. (2012) Doubling of marine N_2 fixation rates based on direct measurements. In press in *Nature*.

- **Manuscript E**

  Großkopf, T., Löscher, C.R., **Schunck, H.**, Lavik, G., Kuypers, M.M.M., Kolber, Z., Friedrich, G., Chavez, F., Schmitz, R.A., LaRoche, J. (2012) High coastal productivity drives Southern Peruvian upwelling system into nitrogen limitation. In preparation for submission.

- **Manuscript F**

  Löscher, C.R., Großkopf, T., Gill, D., **Schunck, H.**, Pinnow, N., Desai, F., Lavik, G., Kuypers, M.M.M., LaRoche, J., Schmitz, R.A. (2012) Niche separation of novel groups of bacterial diazotrophs in the largest oxygen minimum zone of the world's ocean. In preparation for submission.

- **Manuscript G**

  Sperling, M., Roy, A.S., **Schunck, H.**, Gilbert, J.A., LaRoche, J., Engel, A. (2012) Effect of elevated CO_2 on bacterial community structure in the pelagial and in biofilms in an arctic fjord: An *in situ*-mesocosm study. In preparation for submission.

3.1. RNA sampling in oxygen-depleted waters

Influence of the sampling time on the selective decay of abundant β-proteobacterial RNA-sequences from oxygen minimum zone waters off Peru

Harald Schunck[1§], Dhwani Desai[1§], Tobias Großkopf[1], Markus Schilhabel[2], Philip Rosenstiel[2], Julie LaRoche[1]

[1]Helmholtz Centre for Ocean Research Kiel (GEOMAR), Düsternbrooker Weg 20, 24105 Kiel, Germany
[2]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University, Schittenhelmstraße 12, 24105 Kiel, Germany

[§]these authors contributed equally to the work

Abstract

Metatranscriptomics is an emerging technique that provides global analyses of gene expression in natural microbial communities. Rapid sample collection and processing is considered crucial to capture a realistic profile of the transcription activity *in situ*. However, time-dependent changes in transcription profiles after sampling have seldom been documented. Here we present metatranscriptomic data from the oxygen minimum zone off Peru, where replicate subsurface water samples were collected and subsequently filtered after 0, 20, 120 and 300 minutes of delay in order to assess the influence of sample processing time on the transcriptional profile. More than 1 million reads were obtained from total RNA using Roche GS-FLX sequencing technology. The data showed a rapid decrease in the number of β-proteobacteria in the microbial community starting already within a delay of 20 minutes of sample filtration. From the initially detected microbial community 30% of the sequences were affiliated to β-proteobacteria, while after 300 minutes those sequences had decreased to about 1%. This corresponds to a decrease of the β-proteobacterial population by 97%. In contrast, only minor changes were observed in most other bacterial and archaeal taxa within the same time. Our data emphasizes the importance of fast sampling procedures for transcriptomic studies to reduce potential biases in the microbial community structure.

Introduction

Biogeochemical cycling processes in the oceans are mostly carried out by a great number of different microbial taxa (e.g. Brettar and Rheinheimer, 1991; Jorgensen et al., 1991;

Thamdrup et al., 1996; Falkowski et al., 1998; Codispoti et al., 2001). The majority of those microorganisms eludes culturing, and in fact, have never been described (Amann et al., 1995; Pace, 1997; Streit and Schmitz, 2004; Glockner and Joint, 2010). Thus, especially culture-independent methodological approaches have added to our knowledge of microbial diversity and of elemental cycling processes in marine habitats. A broad set of tools is currently used, ranging from (q)PCR and microarray analysis to the sequencing of nucleic acids. In particular, high-throughput sequencing techniques are increasingly common to describe and characterize natural microbial communities.

While DNA-based sequencing efforts (e.g. metagenomics) target relatively stable molecules and provide information about the genetic potential, the analysis of RNA sequences (e.g. metatranscriptomics) display the genetic activity, which is a step closer to the actual metabolic activity of microorganisms *in situ*. However, metatranscriptomic samples are much more subject to biases from collection and handling then metagenomic samples. Due to an often fast regulatory response of organisms towards a change in environmental conditions and the fast (and potentially also selective) degradation of RNA molecules after cell death, the RNA composition of an environmental sample can change within minutes.

In eukaryotic organisms, RNA in defrosted post-mortem tissues and cultures can degrade totally in a time frame from a few minutes to hours, even if the samples are stored on ice (Ibberson et al., 2009). Furthermore, the degradation of eukaryotic RNA can also occur at different paces, depending upon the tissue that was analyzed (Seear and Sweeney, 2008). Generally, intact tissues or cells are much less affected by RNA-degradation, because ribosomal RNA can stabilize messenger RNA (Deana and Belasco, 2005).

In bacterial cultures certain mRNAs can degrade within only 30 seconds (Belasco and Brawerman, 1993). In the cyanobacterium *Prochlorococcus*, the average half-life of mRNAs was only 2.4 minutes (Steglich et al., 2010), while in γ-proteobacterial *Escherichia coli* the half-life of total mRNA was determined to be 6.8 minutes (Selinger et al., 2003). Archaeal mRNAs have similar half-lifes. Two *Sulfolobus spp.* showed an average half-life of about 5 minutes (Andersson et al., 2006) and *Halobacterium salinarum* NRC-1 about 10 minutes (Hundt et al., 2007). A study targeting *Streptococcus pyogenes* has shown that the decay rate of certain mRNAs can also depend on the growth phase of the culture. Messenger RNA (mRNA) extracted from *S. pyrogenes* cultures in the stationary phase showed more than 20-fold greater stability than mRNA from cultures in the late exponential phase (Barnett et al., 2007).

Additionally, the composition of RNAs can also change due to the exposure of the samples to environmental conditions which differ significantly from the *in situ* conditions. A change in the *in situ* conditions (e.g. nutritional availability, oxygen (O_2), temperature, pH, light or pressure) can result in stress for the organisms and thus trigger the expression of transcripts, which encode for enzymes related to stress-response (e.g. heat shock proteins, chaperons or oxygen-detoxification transcripts). In contrast, transcripts encoding for enzymes that play an important regulatory role, can also decrease rapidly in abundance, since these transcripts are especially unstable (Gutierrez et al., 2002; Sharova et al., 2009).

Thus, the stability and composition of RNAs in tissues and single cells cannot be generally determined, but is dependent upon many different factors. Given the time frames of RNA degradation as mentioned above, sampling itself presents a certain stress to the microbial community, since key parameters like temperature, light and pressure may be difficult to control during sample collection or the filtration procedure. Special care has to be taken when working in low O_2 environments, since sample collection and filtration can introduce new O_2 into the samples and hence trigger a response of the microbial community. Due to an increasing interest in metatranscriptomic studies carried out in oxygen minimum zones (OMZ) (e.g. Stewart et al., 2011; Wright et al., 2012), it is crucial to determine how handling artifacts can influence the RNA composition of samples from these very special environments.

To assess the influence of the sampling time on the composition of the microbial community structure in oxygen-depleted subsurface waters from the Peruvian OMZ, we took replicate water samples and subsequently filtered and fixed the samples after 0, 20, 120 and 300 minutes. Massively parallel pyrosequencing was carried out on total RNA.

Results

Description of the sampling site

Samples for high-throughput sequencing (Roche GS-FLX technology) were collected on January 20th, 2009 (around noon) within the OMZ off Peru. The sampling site was located approximately 90 km offshore the city of Pisco (13° 44,993’ S, 077° 1,966’ W) at 75m water depth. The study site shows general features of a pronounced OMZ. Dissolved O_2 concentrations were high in surface waters and fell below the detection limit of our sensor at 37 m (Figure 1). O_2 remained undetectable down to 178 m, but from 179 m to 376 m (maximum deployment depths of the pump CTD), O_2 was sporadically measured at very low concentration.

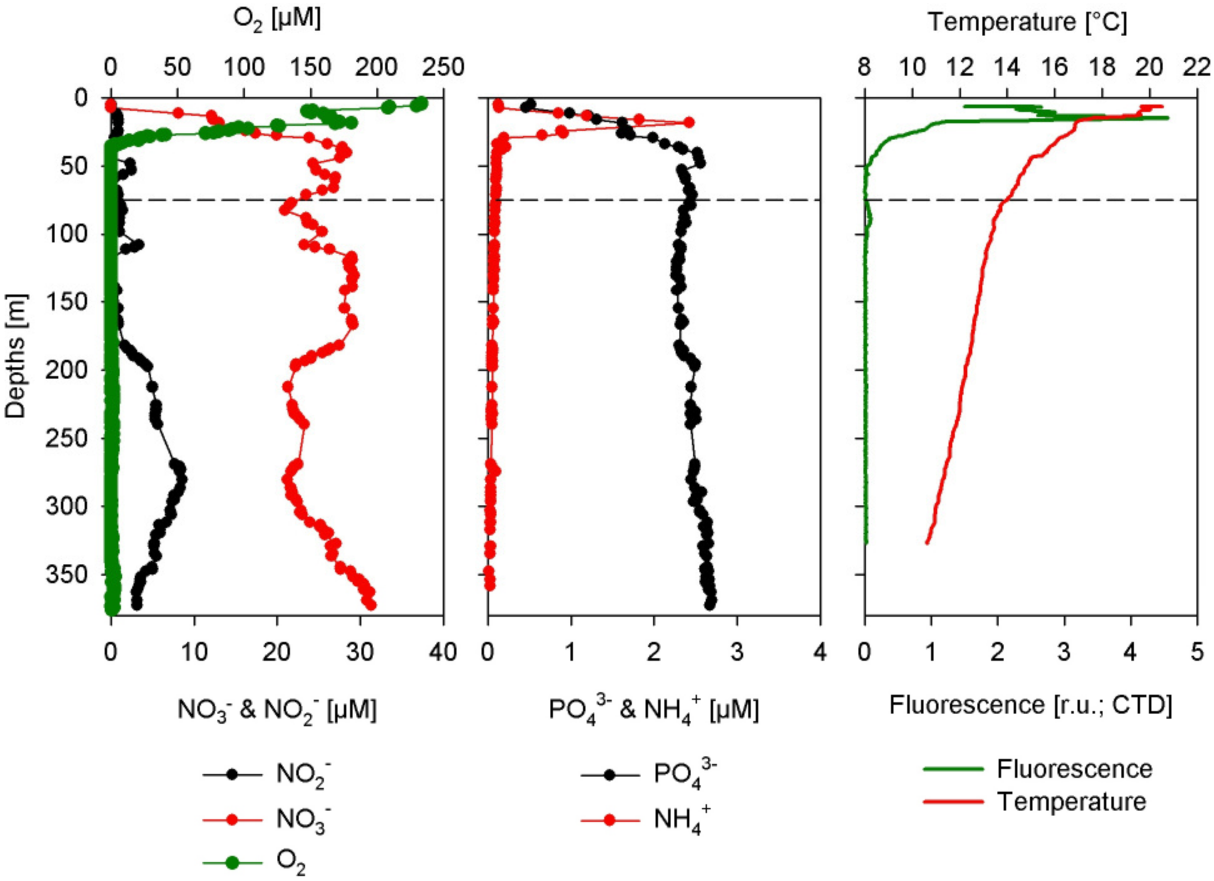


Figure 1: Vertical distribution of physical, chemical and biological water properties. Dashed black lines indicate sampling depths. (A) Concentration of NO_3^-, NO_2^- and O_2. (B) Concentrations of PO_4^{3-} and NH_4^+. (C) Temperature (°C) and Chlorophyll (relative units, measured with the pump CTD).

The concentrations of nitrate (NO_3^-) were inversely correlated with O_2. They were depleted in the surface and increased steadily towards 40 m (reaching 28.4 µM), being relatively stable (20-30 µM) throughout the rest of the water column. In contrast, nitrite (NO_2^-) concentrations were low in the surface (~1µM) and subsurface, but increased towards a deep maximum of 8.5 µM at 273 m. Ammonium (NH_4^+) started low in the surface and formed a pronounced maximum at 19 m (2.4 µM) before dropping to very low concentrations from 40 m downwards (0.03-0.1 µM). Phosphate (PO_4^{3-}) concentrations were lowest at 7 m (0.5 µM), steadily increasing towards 40 m (2.5 µM) and remained almost constant in all our following measurements (~2.5 µM).

Sequencing statistics

From all four samples (time points 0 min, 20 min, 120 min and 300 min) a total of 438,655,148 base pairs were sequenced, resulting in 1,146,516 sequences (Table 1). Each sample contained on average 68,427 unique sequences, when clustered at 98% identity. The vast majority of sequences were assigned a ribosomal origin (1,086,697; 94.8%) where as the remainder accounted for approximately the same numbers of mRNAs (26,730; 2.3%) and unknown sequences (33,089; 2.9%).

Table 1: Sequencing statistics of samples from four time points.

	0 minutes	20 minutes	120 minutes	300 minutes	Total
Basepairs	100,627,992	105,189,393	115,267,782	117,569,981	438,655,148
Raw reads	279,090	274,164	285,959	307,303	1,146,516
Obtained cluster[a]	56,046	66,808	79,898	70,957	68,427[d]
Identified rRNAs[b]	267,201	257,395	268,867	293,234	1,086,697
Non-ribosomal RNAs	11,889	16,769	17,092	14,069	59,819
Identified mRNAs[c]	5,815	6,627	7,201	7,087	26,730
Not identified RNAs	6,074	10,142	9,891	6,982	33,089

[a]as obtained with Cd-hit
[b]as obtained with BLAST-searches against the SILVA database
[c]as obtained with BLAST-searches against NCBI's non-redundant database
[d]average of the four time points

Taxonomic diversity of the microbial community

The microbial community at the sampling site was overall dominated by proteobacteria, which accounted for more than half of our sequences in each time point (Figure 2). Within the proteobacteria, α-, β-, γ- and δ-proteobacteria were abundant, while ε-proteobacteria were only a minor component (~2%). In addition, Bacteroidetes (~3%), Planctomycetes (~3%), Cyanobacteria (~4%) and Deferribacteres (~4%) were also detected in relatively high and stable proportions in all samples. Archaeal sequences were only found in low proportions (~1%), whereas eukaryotic sequences were more common (~6%). The pool of other taxonomic groups ('Unknown and others') was further contributing to considerable proportions (~15%), and stayed relatively stable over time. Although the overall structure of the microbial community was relatively uniform and stable, a major difference between the samples from the subsequent time points was a rapid decrease of β-proteobacterial sequences. In time point 0 min, β-proteobacteria formed the largest taxonomic unit and contributed to about 30% of the total microbial community. Within 20 minutes, the number of β-proteobacteria decreased to 5% (corresponding to a loss of 85% of β-proteobacterial sequences). After 120 minutes of incubations the number further decreased to 3% and after 300 minutes only 1% β-proteobacterial sequences could be detected. Thus, the incubation of 300 minutes reduced the relative abundance of the initially detected β-proteobacterial sequences by 97%. Similarly, the pool of β-proteobacterial mRNA sequences, although much smaller, shows a comparable decrease (26 to 3%, corresponding to a decrease to 12% of the initial population) in abundances over time (Figure S1).

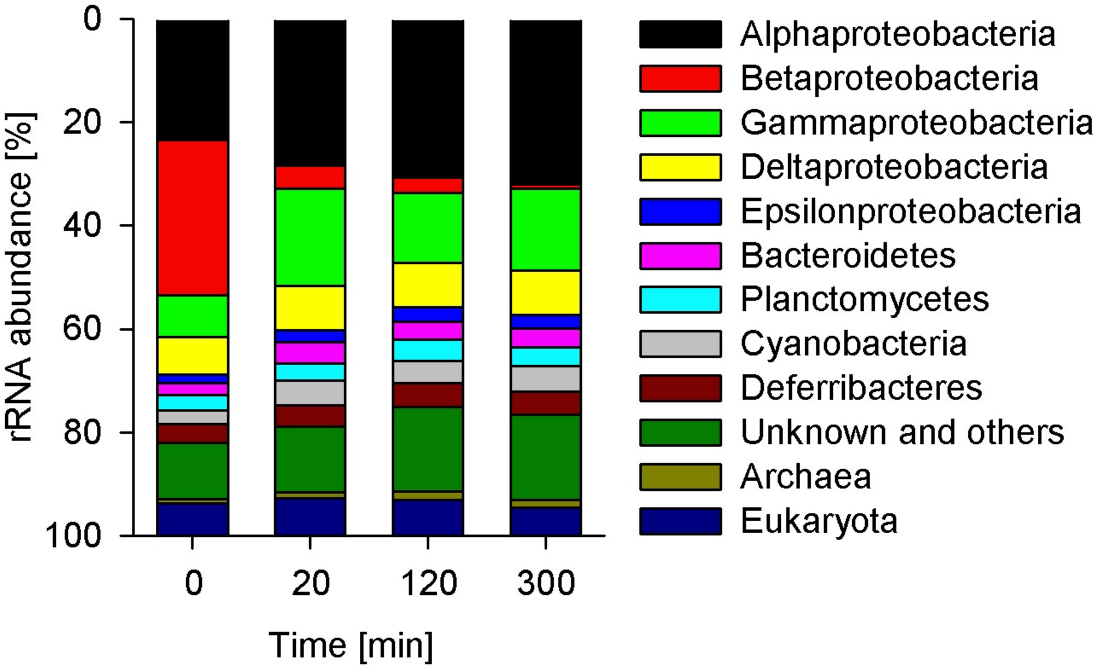


Figure 2: Phylogenetic composition over time. Shown at either domain, phylum or class level in percent of all ribosomal RNA sequences.

A breakdown of the taxonomic assignments to the highest possible level is depicted in Figure 3. The two most abundant taxa in time point 0 min were *Curvibacter spp.* and *Burkholderia spp.*, both belonging to β-proteobacteria. A strong decrease from almost 9% to less than 0.3% occurred within 300 minutes in both groups. Among the 15 most abundant phylogenetic groups, *Delftia spp.*, *Aquabacterium spp.* and *Variovorax spp.* were further identified as β-proteobacterial taxa. Although less abundant, the same trend of decreasing abundance with time was visible in these groups. Apart from β-proteobacteria, especially α-proteobacterial sequences were found in high abundances, too. Sequences similar to Candidatus *Pelagibacter spp.* (~4%), *Roseovarius spp.* (~6%), *Oceanicola spp.* (~2.5%), *Roseobacter* clade CHAB-I-5 lineage (~2%) and unclassified *Rhodobacteraceae* (~2.5%) were retrieved and showed an increase over the incubation time of 5 hours. In addition, not further identifiable ε-proteobacteria (~2%) were found, as well as several unclassified *Deferribacterales* (~3.5) and *Planctomycetes* (~2.5%). Also sequences similar to γ-proteobacterial *Oceanospirillales* ZD0405 (~1.5%) and δ-proteobacterial *Bdellovibrionaceae* OM27 clade (~2%) were among the 15 most abundant taxa in the samples. All taxa, other than β-proteobacteria were either relatively stable over time or increased slightly.

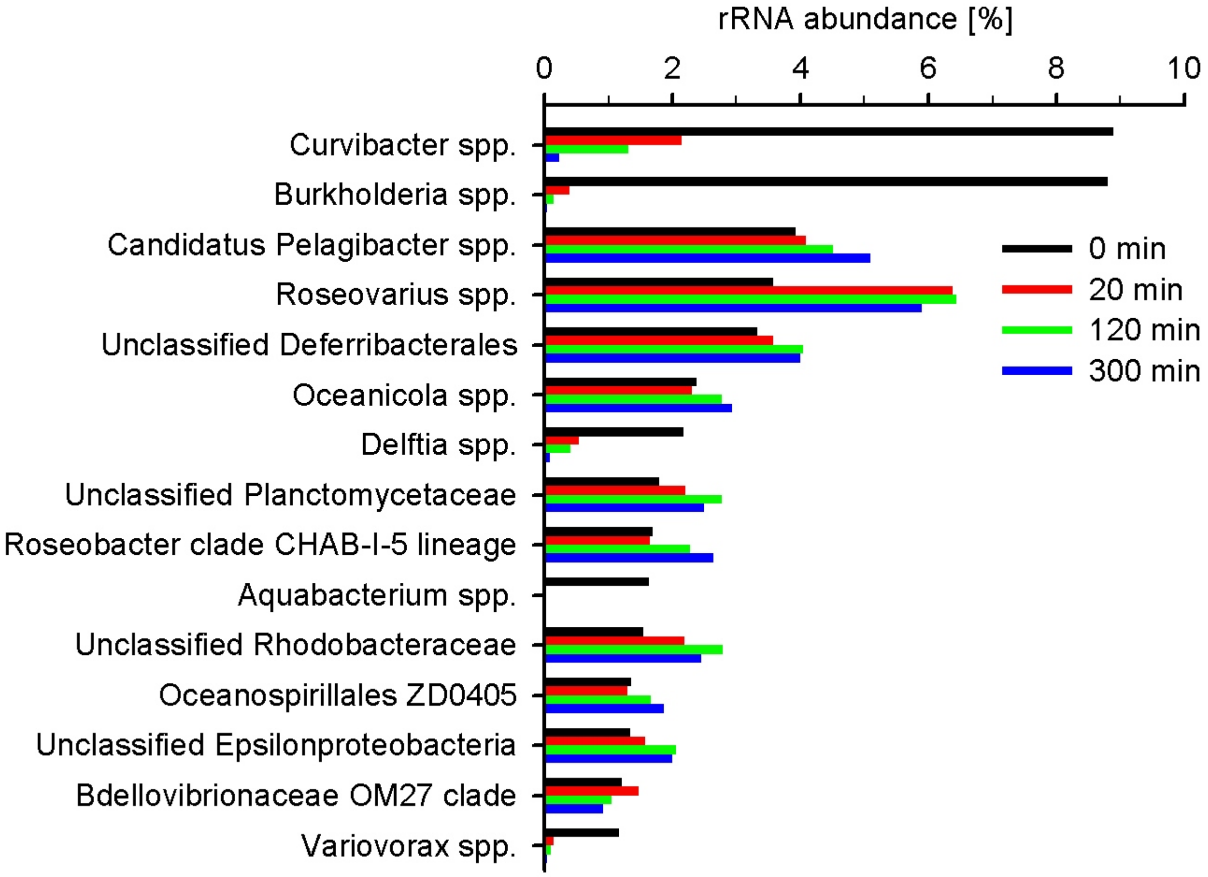


Figure 3: Highest possible taxonomic assignments of rRNA-sequences over time. Shown are the 15 most abundant organisms in percent of all ribosomal RNA sequences, ordered descending according to the counts in the sample 0 min.

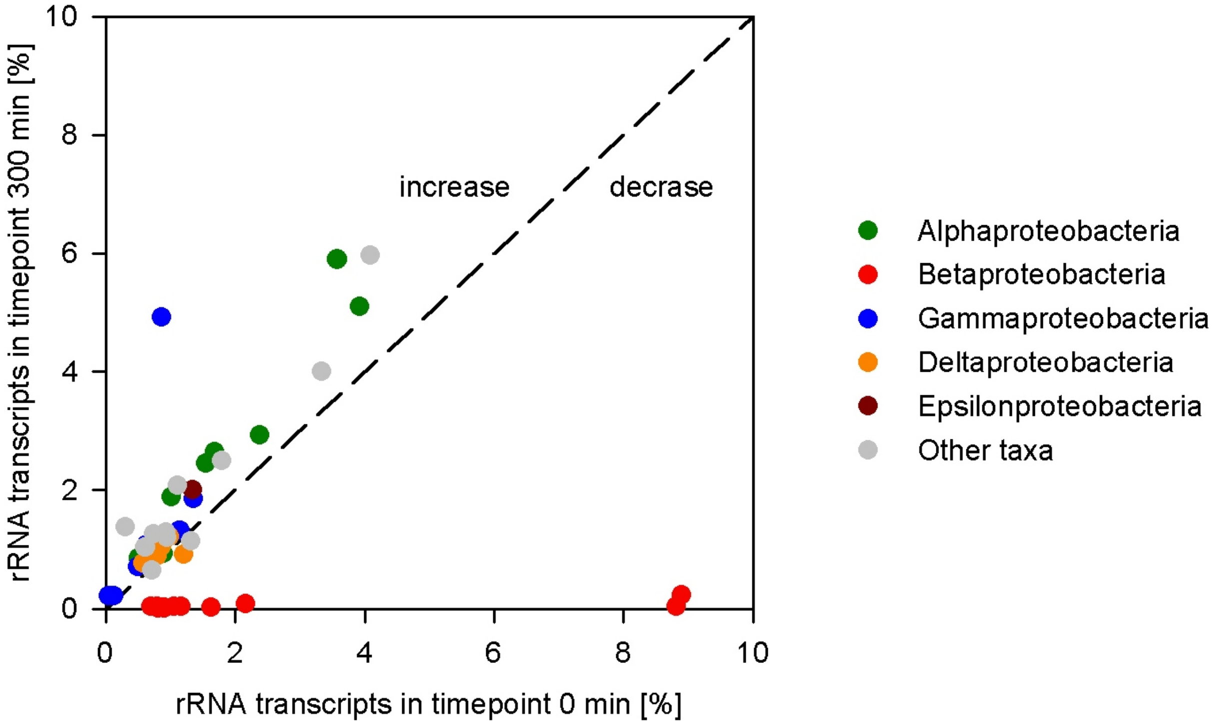


Figure 4: Relative changes of the 45 most abundant taxa over time. Black dashed reference line is indicated (no change over time).

To further investigate the community structure over time, a compilation of the 45 most abundant taxonomic groups (including those from Figure 3) is depicted in Figure 4. Of all presented groups, solely β-proteobacteria showed a decrease over the incubation time. All other taxonomic groups plotted in Figure 4 cluster around or slightly above the reference line, which indicates no or only minor increase in abundance. Notably, the relative abundance of a taxon in the initial sample (time point 0 min) had no influence on the detected changes over time. Thus, the two most abundant taxa, *Curvibacter spp*. and *Burkholderia spp*., showed an equal decrease when compared to the less abundant β-proteobacterial taxa. Similarly, for the groups that stayed constant or slightly increased over time, no correlation between the relative abundance in the first (0 min) and the last time point (300 min) was visible.

Intriguingly, although *Curvibacter spp*. and *Burkholderia spp*. together account for about 18% of all rRNA sequences in time point 0 min, an identification of the ribosomal sequences beyond the level of genus was not possible with BLAST-searches. The same holds true for many other β-proteobacterial taxa, which could not be affiliated to a distinct species or strain. For a better identification and phylogenetic classification a neighbor-joining tree was constructed with assembled ribosomal contigs from time point 0 min (Figure 5).

Some of our ribosomal contigs clustered with sequences of the genus *Ralstonia* and *Aquabacterium*, belonging to either *Burkholderiales* or *Oxalobacteraceae*, respectively. Some other contigs directly clustered to *Burkholderiales* and *Oxalobacteraceae*-groups in the tree, without any further higher taxonomic information. The remainder of the contigs, representing the majority of assembled sequences, had greatest similarity to diverse, not further identifiably groups of unknown β-proteobacteria. Three of these groups (depicted in dark blue) formed monophyletic branches in the tree, with no known β-proteobacterial ribosomal sequence from the SILVA-database clustering along with them.

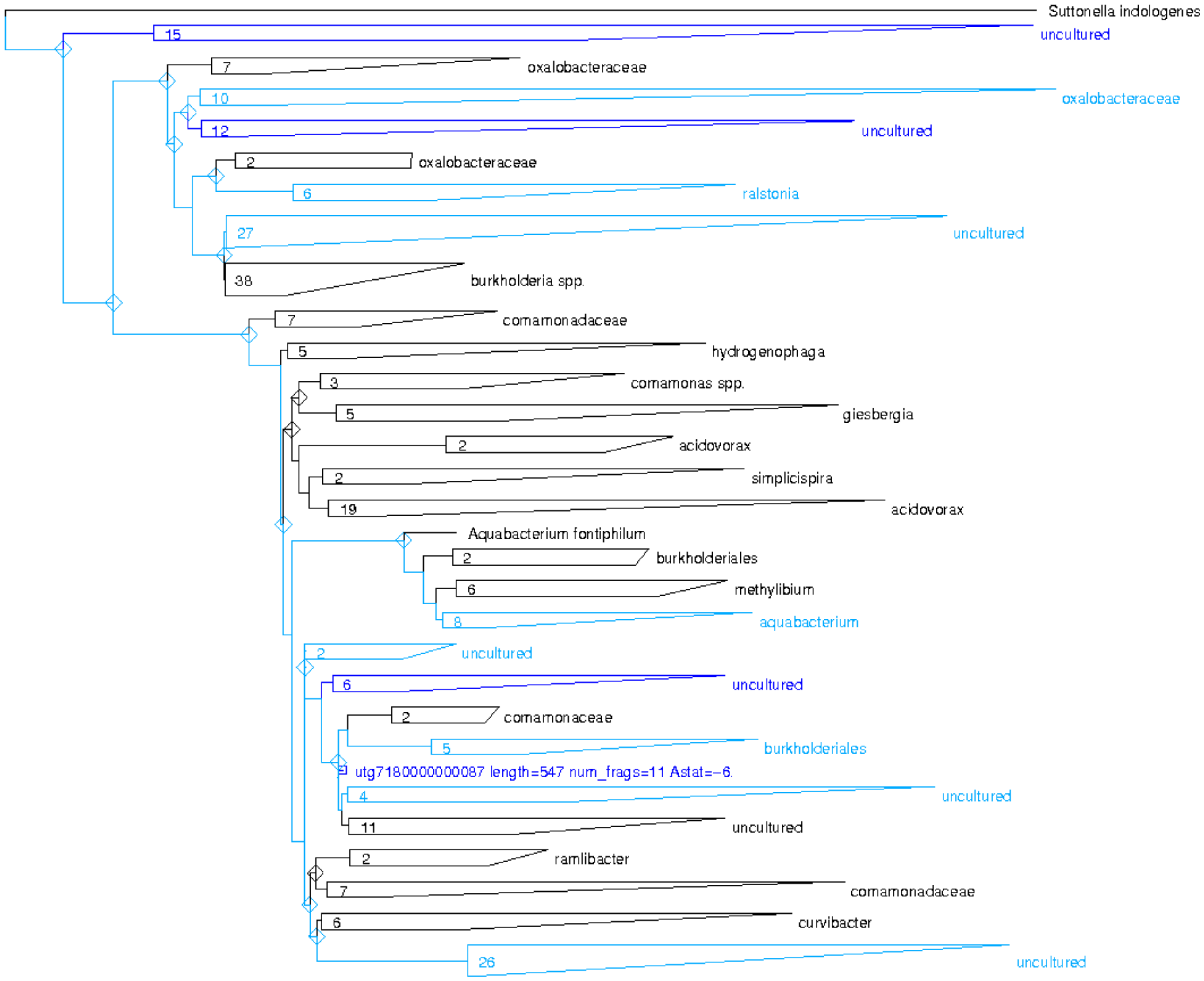


Figure 5: Phylogenetic tree of assembled β-proteobacterial 16S rRNA contigs at time point 0 min. Neighbor-joining tree constructed with ARB, including the closest known relatives of the β-proteobacterial sequences found in this study and a γ-proteobacterial outgroup (*Suttonella indologenes*). Dark blue groups contain only sequences found in this study, black groups only consist of published sequences and light blue groups contain both, sequences found in this study and published sequences. The number of unique contigs in each group is indicated on the left hand side of each branch.

Functional activity of the microbial community

To asses the functional diversity of the microbial community, we used two approaches to investigate our sequence data. Next to the BLAST-searches, we scanned our sequences with profile hidden Markov models of the ModEnzA Enzyme Commission (EC) groups (Desai et al., 2011) and of the Pfam protein families (Finn et al., 2010). A total of 26,730 mRNA sequences could be identified and, if an EC number was assigned, the sequences were mapped onto KEGG metabolic pathways. A compilation of the ten most active pathways over time is shown in Figure 6. In contrast to the phylogenetic assignments of the transcripts, the activity of these KEGG metabolic pathways did not significantly change over the course of 300 minutes. The most active pathways, the purine and pyrimidine metabolisms, both involved in nucleic acid metabolisms, indicated general metabolic activity. This is also reflected by the presence of RNA polymerase (EC 2.7.7.6) and polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) transcripts, which are among the most abundant identified EC numbers (Figure 7).

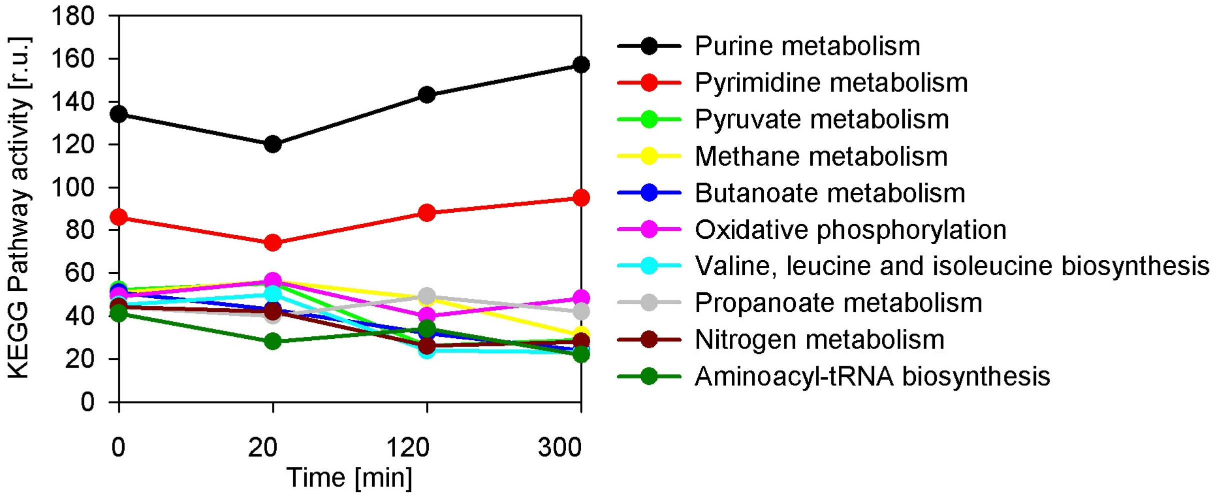


Figure 6: Ten most active (in time point 0 min) KEGG metabolic pathways. Sequences were identified by profile hidden Markov model scans of the ModEnzA EC and Pfam groups and mapped onto the KEGG metabolic pathways. The y-axis depicts the pathway activity in relative units for all four time points.

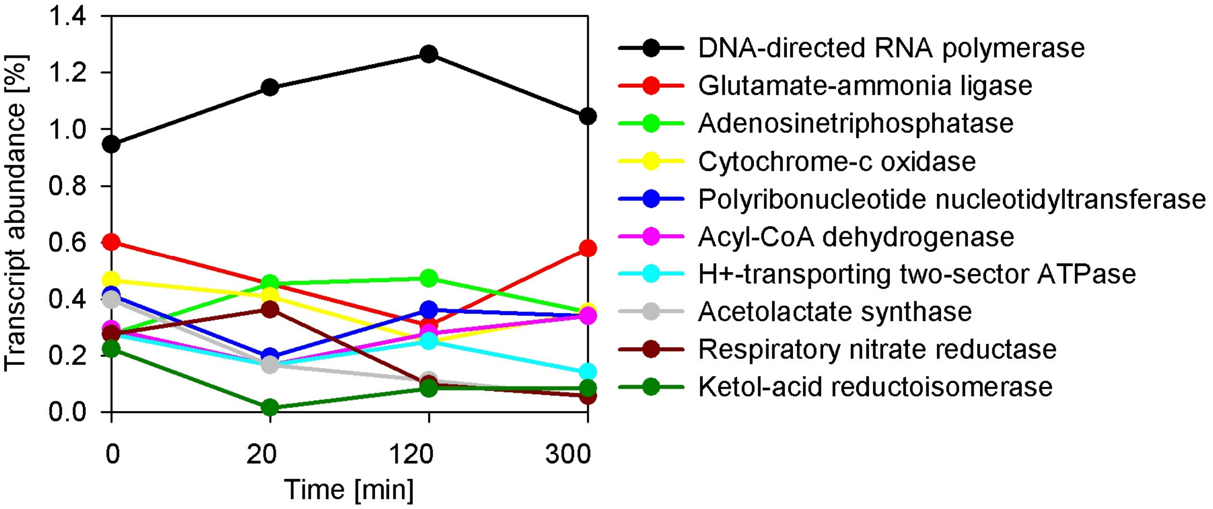


Figure 7: Ten most abundant EC numbers. Identified by profile hidden Markov model scans of the ModEnzA EC and Pfam groups. Show in percent of all protein-coding sequences for all four time points.

More specific KEGG pathways, like the pyruvate, methane, butanoate or nitrogen metabolism were identified among the most active pathways as well, but showed no significant change in activity over time, too. Even transcripts for enzymes which transfer electrons onto terminal electron acceptors during respiration and energy conservation, like the cytochrome c oxidase

(O_2, EC 1.9.3.1) or the respiratory nitrate reductase (NO_3^-, EC 1.7.99.4) only decrease slightly over time.

The relative proportion of transcripts related to the response to environmental stress is depicted in Figure 8. They show a clear increase over the 300 minutes of incubation. In time point 0 min, only about 1% of the non-ribosomal RNA transcripts were related to stress-response. At the end of the experiment, this number increased to over 4%. Transcripts belonging to this grouping include enzymes for general stress-response, for repair and detoxification mechanisms, heat-shock proteins and chaperones. The analysis of mRNA sequences also allowed the detection of sequences affiliated to viruses and phages. The number stayed relatively constant over time, ranging around 1%.

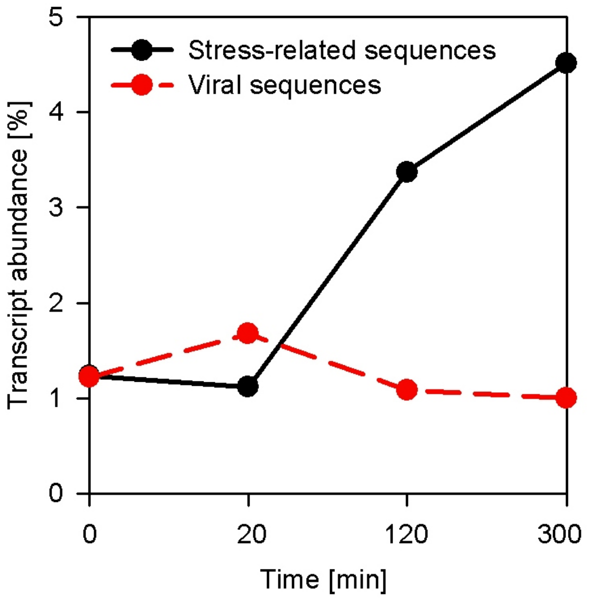


Figure 8: Relative abundance of selected transcripts. Shown in percent of all protein-coding sequences over the four time points. Stress-related sequences include transcripts encoding enzymes for general stress-response, repair and detoxification mechanisms, heat-shock proteins and chaperones.

Functional activity of selected microorganisms

A compilation of selected mRNAs of prominent microorganism as identified with BLAST-searches is depicted in Table S1. *Candidatus* Kuenenia stuttgartiensis-like transcripts encoding nitrate reductases *narG* (0.2% of all protein-coding sequences) and *narH* (0.06%) were among the most abundant in our samples. We further retrieved transcripts encoding a nitrous-oxide reductase (0.07%) from a *Rhodobacterales* bacterium HTCC2654-like organism, as well as a hydroxylamine oxidoreductase (0.05%) expressed by an organism related to *planctomycete* KSU-1. Furthermore, we found several ammonium transporter-transcripts in high abundances (see below). Those were assigned to *Candidatus* Kuenenia stuttgartiensis (0.1%), the uncultured SUP05 cluster bacterium (0.07%), alpha proteobacterium HIMB114 (0.06%) and *Candidatus* Pelagibacter ubique HTCC1002 (0.05%).

Identified β-proteobacterial mRNAs are shown in Table S2. Large proportions of the mRNA sequences were assigned as 'hypothetical proteins' (Table S2). However, it has been speculated that β-proteobacteria from the Peruvian OMZ are potentially linked to NH_4^+ oxidation and thus play an important role in the marine nitrogen cycle (Molina et al., 2007). This has also been found for the pelagic redoxcline of the central Baltic Sea. Solely β-proteobacterial sequences related to *Nitrosomonas* and *Nitrosospira* were identified to encode for ammonia monooxygenase, the key enzyme in aerobic NH_4^+ oxidation in bacteria (Bauer, 2003). However, although ammonium transporters were expressed by a wide range of taxa, we found only very minor fractions of transcripts encoding for ammonia monooxygenase (related to uncultured marine *crenarchaeote* HF4000_ANIW97M7) and none assigned β-proteobacteria. Thus, the functional role of β-proteobacteria at our study site remains unclear.

Discussion

Phylogenetic diversity

The aim of this study was the investigation of changes in the microbial diversity over several, prolonged sampling times, on samples collected within oxygen-depleted waters off the Peruvian coast. Generally, organisms commonly detected in those environments also contributed significantly to the microbial community found in this study. We identified diverse Candidatus *Pelagibacter spp.* and sequences related to the γ-proteobacterial order *Oceanospirillales* (Lavik et al., 2008; Stevens and Ulloa, 2008; Canfield et al., 2010; Stewart et al., 2011). Diverse unclassified *Planctomycetaceae* and cyanobacterial *Prochlorococcus spp.* were also found and are common inhabitants of OMZs (Woebken et al., 2008; Galan et al., 2009; Canfield et al., 2010; Stewart et al., 2011). In contrast, we identified only smaller numbers of archaeal sequences, which were found to be more abundant in other studies (Canfield et al., 2010; Molina et al., 2010; Belmar et al., 2011; Stewart et al., 2011).

Despite these organisms characteristic for OMZs, the detection of β-proteobacterial taxa has been documented only sporadically and in much lower abundances in marine habitats than in our study (Glockner et al., 1999). Nevertheless, β-proteobacterial sequences were obtained from the OMZs off Peru and Chile in minor proportions (Ward et al., 1989; Molina et al., 2007; Lam et al., 2009; Stewart et al., 2011) and from marine sediments (Nold et al., 2000), although it was shown that β-proteobacteria are more commonly found in oxygen-replete and freshwater environments (Nold and Zwart, 1998; Glockner et al., 1999). A contribution of 10% β-proteobacterial of the microbial community from metagenomic rRNA gene sequences were detected in coastal ocean waters (Mou et al., 2008). The authors assigned most sequences to the order *Burkholderiales*, an order which was common in this study as well.

Selective decay of β-proteobacterial sequences

Large changes in the community structure over time were detected for β-proteobacterial taxa and actually mostly restricted to them. All detected β-proteobacterial taxa showed a rapid decrease in their relative abundance. The other taxa were mostly stable or slightly increasing in abundance. The increase, however, reflects an artefact resulting from the decrease of β-proteobacterial sequences, which represented the largest pool of sequences in time point 0 min. The slope of the decrease of β-proteobacterial sequences, of both rRNA and mRNA sequences is nevertheless somewhat surprising. A decay to 15% of the initial population for rRNAs and 21% for mRNAs within only 20 minutes can be considered relatively fast. Taking the abundances for β-proteobacterial rRNAs in the first and the last time point (0 and 300 min), the decrease yielded a total half-life of about 51 minutes. Since the decrease during the first few minutes (between the first and second time point) was only 7 minutes, the idea of an exponential decay with a fixed half-life has to be abandoned.

The reason for the decrease of β-proteobacterial sequences cannot be fully assessed with this dataset, but a sequencing artefact is unlikely, since replicate pyrosequencing runs of the same sample deliver highly similar results (Kauserud et al., 2011). Thus, the selective decrease can be attributed to changes within the samples over time, e.g. to either the lyses of cells or to the degradation of RNA within intact cells. The lyses of β-proteobacterial cells could possibly be caused by phages, which target specific bacterial groups selectively. Although the methodology applied in this study cannot detect free virus particles in the water (due to the use of 0.2µm filters), viral sequences within bacterial cells or attached to them were identified (1%). However, the number stayed relatively constant over time. It has indeed been shown that a high level of stress results in prophage induction (Raynaud et al., 2005; Redon et al., 2005), potentially leading to the lyses of cells. The samples from longer incubations were most likely exposed to environmental stress, which is visible in the increase of transcripts related to stress-response over time, reaching a maximum of more than 2% after 300 minutes.

These genes, e.g. heat-shock proteins and chaperones were actually expressed by the entire microbial community. Chaperones are involved in the general stress response and were highly expressed, as previously observed within microbial communities in milk (Raynaud et al., 2005; Cretenet et al., 2011). Potentially, the β-proteobacteria detected in this study may have been obligate anaerobes, and thus were more sensitive to the introduction of O_2, when compared to aerobic and facultative anaerobic organisms. Depending on the time elapsed during sample collection, strictly anaerobic organisms might die and lyse and thus elude from analysis.

Due to the lack of representative genomes of marine β-proteobacteria, their metabolic strategy cannot be fully assessed. However, β-proteobacterial species, e.g. belonging to *Burkholderiaceae* or *Comamonadaceae* were shown to be extremely versatile in metabolism. Some species can exhibit anaerobic respiration with NO_3^- (denitrification), while some others belonging to *Rhodocyclaceae* can perform anoxygenic photosynthesis (Balows et al., 1992; Brenner et al., 2005). Although many β-proteobacterial species were so far predominantly found in soils, in freshwater and also in clinical samples, other β-proteobacterial like *Polaromonas spp*. (belonging to *Comamonadaceae*) have indeed been detected in marine habitats (Balows et al., 1992; Brenner et al., 2005).

Future sampling campaigns

The minimization of both the decay and the sampling-induced artificial expression of RNAs is commonly addressed by a rapid sample collection and immediate freezing or preservation of the obtained samples. Given the great changes in the microbial community structure we detected within a relatively short time, the handling of RNA samples might be even more important than previously thought. Changes of the magnitude we observed indeed suggest that the sampling time can potentially have a greater impact on the community structure than differences accounted to different environmental parameters (and a different community structure per se). If the sampling time as a parameter is not taken into account, the comparability of samples from different campaigns might be severely biased. For future campaigns, a clear standardization of sample collection, handling and processing is desirable to capture a realistic view of microbial community structure and function.

Conclusions

Our data emphasizes the importance of fast sample processing for all transcriptomic studies to reduce potential biases in the structure and activity of microbial communities. A prolonged sampling time can reduce the identifiable microbial diversity selectively. We have shown that β-proteobacteria decrease in the course of 300 min to 3% of the initial population, whereas only minor changes were observed in most other bacterial and archaeal classes. Most detected β-proteobacteria belonged to either *Comamonadaceae* or *Burkholderiaceae*, taxa which have so far been only rarely attributed to oxygen-depleted marine environments. Along with the change in community structure, we observed an increase in expression of heat shock proteins, chaperones and other stress-response genes, potentially causing prophage induction leading to the lyses of the cells.

Material and Methods

Sample Collection Water samples were collected on January 20th, 2009 at 12:40 a.m. in the Peruvian OMZ in the course of the RV Meteor cruise M77-3. The sampling site (13° 44,993' S, 077° 1,966' W) was located approximately 90 km off the coast (off the city of Pisco) at 75 m depth. The water samples for RNA extraction were pumped from the depth directly on board using a pump-conductivity-temperature-depth water sampler and were filled in 4.5 liter polycarbonate bottles. One of the samples was filtered straight away (time point 0 minutes) using a vacuum pump (Sartorius eJet), the others were incubated at ambient temperatures in the dark for 20, 120, and 300 minutes, respectively and then filtered. For each sample 2 liters of water were prefiltered through 10 µm pore size filters (Millipore/Durapore Membrane filters) and then collected upon 0.2 µm pore size filters (Millipore/Durapore Membrane filters). From the time point the water was pumped on board, or from the end of the incubation, less than 18 minutes elapsed until the filters were put in micro centrifuge reaction tubes and flesh frozen in liquid nitrogen.

RNA-extraction and cDNA-synthesis Total RNA was extracted using the DNA/RNA-Allprep kit (Qiagen) with modifications in the protocol for the lyses step: The frozen filters were crushed with a disposable pestle and incubated with 200 µl lysozym (10µg/µl) and EDTA (1mM) at ambient for 5 minutes. Then, 40 µl of Proteinase K (10µg/µl) was added followed by incubation of 5 minutes at ambient temperatures. After adding 500 µl buffer RLT-Plus (containing 10µl/ml betamercaptoethanol) the manufacturer's instructions were followed. The total RNA was eluted in 50 µl nuclease-free water, followed by a subsequent step of DNA digestion with the Turbo DNA-free kit (Ambion). The DNA-free total RNA was subsequently quantified using nano-litre spectrophotometry (NanoDrop) and checked for degradation with the BioRad Experion (RNA Standard Sense). Highly distinct ribosomal RNA peaks were assumed to resemble intact RNA. cDNA was then synthesized with the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) using random hexameric primers (Qiagen). Leftover reactants and reagents were removed using the PCR Mini Elute Kit (Qiagen), which was followed by another measurement with NanoDrop and Experion to evaluate the quality, quantity and the length of the cDNA. The cDNA was immediately stored at -80°C until pyrosequencing. Throughout the whole procedure nuclease-free plastic consumables and nuclease-free water and reagents were used to hinder any possible degradation or contamination of RNA or cDNA.

Sequencing and data analysis 50 µl of the cDNA-samples (15-20 ng/µl) were sequenced with a GS-FLX pyrosequencer (Roche), each sample was loaded on one quarter of a PicoTiter plate. This resulted in over 450 million base pairs of sequence information, accounting for approximately 1.2 million raw reads. The raw reads were clustered using Cd-hit (Li and Godzik, 2006) with a sequence identity threshold of 98% and word length of 8, delivering about 250.000 cluster representative sequences in total. The ribosomal RNA sequences in these cluster representatives were identified by a BLASTn search (Altschul et al., 1990) against the SILVA database including both prokaryotic and eukaryotic small and large subunit sequences with a bit score cutoff of 86 as described earlier (Urich et al., 2008). This cut-off was also used in MEGAN (Huson et al., 2007) to make taxonomic assignments for the sequences. The remaining unassigned cluster representative sequences (non rRNA-sequences) were translated in all six open reading frames and analyzed against the non-redundant database from NCBI using BLASTx with a bit score cutoff of 35. For the functional assignment of the cluster representatives the top hit of each BLAST-search was used. About 50% of the non-rRNA sequences could be identified as mRNAs, leaving another 50% of sequences as unassigned or novel mRNAs or putative non-coding small RNAs. The non rRNA sequences were also scanned with profile hidden Markov models of the ModEnzA Enzyme Commission (EC) groups (Desai et al., 2011) and of the Pfam protein families (Finn et al., 2010) to asign functions. The clustering information from Cd-hit, BLAST, MEGAN, EC and Pfam assignments for the cluster representative sequences were added to a MySql database for analysis. The Pfam hits were converted to EC numbers and along with the ModEnzA hits, mapped to the KEGG reference pathways and to the EC activity matrix using an in-house java-based pathway mapping and visualization tool. The Assembly of 80390 β-proteobacterial ribosomal sequences identified as ribosomal RNA of β-proteobacterial origin was carried out with the Celera Assembler using parameters to build contigs with 12% error overlaps, and allow 14% error globally. This resulted in 200 contigs covering 27225 reads and 53165 singletons. All contigs were imported to ARB and aligned to the non-redundant silva-nr-rrna-database release 102. 91 of the contigs could be aligned in ARB. These contigs were then used along with 159 sequences representing the closest relatives of the contigs and a γ-proteobacterial outgroup (*Suttonella indologenes*) in the ARB database, to construct a neighbor joining tree. All the contigs were also BLASTed against the Silva database to get the taxonomic distribution of the β-proteobacteria in the sample. The half-life of β-proteobacterial rRNA sequences was calculated based on the fit of an exponential curve between the first and the last time point. The estimated half-life was about 51 minutes. Although the half-life between the first and the second time point was much shorter (7 minutes), we decided to use the longest incubation time (500 min) for the calculation.

References

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.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic Identification and in-Situ Detection of Individual Microbial-Cells without Cultivation. *Microbiol Rev* **59**: 143-169.

Andersson, A.F., Lundgren, M., Eriksson, S., Rosenlund, M., Bernander, R., and Nilsson, P. (2006) Global analysis of mRNA stability in the archaeon Sulfolobus. *Genome Biology* **7**(10): R99. doi: 10.1186/gb-2006-7-10-r99.

Balows, A., Truper, H.G., Dworkin, M., Harder, W., and Schleifer, K.H. (1992) The Prokaryotes: A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications - Volume III. Second Edition, Springer, New York, USA.

Barnett, T.C., Bugrysheva, J.V., and Scott, J.R. (2007) Role of mRNA stability in growth phase regulation of gene expression in the group A streptococcus. *J Bacteriol* **189**: 1866-1873.

Bauer, S. (2003) Structure and function of nitrifying bacterial communities in the eastern Gotland basin. University of Rostock Press, Rostock, Germany.

Belasco, J.G., and Brawerman, G. (1993) Control of messenger RNA stability. Academic Press. San Diego, USA.

Belmar, L., Molina, V., and Ulloa, O. (2011) Abundance and phylogenetic identity of archaeoplankton in the permanent oxygen minimum zone of the eastern tropical South Pacific. *FEMS Microbiol Ecol* **78**: 314-326.

Brenner, D., J., Garrity, G.M., and Bergey, D., H. (2005) Bergey's Manual of Systematic Bacteriology - Volume Two - The Proteobacteria - Part C. Second Edition, Springer, New York, USA.

Brettar, I., and Rheinheimer, G. (1991) Denitrification in the Central Baltic - Evidence for H2s-Oxidation as Motor of Denitrification at the Oxic-Anoxic Interface. *Mar Ecol-Prog Ser* **77**: 157-169.

Canfield, D.E., Stewart, F.J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E.F. et al. (2010) A Cryptic Sulfur Cycle in Oxygen-Minimum-Zone Waters off the Chilean Coast. *Science* **330**: 1375-1378.

Codispoti, L.A., Brandes, J.A., Christensen, J.P., Devol, A.H., Naqvi, S.W.A., Paerl, H.W., and Yoshinari, T. (2001) The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Sci Mar* **65**: 85-105.

Cretenet, M., Laroute, V., Ulve, V., Jeanson, S., Nouaille, S., Even, S. et al. (2011) Dynamic Analysis of the Lactococcus lactis Transcriptome in Cheeses Made from Milk Concentrated by Ultrafiltration Reveals Multiple Strategies of Adaptation to Stresses. *Appl Environ Microbiol* **77**: 247-257.

Deana, A., and Belasco, J.G. (2005) Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Genes Dev* **19**: 2526-2533.

Desai, D.K., Nandi, S., Srivastava, P.K., and Lynn, A.M. (2011) ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities. *Adv Bioinformatics* **2011**: 743782.

Falkowski, P.G., Barber, R.T., and Smetacek, V.V. (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**: 200-207.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E. et al. (2010) The Pfam protein families database. *Nucleic Acids Res* **38**: D211-222.

Galan, A., Molina, V., Thamdrup, B., Woebken, D., Lavik, G., Kuypers, M.M.M., and Ulloa, O. (2009) Anammox bacteria and the anaerobic oxidation of ammonium in the oxygen minimum zone off northern Chile. *Deep-Sea Res Pt II* **56**: 1125-1135.

Glockner, F.O., and Joint, I. (2010) Marine microbial genomics in Europe: current status and perspectives. *Microb Biotechnol* **3**: 523-530.

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

Gutierrez, R.A., Ewing, R.M., Cherry, J.M., and Green, P.J. (2002) Identification of unstable transcripts in Arabidopsis by cDNA microarray analysis: Rapid decay is associated with a group of touch- and specific clock-controlled genes. *Proc Natl Acad Sci U S A* **99**: 11513-11518.

Hundt, S., Zaigler, A., Lange, C., Soppa, J., and Klug, G. (2007) Global analysis of mRNA decay in Halobacterium salinarum NRC-1 at single-gene resolution using DNA Microarrays. *J Bacteriol* **189**: 6936-6944.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377-386.

Ibberson, D., Benes, V., Muckenthaler, M.U., and Castoldi, M. (2009) RNA degradation compromises the reliability of microRNA expression profiling. *Bmc Biotechnol* **9**: 102. doi: 10.1186/1472-6750-9-102.

Jorgensen, B.B., Fossing, H., Wirsen, C.O., and Jannasch, H.W. (1991) Sulfide Oxidation in the Anoxic Black-Sea Chemocline. *Deep-Sea Res* **38**: S1083-S1103.

Kauserud, H., Kumar, S., Brysting, A.K., Norden, J., and Carlsen, T. (2011) High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal plant root samples. *Mycorrhiza* **22**: 309-315.

Lam, P., Lavik, G., Jensen, M.M., van de Vossenberg, J., Schmid, M., Woebken, D. et al. (2009) Revising the nitrogen cycle in the Peruvian oxygen minimum zone. *Proc Natl Acad Sci U S A* **106**: 4752-4757.

Lavik, G., Stuhrmann, T., Bruchert, V., Van der Plas, A., Mohrholz, V., Lam, P. et al. (2008) Detoxification of sulphidic African shelf waters by blooming chemolithotrophs. *Nature* **457**: 581-584.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

Molina, V., Belmar, L., and Ulloa, O. (2010) High diversity of ammonia-oxidizing archaea in permanent and seasonal oxygen-deficient waters of the eastern South Pacific. *Environ Microbiol* **12**: 2450-2465.

Molina, V., Ulloa, O., Farias, L., Urrutia, H., Ramirez, S., Junier, P., and Witzel, K.P. (2007) Ammonia-oxidizing beta-Proteobacteria from the oxygen minimum zone off northern Chile. *Appl Environ Microbiol* **73**: 3547-3555.

Mou, X.Z., Sun, S.L., Edwards, R.A., Hodson, R.E., and Moran, M.A. (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**: 708-711.

Nold, S.C., and Zwart, G. (1998) Patterns and governing forces in aquatic microbial communities. *Aquat Ecol* **32**: 17-35.

Nold, S.C., Zhou, J.Z., Devol, A.H., and Tiedje, J.M. (2000) Pacific northwest marine sediments contain ammonia-oxidizing bacteria in the beta subdivision of the Proteobacteria. *Appl Environ Microbiol* **66**: 4532-4535.

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

Pagani, I., Liolios, K., Jansson, J., Chen, I.M.A., Smirnova, T., Nosrat, B. et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**: D571-D579.

Poretsky, R.S., Hewson, I., Sun, S.L., Allen, A.E., Zehr, J.P., and Moran, M.A. (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11**: 1358-1375.

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

Raynaud, S., Perrin, R., Cocaign-Bousquet, M., and Loubiere, P. (2005) Metabolic and transcriptomic adaptation of Lactococcus lactis subsp lactis biovar diacetylactis in response to autoacidification and temperature downshift in skim milk. *Appl Environ Microbiol* **71**: 8016-8023.

Redon, E., Loubiere, P., and Cocaign-Bousquet, M. (2005) Transcriptome analysis of the progressive adaptation of Lactococcus lactis to carbon starvation. *J Bacteriol* **187**: 3589-3592.

Seear, P.J., and Sweeney, G.E. (2008) Stability of RNA isolated from post-mortem tissues of Atlantic salmon (Salmo salar L.). *Fish Physiol Biochem* **34**: 19-24.

Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M., and Rosenow, C. (2003) Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res* **13**: 216-223.

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M.S.H. (2009) Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Res* **16**: 45-58.

Steglich, C., Lindell, D., Futschik, M., Rector, T., Steen, R., and Chisholm, S.W. (2010) Short RNA half-lives in the slow-growing marine cyanobacterium Prochlorococcus. *Genome Biol* **11**: R54. doi: 10.1186/gb-2010-11-5-r54.

Stevens, H., and Ulloa, O. (2008) Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environ Microbiol* **10**: 1244-1259.

Stewart, F.J., Ulloa, O., and Delong, E.F. (2011) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* **14**: 23-40.

Streit, W.R., and Schmitz, R.A. (2004) Metagenomics - the key to the uncultured microbes. *Curr Opin Microbiol* **7**: 492-498.

Thamdrup, B., Canfield, D.E., Ferdelman, T.G., Glud, R.N., and Gundersen, J.K. (1996) A biogeochemical survey of the anoxic basin Golfo Dulce, Costa Rica. *Rev Biol Trop* **44**: 19-33.

Urich, T., Lanzen, A., Qi, J., Huson, D.H., Schleper, C., and Schuster, S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**(6): e2527. doi: 10.1371/journal.pone.0002527.

Ward, B.B., Glover, H.E., and Lipschultz, F. (1989) Chemoautotrophic Activity and Nitrification in the Oxygen Minimum Zone Off Peru. *Deep-Sea Res* **36**: 1031-1051.

Woebken, D., Lam, P., Kuypers, M.M.M., Naqvi, S.W.A., Kartal, B., Strous, M. et al. (2008) A microdiversity study of anammox bacteria reveals a novel Candidatus Scalindua phylotype in marine oxygen minimum zones. *Environ Microbiol* **10**: 3106-3119.

Wright, J.J., Konwar, K.M., and Hallam, S.J. (2012) Microbial ecology of expanding oxygen minimum zones. *Nat Rev Microbiol* **10**: 381-394.

Youssef, N., Sheik, C.S., Krumholz, L.R., Najar, F.Z., Roe, B.A., and Elshahed, M.S. (2009) Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol* **75**: 5227-5236.

Supplementary material:

Supplementary discussion – Considerations about databases and RNA-types

Taxonomic assignments of ribosomal sequences (using the top hit of BLAST-searches) are still the golden standard and probably deliver the most realistic view of the taxonomic diversity of a microbial community. This is mainly due to a much greater and more diverse number of sequences deposited in ribosomal sequence databases (e.g. SILVA with 2.7 million entries) when compared to a metagenomic or a (rRNA-depleted) metatranscriptomic approach, which mostly relies on fully sequenced genomes (e.g. GOLD with 3173 genomes) (Pruesse et al., 2007; Pagani et al., 2012). Thus, the taxonomic diversity obtained with either of these techniques is likely to differ considerable from each other. As we did not enrich our total RNA for mRNA, our metatranscriptomes were overwhelmingly dominated by rRNA sequences. This could a reason for the unexpected high abundance of β-proteobacterial sequences in our samples and has indeed been shown by Poretsky et al., 2009. They sequenced two separate pools of 16S rRNA genes and mRNAs and found significant differences in the taxonomic composition of the microbial community. Sequences assigned to β-Proteobacteria were among those that showed a much higher count in 16S-pool when compared to the mRNA-pool (Poretsky et al., 2009). Furthermore, the microbial diversity and the estimation of species richness (e.g. by 16S rRNA) also varies quite a lot depending on the read-length the sequences (Youssef et al., 2009), which might be an additional reason for discrepancies in community structure between different samples.

As a result, the detection of certain dominant taxa, which were so far rarely found in marine OMZs with metagenomic and metatranscriptomic approaches, might be due to the lack of available genomes. It might well be that the relatively large percentage of unknown (β-proteobacterial) sequences in metagenomic and metatranscriptomic studies originate from abundant players in the community without a representative genome. Additionally, also the functional characterizations of the β-proteobacteria suffer from the lack of available and curated genomes. Many mRNA sequences were assigned as ‘hypothetical proteins’ (Table S2,) and cannot resolve the functional role of β-proteobacteria in Peruvian OMZ waters (see above).

Supplementary figure and tables:

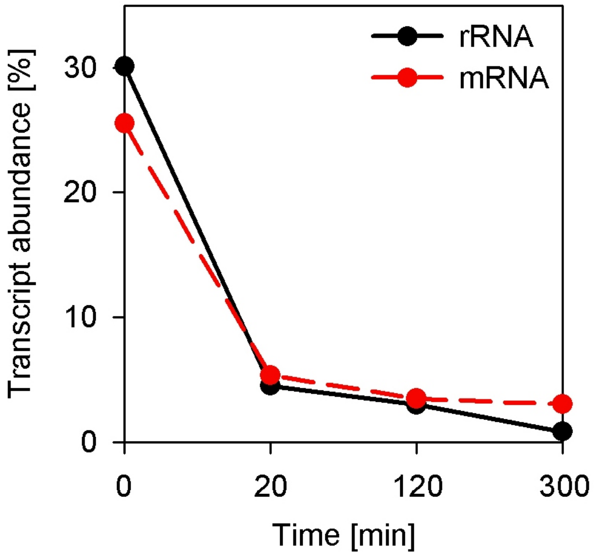


Figure S1: Decay of β-proteobacterial RNA sequence types over time. Shown in percent of all β-proteobacterial rRNA and mRNA sequences, respectively.

Table S1: Selected mRNAs of prominent microorganism with an abundance >10 sequences. Shown is the number of sequences (count), the percentage of all protein-coding sequences, the average of the e-value and the enzyme and organism name (as obtained from the top BLASTx hit). Sequences from all four time points are pooled.

Count	Percentage	E-value	BLASTx hit [organism]
210	0.786	2.0E-05	phasin [alpha proteobacterium HIMB114]
78	0.292	1.7E-06	extracellular solute-binding protein, family 7 [Oceanospirillum sp. MED92]
74	0.277	1.5E-06	putative TCP-1/cpn60 chaperonin family protein [uncultured marine microorganism HF4000_ANIW137I15]
60	0.224	9.0E-08	extracellular solute-binding protein, family 7 [Roseobacter sp. SK209-2-6]
58	0.217	1.3E-04	putative porin [Candidatus Pelagibacter ubique HTCC1002]
57	0.213	3.9E-05	putative porin [Candidatus Pelagibacter ubique HTCC1062]
53	0.198	2.3E-05	similar to nitrate reductase subunit NarG [Candidatus Kuenenia stuttgartiensis]
48	0.180	6.7E-09	putative Sodium:solute symporter family protein [uncultured marine microorganism HF4000_008G09]
41	0.153	3.7E-08	TonB-dependent receptor, putative [Verrucomicrobiae bacterium DG1235]
39	0.146	1.9E-13	binding protein component of ABC sugar transporter [Candidatus Pelagibacter ubique HTCC1062]
39	0.146	1.6E-06	chaperonin GroL [Candidatus Pelagibacter sp. HTCC7211]
38	0.142	7.5E-05	putative porin [Candidatus Pelagibacter sp. HTCC7211]
38	0.142	3.2E-06	similar to hypothetical (di heme) protein [planctomycete KSU-1]
37	0.138	1.1E-12	extracellular solute-binding protein, family 7 [Pseudovibrio sp. JE062]
36	0.135	6.0E-05	10 kDa putative secreted protein [Argas monolakensis]
34	0.127	3.6E-09	ABC transport protein, solute binding component [Agromyces sp. KY5R]
33	0.123	1.7E-05	unnamed protein product [Kluyveromyces lactis]
30	0.112	5.6E-14	putative porin [Oceanicola batsensis HTCC2597]
29	0.108	3.7E-09	peptide ABC superfamily ATP binding cassette transporter [Roseomonas cervicalis ATCC 49957]
27	0.101	2.6E-06	type I secretion target repeat protein [Oceanicola batsensis HTCC2597]
27	0.101	3.7E-06	hypothetical (di heme) protein [Candidatus Kuenenia stuttgartiensis]

27	0.101	1.3E-10	strongly similar to ammonium transporter [Candidatus Kuenenia stuttgartiensis]
25	0.094	4.1E-14	acyl-CoA dehydrogenase family protein [Oceanicola batsensis HTCC2597]
24	0.090	5.3E-20	ABC transporter [Candidatus Pelagibacter ubique HTCC1002]
24	0.090	2.4E-26	co-chaperonin [Syntrophus aciditrophicus SB]
23	0.086	2.5E-07	type I secretion target repeat protein [Oceanicola batsensis HTCC2597]
21	0.079	4.6E-05	PREDICTED: rRNA promoter binding protein-like [Oryctolagus cuniculus]
21	0.079	3.1E-09	extracellular solute-binding protein [Colwellia psychrerythraea 34H]
21	0.079	1.3E-20	chaperonin GroEL [Psychroflexus torquis ATCC 700755]
21	0.079	1.7E-17	TonB-dependent receptor [Alteromonas macleodii ATCC 27126]
20	0.075	2.2E-04	Orf122 [Chlorobaculum tepidum]
19	0.071	2.9E-10	60 kDa chaperonin [Candidatus Pelagibacter ubique HTCC1002]
19	0.071	2.6E-10	TolC family type I secretion outer membrane protein [Polaromonas naphthalenivorans CJ2]
19	0.071	1.8E-26	chaperonin GroEL [Hahella chejuensis KCTC 2396]
18	0.067	7.1E-05	HSP60 [Candidatus Pelagibacter ubique]
18	0.067	5.0E-57	nitrous-oxide reductase precurser [Rhodobacterales bacterium HTCC2654]
18	0.067	7.5E-40	Oligopeptide/dipeptide ABC transporter, ATP-binding protein [Agromyces sp. KY5R]
18	0.067	1.9E-17	acetyl-CoA acetyltransferase (Acetoacetyl-CoA thiolase) [alpha proteobacterium HIMB114]
18	0.067	3.8E-15	ammonia permease [uncultured SUP05 cluster bacterium]
18	0.067	3.0E-10	hydrazine oxidoreductase [uncultured anaerobic ammonium-oxidizing bacterium]
18	0.067	1.5E-27	chaperone protein DnaK [Solibacter usitatus Ellin6076]
17	0.064	8.3E-07	probable ammonium transporter, marine subtype [alpha proteobacterium HIMB114]
17	0.064	7.8E-48	ribosomal protein S3 [Burkholderia sp. Ch1-1]
17	0.064	3.0E-06	immunoreactive 62 kDa antigen PG96 [Porphyromonas gingivalis]
16	0.060	8.2E-17	putative solute-binding periplasmic protein [Psychroflexus torquis ATCC 700755]
16	0.060	1.4E-11	translation elongation factor Tu [Burkholderia sp. CCGE1002]
16	0.060	1.2E-06	extracellular solute-binding protein family 7 [Vibrio furnissii CIP 102972]
16	0.060	1.4E-30	porin [Variovorax paradoxus S110] ┌ gb|ACS17739.1| porin [Variovorax paradoxus S110]
16	0.060	1.4E-25	strongly similar to nitrate reductase (NarH) [Candidatus Kuenenia stuttgartiensis]
16	0.060	1.0E-09	NLP/P60 protein [Acidovorax delafieldii 2AN]
16	0.060	5.0E-47	translation elongation factor Tu [Alcanivorax sp. DG881]
16	0.060	2.0E-76	nicotinate phosphoribosyltransferase [Leeuwenhoekiella blandensis MED217]
16	0.060	4.0E-13	flagellin domain protein [Nitrosomonas sp. AL212]
16	0.060	1.1E-34	chaperone protein [Candidatus Pelagibacter ubique HTCC1062]
15	0.056	1.0E-12	strongly similar to nitrogen regulatory protein P-II [Candidatus Kuenenia stuttgartiensis]
15	0.056	2.0E-35	extracellular solute-binding protein family 3 [Rhizobium leguminosarum bv. trifolii WSM2304]
15	0.056	7.0E-45	heat shock protein HslVU, ATPase subunit HslU [Burkholderia sp. CCGE1002]
15	0.056	4.0E-72	actin [Hyperamoeba sp. ATCC PRA-39]
15	0.056	1.0E-19	aminotransferase [Brucella ceti str. Cudo]
15	0.056	3.0E-39	transcriptional regulator, LysR family protein [Marinobacter sp. ELB17]
14	0.052	9.4E-15	hydroxylamine oxidoreductase [planctomycete KSU-1]
14	0.052	9.0E-13	extracellular solute-binding protein [Marinomonas sp. MWYL1]
14	0.052	1.0E-27	ribosomal protein S8 [Oceanicola batsensis HTCC2597]
14	0.052	1.4E-20	T4-like prohead core scaffold protein [Prochlorococcus phage P-SSM2]

14	0.052	7.1E-05	exocyst complex subunit 4 [Polysphondylium pallidum PN500]
14	0.052	1.1E-45	isocitrate lyase [Alteromonas macleodii ATCC 27126]
14	0.052	1.5E-43	chaperonin GroEL [Marinobacter sp. ELB17]
14	0.052	2.0E-12	extracellular solute-binding protein [Thermotoga sp. RQ2]
14	0.052	3.0E-89	putative trehalose synthase protein [Pseudomonas fluorescens SBW25]
14	0.052	1.7E-15	chaperone protein DnaK [uncultured marine bacterium 583]
14	0.052	3.6E-06	chaperone protein DnaK [Candidatus Pelagibacter sp. HTCC7211]
13	0.049	2.1E-07	putative PKD domain protein [uncultured marine crenarchaeote HF4000_ANIW93J19]
13	0.049	4.3E-55	MraZ protein [uncultured SUP05 cluster bacterium]
13	0.049	1.0E-07	cysteine protease 1 [Noctiluca scintillans]
13	0.049	1.5E-12	elongation factor Tu [Candidatus Koribacter versatilis Ellin345]
13	0.049	7.8E-10	chaperone protein DnaK [uncultured marine bacterium 560]
13	0.049	2.6E-33	Co-chaperonin GroES (HSP10) [Hahella chejuensis KCTC 2396]
13	0.049	4.4E-08	flagellin-like protein [Roseobacter sp. CCS2]
13	0.049	1.1E-46	chaperonin GroEL [Opitutus terrae PB90-1]
13	0.049	2.7E-17	molecular chaperone, HSP90 family [uncultured SUP05 cluster bacterium]
12	0.045	8.9E-18	putative NLPA lipoprotein [uncultured marine microorganism HF4000_007I05]
12	0.045	5.0E-28	ribosomal protein L13 [Variovorax paradoxus S110]
12	0.045	3.4E-05	flagellin domain protein [Nitrosococcus halophilus Nc4]
12	0.045	3.0E-36	ammonium transporter [Candidatus Pelagibacter ubique HTCC1002]
12	0.045	2.2E-14	chaperonin Cpn10 [Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)]
12	0.045	3.6E-27	chaperonin GroEL [bacterium Ellin514]
12	0.045	1.0E-27	chaperonin-60 [uncultured bacterium]
11	0.041	1.5E-49	multidrug efflux system transmembrane protein [Pseudomonas fluorescens SBW25]
11	0.041	5.7E-19	chaperonin GroL [alpha proteobacterium HIMB114]
11	0.041	4.0E-26	twin-arginine translocation pathway signal [Silicibacter lacuscaerulensis ITI-1157]
11	0.041	2.5E-08	putative SPFH domain / Band 7 family protein [uncultured marine microorganism HF4000_133G03]
11	0.041	4.2E-31	ABC-type sugar transport system, periplasmic component [Hahella chejuensis KCTC 2396]
11	0.041	6.7E-19	TonB-dependent receptor, plug [gamma proteobacterium NOR5-3]
11	0.041	2.5E-40	membrane protein, putative [uncultured marine bacterium 314]
11	0.041	1.5E-53	30S ribosomal protein S6 [Roseobacter sp. AzwK-3b]
11	0.041	6.0E-05	structural maintenance of chromosome seggregation ATPase protein [Candidatus Kuenenia stuttgartiensis]
11	0.041	1.0E-15	aldehyde dehydrogenase family protein [Oceanicola batsensis HTCC2597]
11	0.041	8.8E-07	putative Permease family protein [uncultured marine microorganism HF4000_133G03]
11	0.041	9.1E-06	50S ribosomal protein L10 [Alteromonas macleodii ATCC 27126]
11	0.041	6.7E-45	actin [Salpingoeca amphoridium]
11	0.041	1.0E-14	Chain A, Structure Of 6-Aminohexanoate Cyclic Dimer Hydrolase Complexed With Substrate
11	0.041	6.7E-33	chaperone protein DnaK [Nitrosococcus halophilus Nc4]
11	0.041	4.0E-40	Formamidase [Methylophilus methylotrophus]
11	0.041	5.6E-07	metalloprotease FtsH [Psychroflexus torquis ATCC 700755]
11	0.041	3.0E-62	aconitate hydratase [Flavobacteria bacterium BAL38]
11	0.041	8.3E-14	heat shock protein Hsp20 [Roseiflexus castenholzii DSM 13941]
11	0.041	8.8E-32	chaperonin GroEL [Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)]
11	0.041	2.0E-57	thiamine biosynthesis protein ThiC [Synechococcus sp. RS9916]

Table S2: All identified β-proteobacterial mRNAs with an abundance ≥5 sequences. Shown is the number of sequences (count), the percentage of all protein-coding sequences from the respective time points, the time point, the average e-value and the enzyme and organism name (as obtained from the top BLASTx hit).

Count	Percentage	Time point	e-value	BLASTx hit [organism]
27	0.46	0 min	2.3E-07	hypothetical protein Veis_0676 [Verminephrobacter eiseniae EF01-2]
22	0.38	0 min	4.4E-05	hypothetical protein [Curvibacter putative symbiont of Hydra magnipapillata]
19	0.33	0 min	2.6E-10	TolC family type I secretion outer membrane protein [Polaromonas naphthalenivorans CJ2]
16	0.28	0 min	1.0E-47	ribosomal protein S3 [Burkholderia sp. Ch1-1]
16	0.24	20 min	1.0E-09	NLP/P60 protein [Acidovorax delafieldii 2AN]
16	0.22	120 min	4.0E-13	flagellin domain protein [Nitrosomonas sp. AL212]
15	0.26	0 min	1.7E-11	translation elongation factor Tu [Burkholderia sp. CCGE1002]
15	0.26	0 min	3.0E-06	hypothetical protein Aave_4148 [Acidovorax avenae subsp. citrulli AAC00-1]
15	0.26	0 min	7.0E-45	heat shock protein HslVU, ATPase subunit HslU [Burkholderia sp. CCGE1002]
12	0.21	0 min	1.5E-30	porin [Variovorax paradoxus S110]
12	0.21	0 min	5.0E-28	ribosomal protein L13 [Variovorax paradoxus S110]
12	0.17	300 min	1.0E-26	hypothetical protein Tbd_2284 [Thiobacillus denitrificans ATCC 25259]
11	0.17	20 min	5.3E-05	hypothetical protein [Curvibacter putative symbiont of Hydra magnipapillata]
11	0.17	20 min	4.0E-40	Formamidase [Methylophilus methylotrophus]
10	0.17	0 min	1.3E-57	ribosomal protein L2 [Burkholderia sp. CCGE1002]
10	0.15	20 min	2.0E-12	conserved hypothetical protein [Acidovorax avenae subsp. avenae ATCC 19860]
9	0.15	0 min	8.0E-52	porin Gram-negative type [Burkholderia sp. Ch1-1]
9	0.15	0 min	6.7E-11	ribosomal protein S3 [Acidovorax delafieldii 2AN]
9	0.15	0 min	3.3E-41	ribosomal protein S8 [Acidovorax delafieldii 2AN]
9	0.15	0 min	2.0E-21	porin [Variovorax paradoxus S110]
8	0.14	0 min	1.7E-09	preprotein translocase, SecY subunit [Burkholderia sp. Ch1-1]
8	0.14	0 min	5.0E-54	ribosomal protein S4 [Variovorax paradoxus S110]
8	0.12	20 min	2.3E-15	hypothetical protein Veis_0676 [Verminephrobacter eiseniae EF01-2]
8	0.11	120 min	3.0E-04	ribonuclease E [Acidovorax sp. JS42]
8	0.11	300 min	1.7E-34	6-aminohexanoate-cyclic-dimer hydrolase [Pseudomonas strain NK87]
7	0.12	0 min	2.5E-46	ribosomal protein L15 [Ralstonia pickettii 12J]
7	0.12	0 min	2.0E-81	translation elongation factor G [Delftia acidovorans SPH-1]
7	0.12	0 min	3.5E-44	30S ribosomal protein S7 [Curvibacter putative symbiont of Hydra magnipapillata]
7	0.12	0 min	1.7E-38	ribosomal protein L4/L1e [Acidovorax avenae subsp. avenae ATCC 19860]
7	0.12	0 min	2.3E-23	ribosomal protein S13 [Acidovorax delafieldii 2AN]
7	0.11	20 min	4.0E-38	methyltransferase type 12 [Polaromonas sp. JS666]
7	0.11	20 min	4.0E-57	cell shape determining protein, MreB/Mrl family [Burkholderia graminis C4D1M]
7	0.11	20 min	1.0E-24	transposase for IS1655 [Neisseria meningitidis 053442]
6	0.10	0 min	1.3E-13	glutamine synthetase, type I [Burkholderia phymatum STM815]
6	0.10	0 min	5.5E-47	50S ribosomal protein L23P [Acidovorax avenae subsp. citrulli AAC00-1]
6	0.10	0 min	5.0E-07	hypothetical protein AcavDRAFT_4806 [Acidovorax avenae

				subsp. avenae ATCC 19860]
6	0.10	0 min	3.5E-04	hypothetical protein BMULJ_05092 [Burkholderia multivorans ATCC 17616]
6	0.10	0 min	2.0E-13	acyltransferase [Azoarcus sp. BH72]
6	0.10	0 min	6.0E-56	outer membrane porin [Burkholderia pseudomallei 1710b]
6	0.09	20 min	8.0E-48	MarR family transcriptional regulator [Burkholderia xenovorans LB400]
6	0.09	20 min	3.0E-44	nitrite reductase, copper-containing [Methylotenera mobilis JLW8]
6	0.08	120 min	2.5E-24	poly(R)-hydroxyalkanoic acid synthase, class I [Thauera sp. MZ1T]
6	0.08	300 min	5.0E-05	hypothetical protein [Curvibacter putative symbiont of Hydra magnipapillata]
6	0.08	300 min	1.7E-05	hypothetical protein AcavDRAFT_4806 [Acidovorax avenae subsp. avenae ATCC 19860]]
6	0.08	300 min	5.0E-17	5'-methylthioadenosine phosphorylase [Rhodoferax ferrireducens T118]
5	0.09	0 min	1.0E-04	50S ribosomal protein L1 [Acidovorax sp. JS42]
5	0.09	0 min	4.0E-24	OmpA/MotB domain protein [Burkholderia sp. Ch1-1]
5	0.09	0 min	3.5E-11	ketol-acid reductoisomerase [Burkholderia sp. CCGE1001]
5	0.09	0 min	2.0E-77	acetolactate synthase, large subunit, biosynthetic type [Acidovorax delafieldii 2AN]
5	0.09	0 min	9.0E-32	glycine cleavage system T protein [Polaromonas naphthalenivorans CJ2]
5	0.09	0 min	1.5E-64	3-hydroxyacyl-CoA dehydrogenase NAD-binding [Burkholderia graminis C4D1M]
5	0.09	0 min	1.3E-22	uridylate kinase [Rhodoferax ferrireducens T118]
5	0.09	0 min	5.0E-41	ATP-dependent protease peptidase subunit [Verminephrobacter eiseniae EF01-2]
5	0.09	0 min	2.7E-09	flagellin domain protein [Burkholderia sp. H160]
5	0.09	0 min	3.0E-27	cytochrome-c oxidase [Ralstonia eutropha JMP134]
5	0.09	0 min	3.0E-58	Outer membrane protein (porin) [Burkholderia xenovorans LB400]
5	0.08	20 min	6.1E-05	conserved hypothetical protein [uncultured beta proteobacterium CBNPD1 BAC clone 578]
5	0.08	20 min	1.7E-04	hypothetical protein BMULJ_05092 [Burkholderia multivorans ATCC 17616]
5	0.07	120 min	4.2E-04	hypothetical protein BMULJ_05092 [Burkholderia multivorans ATCC 17616]
5	0.07	300 min	3.5E-04	hypothetical protein BMULJ_05092 [Burkholderia multivorans ATCC 17616]

3.2. Bioinformatic analysis of high-throughput sequence data

Application Notes

Fragment Recruitment on Metabolic Pathways (FROMP): Comparative metabolic profiling of metagenomes and metatranscriptomes

Dhwani K Desai[1,§], Harald Schunck[1,§], Johannes W Löser[1,2], Markus Lommer[1], Julie LaRoche[1]

[1]GEOMAR | Helmholtz-Zentrum für Ozeanforschung Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany

[2]Institut für Informatik, Christian-Albrechts-Universität zu Kiel, Ludewig-Meyn-Str. 4, 24118 Kiel, Germany

[§]Authors contributed equally to this work

Received on 15.05.2012

Abstract

Motivation: The sheer scale of the Meta-omic (metagenomic and metatranscriptomic) datasets that are now available warrants the development of automated protocols for organizing, annotating and comparing the samples in terms of their metabolic profiles. We describe a user-friendly java program FROMP (Fragment Recruitment on Metabolic Pathways) for mapping and visualizing enzyme annotations onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways and comparing the samples in terms of either their Pathway Completeness Scores or their relative Activity Scores. This program along with our fully-configurable PERL based annotation organization pipeline Meta2Pro (METAbolic PROfiling of META-omic data) offers a quick and accurate standalone solution for metabolic profiling of environmental samples. Apart from pictorial comparisons, FROMP can also generate score matrices for multiple meta-omics samples which can be used directly by other statistical programs.

Availability: The source code and documentation for FROMP can be downloaded from https://sites.google.com/site/dhwanidesai/home/software along with the Meta2Pro collection of PERL scripts.

Supplementary data: All supplementary data is available on *Bioinformatics* online and on https://sites.google.com/site/dhwanidesai/home/fromp_suppl.

Contact: ddesai@geomar.de, hschunck@geomar.de, jlo@informatik.uni-kiel.de

Introduction

The rapidly accumulating environmental meta-omic projects resulting from high-throughput next-generation sequencing techniques warrant the development of new protocols which can provide a quick overview of the microbial metabolic activity. There has been some effort towards management of such data (Sun, et al., 2011), its taxonomic and metabolic profiling (Bork, et al., 2010; Huson, et al., 2007; Meyer, et al., 2008), visualization of metabolic pathways (Bork, et al., 2011) and statistical analyses of community differences (Beiko and Parks, 2010). In most cases, the tools are web-based and the primary method for annotation is BLAST (Altschul, et al., 1990). We describe here a standalone set of tools to get a rapid and accurate overview of the metabolic functions of the resident microbial community. The enzyme identification component of this pipeline, based on the ModEnzA Enzyme Commission (Desai, et al., 2011) and Pfam (Bateman, et al., 2004) profile hidden Markov models (HMMs), provides a quick and accurate EC number identification. The standout feature is the FROMP pathway mapping and comparative visualization tool which maps EC numbers and Pfam annotations onto the KEGG (Kanehisa, et al., 2010) reference metabolic pathways based on either a Pathway Completeness Score modified from Inskeep and colleagues (2010), a Pathway Activity Score or an odds-ratio for gene enrichment (Gill, et al., 2006).

Methods and features

The java program FROMP is a part of the Meta2Pro pipeline (supplementary figure S1). It maps the EC numbers from ModEnzA directly onto the KEGG pathways, whereas the Pfam hits are first mapped to the corresponding Gene Ontology IDs (Ashburner, et al., 2000), (using the conversion files pfam2go, kegg2go and ec2go downloaded from http://www.geneontology.org/external2go/) which are then mapped to KEGG reaction IDs or EC numbers.

Pathway Completeness Score: FROMP uses a weighting scheme for each EC number as described in (Inskeep, et al., 2010) which weighs down ECs present in multiple pathways. We have modified the weight by adding a term for the presence of continuous, unbranched chains of reactions with the logic that a pathway is more likely to be functional in a sample if two or more links in an unbranched chain of reactions are detected.

For each EC *i* the weight

$$W_i = \left((N_{(T,i)}/N_{(U,i)})/N_{(P,i)}\right) * \sqrt{L_{(UBC,r)}}$$

where $N_{(T,i)}$ is the total number of ECs in all pathways that have EC *i*, $N_{(U,i)}$ is the number of unique ECs in all the pathways that have EC *i* and $N_{(P,i)}$ is the total number of pathways where EC *i* is present and $L_{(UBC,r)}$ is the total edge-length of the unbranched chain containing EC *i* in the reference pathway.

The pathway completeness score for a pathway *p* is then

$$C_P = \left(\frac{\sum\limits_{i \in EC_p} W_i * I_i * \sqrt{L_{(UBC,s)}}}{\sum\limits_{i \in EC_p} W_i} \right)$$

where W_i is the specificity weight of each EC i in pathway p, and I_i is 1 if the EC number is detected in the sample and $L_{(UBC,s)}$ is the edge-length of the unbranched chain containing EC i in the sample.

Odds-ratio for gene enrichment: As described in (Gill, et al., 2006) if A and C are the occurrence counts of a given EC in sample i and all other comparison samples j, respectively, and B and D are occurrence counts of all other ECs in sample i and comparison samples j respectively, then the odds ratio for the given EC in sample i is (A/B)/ (C/D).

Pathway Activity Score: This is simply the sum of counts for the ECs in a given pathway multiplied by the EC weight.

The user can standardize unequal sample sizes to the smallest sample by randomly selecting equal numbers from the other samples, before calculating the odds ratio or the pathway scores.

Input: Apart from reading the output of the *hmmscan* program (Eddy et al., 1998), FROMP can also read in tab or comma separated list of EC numbers and Pfams (one-column), ECs and Pfams with counts (two-column) and ECs and Pfams with counts and sequence Ids (three-column) of the meta-omic sequences.

Output: The comparative recruitment of various samples on the reference pathways can be exported as PNG files. The various score matrices (including the EC count matrix) for the samples and the sequence IDs of the fragments mapping onto each EC or pathway can also be exported as text files.

Comparative analysis of *Thalassiosira oceanica* transcriptomes

Transcriptomes of *T. oceanica* obtained under iron starvation and supplementation were compared using FROMP. Figure S2 shows the FROMP output comparing the transcriptomes in terms of the oxidative phosphorylation and photosynthesis pathways. The EC Activity Matrix for the transcriptomes was also analyzed using the Statistical Analysis of *Metagenomic* Profiles (STAMP) program (Beiko and Parks, 2010). Supplementary Table ST1 shows ECs with significantly different occurrences in the two conditions (Fisher's exact test calculated from STAMP, $p\text{-value} < 0.005$) and their corresponding enrichment factors in FROMP. Upregulated functions belong to the amino acid and lipid core metabolism and degradation of organic matter (chitinase). The photosynthetic pathway (RubisCo, protochlorophyllide reductase, phosphoglycerate kinase) appears to be downregulated. The results from the screening obtained with FROMP are in line with the more complex analysis presented for iron starved diatoms (Allen, et al., 2008) and confirms that FROMP detects the important differences in gene expression patterns between transcriptomes of different origins. Supplementary figures S3, S4 and S5 show a comparative analysis of FROMP on multiple metatranscriptomic samples from marine pelagic microbial communities obtained from the permanent oxygen minimum zone off Peru.

Conclusions

We present here a set of tools for accurate standalone metabolic profiling of meta-omic data. The FROMP program takes as input results generated from HMM scans of the meta-omic data with models of EC numbers or Pfams and generates metabolic profiles in terms of KEGG Pathway Completeness Score, Pathway Activity Score or the EC counts. It also provides user-friendly pathway visualization capabilities for comparing multiple samples. The use of EC numbers instead of BLAST hits for metabolic annotation provides a direct link to the reference pathways.

Acknowledgments

Funding: This work is funded by the WGL-pact project "REAL" (Leibniz Association) and is a contribution of the Collaborative Research Centre 754 "Climate – Biogeochemistry Interactions in the Tropical Oceans" (www.sfb754.de), which is supported by the German Research Association.

References

Allen, A.E. *et al.* (2008) Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation, *Proc Natl Acad Sci U S A*, **105**, 10438-10443.

Altschul, S.F. *et al.* (1990) Basic Local Alignment Search Tool, *Journal of Molecular Biology*, **215**, 403-410.

Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bateman, A. *et al.* (2004) The Pfam protein families database, *Nucleic Acids Res*, **32**, D138-141.

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

Bork, P. *et al.* (2010) SmashCommunity: a metagenomic annotation and analysis tool, *Bioinformatics*, **26**, 2977-2978.

Bork, P. *et al.* (2011) iPath2.0: interactive pathway explorer, *Nucleic Acids Research*, **39**, W412-W415.

Desai, D.K. *et al.* (2011) ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities, *Adv Bioinformatics*, **2011**, 743782.

Eddy, S.R. (1998) HMMER: biological sequence analysis using profile hidden Markov models. http://hmmer.org/.

Gill, S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome, *Science*, **312**, 1355-1359.

Huson, D.H. *et al.* (2007) MEGAN analysis of metagenomic data, *Genome Research*, **17**, 377-386.

Inskeep, W.P. *et al.* (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function, *PLoS One*, **5**, e9773.

Kanehisa, M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res,* **38**, D355-D360

Meyer, F. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes, *Bmc Bioinformatics*, **9**.

Sun, S.L. *et al.* (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource, *Nucleic Acids Research*, **39**, D546-D551.

Supplementary information:

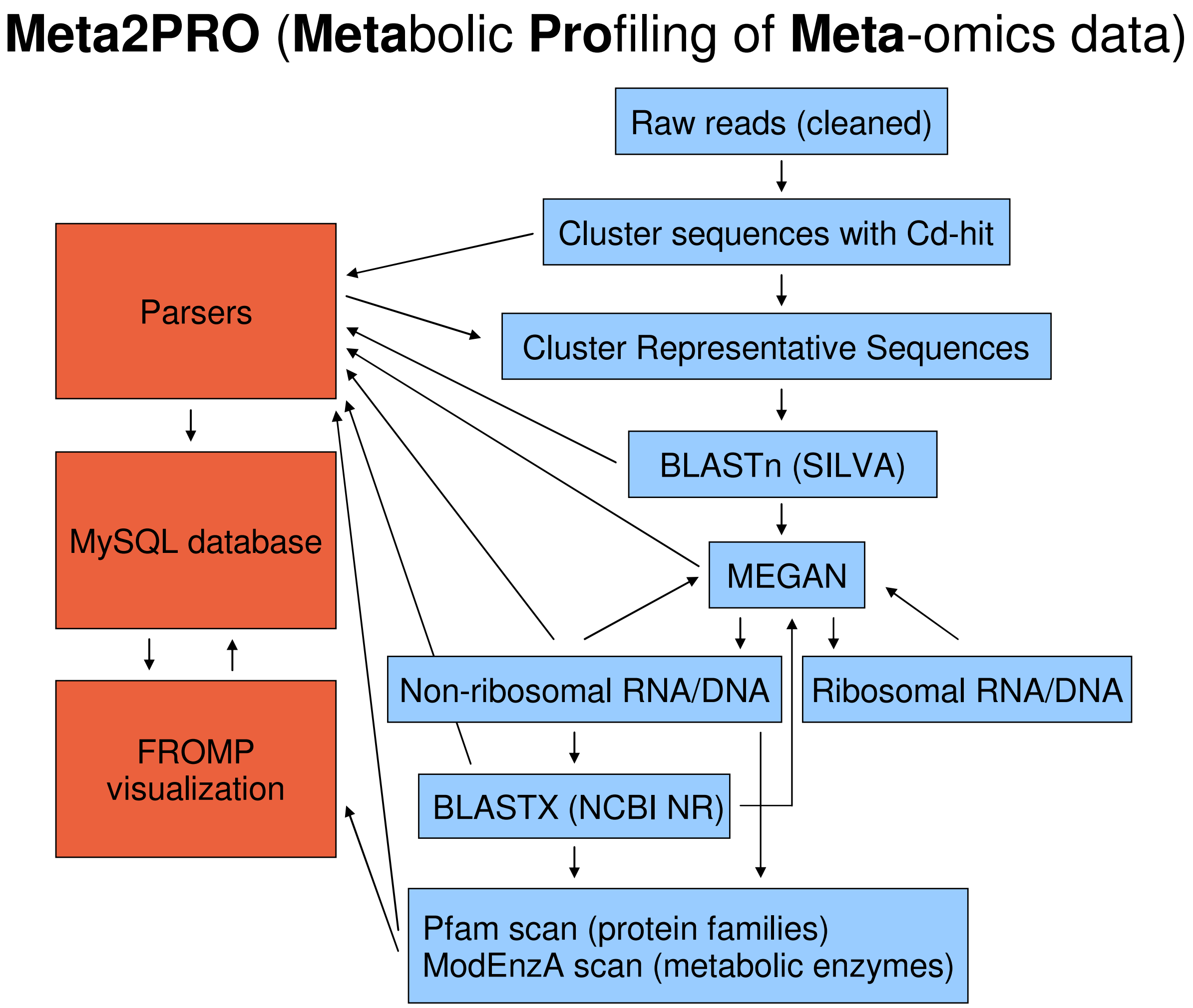
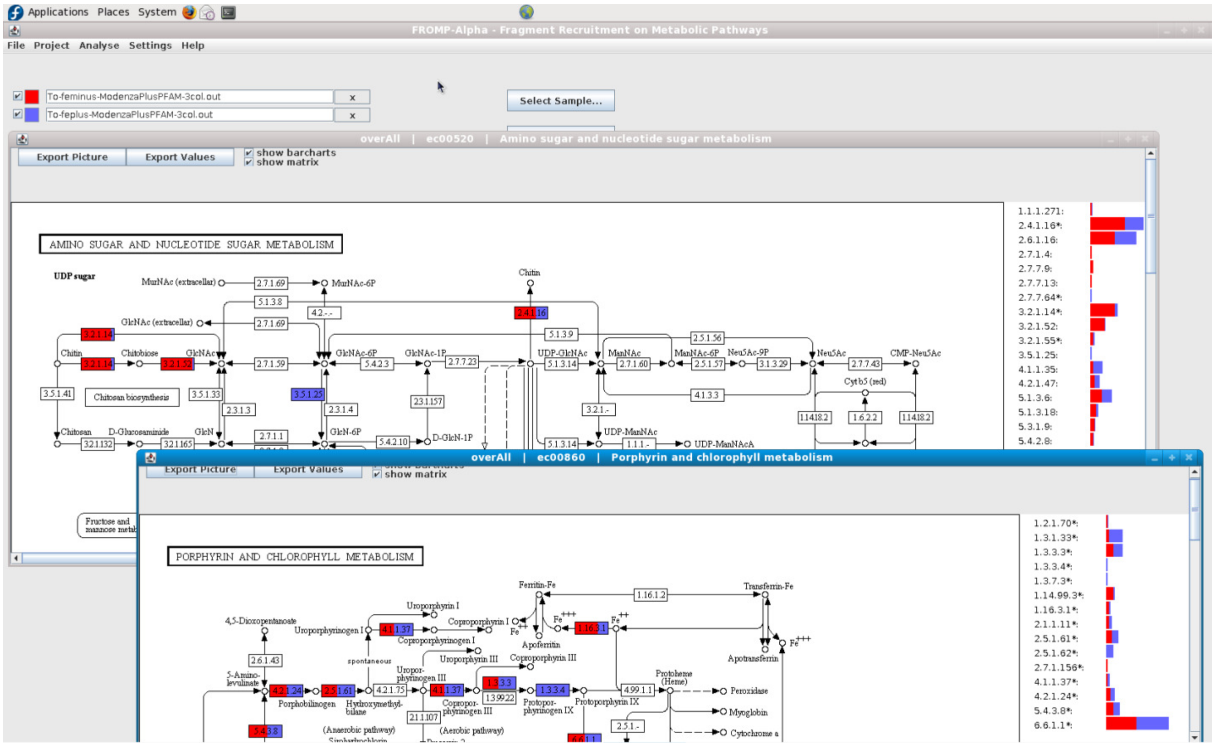
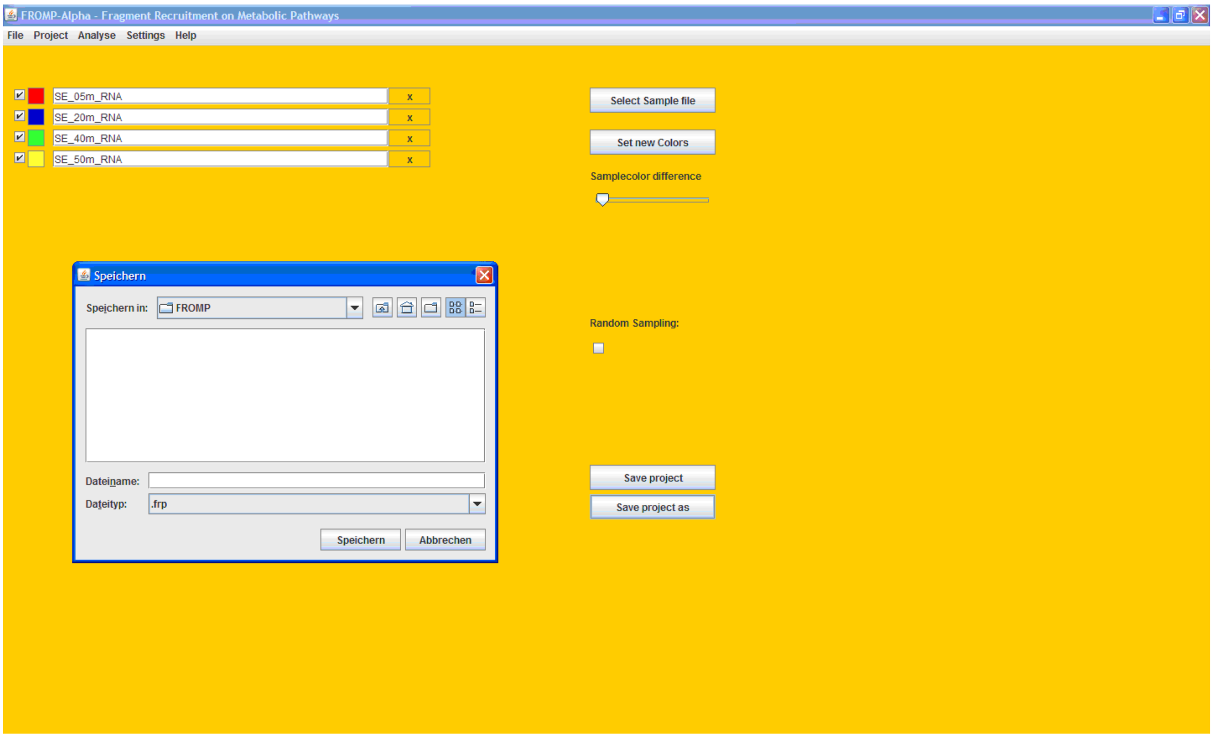

Meta2PRO (Metabolic Profiling of Meta-omics data)
Raw reads (cleaned)
Cluster sequences with Cd-hit
Cluster Representative Sequences
BLASTn (SILVA)
MEGAN
Non-ribosomal RNA/DNA
Ribosomal RNA/DNA
BLASTX (NCBI NR)
Pfam scan (protein families)
ModEnzA scan (metabolic enzymes)
Parsers
MySQL database
FROMP visualization

Figure S1: Schematic diagram of the Meta2Pro annotation workflow. Meta2Pro includes sequence clustering, BLAST-searches, profile hidden Markov models of the ModEnzA Enzyme Commission groups and of the Pfam protein families, the profiling of EC number counts with FROMP and the storage of the data and the metadata in a web browser-based MySQL database (to be accessed with phpMyAdmin).


Applications Places System
FROMP-Alpha - Fragment Recruitment on Metabolic Pathways
File Project Analyse Settings Help
Select Sample...
overAll | ec00520 | Amino sugar and nucleotide sugar metabolism
Export Picture
Export Values
show barcharts
show matrix
AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM
overAll | ec00860 | Porphyrin and chlorophyll metabolism
PORPHYRIN AND CHLOROPHYLL METABOLISM

☑ 🟥 To-feminus-ModenzaPlusPFAM-3col.out x
☑ 🟦 To-feplus-ModenzaPlusPFAM-3col.out x

Figure S2: Screen shot of FROMP demonstrating the mapping of EC numbers on two KEGG reference pathways of transcriptomic samples obtained from T. oceanica cultured under iron stress compared to normal growth conditions (Table S1). Shown are bar chart visualizations of the EC number counts.



☑ 🟥 SE_05m_RNA x
☑ 🟦 SE_20m_RNA x
☑ 🟩 SE_40m_RNA x
☑ 🟨 SE_50m_RNA x

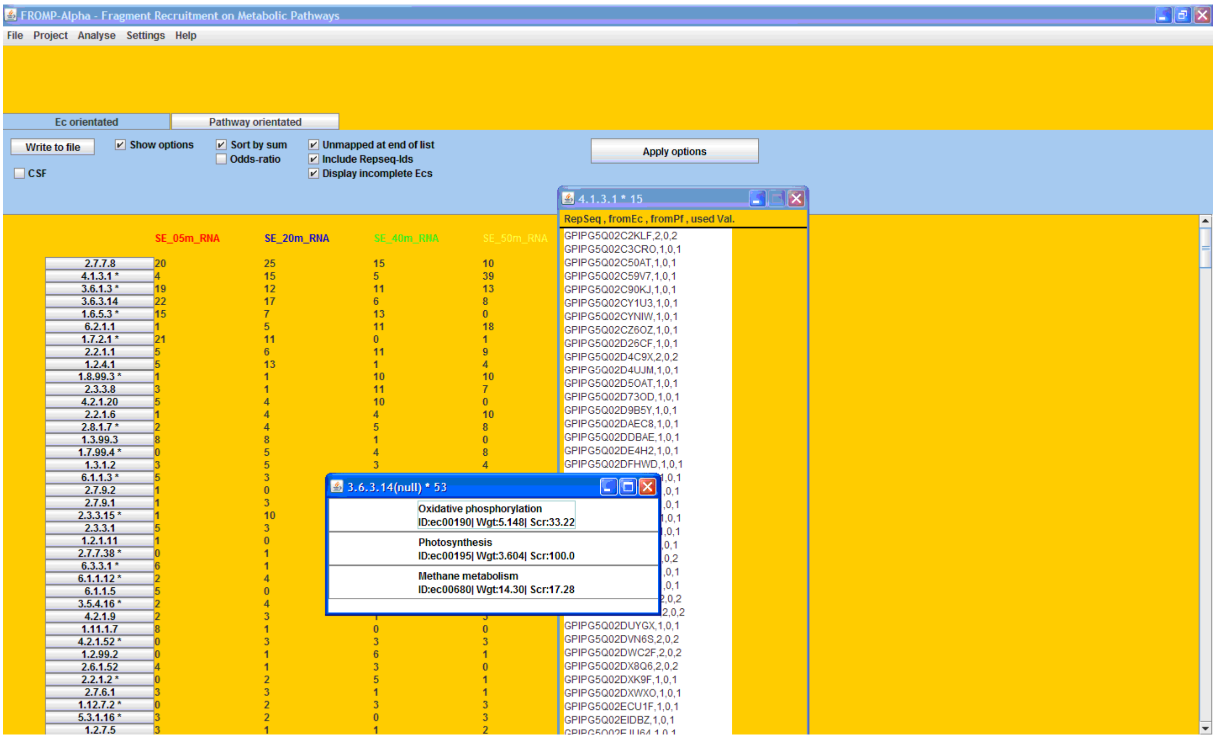
Random Sampling: □

Speichern in: FROMP

Dateiname:

Dateityp: .frp

Figure S3: Screen shot demonstrating the sample uploading, editing and saving options of FROMP.



☑ Show options
☑ Sort by sum
□ Odds-ratio
☑ Unmapped at end of list
☑ Include Repseq-Ids
☑ Display incomplete Ecs
□ CSF

	SE_05m_RNA	SE_20m_RNA	SE_40m_RNA	SE_50m_RNA
2.7.7.8	20	25	15	10
4.1.3.1 *	4	15	5	39
3.6.1.3 *	19	12	11	13
3.6.3.14	22	17	6	8
1.6.5.3 *	15	7	13	0
6.2.1.1	1	5	11	18
1.7.2.1 *	21	11	0	1
2.2.1.1	5	6	11	9
1.2.4.1	5	13	1	4
1.8.99.3 *	1	1	10	10
2.3.3.8	3	1	11	7
4.2.1.20	5	4	10	0
2.2.1.6	1	4	4	10
2.8.1.7 *	2	4	5	8
1.3.99.3	8	8	1	0
1.7.99.4 *	0	5	4	8
1.3.1.2	3	5	3	4
6.1.1.3 *	5	3		
2.7.9.2	1	0		
2.7.9.1	1	3		
2.3.3.15 *	1	10		
2.3.3.1	5	3		
1.2.1.11	1	0		
2.7.7.38 *	0	1		
6.3.3.1 *	6	1		
6.1.1.12 *	2	4		
6.1.1.5	5	0		
3.5.4.16 *	2	4		
4.2.1.9	2	3		
1.11.1.7	8	1	0	0
4.2.1.52 *	0	3	3	3
1.2.99.2	0	1	6	1
2.6.1.52	4	1	3	0
2.2.1.2 *	0	2	5	1
2.7.6.1	3	3	1	1
1.12.7.2 *	0	2	3	3
5.3.1.16 *	3	2	0	3
1.2.7.5	3	1	1	2

RepSeq , fromEc , fromPf , used Val.

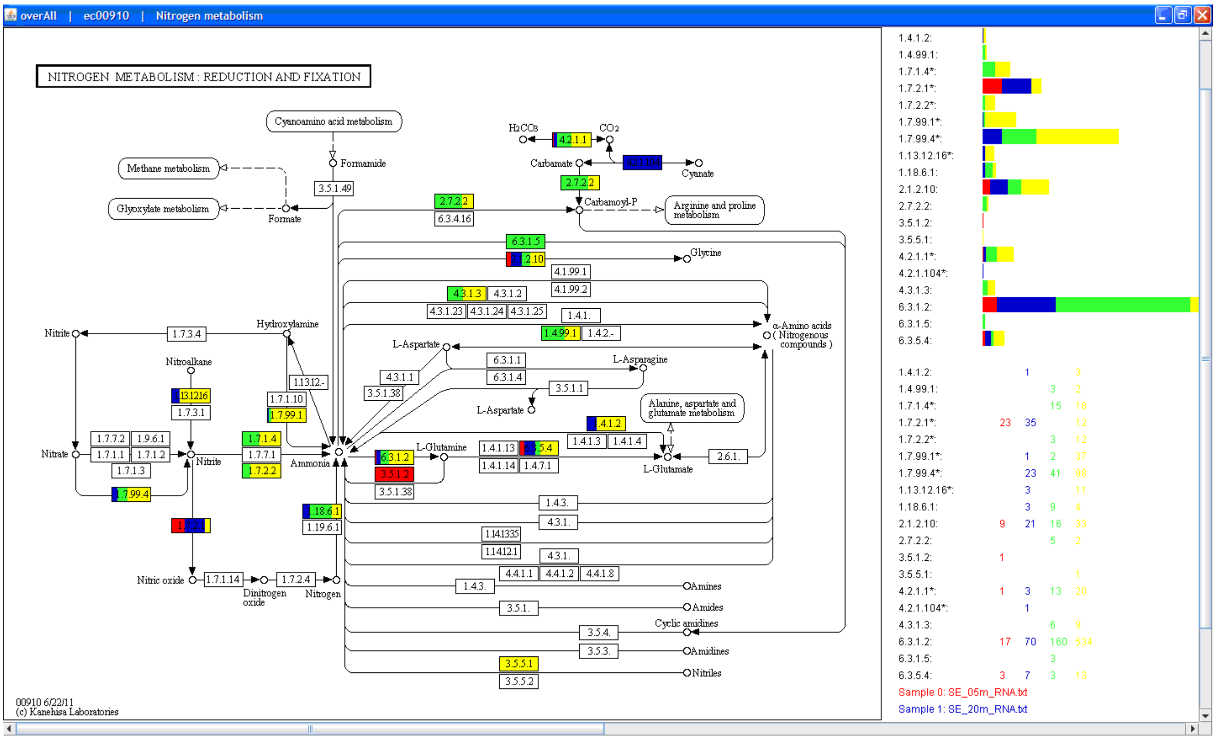
```
GPIPG5Q02C2KLF,2,0,2
GPIPG5Q02C3CRO,1,0,1
GPIPG5Q02C50AT,1,0,1
GPIPG5Q02C59V7,1,0,1
GPIPG5Q02C90KJ,1,0,1
GPIPG5Q02CY1U3,1,0,1
GPIPG5Q02CYNIW,1,0,1
GPIPG5Q02CZ6OZ,1,0,1
GPIPG5Q02D26CF,1,0,1
GPIPG5Q02D4C9X,2,0,2
GPIPG5Q02D4UJM,1,0,1
GPIPG5Q02D5OAT,1,0,1
GPIPG5Q02D73OD,1,0,1
GPIPG5Q02D9B5Y,1,0,1
GPIPG5Q02DAEC8,1,0,1
GPIPG5Q02DDBAE,1,0,1
GPIPG5Q02DE4H2,1,0,1
GPIPG5Q02DFHWD,1,0,1
1,0,1
,0,1
,0,1
1,0,1
1,0,1
,0,1
,0,2
,0,1
,0,1
2,0,2
2,0,2
GPIPG5Q02DUYGX,1,0,1
GPIPG5Q02DVN6S,2,0,2
GPIPG5Q02DWC2F,2,0,2
GPIPG5Q02DX8Q6,2,0,2
GPIPG5Q02DXK9F,1,0,1
GPIPG5Q02DXWXO,1,0,1
GPIPG5Q02ECU1F,1,0,1
GPIPG5Q02EIDBZ,1,0,1
GPIPG5Q02EJU64,1,0,1
```

Oxidative phosphorylation
ID:ec00190| Wgt:5.148| Scr:33.22

Photosynthesis
ID:ec00195| Wgt:3.604| Scr:100.0

Methane metabolism
ID:ec00680| Wgt:14.30| Scr:17.28

Figure S4: Screen shot of FROMP demonstrating the comparative analysis of (randomly sampled) EC number counts of four different metatranscriptomic samples. Shown are pop-up windows indicating KEGG-pathways and information on the uploaded sequences (RepSeq (sequence identifier), EC number- and Pfam-assignments and the respective counts).



1.4.1.2:		1		3
1.4.99.1:			3	2
1.7.1.4*:			15	18
1.7.2.1*:	23	35		12
1.7.2.2*:			3	12
1.7.99.1*:		1	2	37
1.7.99.4*:		23	41	98
1.13.12.16*:		3		11
1.18.6.1:		3	9	4
2.1.2.10:	9	21	16	33
2.7.2.2:			5	2
3.5.1.2:	1			
3.5.5.1:				1
4.2.1.1*:	1	3	13	20
4.2.1.104*:		1		
4.3.1.3:			6	9
6.3.1.2:	17	70	160	534
6.3.1.5:			3	
6.3.5.4:	3	7	3	13

Sample 0: SE_05m_RNA.txt
Sample 1: SE_20m_RNA.txt

Figure S5: Screen shot demonstrating the mapping of EC numbers from four metatranscriptomic samples on the ‘Nitrogen Metabolism’ as denoted in KEGG. Shown are bar chart visualizations and a matrix of the EC number counts.

Table S1: Enriched transcripts of iron supplemented and iron-starved *T. oceanica* cultures. The samples were equalized so that an equal number of hits were processed for both the samples (the size of sample with the smaller number of hits size was fixed as the sample size and an equal number of hits were randomly chosen from the other sample).

		Iron Sufficient (Control)		Iron starvation		
EC number	Enzyme name	Transcripts	Enrichment	Transcripts	Enrichment	p-values
4.1.2.13	Fructose-bisphosphate aldolase	0	NA	38	999	6E−16
3.6.1.3	Adenosinetriphosphatase	51	0.44	113	2.26	1E−09
4.1.1.39	Ribulose-bisphosphate carboxylase	68	3.84	18	0.25	2E−06
5.1.1.11	Phenylalanine racemase	8	0.22	36	4.54	8E−06
1.3.1.33	Protochlorophyllide reductase	19	9.55	2	0.1	0.0002
6.3.1.2	Glutamine synthetase	42	3.03	14	0.33	0.0014
4.2.1.11	Carbonate dehydratase	15	3.76	4	0.26	0.0023
3.2.1.14	Chitinase	4	0.23	17	4.26	0.0042
2.1.2.1	Glycine hydroxyl-methyltransferase	8	0.36	22	2.76	0.0044
1.4.4.2	Glycine decarboxylase	18	0.57	31	1.73	0.0070
2.7.2.3	Phosphoglycerate kinase	15	5.03	3	0.19	0.0074
4.1.1.49	Phosphoenolpyruvate carboxylase	44	2.77	16	0.39	0.0085
3.1.1.4	Phospholipase A2	18	0.51	35	1.95	0.0098

FROMP-v1.0 User's Guide

Dhwani K Desai, Harald Schunck, Johannes Löser and Julie LaRoche

1 Aim

FROMP aims at profiling and visualizing metabolic categories of high-throughput sequence data through:

- Mapping Enzyme Commission (EC) and Protein Family (Pfam) assignments to Kyoto Encyclopedia of Genes and Genomes (KEGG) reference metabolic pathways using weights for ECs and a Pathway Completeness score, Pathway Activity Score or an odds-ratio for gene enrichment.
- Displaying the KEGG reference pathways with proportionate contributions from each sample to each EC in the pathways.
- Exporting the metabolic profiles for the samples in a project in the form of matrices of either Pathway Completeness or Pathway Activity scores or Odds-ratios (for individual ECs) as text files for further analysis.

Abstract

Motivation: The sheer scale of the Meta-omic (metagenomic and metatranscriptomic) datasets that are now available warrants the development of automated protocols for organizing, annotating and comparing the samples in terms of their metabolic profiles. We describe a user-friendly java program FROMP (Fragment Recruitment on Metabolic Pathways) for mapping and visualizing enzyme annotations onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways and comparing the samples in terms of either their Pathway Completeness Scores or their relative Activity Scores. This program along with our fully-configurable PERL based annotation organization pipeline Meta2Pro (METAbolic PROfiling of META-omic data) offers a quick and accurate standalone solution for metabolic profiling of environmental samples. Apart from pictorial comparisons, FROMP can also generate score matrices for multiple meta-omics samples which can be used directly by other statistical programs.

Availability: The source code and documentation for FROMP along with the Meta2Pro collection of PERL scripts can be downloaded from https://sites.google.com/site/dhwanidesai/home/software.

Supplementary data: All supplementary data is available at https://sites.google.com/site/dhwanidesai/home/software or at *Bioinformatics* online.

Contact: ddesai@geomar.de, hschunck@geomar.de, jlo@informatik.uni-kiel.de

1.1 What FROMP is NOT!

- FROMP is NOT a "black box" solution for a complete analysis of metagenomes or metatranscriptomes which can work on raw sequences: It is meant to be used as an analytical tool for comparing multiple meta-omic samples in terms of their PFAM and ModEnzA EC annotations.
- FROMP is NOT a statistical comparison tool: The matrices output by FROMP could be used as input for other statistical comparison program like STAMP, but there is no provision for statistical differences between groups of samples in FROMP.

1.2 Input Format

Apart from reading the output of the hmmscan program from the HMMER package, FROMP can also read in tab or comma separated list of EC numbers and Pfams (one-column), ECs and Pfams with counts (two-column) and ECs and Pfams with counts and sequence Ids (three-column) of the meta-omic sequences. Some example input files are provided in the Real-Samples folder. These include 10 metagenomes from chimneys and hydrothermal vents (located in /Real-Samples/chimney-data) and the transcriptomes of *T. oceanica* obtained under iron limitation and iron sufficiency (located in /Real-Samples/T-oceanica). These files have been prepared by concatenating the output of the program hmmscan run with the PFAM and the ModEnzA EC profiles on the amino acid sequences translated from the metagenome or metatranscriptome sequences.

1.3 Output

The comparative recruitment of various samples on the reference pathways can be exported as PNG files. The various score matrices (including the EC count matrix) for the samples and the sequence IDs of the fragments mapping onto each EC or pathway can also be exported as text files.

2 Running Instructions

FROMP is a Java program and hence requires an installed and working Java version. You can get the latest java version from http://java.com/en/.

2.1 Instructions for Windows

Unzip the .zip file to extract the Fromp-v1.0 folder and double click on the "dnarna.jar" file. This will create a FROMP.bat file with instructions to increase the java heap size which is required when dealing with very large input files. After the first run, you can use the .bat file for subsequent runs of the program.

2.2 Instructions for Linux

The default java environment in Linux is the OpenJDK. However, if you are not sure which java environment came with your version of Linux, you should get the latest Sun Java JDK and the Java Runtime Environment (JRE) and then configure your Linux so that it uses the Sun java instead of the default. You can find tips on how to do that for Fedora, RHEL and CentOS at http://www.freetechie.com/blog/installing-sun-java-on-fedora-12/ and for Ubuntu at https://help.ubuntu.com/community/Java

Unzip the .zip file to extract the Fromp-v1.0 folder. Change directory to this folder and type:

```
java -Xms128m -Xmx256m -jar dnarna.jar start
```

This fixes the heap space (or memory required to run FROMP). In case you get a "java.lang.OutOfMemoryError: Java heap space" error while running huge samples, you can try changing the -Xmx option to -Xmx512m depending upon the RAM capacity of your machine.

3 Using FROMP

All data in FROMP is organized in the form of projects. A project can have any number of samples. All projects are saved as .frp files which can be exchanged easily between different computers and users.

3.1 File Menu

The File Menu can be used to create a new project, open an existing project or save an open project.

New Project: `File->New Project->Enter Project Name->Save`

Open Project (.frp file): `File->Open Project->Browse files on computer`

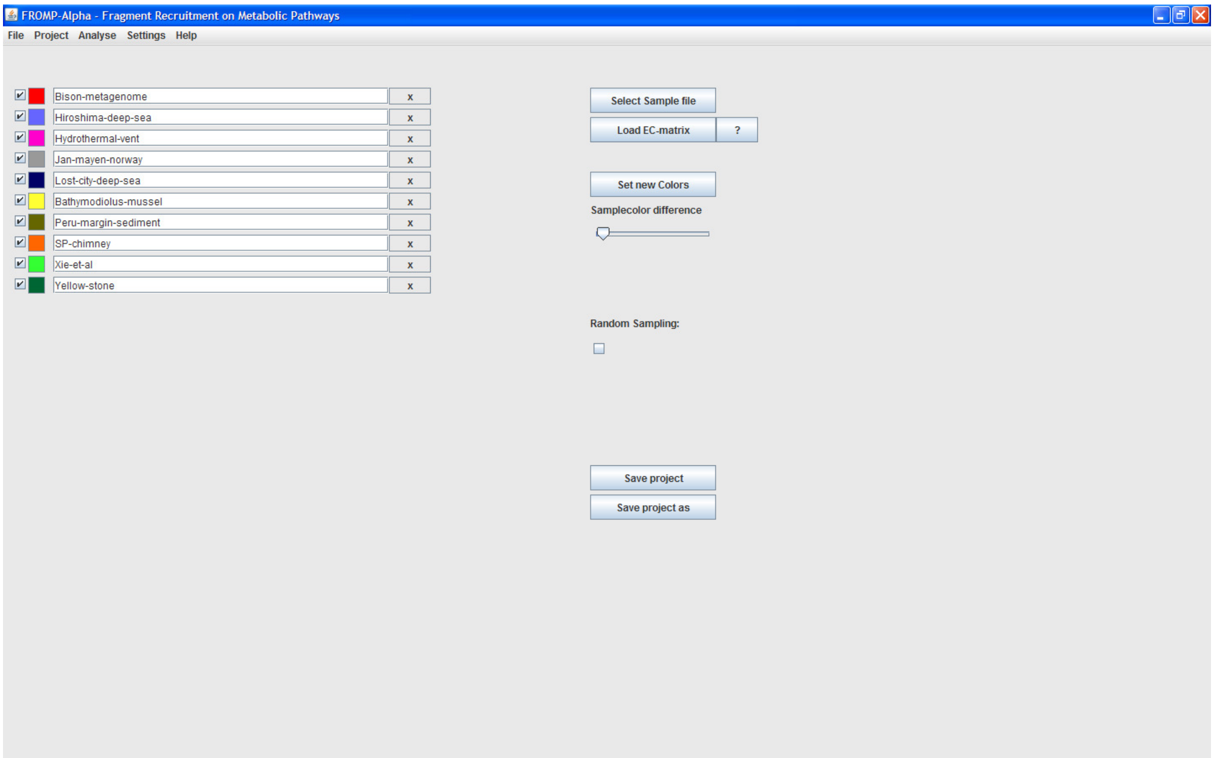
Save project (.frp file): `File->Save Project/Save Project As`

3.2 Project Menu

Here, you can manage samples in a given project (add/remove them to a project, change colors for each sample etc).

3.2.1 Edit Samples

Add samples: Click on the `Select sample` button (Figure 1), browse for the sample files (Check the Input formats 1.2).



☑ Bison-metagenome
☑ Hiroshima-deep-sea
☑ Hydrothermal-vent
☑ Jan-mayen-norway
☑ Lost-city-deep-sea
☑ Bathymodiolus-mussel
☑ Peru-margin-sediment
☑ SP-chimney
☑ Xie-et-al
☑ Yellow-stone

Random Sampling: □

Figure 1: The "`Edit Samples`" window in FROMP.

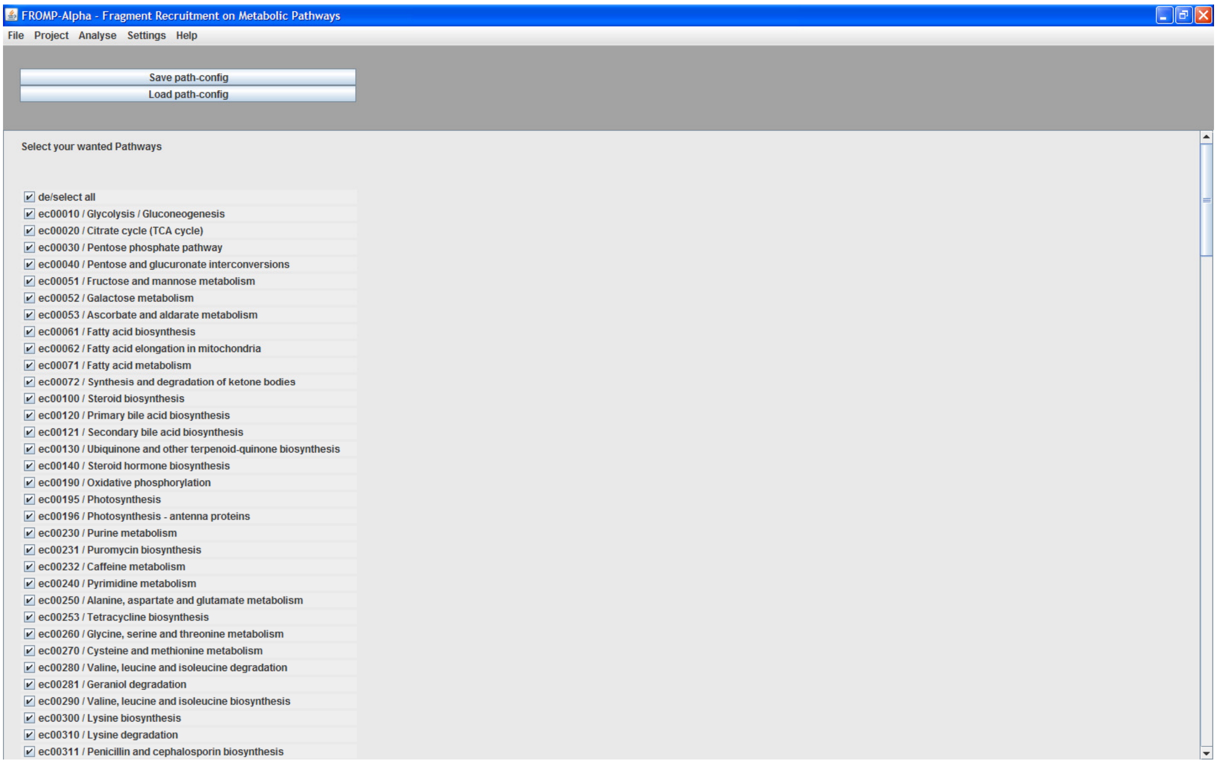
Edit sample names: Click on the text boxes containing the sample names to edit the sample names (Figure 1). These names will be displayed in all results.

Set colors for samples: By default FROMP automatically associates each sample with a varying shade of red 255,0,0 => 200,0,0 => 150,0,0 etc. for each successive sample added. You can increase the difference by changing the `Sample Color difference` meter and then clicking `Set New Colors`. Alternatively, you can click on the color square next to the sample name text box and choose color for each sample individually (Figure 1).

Equalize sample sizes: Clicking the checkbox for `Random Sampling` equalizes the sample sizes. FROMP fixes the size of the smallest sample as the sample size and randomly samples equal numbers of hits from the other samples.

3.2.2 Select Pathways

You can select just a subset of interesting pathways that you want to analyze using this option. The selected pathways can be saved as a text file with extension. cg (`Save path config` button) and a saved selection can be loaded into the program using the `Load path config` button (Figure 2).



Select your wanted Pathways

- ☑ de/select all
- ☑ ec00010 / Glycolysis / Gluconeogenesis
- ☑ ec00020 / Citrate cycle (TCA cycle)
- ☑ ec00030 / Pentose phosphate pathway
- ☑ ec00040 / Pentose and glucuronate interconversions
- ☑ ec00051 / Fructose and mannose metabolism
- ☑ ec00052 / Galactose metabolism
- ☑ ec00053 / Ascorbate and aldarate metabolism
- ☑ ec00061 / Fatty acid biosynthesis
- ☑ ec00062 / Fatty acid elongation in mitochondria
- ☑ ec00071 / Fatty acid metabolism
- ☑ ec00072 / Synthesis and degradation of ketone bodies
- ☑ ec00100 / Steroid biosynthesis
- ☑ ec00120 / Primary bile acid biosynthesis
- ☑ ec00121 / Secondary bile acid biosynthesis
- ☑ ec00130 / Ubiquinone and other terpenoid-quinone biosynthesis
- ☑ ec00140 / Steroid hormone biosynthesis
- ☑ ec00190 / Oxidative phosphorylation
- ☑ ec00195 / Photosynthesis
- ☑ ec00196 / Photosynthesis - antenna proteins
- ☑ ec00230 / Purine metabolism
- ☑ ec00231 / Puromycin biosynthesis
- ☑ ec00232 / Caffeine metabolism
- ☑ ec00240 / Pyrimidine metabolism
- ☑ ec00250 / Alanine, aspartate and glutamate metabolism
- ☑ ec00253 / Tetracycline biosynthesis
- ☑ ec00260 / Glycine, serine and threonine metabolism
- ☑ ec00270 / Cysteine and methionine metabolism
- ☑ ec00280 / Valine, leucine and isoleucine degradation
- ☑ ec00281 / Geraniol degradation
- ☑ ec00290 / Valine, leucine and isoleucine biosynthesis
- ☑ ec00300 / Lysine biosynthesis
- ☑ ec00310 / Lysine degradation
- ☑ ec00311 / Penicillin and cephalosporin biosynthesis

Figure 2: The Select Pathways option in the Edit Menu.

3.2.3 Search Pathways

The Pathways can be searched by either the Pathway ID (e.g. ec00010), the Pathway name (e.g. Glycolysis) or an EC number present within the pathway (e.g. 1.1.1.100). Inputting just a part of the name or EC number also works.

3.3 Analysis Menu

Using this menu you can analyze the samples for their Pathway Completeness Scores (what percentage of a given pathway is recoverable in the sample), Pathway Activity Score (the total hits to all ECs in a given pathway) and the EC Activity (total hits to all ECs in the samples).

3.3.1 Pathway Completeness Score

This refers to the extent to which a pathway can be recovered in a given sample. The purpose of FROMP is to map the EC numbers to Pathways. Some EC numbers participate in multiple pathways and hence are not good indicators of the pathway being actually present in the sample. So we assign weights to ECs based on the number of pathways that they participate in. The weights for the ECs for a given pathway are then summed up for each EC i the weight

$$W_i = \left((N_{(T,i)}/N_{(U,i)})/N_{(P,i)}\right) * \sqrt{L_{(UBC,r)}}$$

where $N_{(T,i)}$ is the total number of ECs in all pathways that have EC i, $N_{(U,i)}$ is the number of unique ECs in all the pathways that have EC i and $N_{(P,i)}$ is the total number of pathways where EC i is present and $L_{(UBC,r)}$ is the total edge-length of the unbranched chain containing EC i in the reference pathway. The pathway completeness score for a pathway p is then

$$C_P = \left(\frac{\sum_{i \in EC_p} W_i * I_i * \sqrt{L_{(UBC,s)}}}{\sum_{i \in EC_p} W_i} \right)$$

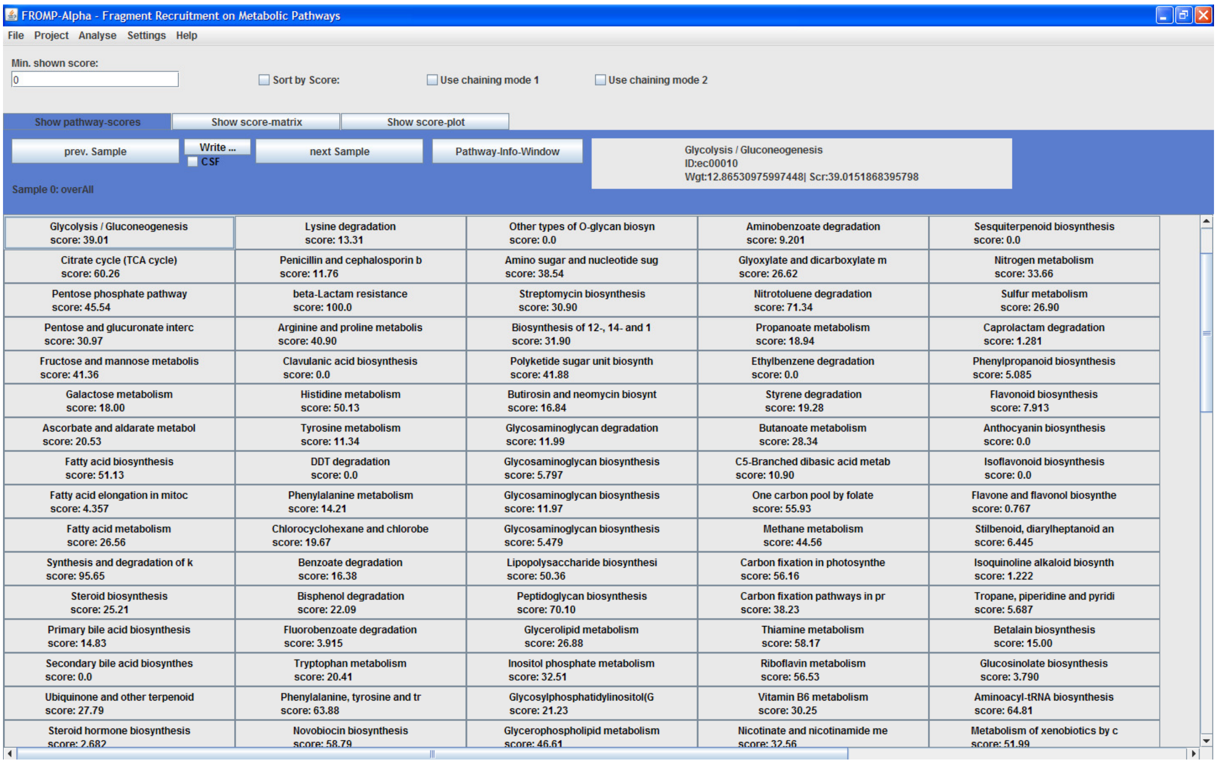
where Wi is the specificity weight of each EC i in pathway p, and Ii is 1 if the EC number is detected in the sample and $L_{(UBC,s)}$ is the edge-length of the unbranched chain containing EC i in the sample. There are three options for Pathway Completeness score calculation:

1. The score without using the unbranched chain information (default).
2. The chain information is used while calculating the weights, which means the unbranched chains of the reference pathways are identified and if an EC in such a chain is present in the sample the weight is multiplied by $\sqrt{(L_{(U\,B\,C,r)})}$. This can be achieved by clicking the `Use chaining mode 1` check-box in the Pathway Completeness Analysis window (3).

3. The chain information is used both in the weights as well as the score calculations. (click the `Use chaining mode 2` check box in 3).

Show Pathway Scores: The `Show pathway scores` tab shows the Pathway Completeness Scores for individual samples. The `next` or `prev.` Sample buttons can be used to navigate between the samples. Clicking on a pathway button for any sample opens the mapping display of the ECs in the sample mapped to this pathway along with the number of hits for each EC. This tab also shows an overall Pathway completeness score where all the pathways in all samples are pooled together. Clicking any pathway button in the Overall window displays the mapping of all the samples on the pathway. The EC boxes in the KEGG pathway maps are colored by the sample colors proportionate to the number of hits for the EC in each sample. The number of hits in all samples for each EC is also displayed as bar-charts and number matrices (5).

The pathways can be sorted according to the Pathway Completeness Scores by checking the `Sort by score` box. A minimum score limit can be set by entering the cut off score in the `Min shown score` box. For example, entering 30 in this box will display only those pathway buttons where the score is ≥ 30 (Figure 3).



Min. shown score: 0

□ Sort by Score:

□ Use chaining mode 1

□ Use chaining mode 2

□ CSF

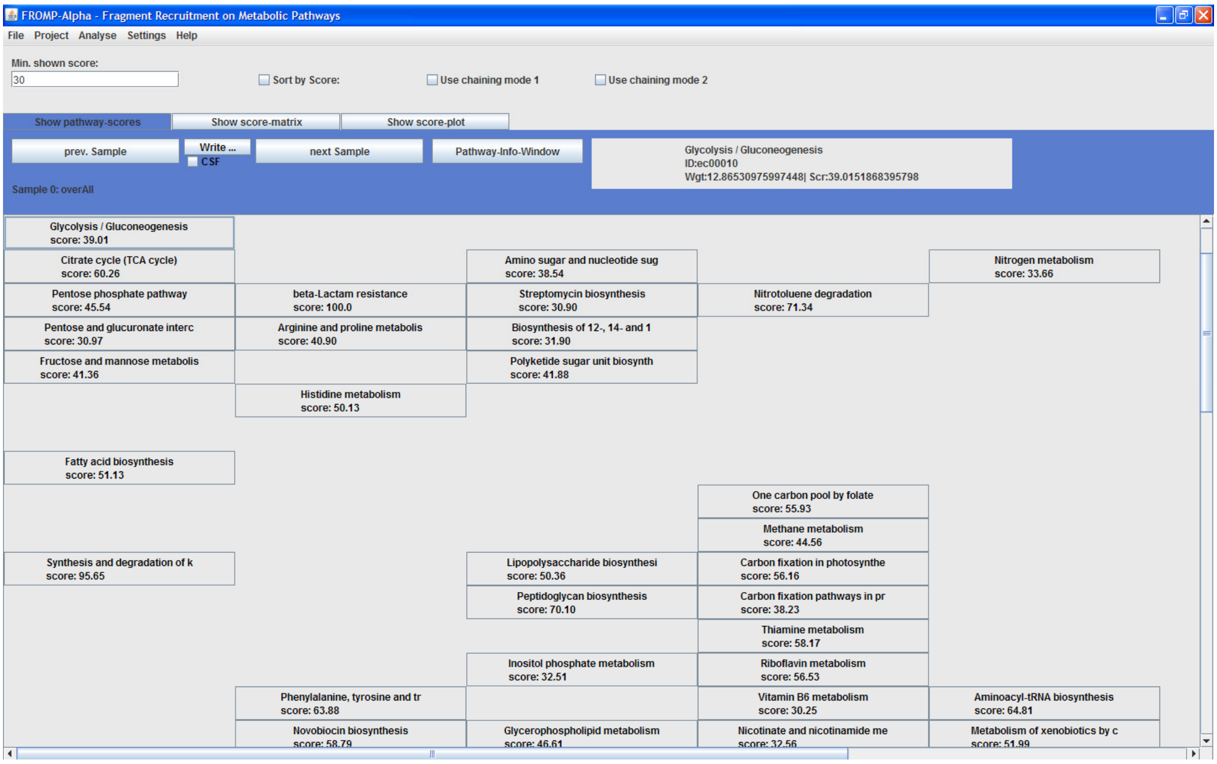
Sample 0: overAll

Glycolysis / Gluconeogenesis
ID:ec00010
Wgt:12.86530975997448| Scr:39.0151868395798

Glycolysis / Gluconeogenesis score: 39.01	Lysine degradation score: 13.31	Other types of O-glycan biosyn score: 0.0	Aminobenzoate degradation score: 9.201	Sesquiterpenoid biosynthesis score: 0.0
Citrate cycle (TCA cycle) score: 60.26	Penicillin and cephalosporin b score: 11.76	Amino sugar and nucleotide sug score: 38.54	Glyoxylate and dicarboxylate m score: 26.62	Nitrogen metabolism score: 33.66
Pentose phosphate pathway score: 45.54	beta-Lactam resistance score: 100.0	Streptomycin biosynthesis score: 30.90	Nitrotoluene degradation score: 71.34	Sulfur metabolism score: 26.90
Pentose and glucuronate interc score: 30.97	Arginine and proline metabolis score: 40.90	Biosynthesis of 12-, 14- and 1 score: 31.90	Propanoate metabolism score: 18.94	Caprolactam degradation score: 1.281
Fructose and mannose metabolis score: 41.36	Clavulanic acid biosynthesis score: 0.0	Polyketide sugar unit biosynth score: 41.88	Ethylbenzene degradation score: 0.0	Phenylpropanoid biosynthesis score: 5.085
Galactose metabolism score: 18.00	Histidine metabolism score: 50.13	Butirosin and neomycin biosynt score: 16.84	Styrene degradation score: 19.28	Flavonoid biosynthesis score: 7.913
Ascorbate and aldarate metabol score: 20.53	Tyrosine metabolism score: 11.34	Glycosaminoglycan degradation score: 11.99	Butanoate metabolism score: 28.34	Anthocyanin biosynthesis score: 0.0
Fatty acid biosynthesis score: 51.13	DDT degradation score: 0.0	Glycosaminoglycan biosynthesis score: 5.797	C5-Branched dibasic acid metab score: 10.90	Isoflavonoid biosynthesis score: 0.0
Fatty acid elongation in mitoc score: 4.357	Phenylalanine metabolism score: 14.21	Glycosaminoglycan biosynthesis score: 11.97	One carbon pool by folate score: 55.93	Flavone and flavonol biosynthe score: 0.767
Fatty acid metabolism score: 26.56	Chlorocyclohexane and chlorobe score: 19.67	Glycosaminoglycan biosynthesis score: 5.479	Methane metabolism score: 44.56	Stilbenoid, diarylheptanoid an score: 6.445
Synthesis and degradation of k score: 95.65	Benzoate degradation score: 16.38	Lipopolysaccharide biosynthesi score: 50.36	Carbon fixation in photosynthe score: 56.16	Isoquinoline alkaloid biosynth score: 1.222
Steroid biosynthesis score: 25.21	Bisphenol degradation score: 22.09	Peptidoglycan biosynthesis score: 70.10	Carbon fixation pathways in pr score: 38.23	Tropane, piperidine and pyridi score: 5.687
Primary bile acid biosynthesis score: 14.83	Fluorobenzoate degradation score: 3.915	Glycerolipid metabolism score: 26.88	Thiamine metabolism score: 58.17	Betalain biosynthesis score: 15.00
Secondary bile acid biosynthes score: 0.0	Tryptophan metabolism score: 20.41	Inositol phosphate metabolism score: 32.51	Riboflavin metabolism score: 56.53	Glucosinolate biosynthesis score: 3.790
Ubiquinone and other terpenoid score: 27.79	Phenylalanine, tyrosine and tr score: 63.88	Glycosylphosphatidylinositol(G score: 21.23	Vitamin B6 metabolism score: 30.25	Aminoacyl-tRNA biosynthesis score: 64.81
Steroid hormone biosynthesis score: 2.682	Novobiocin biosynthesis score: 58.79	Glycerophospholipid metabolism score: 46.61	Nicotinate and nicotinamide me score: 32.56	Metabolism of xenobiotics by c score: 51.99

Figure 3: The "Pathway Completeness Analysis" window in FROMP.

Show Score Matrix: The `Show score matrix` tab shows the pathway completeness scores of all pathways for all samples as a matrix (Figure 6). Again, clicking any pathway button in any sample here will open up the pathway map for that sample. Clicking a pathway in the overall column (the first column, black color by default) will open up the mapping of all the samples on that pathway (similar to Figure 5).



Min. shown score: 30

□ Sort by Score:

□ Use chaining mode 1

□ Use chaining mode 2

□ CSF

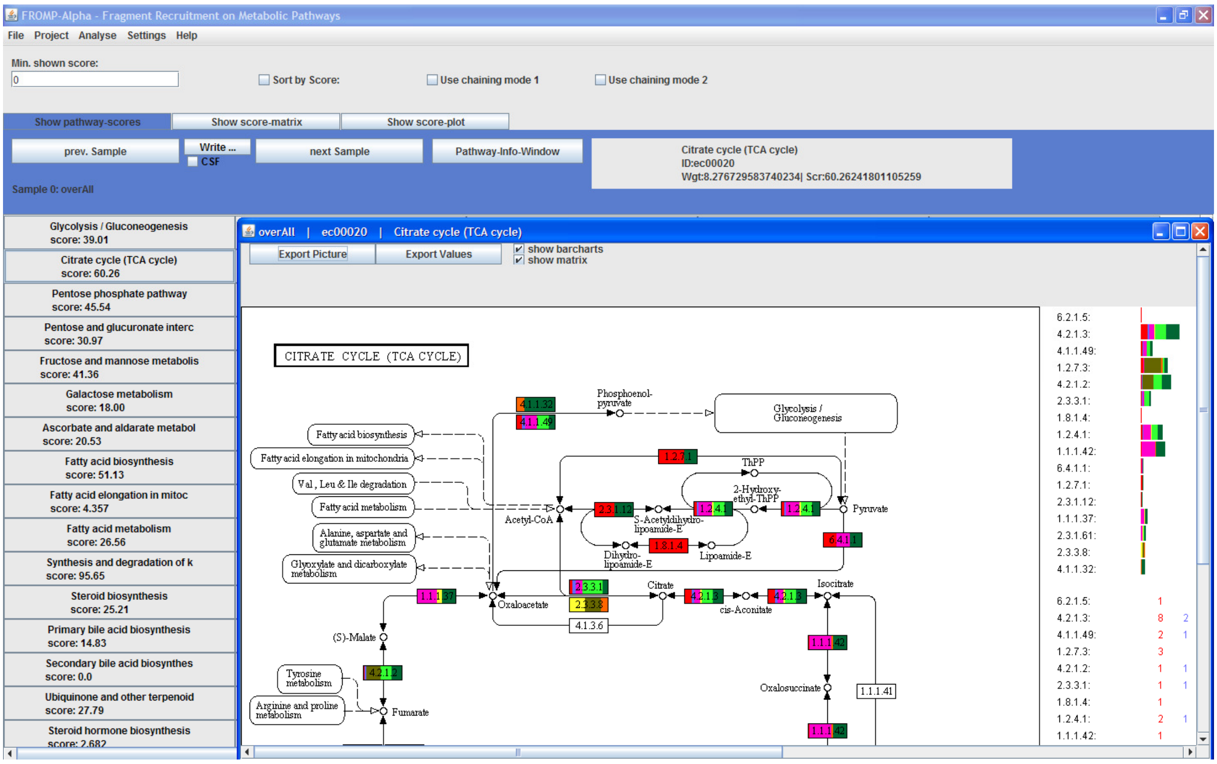
Sample 0: overAll

Glycolysis / Gluconeogenesis
ID:ec00010
Wgt:12.86530975997448| Scr:39.0151868395798

Glycolysis / Gluconeogenesis score: 39.01				
Citrate cycle (TCA cycle) score: 60.26		Amino sugar and nucleotide sug score: 38.54		Nitrogen metabolism score: 33.66
Pentose phosphate pathway score: 45.54	beta-Lactam resistance score: 100.0	Streptomycin biosynthesis score: 30.90	Nitrotoluene degradation score: 71.34	
Pentose and glucuronate interc score: 30.97	Arginine and proline metabolis score: 40.90	Biosynthesis of 12-, 14- and 1 score: 31.90		
Fructose and mannose metabolis score: 41.36		Polyketide sugar unit biosynth score: 41.88		
	Histidine metabolism score: 50.13			
Fatty acid biosynthesis score: 51.13				
			One carbon pool by folate score: 55.93	
			Methane metabolism score: 44.56	
Synthesis and degradation of k score: 95.65		Lipopolysaccharide biosynthesi score: 50.36	Carbon fixation in photosynthe score: 56.16	
		Peptidoglycan biosynthesis score: 70.10	Carbon fixation pathways in pr score: 38.23	
			Thiamine metabolism score: 58.17	
		Inositol phosphate metabolism score: 32.51	Riboflavin metabolism score: 56.53	
	Phenylalanine, tyrosine and tr score: 63.88		Vitamin B6 metabolism score: 30.25	Aminoacyl-tRNA biosynthesis score: 64.81
	Novobiocin biosynthesis score: 58.79	Glycerophospholipid metabolism score: 46.61	Nicotinate and nicotinamide me score: 32.56	Metabolism of xenobiotics by c score: 51.99

Figure 4: The Minimum score cut off feature in "Pathway Completeness Analysis".

On both the `Show Pathway Scores` and the `Show score matrix` tabs, there is an option (clicking the `Write to file` button) to write out the corresponding matrices as TAB separated (default) or Comma separated (by clicking the CSF check-box) text files.



Min. shown score:
0

☐ Sort by Score: ☐ Use chaining mode 1 ☐ Use chaining mode 2

Citrate cycle (TCA cycle)
ID:ec00020
Wgt:8.276729583740234| Scr:60.26241801105259

Sample 0: overAll

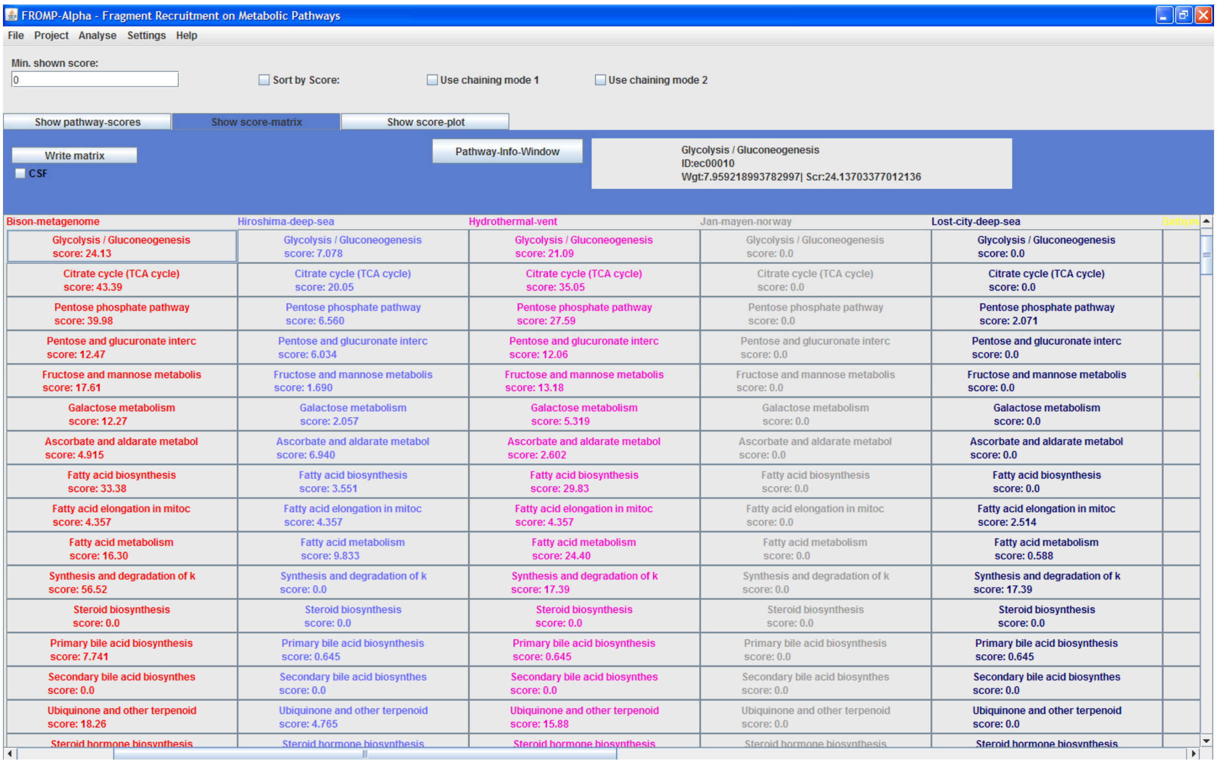
Glycolysis / Gluconeogenesis score: 39.01
Citrate cycle (TCA cycle) score: 60.26
Pentose phosphate pathway score: 45.54
Pentose and glucuronate interc score: 30.97
Fructose and mannose metabolis score: 41.36
Galactose metabolism score: 18.00
Ascorbate and aldarate metabol score: 20.53
Fatty acid biosynthesis score: 51.13
Fatty acid elongation in mitoc score: 4.357
Fatty acid metabolism score: 26.56
Synthesis and degradation of k score: 95.65
Steroid biosynthesis score: 25.21
Primary bile acid biosynthesis score: 14.83
Secondary bile acid biosynthes score: 0.0
Ubiquinone and other terpenoid score: 27.79
Steroid hormone biosynthesis score: 2.682

☑ show barcharts
☑ show matrix

6.2.1.5:
4.2.1.3:
4.1.1.49:
1.2.7.3:
4.2.1.2:
2.3.3.1:
1.8.1.4:
1.2.4.1:
1.1.1.42:
6.4.1.1:
1.2.7.1:
2.3.1.12:
1.1.1.37:
2.3.1.61:
2.3.3.8:
4.1.1.32:

6.2.1.5:	1	
4.2.1.3:	8	2
4.1.1.49:	2	1
1.2.7.3:	3	
4.2.1.2:	1	1
2.3.3.1:	1	1
1.8.1.4:	1	
1.2.4.1:	2	1
1.1.1.42:	1	

Figure 5: Pathway mapping visualization in "Pathway Completeness Analysis".



Min. shown score:
0

☐ Sort by Score: ☐ Use chaining mode 1 ☐ Use chaining mode 2

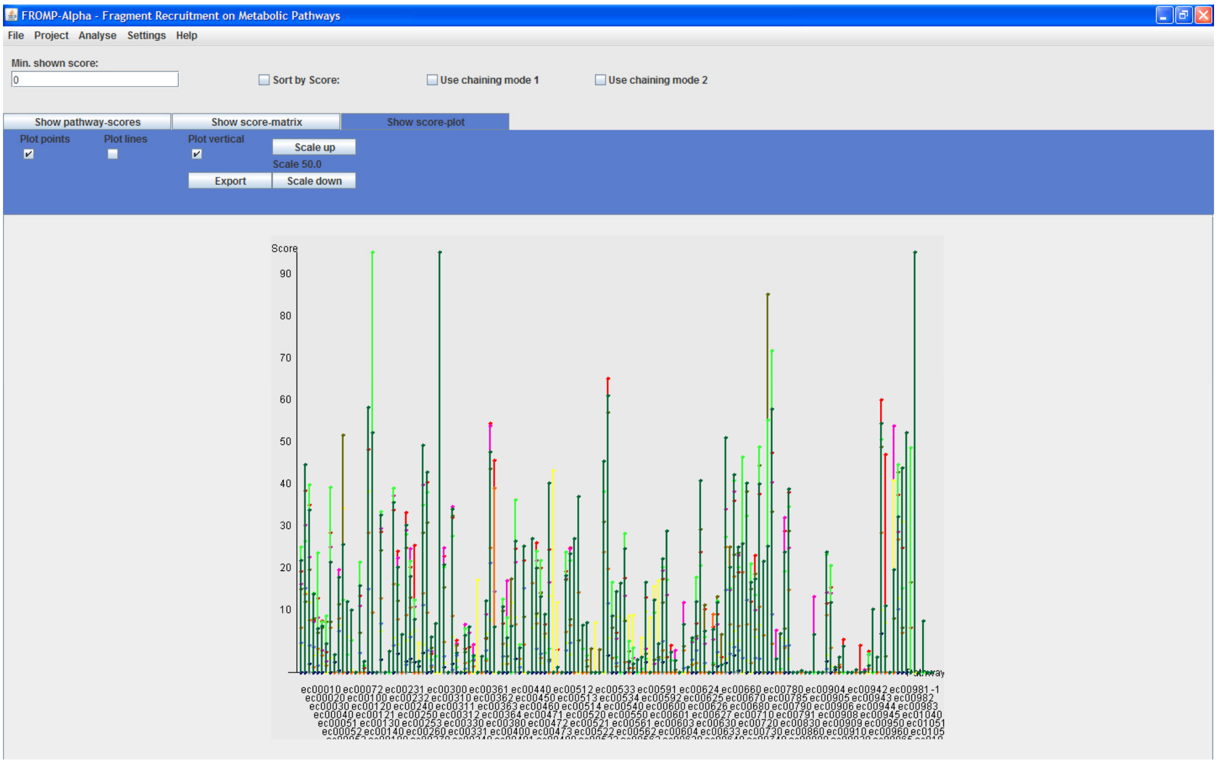
☐ CSF

Glycolysis / Gluconeogenesis
ID:ec00010
Wgt:7.959218993782997| Scr:24.13703377012136

Bison-metagenome	Hiroshima-deep-sea	Hydrothermal-vent	Jan-mayen-norway	Lost-city-deep-sea	Bathym
Glycolysis / Gluconeogenesis score: 24.13	Glycolysis / Gluconeogenesis score: 7.078	Glycolysis / Gluconeogenesis score: 21.09	Glycolysis / Gluconeogenesis score: 0.0	Glycolysis / Gluconeogenesis score: 0.0	
Citrate cycle (TCA cycle) score: 43.39	Citrate cycle (TCA cycle) score: 20.05	Citrate cycle (TCA cycle) score: 35.05	Citrate cycle (TCA cycle) score: 0.0	Citrate cycle (TCA cycle) score: 0.0	
Pentose phosphate pathway score: 39.98	Pentose phosphate pathway score: 6.560	Pentose phosphate pathway score: 27.59	Pentose phosphate pathway score: 0.0	Pentose phosphate pathway score: 2.071	
Pentose and glucuronate interc score: 12.47	Pentose and glucuronate interc score: 6.034	Pentose and glucuronate interc score: 12.06	Pentose and glucuronate interc score: 0.0	Pentose and glucuronate interc score: 0.0	
Fructose and mannose metabolis score: 17.61	Fructose and mannose metabolis score: 1.690	Fructose and mannose metabolis score: 13.18	Fructose and mannose metabolis score: 0.0	Fructose and mannose metabolis score: 0.0	
Galactose metabolism score: 12.27	Galactose metabolism score: 2.057	Galactose metabolism score: 5.319	Galactose metabolism score: 0.0	Galactose metabolism score: 0.0	
Ascorbate and aldarate metabol score: 4.915	Ascorbate and aldarate metabol score: 6.940	Ascorbate and aldarate metabol score: 2.602	Ascorbate and aldarate metabol score: 0.0	Ascorbate and aldarate metabol score: 0.0	
Fatty acid biosynthesis score: 33.38	Fatty acid biosynthesis score: 3.551	Fatty acid biosynthesis score: 29.83	Fatty acid biosynthesis score: 0.0	Fatty acid biosynthesis score: 0.0	
Fatty acid elongation in mitoc score: 4.357	Fatty acid elongation in mitoc score: 4.357	Fatty acid elongation in mitoc score: 4.357	Fatty acid elongation in mitoc score: 0.0	Fatty acid elongation in mitoc score: 2.514	
Fatty acid metabolism score: 16.30	Fatty acid metabolism score: 9.833	Fatty acid metabolism score: 24.40	Fatty acid metabolism score: 0.0	Fatty acid metabolism score: 0.588	
Synthesis and degradation of k score: 56.52	Synthesis and degradation of k score: 0.0	Synthesis and degradation of k score: 17.39	Synthesis and degradation of k score: 0.0	Synthesis and degradation of k score: 17.39	
Steroid biosynthesis score: 0.0	Steroid biosynthesis score: 0.0	Steroid biosynthesis score: 0.0	Steroid biosynthesis score: 0.0	Steroid biosynthesis score: 0.0	
Primary bile acid biosynthesis score: 7.741	Primary bile acid biosynthesis score: 0.645	Primary bile acid biosynthesis score: 0.645	Primary bile acid biosynthesis score: 0.0	Primary bile acid biosynthesis score: 0.645	
Secondary bile acid biosynthes score: 0.0	Secondary bile acid biosynthes score: 0.0	Secondary bile acid biosynthes score: 0.0	Secondary bile acid biosynthes score: 0.0	Secondary bile acid biosynthes score: 0.0	
Ubiquinone and other terpenoid score: 18.26	Ubiquinone and other terpenoid score: 4.765	Ubiquinone and other terpenoid score: 15.88	Ubiquinone and other terpenoid score: 0.0	Ubiquinone and other terpenoid score: 0.0	
Steroid hormone biosynthesis	Steroid hormone biosynthesis	Steroid hormone biosynthesis	Steroid hormone biosynthesis	Steroid hormone biosynthesis	

Figure 6: The matrix display in "Pathway Completeness Analysis".

Show Score Plot: The `Show Score plot` tab shows a graphical plot of the Pathway scores. This plot can be enlarged or minimized (`Scale up` or `Scale down` buttons) and exported as a PNG file (Figure 7).



FROMP-Alpha - Fragment Recruitment on Metabolic Pathways

Min. shown score: 0

□ Sort by Score:

□ Use chaining mode 1

□ Use chaining mode 2

Plot points ☑

Plot lines □

Plot vertical ☑

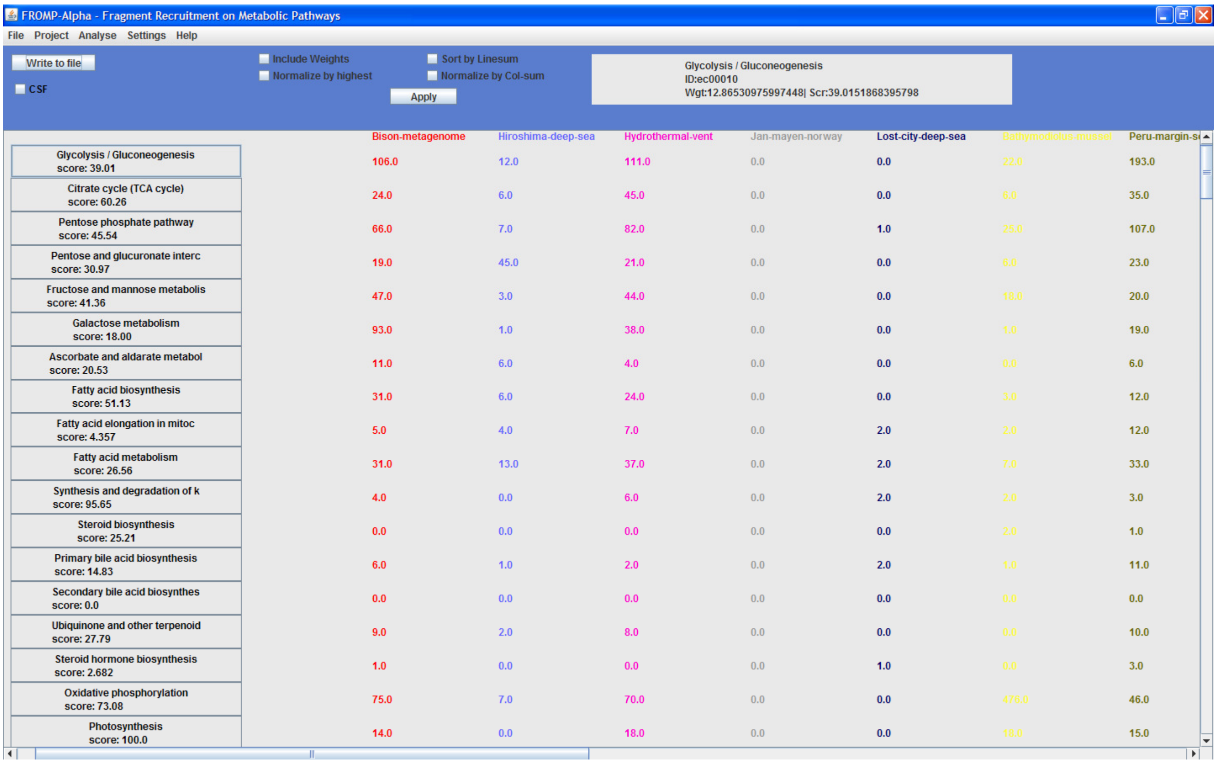
Scale 50.0

File
Project
Analyse
Settings
Help
Show pathway-scores
Show score-matrix
Show score-plot
Scale up
Export
Scale down
Score
90
80
70
60
50
40
30
20
10
Pathway
ec00010 ec00072 ec00231 ec00300 ec00361 ec00440 ec00512 ec00533 ec00591 ec00624 ec00660 ec00780 ec00904 ec00942 ec00981 -1
ec00020 ec00100 ec00232 ec00310 ec00362 ec00450 ec00513 ec00534 ec00592 ec00625 ec00670 ec00785 ec00905 ec00943 ec00982
ec00030 ec00120 ec00240 ec00311 ec00363 ec00460 ec00514 ec00540 ec00600 ec00626 ec00680 ec00790 ec00906 ec00944 ec00983
ec00040 ec00121 ec00250 ec00312 ec00364 ec00471 ec00520 ec00550 ec00601 ec00627 ec00710 ec00791 ec00908 ec00945 ec01040
ec00051 ec00130 ec00253 ec00330 ec00380 ec00472 ec00521 ec00561 ec00603 ec00630 ec00720 ec00830 ec00909 ec00950 ec01051
ec00052 ec00140 ec00260 ec00331 ec00400 ec00473 ec00522 ec00562 ec00604 ec00633 ec00730 ec00860 ec00910 ec00960 ec0105

Figure 7: The Pathway Completeness Score plot.

3.3.2 Pathway Activity

This is simply the sum of counts for the ECs in a given pathway. There is also an option to multiply this score by the EC weight (click on the `Include weights` check-box) to get a weighted Pathway activity score. The Pathway Activity matrix can be sorted by the total Pathway Activity for each pathway (i.e the sum of the rows) by clicking the `Sort by Linesum` check-box (Figure 8).



□ CSF

□ Include Weights

□ Normalize by highest

□ Sort by Linesum

□ Normalize by Col-sum

Glycolysis / Gluconeogenesis
ID:ec00010
Wgt:12.86530975997448| Scr:39.0151868395798

	Bison-metagenome	Hiroshima-deep-sea	Hydrothermal-vent	Jan-mayen-norway	Lost-city-deep-sea	Bathymodiolus-mussel	Peru-margin-s
Glycolysis / Gluconeogenesis score: 39.01	106.0	12.0	111.0	0.0	0.0	22.0	193.0
Citrate cycle (TCA cycle) score: 60.26	24.0	6.0	45.0	0.0	0.0	8.0	35.0
Pentose phosphate pathway score: 45.54	66.0	7.0	82.0	0.0	1.0	25.0	107.0
Pentose and glucuronate interc score: 30.97	19.0	45.0	21.0	0.0	0.0	8.0	23.0
Fructose and mannose metabolis score: 41.36	47.0	3.0	44.0	0.0	0.0	18.0	20.0
Galactose metabolism score: 18.00	93.0	1.0	38.0	0.0	0.0	1.0	19.0
Ascorbate and aldarate metabol score: 20.53	11.0	6.0	4.0	0.0	0.0	0.0	6.0
Fatty acid biosynthesis score: 51.13	31.0	6.0	24.0	0.0	0.0	3.0	12.0
Fatty acid elongation in mitoc score: 4.357	5.0	4.0	7.0	0.0	2.0	2.0	12.0
Fatty acid metabolism score: 26.56	31.0	13.0	37.0	0.0	2.0	7.0	33.0
Synthesis and degradation of k score: 95.65	4.0	0.0	6.0	0.0	2.0	2.0	3.0
Steroid biosynthesis score: 25.21	0.0	0.0	0.0	0.0	0.0	2.0	1.0
Primary bile acid biosynthesis score: 14.83	6.0	1.0	2.0	0.0	2.0	1.0	11.0
Secondary bile acid biosynthes score: 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ubiquinone and other terpenoid score: 27.79	9.0	2.0	8.0	0.0	0.0	0.0	10.0
Steroid hormone biosynthesis score: 2.682	1.0	0.0	0.0	0.0	1.0	0.0	3.0
Oxidative phosphorylation score: 73.08	75.0	7.0	70.0	0.0	0.0	476.0	46.0
Photosynthesis score: 100.0	14.0	0.0	18.0	0.0	0.0	10.0	15.0

Figure 8: The Pathway Activity Analysis.

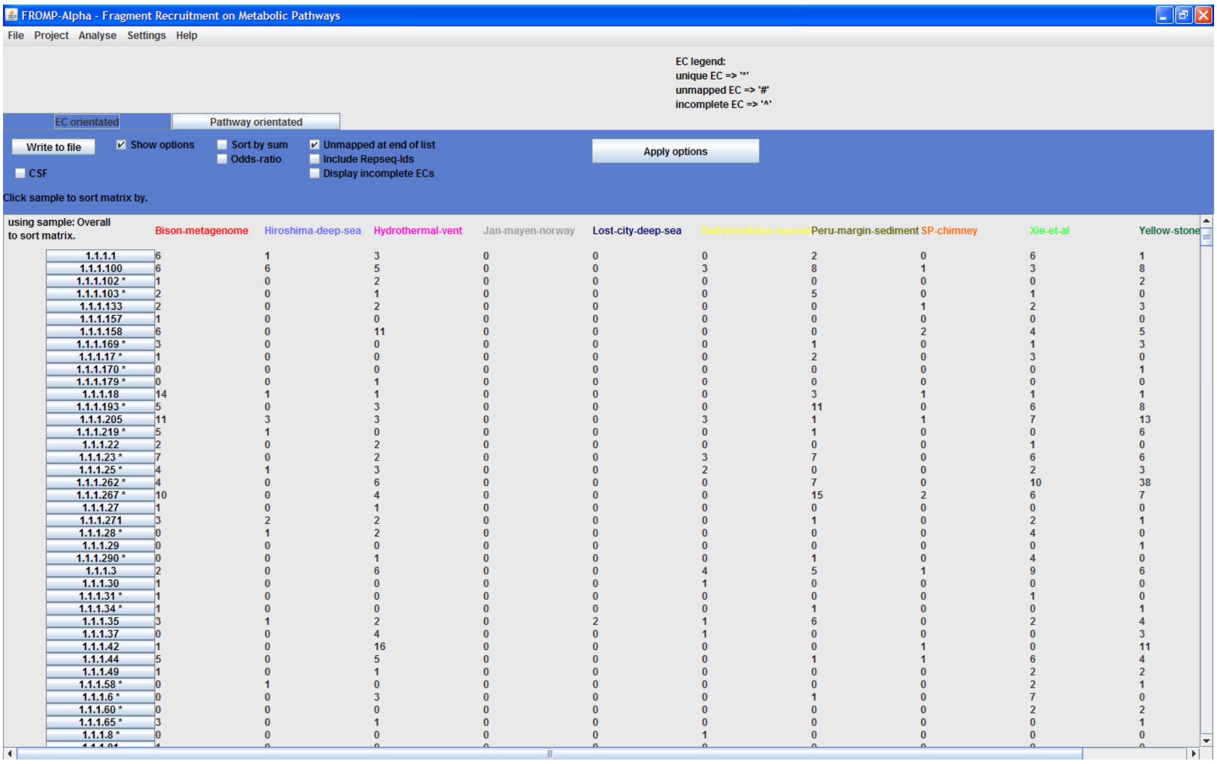
Using the `Normalize by column` or the `Normalize by highest` options, the score can be normalized by the Sample total (total for each column) or the Pathway total (row total) to get a percentage contribution of each pathway to the Sample or a percentage contribution of each sample to the overall pathway activity respectively (Figure 8).

Like the Pathway Completeness analysis this activity matrix can also be written out as a TAB separated (default) or Comma separated (by clicking the `Write to file` button).

3.3.3 EC Activity

This analysis window displays the hit counts for each EC number in each of the samples.

EC orientated view: This is the default view in the EC activity analysis where the hit counts for each EC are displayed for each sample. By default the matrix is sorted by EC number (Figure 9).



EC legend:
unique EC => '*'
unmapped EC => '#'
incomplete EC => '^'

☑ Show options
□ Sort by sum
□ Odds-ratio
☑ Unmapped at end of list
□ Include Repseq-Ids
□ Display incomplete ECs
□ CSF

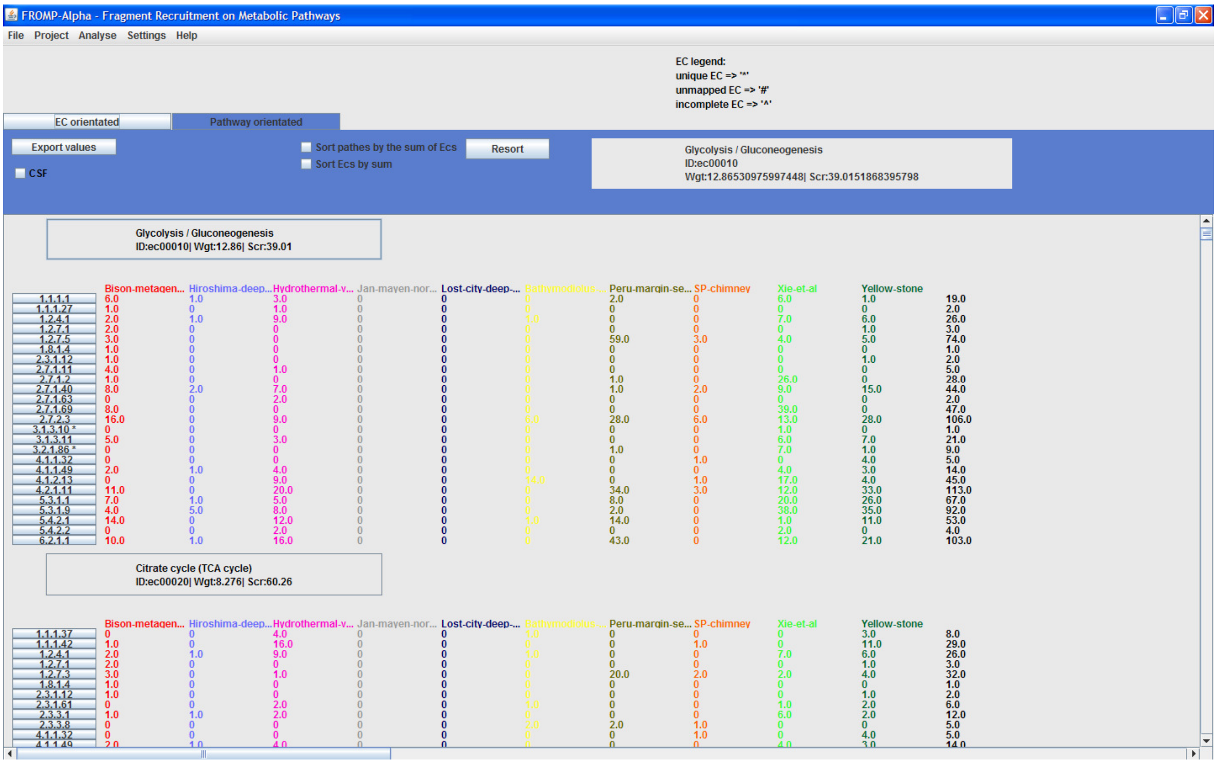
Click sample to sort matrix by.

using sample: Overall to sort matrix.	Bison-metagenome	Hiroshima-deep-sea	Hydrothermal-vent	Jan-mayen-norway	Lost-city-deep-sea	Bathymodiolus-mussel	Peru-margin-sediment	SP-chimney	Xie-et-al	Yellow-stone
1.1.1.1	6	1	3	0	0	0	2	0	6	1
1.1.1.100	6	6	5	0	0	3	8	1	3	8
1.1.1.102 *	1	0	2	0	0	0	0	0	0	2
1.1.1.103 *	2	0	1	0	0	0	5	0	1	0
1.1.1.133	2	0	2	0	0	0	0	1	2	3
1.1.1.157	1	0	0	0	0	0	0	0	0	0
1.1.1.158	6	0	11	0	0	0	0	2	4	5
1.1.1.169 *	3	0	0	0	0	0	1	0	1	3
1.1.1.17 *	1	0	0	0	0	0	2	0	3	0
1.1.1.170 *	0	0	0	0	0	0	0	0	0	1
1.1.1.179 *	0	0	1	0	0	0	0	0	0	0
1.1.1.18	14	1	1	0	0	0	3	1	1	1
1.1.1.193 *	5	0	3	0	0	0	11	0	6	8
1.1.1.205	11	3	3	0	0	3	1	1	7	13
1.1.1.219 *	5	1	0	0	0	0	1	0	0	6
1.1.1.22	2	0	2	0	0	0	0	0	1	0
1.1.1.23 *	7	0	2	0	0	3	7	0	6	6
1.1.1.25 *	4	1	3	0	0	2	0	0	2	3
1.1.1.262 *	4	0	6	0	0	0	7	0	10	38
1.1.1.267 *	10	0	4	0	0	0	15	2	6	7
1.1.1.27	1	0	1	0	0	0	0	0	0	0
1.1.1.271	3	2	2	0	0	0	1	0	2	1
1.1.1.28 *	0	1	2	0	0	0	0	0	4	0
1.1.1.29	0	0	0	0	0	0	0	0	0	1
1.1.1.290 *	0	0	1	0	0	0	1	0	4	0
1.1.1.3	2	0	6	0	0	4	5	1	9	6
1.1.1.30	1	0	0	0	0	1	0	0	0	0
1.1.1.31 *	1	0	0	0	0	0	0	0	1	0
1.1.1.34 *	1	0	0	0	0	0	1	0	0	1
1.1.1.35	3	1	2	0	2	1	6	0	2	4
1.1.1.37	0	0	4	0	0	1	0	0	0	3
1.1.1.42	1	0	16	0	0	0	0	1	0	11
1.1.1.44	5	0	5	0	0	0	1	1	6	4
1.1.1.49	1	0	1	0	0	0	0	0	2	2
1.1.1.58 *	0	1	0	0	0	0	0	0	2	1
1.1.1.6 *	0	0	3	0	0	0	1	0	7	0
1.1.1.60 *	0	0	0	0	0	0	0	0	2	2
1.1.1.65 *	3	0	1	0	0	0	0	0	0	1
1.1.1.8 *	0	0	0	0	0	1	0	0	0	0

Figure 9: The EC orientated tab in EC Activity Analysis.

The following features are available here:

- The `sort by sum` option sorts the entire matrix by the row sum.
- The `Unmapped at end of list` option is always checked by default. This keeps the ECs which cannot be mapped to any pathway at the end of the EC activity matrix. Un-checking this box results in the mapped as well as the unmapped ECs being sorted by EC number.
- `Odds Ratio` calculates the Odds ratio for enrichment of EC numbers in samples as described in (Gill, et al., 2006). If A and C are the occurrence counts of a given EC in sample *i* and all other comparison samples *j*, respectively, and B and D are occurrence counts of all other ECs in sample *i* and comparison samples *j* respectively, then the odds ratio for the given EC in sample *i* is ((A/B)/(C/D)).
- The hit counts can be linked to the corresponding sequence IDs of the hits by checking the `Include Repseq IDs` option.
- To display the hits to incomplete EC numbers (those ECs where all 4 digits in the EC number are not present) in the EC activity matrix, click on the `Display incomplete ECs` check-box.
- Clicking on a Sample Name (the column heading) will sort the matrix by the hit counts in that sample.


FROMP-Alpha - Fragment Recruitment on Metabolic Pathways
File Project Analyse Settings Help
EC orientated
Pathway orientated
Export values
Resort

EC legend:
unique EC => '*'
unmapped EC => '#'
incomplete EC => '^'

□ CSF

□ Sort pathes by the sum of Ecs
□ Sort Ecs by sum

Glycolysis / Gluconeogenesis
ID:ec00010
Wgt:12.86530975997448| Scr:39.0151868395798

Glycolysis / Gluconeogenesis
ID:ec00010| Wgt:12.86| Scr:39.01

	Bison-metagen...	Hiroshima-deep...	Hydrothermal-v...	Jan-mayen-nor...	Lost-city-deep-...	Bathymodiolus-...	Peru-margin-se...	SP-chimney	Xie-et-al	Yellow-stone	
1.1.1.1	6.0	1.0	3.0	0	0	0	2.0	0	6.0	1.0	19.0
1.1.1.27	1.0	0	1.0	0	0	0	0	0	0	0	2.0
1.2.4.1	2.0	1.0	9.0	0	0	1.0	0	0	7.0	6.0	26.0
1.2.7.1	2.0	0	0	0	0	0	0	0	0	1.0	3.0
1.2.7.5	3.0	0	0	0	0	0	59.0	3.0	4.0	5.0	74.0
1.8.1.4	1.0	0	0	0	0	0	0	0	0	0	1.0
2.3.1.12	1.0	0	0	0	0	0	0	0	0	1.0	2.0
2.7.1.11	4.0	0	1.0	0	0	0	0	0	0	0	5.0
2.7.1.2	1.0	0	0	0	0	0	1.0	0	26.0	0	28.0
2.7.1.40	8.0	2.0	7.0	0	0	0	1.0	2.0	9.0	15.0	44.0
2.7.1.63	0	0	2.0	0	0	0	0	0	0	0	2.0
2.7.1.69	8.0	0	0	0	0	0	0	0	39.0	0	47.0
2.7.2.3	16.0	0	9.0	0	0	6.0	28.0	6.0	13.0	28.0	106.0
3.1.3.10 *	0	0	0	0	0	0	0	0	1.0	0	1.0
3.1.3.11	5.0	0	3.0	0	0	0	0	0	6.0	7.0	21.0
3.2.1.86 *	0	0	0	0	0	0	1.0	0	7.0	1.0	9.0
4.1.1.32	0	0	0	0	0	0	0	1.0	0	4.0	5.0
4.1.1.49	2.0	1.0	4.0	0	0	0	0	0	4.0	3.0	14.0
4.1.2.13	0	0	9.0	0	0	14.0	0	1.0	17.0	4.0	45.0
4.2.1.11	11.0	0	20.0	0	0	0	34.0	3.0	12.0	33.0	113.0
5.3.1.1	7.0	1.0	5.0	0	0	0	8.0	0	20.0	26.0	67.0
5.3.1.9	4.0	5.0	8.0	0	0	0	2.0	0	38.0	35.0	92.0
5.4.2.1	14.0	0	12.0	0	0	1.0	14.0	0	1.0	11.0	53.0
5.4.2.2	0	0	2.0	0	0	0	0	0	2.0	0	4.0
6.2.1.1	10.0	1.0	16.0	0	0	0	43.0	0	12.0	21.0	103.0

Citrate cycle (TCA cycle)
ID:ec00020| Wgt:8.276| Scr:60.26

	Bison-metagen...	Hiroshima-deep...	Hydrothermal-v...	Jan-mayen-nor...	Lost-city-deep-...	Bathymodiolus-...	Peru-margin-se...	SP-chimney	Xie-et-al	Yellow-stone	
1.1.1.37	0	0	4.0	0	0	1.0	0	0	0	3.0	8.0
1.1.1.42	1.0	0	16.0	0	0	0	0	1.0	0	11.0	29.0
1.2.4.1	2.0	1.0	9.0	0	0	1.0	0	0	7.0	6.0	26.0
1.2.7.1	2.0	0	0	0	0	0	0	0	0	1.0	3.0
1.2.7.3	3.0	0	1.0	0	0	0	20.0	2.0	2.0	4.0	32.0
1.8.1.4	1.0	0	0	0	0	0	0	0	0	0	1.0
2.3.1.12	1.0	0	0	0	0	0	0	0	0	1.0	2.0
2.3.1.61	0	0	2.0	0	0	1.0	0	0	1.0	2.0	6.0
2.3.3.1	1.0	1.0	2.0	0	0	0	0	0	6.0	2.0	12.0
2.3.3.8	0	0	0	0	0	2.0	2.0	1.0	0	0	5.0
4.1.1.32	0	0	0	0	0	0	0	1.0	0	4.0	5.0
4.1.1.49	2.0	1.0	4.0	0	0	0	0	0	4.0	3.0	14.0

Figure 10: The Pathway orientated tab in EC Activity Analysis.

Pathway orientated view: In this view the EC hits are arranged according to the pathways (Figure 10). Here one can sort the Pathways according to their pathway activity scores (Sort `pathes by sum`) and within each pathway table the EC sub-matrix can be sorted according to the row sum (`Sort ECs by sum`**).**

The EC matrix in both these tabs can be exported out using the `Write to file` button as described earlier.

3.3. Microbial communities in sulfidic ocean waters

Giant hydrogen sulfide plume in the oxygen minimum zone off Peru stimulates high chemoautotrophic carbon dioxide fixation

Harald Schunck[1§] (hschunck@geomar.de)
Gaute Lavik[2§] (glavik@mpi-bremen.de)
Dhwani K Desai[1] (ddesai@geomar.de)
Tobias Großkopf[1] (tgrosskopf@geomar.de)
Tim Kalvelage[2] (tkalvela@mpi-bremen.de)
Carolin R Loescher[3] (cloescher@ifam.uni-kiel.de)
Aurélien Paulmier[2,4] (aurelien.paulmier@gmail.com)
Marc Mußmann[2] (mmussman@mpi-bremen.de)
Moritz Holtappels[2] (mholtapp@mpi-bremen.de)
Sergio Contreras[2] (scontrer@d.umn.edu)
Herbert Siegel[5] (herbert.siegel@io-warnemuende.de)
Philip Rosenstiel[6] (p.rosenstiel@mucosa.de)
Markus B Schilhabel[6] (m.schilhabel@ikmb.uni-kiel.de)
Michelle Graco[7] (mgraco@imarpe.gob.pe)
Ruth A Schmitz[3] (rschmitz@ifam.uni-kiel.de)
Marcel MM Kuypers[2] (mkuypers@mpi-bremen.de)
Julie LaRoche[1,8] (jlaroche@geomar.de)

[1]Helmholtz Centre for Ocean Research Kiel (GEOMAR), Düsternbrooker Weg 20, 24105 Kiel, Germany

[2]Max-Planck-Institute for Marine Microbiology (MPIMM), Celsiusstraße 1, 28359 Bremen, Germany

[3]Institute for General Microbiology (IFAM), Christian-Albrechts-University, Am Botanischen Garten 1-9, 24118 Kiel, Germany

[4]Laboratory of Studies in Geophysics and Space Oceanography, Institute of Research for Development (LEGOS/IRD), 18 Av. Ed. Belin, 31400 Toulouse, France

[5]Leibniz Institute for Baltic Sea Research Warnemünde (IOW), Seestraße 15, 18119 Rostock, Germany

[6]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University, Schittenhelmstraße 12, 24105 Kiel, Germany

[7]Instituto del Mar Perú (IMARPE), Av Gamarra y Gral. Valle s/n, Chucuito, Callao, Peru

[8]Dalhousie University, Department of biology, 1355 Oxford Street, Halifax, Canada

[§]authors contributed equally to the work

Abstract

In Eastern Boundary Upwelling Systems nutrient-rich waters are transported to the ocean surface, fuelling high photoautotrophic primary production. The subsequent heterotrophic decomposition of the produced biomass leads to oxygen-depletion at intermediate water depths, resulting in the formation of oxygen minimum zones (OMZ). These OMZs can sporadically accumulate substantial amounts of hydrogen sulfide, which is toxic to multicellular organisms and have been evoked for massive fish kills.

During a cruise to the OMZ off Peru we found a >8000 square kilometer covering sulfidic plume in shelf waters, which contained ~3.5 x 10^4 tons of hydrogen sulfide. To our knowledge, this is the first time that hydrogen sulfide was measured in the Peruvian OMZ and the largest plume ever reported for ocean waters. To assess the phylogenetic and functional diversity of the inhabiting microbial community, we applied high-throughput sequencing of community DNA and RNA, and analyzed the sequence information in the context of group-specific microbial cell counts, as well as of rate measurements of carbon dioxide fixation and nitrogen transformation processes. Some of the microorganisms previously detected in high abundances in oxygen minimum zone waters were very scarce. Instead, the waters were dominated by several distinct α-, γ-, δ- and ε-proteobacterial taxa associated with either sulfur oxidation or sulfate reduction. Our combined results indicated these chemolithoautotrophic bacteria utilized several oxidants (oxygen, nitrate, nitrite, nitrous oxide and nitric oxide) to detoxify the waters well below the oxic surface. The chemolithoautotrophic activity led to high dark inorganic carbon fixation, representing ~30% of the photoautotrophic carbon fixation.

Postulated changes such as eutrophication and global warming, which may lead to an expansion and intensification of oxygen-depletion, might also increase the frequency of sulfidic waters. The chemoautotrophically fixed carbon could fuel further sulfate reduction and thus potentially stabilize the sulfidic OMZ waters.

Author Summary

Oxygen production through photosynthesis is limited to the light-penetrated ocean surface, while oxygen consumption through respiration occurs at all depths. Waters, in which oxygen consumption exceeds the production, become depleted in oxygen and microorganisms need to switch to metabolic strategies that do not rely on oxygen, e.g. the respiration of nitrate, nitrite or sulfate. The respiration of sulfate results in the formation of hydrogen sulfide, which is highly toxic to multicellular organisms and can have negative impacts on aquatic ecosystems, including large kills of commercially important fish species. Global climate change scenarios predict that oxygen-depletion will intensify in the future and thus the frequency and persistence of hydrogen sulfide plumes may also increase in the future. We analysed the microbial community in a very large sulfidic plume in shelf waters off the Peruvian coast. The dominant microbes were related to those present in hydrothermal vent systems. These abundant microbes gained chemical energy from detoxifying the sulfidic waters. They were further responsible for considerable light-independent carbon dioxide fixation, and were thus producing new biomass within the sulfidic waters. The retention of this fixed carbon might enhance the duration of sulfidic events in addition to the suggested intensification due to anthropogenic global change.

Introduction

Eastern Boundary Upwelling Systems are found along the westward shelfs of the continents in both the Atlantic and the Pacific Ocean. They are characterized by high primary production through photoautotrophy, which is driven by the upwelling of nutrient-rich waters [1]. The produced biomass supports large fish populations in these regions, underlining the importance of Eastern Boundary Upwelling Systems in providing a source of food for mankind [2-6]. However, a significant proportion of the produced biomass also sinks through the water column and is remineralized in subsurface waters, leading to oxygen (O_2) depletion at intermediate depths [7,8]. These oxygen-depleted waters, also referred to as oxygen minimum zones (OMZs), are found in the Eastern tropical North and South Pacific, and to a lesser extent in the Eastern tropical North and South Atlantic [9]. In addition to OMZs found in regions of strong upwelling, oxygen-depleted waters are also present in enclosed basins like the Baltic and the Black Sea, as well as in the northern Indian Ocean [10].

The OMZ off Peru, Chile and Ecuador in the South Pacific Ocean is the largest oceanic area where O_2 concentrations are reported to fall below the detection limit of oxygen sensors (~10-100 nM) [11-14]. In the absence of O_2, organic carbon degradation has been primarily attributed to heterotrophic denitrification, the reduction of nitrate (NO_3^-) to dinitrogen gas (N_2) [15-17]. However, *in situ* experiments show only minor evidence for active heterotrophic denitrification in most OMZ waters. Instead, numerous studies have demonstrated that anammox, the anaerobic oxidation of ammonium (NH_4^+) with nitrite (NO_2^-) to N_2, is responsible for the major loss of fixed nitrogen from OMZ waters off Namibia [18], off Oman [13], off Peru [19,20] and off Chile [21]. As anammox is an autotrophic process, its dominance over heterotrophic denitrification questions our understanding of organic matter remineralization in OMZ waters. This is in line with the hypothesis that O_2-depletion leads to a shift of the microbial community from organoheterotrophs to chemolithotrophs [22,23].

Hydrogen sulfide (H_2S) is formed upon microbial SO_4^{2-} reduction, which is commonly occurring in anoxic marine sediments, where it is considered to be the main heterotrophic process for the degradation of organic carbon [24]. The build-up of high concentrations in the water column was mainly explained with the release of large quantities of H_2S from underlying sediments [25-28]. However, it has also been suggested that H_2S build-up in ocean waters could be caused by SO_4^{2-} reduction within the water column [29,30]. A recent study suggested that an active, but cryptic sulfur cycle is present in non-sulfidic subsurface waters in the eastern tropical South Pacific OMZ off northern Chile [12]. According to this hypothesis, sulfate (SO_4^{2-}) reduction and consequently H_2S formation takes place in the water column, even when the pool of thermodynamically more favourable electron acceptors like NO_3^- or NO_2^- is not yet depleted. However, the resulting H_2S would be rapidly re-oxidized to elemental sulfur (S^0) or SO_4^{2-}, such that the two processes are in steady-state and H_2S does not accumulate [26,31].

Indeed, chemolithoautotrophic γ-proteobacteria involved in sulfur cycling (e.g. related to the uncultured SUP05 cluster bacterium (SUP05)) were detected in non-sulfidic OMZ waters with high-throughput sequencing, although in low abundances [12,32]. Studies conducted during occurrences of sulfidic events in the Benguela Current upwelling OMZ and in a seasonally anoxic fjord in Canada suggested that these γ-proteobacteria are actually much more abundant in sulfidic waters and that they detoxified the waters via the chemolithotrophic oxidation of H_2S coupled to the reduction of NO_3^- [26,33], a reaction termed sulfur-driven autotrophic denitrification. In addition, chemolithoautotrophic α- and ε-proteobacteria have been detected in the sulfidic waters in the Benguela Current upwelling OMZ, which could therefore be important members of microbial communities in sulfidic plumes [26].

The initiation, termination and frequency of sulfidic events in oceanic OMZs are so far poorly understood, and H_2S in the water column is so far mostly known from isolated and enclosed basins like the Baltic Sea [34-36], the Black Sea [37-39], the Cariaco Basin off Venezuela [40,41] and the Saanich Inlet in Canada [33,42]. In the oceans, sulfidic waters have rarely been measured [26,30], although there are studies mentioning the characteristic odor of H_2S

[29] and anecdotal reports of Peruvian fishermen on ‘black’ fishing gear in relation to the so-called ‘aguajes’ conditions for the OMZ off Peru.

Potential negative consequences on fish stocks and quality of life along the populated coastal upwelling regions are severe, because H_2S is highly toxic to animals and humans and has already been invoked as the cause for occasional massive fish kills in African shelf waters [43-45]. The anticipated decrease in O_2 concentrations and the increase in water column stratification, as predicted from global change [9], as well as local eutrophication [22,30], might lead to more frequent and intense depletion of O_2 and of alternate electron acceptors (e.g. NO_3^- and NO_2^-) and thus favour the development of sulfidic waters within OMZs [26]. Given that the detoxification of sulfidic water is a microbial process, it is important to assess the metabolic response of the endemic microbial community to the accumulation of H_2S.

We explored the microbial community structure and its transcriptional activity with high-throughput metagenomic and metatranscriptomic sequencing and present rate measurements of carbon dioxide (CO_2) fixation and nitrogen-transformation processes. Total bacterial cell counts with flow-cytometry and counts of specific phylogenetic groups performed with fluorescence *in situ* hybridization (FISH) were used to guide the interpretation of the sequence data and the rate measurements. Several of the proteobacterial taxa that were dominant in the sulfidic waters were expressing genes involved in the sulfur cycle, which reflected several different metabolic strategies for H_2S oxidation. We further show that these organisms were responsible for considerable light-independent CO_2 fixation.

Results and discussion

Description of the sampling site

During RV Meteor cruise M77/3 on the Peruvian shelf (December 27th to January 24th, 2009) we found sulfidic waters stretching from Lima to Paracas Natural Reserve southwest of Pisco (Figure 1). O_2 concentration in shelf waters in the study area were generally below the detection limit at water depth below 20 m (from 12°S to 14°S; Figure 1B), while NO_x (the sum of NO_3^- and NO_2^-) was heavily depleted in the water column throughout the transect from 20-60 m to the bottom (from 12° 20’S to 13° 30’S; Figure 1C), mirror-imaging the distribution of H_2S (Figure 1D). Sulfidic waters were first detected on January 9th south of Lima and seemed to have persisted until the end of the cruise, when H_2S-containing waters covered ~8000 km² of the shelf. The thickness of the sulfidic layer was on average 80 m (Figure 1D), yielding by far the largest sulfidic plume (~640 km³) ever reported for oceanic waters. In total, we calculated a H_2S content of approximately 3.5 x 10^4 tons. The total area affected by H_2S may have been even larger, as we did not map the extent of the sulfidic plume into the inner territorial waters of Peru and into the protected area of the Paracas Natural Reserve.

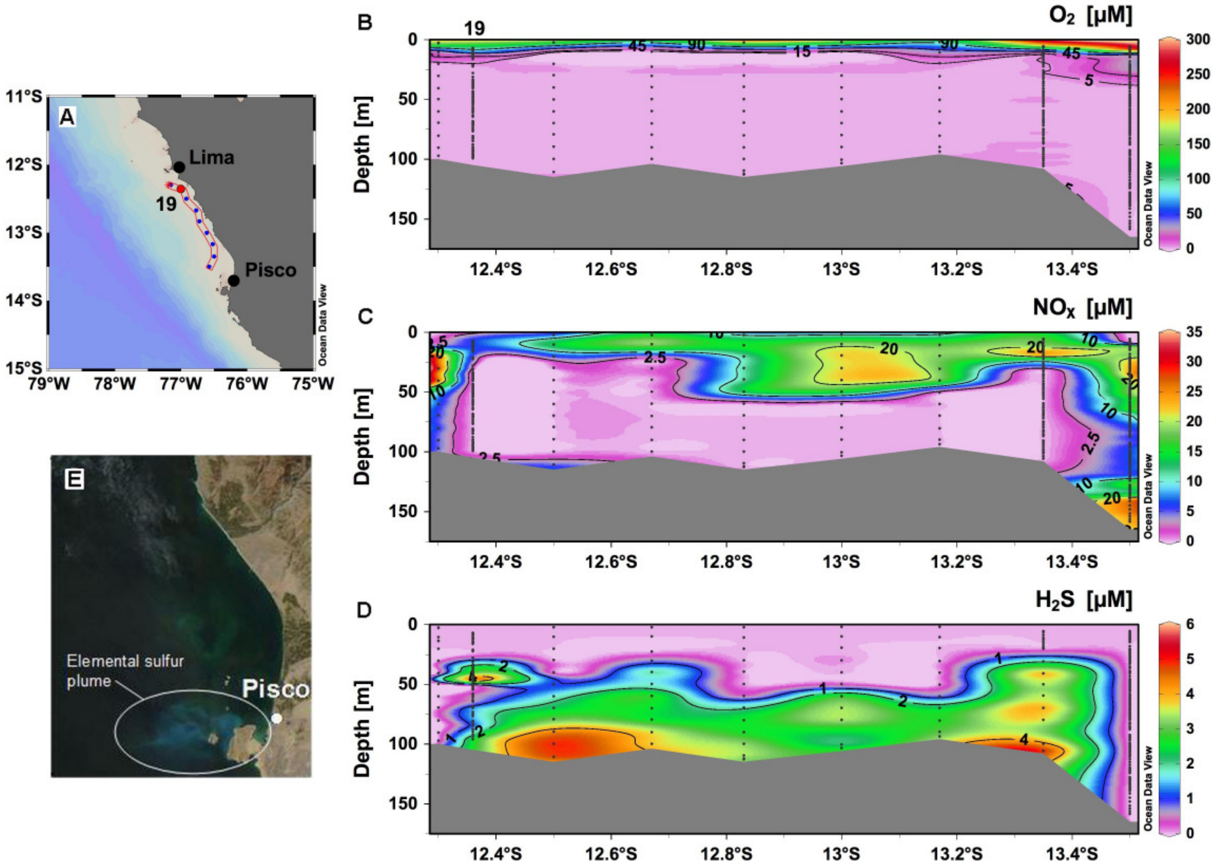


Figure 1: Extent of the sulfidic plume off the Peruvian coast. (A) Areal view of stations sampled within the plume. The station (19) for the detailed analysis is marked with a red dot. (B) Vertical distribution of O_2 concentrations. (C) Vertical distribution of NO_x (the sum of NO_3^- and NO_2^-) concentrations. (D) Vertical distribution of H_2S concentrations. (E) Satellite image (MODIS) showing the elemental sulfur path on May 7th, 2009.

Remote satellite sensing revealed large patches (50-150 km²) of turquoise discoloured surface-waters, attributable to the formation of colloidal sulfur upon H_2S oxidation [46,47] in Paracas Natural Reserve as well as off Lima during our sampling campaign (Figure S1). The larger extension of H_2S in deeper waters when compared to the colloidal sulfur in the surface indicated that most of the sulfur was oxidized in subsurface waters, similar to the observations from the Benguela upwelling system [26]. Colloidal sulfur clouds measuring up to 1000 km² were observed in the same region in May 2009, indicating that the sulfidic event detected in January 2009 either persisted for several months or recurred (Figure 1E). This suggests that the occurrence or duration of sulfidic waters in the OMZ off Peru might be more widespread and frequent than originally thought.

A vertical profile of the sulfidic water column (station 19), sampled during the upcast with a pump CTD system on January 9th, 2009 at a site located approximately 15 km offshore Lima (12° 21.88'S, 77° 0.00'W, 100 m water depth, Figure 1A) was the target of a detailed analysis. The surface mixed layer was shallow with the thermocline at about 10 m water depth (Figure 2D).

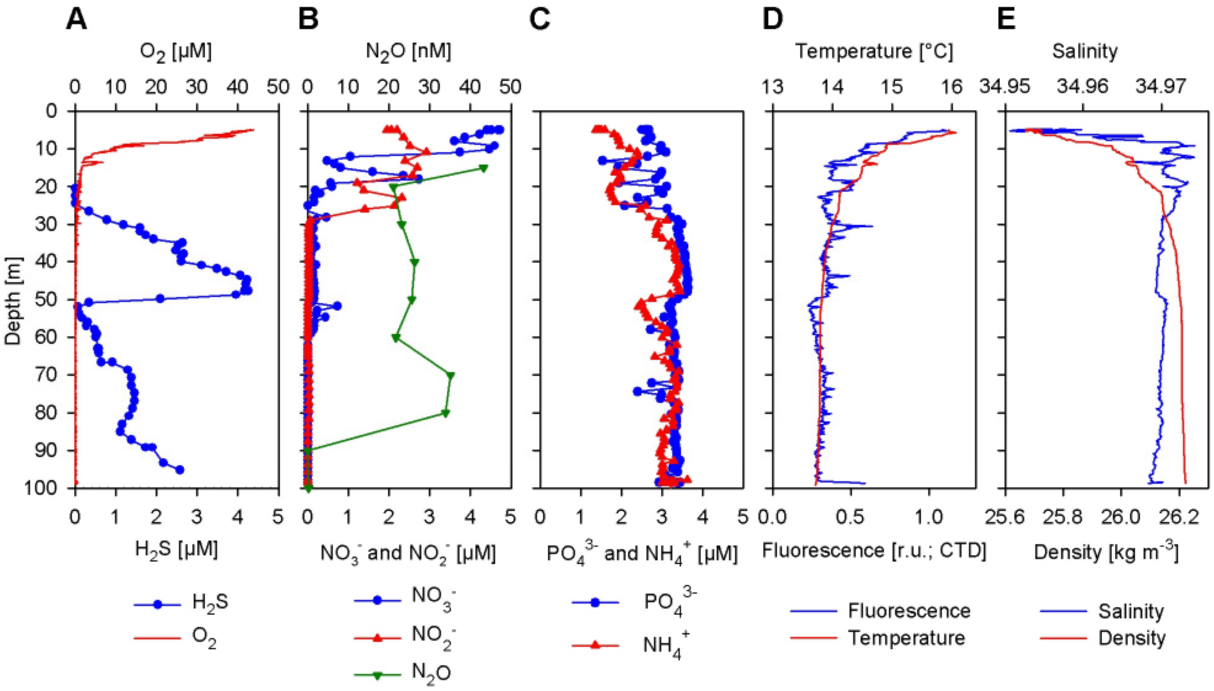


Figure 2: Vertical distribution of physical, chemical and biological water properties. (A) Concentrations of H_2S and O_2. (B) Concentrations of NO_3^-, NO_2^- and N_2O. (C) Concentrations of PO_4^{3-} and NH_4^+. (D) *In situ* fluorescence (chlorophyll, relative units as measured with the pump CTD) and temperature. (E) Salinity and density.

The surface temperature (>16°C) was only ~2°C warmer than the bottom waters and the salinity (34.95-34.97) changed merely slightly with depths, which indicated an active upwelling of subsurface waters. Even the surface waters were characterized by low O_2 conditions, which is reflected in 40 µM O_2 (or ~15% saturation, Figure 2A). During the downcast O_2 decreased at the thermocline and dropped to about the detection limit (0.5-1 µM) of our microsensor at about 20 m. Nevertheless, trace amounts of O_2 (< 1 µM) were still detected with some variability down to ~40 m water depth. These low concentrations of O_2 were close to detection limit of our sensor and we cannot rule out that this was due to water advection caused by the CTD rosette or a memory effect of the sensor. However, using a highly-sensitive self calibrating STOX sensor (Switchable Trace amount OXygen, detection limit ~50 nM) [11,14] during the upcast of the CTD rosette, O_2 was undetectable below 20 m, and therefore we defined this zone as anoxic. However, we detected large vertical movement in the oxycline from 5-18 m due to internal waves, which may have caused non steady state conditions and induced a flux of O_2 down into the anoxic waters.

Phosphate (PO_4^{3-}) and NH_4^+ concentrations remained high (both around 3 µM) and relatively stable throughout the water column, with NH_4^+ only having a minor drop in concentrations near 50 m (Figure 2C). NO_3^- concentrations (detection limit 0.1 µM) in the water column were lower as expected from upwelling regions and also compared to the southern part of the study area. Highest concentrations were measured in surface waters with 5 µM, but then rapidly dropping below 1 µM at the oxycline (19 m; Figure 2B). Detectable concentrations of NO_3^- (ranging from 0.1-0.2 µM) were measured down to 59 m with the exception of 52-55 m, where a small increase up to 0.7 µM was observed. NO_2^- (detection limit 0.01 µM) was high (1.5-3 µM) from surface waters down to 26 m. Trace concentrations (~25 nM) were measurable down to 50 m and again from 67-81 m. N_2O concentrations ranged between ~20-40 nM from 15-80 m and dropped below the detection limit closer to the bottom of the water column (Figure 2B). H_2S was first detected (with both microsensor and wet chemistry) at 26 m and increased steadily, reaching a concentration of 4.2 µM at 48 m (Figure 2A). This maximum was followed by a rapid drop in concentrations to below 0.1 µM at 52-53 m, before they increased again to about 2.6 µM at 95 m, approximately 5 m above the sediment. H_2S concentrations directly at the sediment-water interface were probably even greater, but have not been measured in this study.

Phylogenetic diversity of the microbial community

Based on the monitoring of O_2, NO_2^- and H_2S concentrations during the upcast, we defined three zones within the water column, where we carried out a detailed sampling: the oxic surface (5 m sample, where sampling was stopped when internal waves decreased O_2 concentrations to below 30 µM), the upper boundary of the anoxic zone (15 and 20 m samples) and the sulfidic zone (30, 40, 50, 60, 80 and 100 m samples). Both a hierarchical clustering approach and a statistical analysis of the taxonomic assignments indicated that the selected sample groups were justified (an ANOSIM test using a Bray-Curtis distance measure showed a Global R value of 0.83 and a significance level of 0.1%, Figure S2).

Using a 98% similarity cut off, the metagenomes and metatranscriptomes accounted for an average of 263,606 (DNA) and 98,785 (RNA) unique sequences per depths (Table S1). A total of 4809 (DNA) and 3872 (RNA) different taxa were identified using BLAST-searches, revealing a highly diverse microbial community over all depths. However, a large percentage of all non-ribosomal sequences found in both the metagenomes and metatranscriptomes had no significant match against the non-redundant database of NCBI (Figure 3). On average, 49% of the sequences remained unidentified, which is comparable to other studies that utilized high-throughput sequencing technologies in marine habitats [32,48-50].

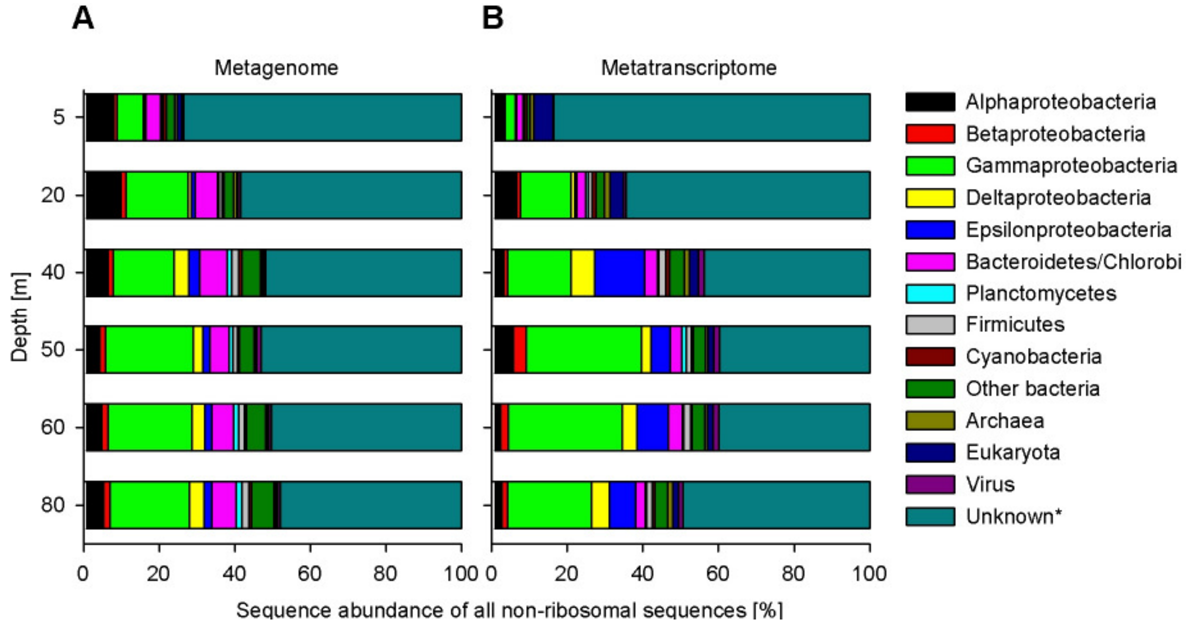


Figure 3: Vertical distribution of taxonomic assignments of the sequence data. Shown on either domain, phylum or class level in percent of all non-ribosomal (A) DNA and (B) RNA sequences. 'Other Bacteria' include Acidobacteria, Actinobacteria, Aquificae, Chlamydiae, Chloroflexi, Deferribacteres, Deinococcus-Thermus, Dictyoglomi, Elusimicrobia, Fibrobacteres, Fusobacteria, Gemmatimonadetes, Lentisphaerae, Nitrospirae, Spirochaetes, Synergistetes, Tenericutes, Thermotogae and Verrucomicrobia.

***Sequences with no significant match against the non-redundant database of NCBI or sequences with a match that lacks taxonomic information**

The community structure presented a relatively stable and uniform distribution at the phylum-level, especially within the sulfidic zone (Figure 3). The metagenome data suggested that the microbial community was overall dominated by proteobacteria (16.6-34.1% of all non-ribosomal sequences (including the unidentified sequences)), while the *Bacteroidetes*/*Chlorobi*-group was the second largest group we could identify (3.9-7.4%). In the oxic and anoxic waters, both α- and γ-proteobacterial sequences were abundant (6.9-16.4%), similar to previous findings in the OMZ off northern Chile [51]. In sulfidic waters, γ-proteobacteria were clearly dominating (up to 23.2 %) and we further found a significant increase in the frequencies of δ- and ε-proteobacterial sequences (1.9-4.0%), which were much less abundant in the 5 and 20 m samples.

The metatranscriptomes showed a more variable picture of the microbial community. In surface waters eukaryotic sequences displayed the largest identifiable group (5%), while in all other depths γ-proteobacteria were dominating (13.3-30.5%). In sulfidic waters ε-proteobacterial transcripts were further identified in high, but also variable proportions (5.0-13.2%), when compared to the metagenome. Notably, "other bacteria", summarizing 19 bacterial phyla, were present at all depth, but never exceeded 5.8% of the sequences, in both the metagenome and the metatranscriptome datasets (Figure 3).

A more detailed analysis of the metagenomes (on species level) revealed that the oxic surface waters harboured several different phototrophic organisms (Figure S3). Prokaryotes similar to *Candidatus* Pelagibacter sp. HTCC7211 accounted for 1.6% and relatives of *Candidatus* Pelagibacter sp. HTCC1002 made up 0.4% of all DNA-sequences, which is in agreement with other studies conducted in OMZs [12,32]. *Synechococcus spp.*, which are also known to be present in OMZ-waters [52] accounted for about 0.4% of the sequences. The most abundant single taxon identified in the oxic surface metagenome had high similarity to the (non-phototrophic) uncultured SUP05 cluster bacterium (1.9%), a chemolithoautotrophic sulfur-oxidizer, which has previously been detected in sulfidic waters [26,33,51,53]. In the metagenome sample from the anoxic zone (20 m), SUP05 was also the dominant taxon with 6.6% of all DNA-sequences. At 20 m and below, also other γ-proteobacterial sulfur oxidizers (GSO) became increasingly abundant. Further common GSOs were related to *Candidatus* Ruthia magnifica str. Cm (2.6%) and to *Candidatus* Vesicomyosocius okutanii HA (1.4%), which are both gill symbionts of deep-sea hydrothermal-vent clams [54,55]. This indicated that the GSO-community at our sampling site was composed of at least three separate taxa. In the sulfidic zone, the dominance of the GSO-group was even higher, reaching a maximum of 17% of all DNA-sequences at 50 m (Figure S3). Other common microorganisms in the metagenome were similar to the ε-proteobacterium *Sulfurovum* sp. NBC37- 1 (up to 1.7%) and to the δ-proteobacterium *Desulfobacterium autotrophicum* HRM2 (up to 1.4%) [56,57]. In all sulfidic depths, organisms related to SUP05, *R. magnifica*, *V. okutanii*, *Sulfurovum* and *D. autotrophicum* were the five most abundant organisms as identified from BLAST-searches.

In contrast to the metagenomes, the metatranscriptomes reflect the suite of genes that were expressed in the entire microbial community and therefore displayed a much more variable picture of the microbial community, reflecting to some extent its actual metabolic activity. In the oxic surface waters, we found RNA-sequences similar to the archaeal ammonia-oxidizer *Nitrosopumilus maritimus* SCM1 and the α-proteobacterium *Magnetospirillum gryphiswaldense* to comprise the largest single taxa (0.7 and 0.4%). *N. maritimus* is considered one of the 'classical' OMZ-inhabitants and has been detected in large proportions in OMZ-waters before [12,32]. At 20 m, RNA-sequences similar to the γ-proteobacteria *Marinomonas* sp. MWYL1 and *Neptuniibacter caesariensis* were most abundant (1.3 and 1.2%).

Similar to the metagenome assignments, the GSO-group was predominant in the metatranscriptomes throughout the sulfidic zone. However, the composition of the GSO-group was variable and changed with depths. While SUP05 was virtually undetectable using BLAST-searches in the 5, 20 and 40 m metatranscriptomes (0-0.1 %; Figure S3), GSO-related RNA-sequences at these depths could be almost exclusively assigned to the relatives of *R. magnifica* and *V. okutanii* (more than 80% of all GSO-sequences). In all other sulfidic metatranscriptomes (50, 60 and 80 m), sequences similar to SUP05 presented the most abundant single taxon we could identify (5.5-11%).

Overall and similar to the metagenomes, organisms related to SUP05, *R. magnifica*, *V. okutanii*, *Sulfurovum* and *D. autotrophicum* were the five most abundant organisms detected in the metatranscriptomes within sulfidic waters (except at 40 m). We also identified several other taxa in the RNA-dataset, which were less abundant in the metagenomes, whereby some of these are known to inhabit hydrothermal vent systems. At 40 m, organisms related to the ε-proteobacterium *Arcobacter butzleri* RM4018 accounted for 2% of all RNA-sequences and at 50 m the γ-proteobacterium *Colwellia psychrerythraea* 34H for 3%. In deeper part of the water column (60 and 80 m), organisms similar to the ε-proteobacteria *Campylobacterales bacterium* GD 1 (1.1-1.3%) and *Arcobacter nitrofigilis* DSM 7299 (0.9-1%) were detected. Although sequences similar to the anammox-planctomycete *Candidatus* Kuenenia stuttgartiensis were found in relatively high abundances in OMZ waters [12,32], they were relatively rare within our samples, never exceeding 0.7% for DNA and 0.8% for RNA.

We further conducted microscopic cell-counts over the vertical profile with FISH-probes targeting specific proteobacterial groups. A comparison of the FISH counts to the sequence estimates is shown in Figure 4 and S4. The most abundant group identified in the anoxic and sulfidic depths with FISH were γ-proteobacteria (34-40% of all DAPI-stained cells; Figure 4A). This is in good agreement with our metagenomic assignments, which also found γ-proteobacteria to be the dominating group (16-23.2%). Hence, with the metagenomes, we found about the equivalent of 50% of the γ-proteobacteria detected with FISH-counts.

Targeting only the GSO-group (mostly SUP05, *R. magnifica* and *V. okutanii*), both approaches yielded very similar estimates. The metagenome assignments range from 9.3-17.4% and the FISH-counts from 11.7-15.2% (Figure 4B and 4C). On the other hand, ε-proteobacteria seemed to be underrepresented in the metagenomes (0.9-2.9%) relative to the FISH-counts (7.5-11.3%; Figure 4D and 4E). Likewise, the α-proteobacterial *Roseobacter*-clade was negligible in the metagenome datasets in numbers (0.4-0.9%), but was found to be much more abundant in the FISH-counts (4.3-8.4%; Figure S4B).

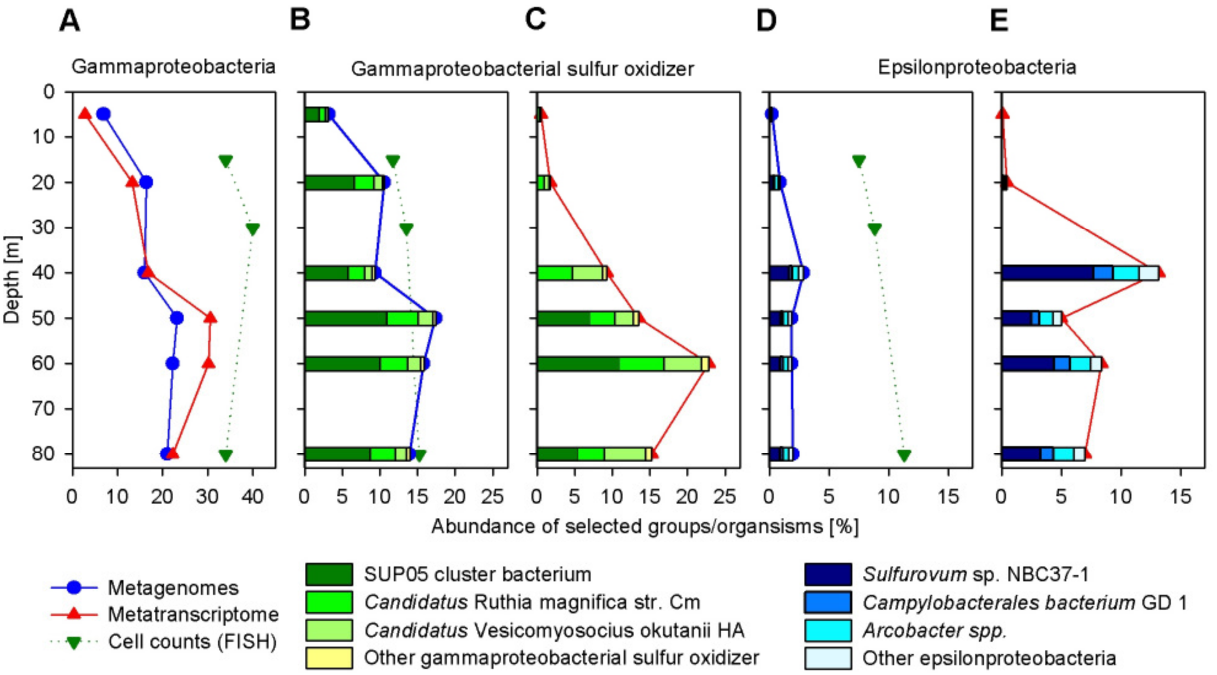


Figure 4: Vertical distribution of FISH counts and of taxonomic assignments of the sequence data. Sequence data shown in percent of all non-ribosomal sequences and FISH counts in percent of all DAPI-stained cells, respectively. (A) DNA and RNA estimates and FISH counts for γ-proteobacteria. (B) DNA estimates (including taxonomic assignments) and FISH counts for γ-proteobacterial sulfur oxidizer (GSO). (C) RNA estimates (including taxonomic assignments) for γ-proteobacterial sulfur oxidizers. (D) DNA estimates (including taxonomic assignments) and FISH counts for ε-proteobacteria. (E) RNA (including taxonomic assignments) estimates for ε-proteobacteria.

The difference in counts encountered with the two methods may be in parts due to the fact that some of the dominant (proteo-) bacteria in the marine environment (e.g. *Arcobacter spp.*, Figure S4A) have only few sequenced genomes and are probably underrepresented in the non-redundant database of NCBI. Our results point to the need of more representative genomes for these groups.

However, in general both metagenome estimates and FISH-counts are complementary and in mutual agreement. Both GSO and ε-proteobacteria together contributed in the metagenomic estimates to 11.5-19.3% and with FISH-counts to 19-27% of the whole bacterioplankton community in sulfidic waters. This number even exceeds the reported 20% (using FISH-counts) in the Benguela Current upwelling system [26].

Metabolic activity and functional diversity of the microbial community

Sequence analysis and general activity pattern

To asses the functional diversity of the microbial community in detail, we used three different approaches to investigate our sequence data. The BLAST-searches were supplemented by scans of our sequences with profile hidden Markov models of the ModEnzA Enzyme Commission (EC) groups [58] and of the Pfam protein families [59]. Furthermore, we recruited the DNA- and RNA-sequences onto the (meta-) genomes of the five organisms most often recognized by our BLAST-searches, SUP05, *R. magnifica*, *V. okutanii*, *Sulfurovum* and *D. autotrophicum* [33,54-57]. For these genome recruitments, we calculated the expression-ratio, a measure of the enrichment of selected transcripts over the corresponding genes, normalized to the pool of all protein-coding sequences (Figure 5 and S5).

Overall, the genome recruitment plots indicated high expression ratios for ribosomal proteins, DNA- and RNA-polymerases, cell division proteins and transcription and translation factors indicating a growing microbial community (data not shown). Similarly, the collection of all abundant EC numbers also suggested an overall active microbial community (Figure S6). Sequences encoding for ubiquitous proteins related to general metabolic activity like RNA- and DNA-polymerases, DNA-topoisomerases and adenosinetriphosphatases (a general ATP-binding and hydrolyzing motif in sequences which lack further specific functional information) were among the most abundant at all depths.

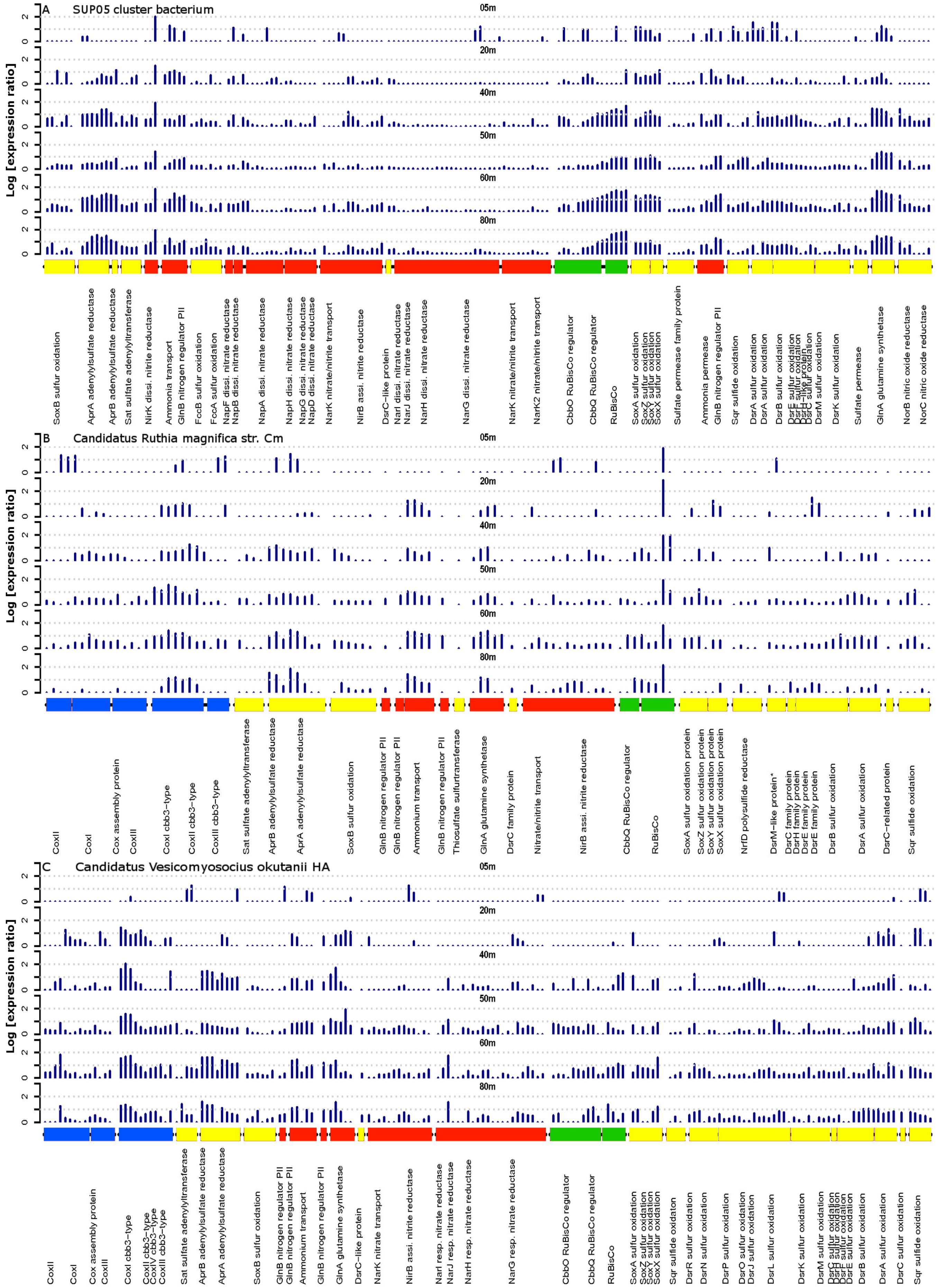
Hydrogen sulfide sources

H_2S is formed through microbial SO_4^{2-} reduction and is commonly present in anoxic marine sediments, where it is considered to be the main heterotrophic process degrading organic carbon [24]. The sedimentary flux has been shown to be the main source of H_2S in the water column during sulfidic events in the Benguela Current upwelling system [25,26]. Although this sedimentary flux is usually considered to be the main source of H_2S diffusing into the water column, water column SO_4^{2-} reduction has been suggested to contribute at times significantly to the H_2S accumulation in oceanic waters [29,30]. This has been demonstrated in the ~2000 m thick anoxic water column in the Black Sea, where water column SO_4^{2-} reduction can be significant (0.02-3.5 nmol l^{-1} d^{-1}) [60]. H_2S formation from SO_4^{2-} was measured even in the presence of more favourable electron acceptors (NO_x) in the OMZ waters off northern Chile after preincubations with H_2S [12].

We investigated the sulfur cycling by identifying specific metabolic functions related to sulfur transformation processes in our sequence datasets and used flux calculations to estimate the sedimentary source of H_2S. The genome recruitment plots of *D. autotrophicum* (Figure S5) show regions of the genome mostly related to energy metabolism and nutrient cycling. *D. autotrophicum* is a metabolically versatile SO_4^{2-} reducing marine δ-proteobacterium, which can completely oxidize organic carbon compounds to CO_2, but was also shown to be capable of growing autographically on hydrogen (H_2) [57]. High expression ratios for key sulfur-metabolizing enzymes (dissimilatory sulfite reductase (*dsrABD*), adenylylsulfate reductase (*aprAB*) and sulfate adenylyltransferase (*sat2*)) suggest that *D. autotrophicum* could have been reducing SO_4^{2-} and thus contributed to the formation of H_2S at our sampling site. Nevertheless, many sulfur cycling proteins (e.g. *dsr* and *apr*) can function in both the oxidation and reduction of sulfur species [61,62]. Thus, also other chemoautotrophic metabolic strategies, e.g. the disproportionation of sulfur compounds (resulting in the simultaneous formation of H_2S and SO_4^{2-}) could be catalyzed by these enzymes [63]. The presence of large plumes of elemental sulfur (Figure 1F) in the study area would have supported this chemoautotrophic reaction.

Figure 5 (next page): Vertical distribution of sequences recruited onto the (meta-) genomes of prominent γ-proteobacterial sulfur oxidizers. Shown are selected genes (in the corresponding order of the (meta-) genomes), mostly involved in oxygen (blue), sulfur (yellow), nitrogen (red) and carbon metabolism (green). The y-axis depicts the log of the expression-ratio, a measure for the selective enrichment of transcripts over the corresponding gene, normalized to the total pool of protein-coding sequences. A list of the start and end position of all genes and the full names of the corresponding enzymes are shown in Table S3. (A) SUP05 cluster bacterium. (B) *Candidatus* Ruthia magnifica str. Cm. (C) *Candidatus* Vesicomyosocius okutanii HA.

[*]This DsrM-like protein has also high similarity to a *narG* respiratory nitrate reductase.


A SUP05 cluster bacterium
B Candidatus Ruthia magnifica str. Cm
C Candidatus Vesicomyosocius okutanii HA
Log [expression ratio]
05m
20m
40m
50m
60m
80m
0
2
SoxB sulfur oxidation
AprA adenylylsulfate reductase
AprB adenylylsulfate reductase
Sat sulfate adenylyltransferase
NirK dissi. nitrite reductase
Ammonia transport
GlnB nitrogen regulator PII
FccB sulfur oxidation
FccA sulfur oxidation
NapF dissi. nitrate reductase
NapB dissi. nitrate reductase
NapA dissi. nitrate reductase
NapH dissi. nitrate reductase
NapG dissi. nitrate reductase
NapD dissi. nitrate reductase
NarK nitrate/nitrite transport
NirB assi. nitrite reductase
DsrC−like protein
NarI dissi. nitrate reductase
NarJ dissi. nitrate reductase
NarH dissi. nitrate reductase
NarG dissi. nitrate reductase
NarK nitrate/nitrite transport
NarK2 nitrate/nitrite transport
CbbO RuBisCo regulator
CbbQ RuBisCo regulator
RuBisCo
SoxA sulfur oxidation
SoxZ sulfur oxidation
SoxY sulfur oxidation
SoxX sulfur oxidation
Sulfate permease family protein
Ammonia permease
GlnB nitrogen regulator PII
Sqr sulfide oxidation
DsrA sulfur oxidation
DsrA sulfur oxidation
DsrB sulfur oxidation
DsrE sulfur oxidation
DsrF sulfur oxidation
DsrH−like protein
DsrC sulfur oxidation
DsrM sulfur oxidation
DsrK sulfur oxidation
Sulfate permease
GlnA glutamine synthetase
NorB nitric oxide reductase
NorC nitric oxide reductase
CoxII
CoxI
Cox assembly protein
CoxIII
CoxI cbb3−type
CoxII cbb3−type
CoxIII cbb3−type
Sat sulfate adenylyltransferase
AprB adenylylsulfate reductase
AprA adenylylsulfate reductase
SoxB sulfur oxidation
GlnB nitrogen regulator PII
GlnB nitrogen regulator PII
Ammonium transport
GlnB nitrogen regulator PII
Thiosulfate sulfurtransferase
GlnA glutamine synthetase
DsrC family protein
Nitrate/nitrite transport
NirB assi. nitrite reductase
CbbQ RuBisCo regulator
RuBisCo
SoxA sulfur oxidation protein
SoxZ sulfur oxidation protein
SoxY sulfur oxidation protein
SoxX sulfur oxidation protein
NrfD polysulfide reductase
DsrM−like protein*
DsrC family protein
DsrH family protein
DsrE family protein
DsrE family protein
DsrB sulfur oxidation
DsrA sulfur oxidation
DsrC−related protein
Sqr sulfide oxidation
CoxII
CoxI
Cox assembly protein
CoxIII
CoxI cbb3−type
CoxII cbb3−type
CoxIV cbb3−type
CoxIII cbb3−type
Sat sulfate adenylyltransferase
AprB adenylylsulfate reductase
AprA adenylylsulfate reductase
SoxB sulfur oxidation
GlnB nitrogen regulator PII
GlnB nitrogen regulator PII
Ammonium transport
GlnB nitrogen regulator PII
GlnA glutamine synthetase
DsrC−like protein
NarK nitrate transport
NirB assi. nitrite reductase
NarI resp. nitrate reductase
NarJ resp. nitrate reductase
NarH resp. nitrate reductase
NarG resp. nitrate reductase
CbbO RuBisCo regulator
CbbQ RuBisCo regulator
RuBisCo
SoxA sulfur oxidation
SoxZ sulfur oxidation
SoxY sulfur oxidation
SoxX sulfur oxidation
Sqr sulfide oxidaton
DsrR sulfur oxidation
DsrN sulfur oxidation
DsrP sulfur oxidation
DsrO sulfur oxidation
DsrJ sulfur oxidation
DsrL sulfur oxidation
DsrK sulfur oxidation
DsrM sulfur oxidation
DsrC sulfur oxidation
DsrH sulfur oxidation
DsrF sulfur oxidation
DsrE sulfur oxidation
DsrB sulfur oxidation
DsrA sulfur oxidation
DsrC sulfur oxidation
Sqr sulfide oxidation

A numerical compilation of all sequences affiliated to known SO_4^{2-} reducers and sulfur oxidizers [12,61,62] shows that the abundance of SO_4^{2-} reducers throughout sulfidic waters was not correlated with the H_2S concentrations, and that the sequence abundance of SO_4^{2-} reducers was low compared to sulfur oxidizing organisms (Figure 6). Although SO_4^{2-} reduction probably occurred in the water column at the time of sampling, the lower occurrence of SO_4^{2-} reducers compared to sulfur oxidizers suggest that SO_4^{2-} reduction in water column was too small to account for the observed accumulation of 4.2 µM H_2S.

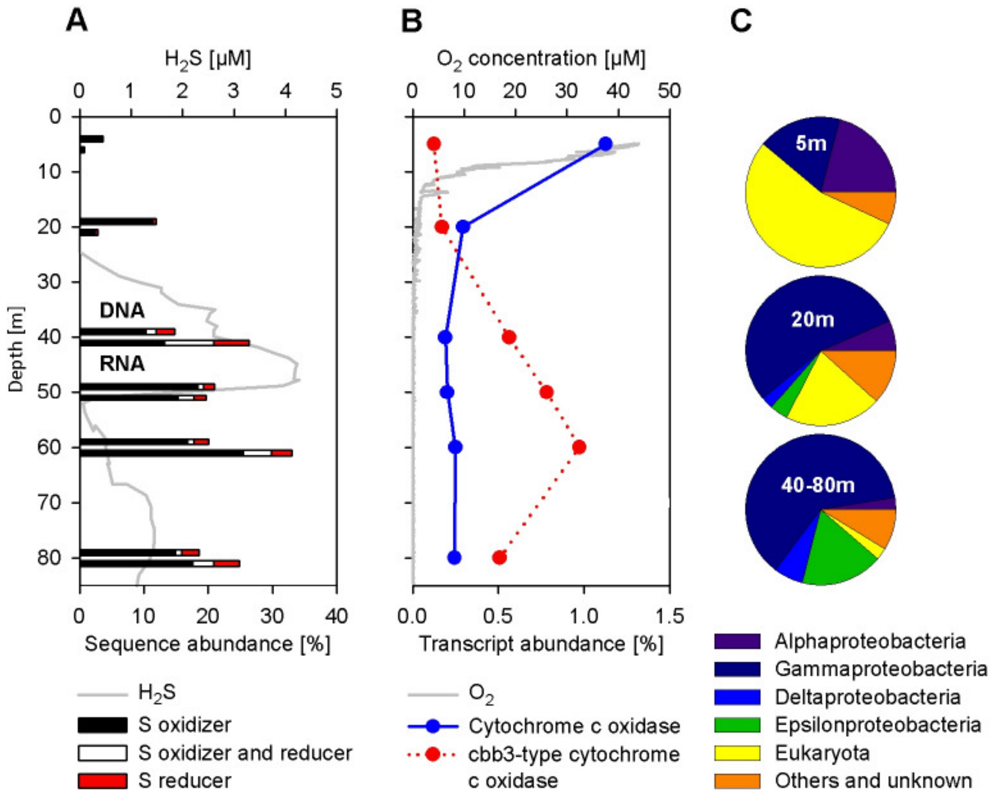


Figure 6: Vertical distribution of H_2S and O_2 and functional assignments of sequences. (A) H_2S concentration and abundance of sequences affiliated to organisms either capable of oxidizing or reducing inorganic sulfur species. Shown in percent of all non-ribosomal sequences and summed according to their metabolic potentials. Top bars depict DNA and bottom bars RNA datasets. (B) O_2 concentrations and transcript abundance of the cytochrome c oxidase and the cbb3-type cytochrome c oxidase (both 1.9.3.1). Shown in percent of all protein-coding RNA sequences. (C) Phylogenetic affiliation of the transcripts encoding for both types of the cytochrome c oxidase.

Moreover, assuming the maximum SO_4^{2-} reduction rates (1.3-12 nmol l^{-1} d^{-1}) as reported from northern Chile [12], it would take more than one year (in the absence of any oxidation) to accumulate the H_2S concentrations reported in our study. In contrast, SO_4^{2-} reduction is generally very high in sediments (10-30 mmol m^{-2} d^{-1}) underlying the eastern tropical South Pacific OMZ [64].

Assuming steady state conditions, we used the density structure and the H_2S concentration gradient (Figure 2A) in the bottom water to estimate a turbulent diffusion of ~10^{-4} m^2 s^{-1} and, subsequently, a sedimentary efflux of H_2S of ~2 mmol m^{-2} d^{-1}. This is well within the estimates based on sedimentary flux calculations and SO_4^{2-} reduction rate measurements (1-11 mmol m^{-2} d^{-1}) from sediments directly underlying sulfidic events [65]. Moreover, the repeated observation of bottom water H_2S maxima along the transect (Figure 1D), point towards the sediment as the main H_2S source.

We observed a second distinct H_2S maximum in the water column at ~48 m with no direct contact to the sediment. Nevertheless, the salinity and the corresponding PO_4^{3-} and NH_4^+ concentrations (Figure 2C and 2E) indicated that the upper H_2S layer was laterally advected from nearby bottom waters that had recently been in contact with H_2S-bearing sediments.

Sulfur oxidation

The largest functional group of microbes we detected in the sulfidic waters were γ–proteobacterial sulfur oxidizers. Figure 5 shows the expression-ratio for selected genome regions of the three most abundant GSO-representatives (similar to SUP05, *R. magnifica* and *V. okutanii*). Figure S5 further depicts the sequences recruited to genome of ε-proteobacterial *Sulfurovum*. The recruitment of the sequences onto separate (meta-) genomes provided evidence that at least three distinct taxa within the GSO-community (similar to SUP05, *R. magnifica* and *V. okutanii*) were present at our study site and actively growing and metabolizing. Recruitment of sequences onto the genome of *Sulfurovum* was much less extensive than for the GSO-group, and showed low coverage especially in the oxic and anoxic depths (5 and 20 m). A large number of transcripts was recruited to genes (if present in the genomes) of the reverse dissimilatory sulfite reduction (*dsr* – oxidation of intracellular S^0) and the periplasmic sulfur oxidation (*sox* - $S_2O_3^{2-}$ oxidation) pathways, both encoding for genes involved in oxidation of reduced sulfur compounds. Additionally, we also found high expression-ratios for the sulfide:quinone oxidoreductase (*sqr* - H_2S oxidation to S^0) and adenylylsulfate reductase (*apr*) and sulfate adenylyltransferase (*sat* - latter both SO_3^{2-} oxidation to SO_4^{2-}). The abundance and coverage of transcripts matching these clusters for the GSO-group and *Sulfurovum* generally increased with depth, delivering the highest expression-ratios of sulfur oxidizing genes within sulfidic waters. For *Sulfurovum*, which is thought to be capable of both sulfur oxidation and SO_4^{2-} reduction [56,66,67], the transcript coverage for its *sox* cluster and *sqr*-gene suggested that it was likely acting as a sulfur oxidizer at the time of sampling.

R. magnifica, *V. okutanii* and *Sulfurovum* genomes harbour genes for different cytochrome c oxidases which are used in oxic respiration and are absent in the currently annotated version of the SUP05 metagenome [33,54-56]. Of the transcripts recruiting to the *R. magnifica* genome, the cytochrome c oxidase was among the most abundant in the oxic surface. In anoxic and sulfidic waters, however, expression patterns changed and instead transcripts for a cbb3-type cytochrome c oxidase became dominant. For *V. okutanii* and *Sulfurovum* transcripts for the cbb3-type cytochrome c oxidase were also among the most highly expressed in sulfidic waters.

The cbb3-type cytochrome c oxidase is thought to enable the performance of a specialized microaerobic respiration. This enzyme has an extremely high affinity to O_2, allowing certain proteobacteria to colonize oxygen-limited environments or environments where O_2 is below the detection limit of currently available oxygen sensors [68]. The Km value of the cbb3-type cytochrome c oxidase for O_2 can in membranes be as low as 7 nM [69], which is well below the detection limit (50 nM) of the sensors we used in this study. Furthermore, it has been shown that *Escherichia coli* cultures actively grew and respired O_2 even below the detection limit of a STOX sensor (≤3 nM), probably using a high-affinity cytochrome bd oxidase [70].

A clear separation between the oxic and the anoxic/sulfidic zone for cytochrome c oxidase expression was visible, directly reflecting the availability of O_2 (Figure 6B). In the oxic surface waters, where the low-affinity cytochrome c oxidase was dominant over the cbb3-type, it was assigned mainly to eukaryotic sequences, as well as to diverse bacterial groups belonging to α- and γ-proteobacteria (Figure 6C). In contrast, in sulfidic waters as much as 80% of the cytochrome c oxidases (compilation of both types) could be assigned to either γ- and ε-proteobacteria. For the *R. magnifica*-like organism, which possesses both types of cytochrome oxidases, the switch in the expression from the low affinity type in oxic surface waters (5 m) to the high affinity type in anoxic and sulfidic waters (20-80 m) can be observed in the genome recruitment plots (Figure 5B).

Despite the presence of H_2S, the oxygen microsensor showed trace amounts of O_2 down to 40m water depth during the downcast, which might be an artefact from water advection caused by the CTD rosette (see above). Assuming though that the O_2 concentrations were not an artifact, we calculated a vertical down flux of O_2 of 0.07 mmol m^{-2} d^{-1} between 17 and 27 m. Although these concentration could only account for the oxidation of ~6% of the sedimentary H_2S flux of 2 mmol m^{-2} d^{-1}, it could partly sustain microaerobic activity in the sulfidic waters as suggested by the presence of the cbb3-type cytochrome c oxidase transcripts. Lateral advection of oxic waters and water exchange induced by internal waves may also have supported microaerobic activity at O_2 concentrations that would remain below the detection limit of our STOX sensor (~50nM).

However, mixing due to internal waves would have only influenced the upper part of the water column and the widespread anoxic conditions below 30 m gave little indication of lateral advection of oxic waters from 30-100 m, where the cbb3-type cytochrome c oxidase was nevertheless expressed. Either, no steady-state conditions were present, or the O_2 was consumed shortly before our sampling campaign. An alternative explanation for the high cbb3-type cytochrome c oxidase expression in the sulfidic waters could also be the use of nitric oxide (NO) as an alternate substrate instead of O_2, as it has been hypothesized for a γ-proteobacterial species [71]. This hypothesis is supported by structural similarities of cbb3-type cytochrome c oxidases to bacterial nitric oxide reductases [72]. Although we did not measure NO in our study, a relatively high expression-ratio of the genes encoding nitric oxide reductases (*norBC*) were found in the metagenome recruitments for SUP05 (these genes are absent from currently available versions of the *R. magnifica* and *V. okutanii* genomes) in sulfidic depths and at 60 m also for *Sulfurovum*. Furthermore, SUP05-like organisms also expressed the dissimilatory nitrite reductase (*nirK)* throughout all depths in high numbers. Since this enzyme reduces NO_2^- to NO, minor concentrations of NO might have been present and could have been used for respiration processes linked to sulfur oxidation.

Alternative to the use of O_2 (and NO), SUP05, *V. okutanii* and *Sulfurovum* and several other ε-proteobacteria [73,74] are also thought to couple the oxidation of H_2S to the reduction of NO_x. Low, but persistent concentrations of NO_x were measurable at most depths down to 60 m, and nitrate as well as nitrite reductase (*nar*-, *nap- and nir*-genes) transcripts were expressed in all anoxic and sulfidic depths (Figure 5). The genome recruitment plots of SUP05 (the published genome contains all three clusters) were showing expression of *nap*- and especially *nir*-genes, while *V. okutanii* showed high expression ratios mostly for *nar*-, and *Sulfurovum* for *nap*-genes (Figure 5 and S5). In *R. magnifica* on the other hand, most of the above mentioned genes are absent from the currently available version of its genome and it is thought that *R. magnifica* reduces NO_x only for assimilatory reasons [55].

We cannot calculate fluxes from the trace amounts of NO_x below 30 meters. However, at the upper boundary of the second H_2S peak (25-29m) the NO_2^- and H_2S profiles overlap. Although the lateral intrusion of the sulfidic water mass suggested that these profiles were mainly caused by water mass mixing, the ratio of the opposing concentration gradients of

NO_2^- (-0.53 µM m^{-1}) and H_2S (0.22 µM m^{-1}) between 25m and 29m depth indicated that the NO_2^- flux was sufficient to oxidize the upward flux of H_2S. A diffusion coefficient of ~2 x 10^{-5} m^2 s^{-1} can be estimated from the density gradient at this depth, resulting in an upward flux of 0.38 mmol H_2S m^{-2} d^{-1} and an average oxidation rate of ~100 nmol H_2S l^{-1} d^{-1} within the ~4 m thick overlapping layer. This is in the same range as the experimentally measured reduction of NO_2^- to N_2 (126 nmol N l^{-1} d^{-1}) at 30 m depth (see section below), suggesting H_2S oxidation via the sulfur-driven autotrophic denitrification. The removal rate of NO_2^- calculated from the downward flux of NO_2^- was 230 nmol N l^{-1} d^{-1} and matched the experimentally measured NO_2^- removal of 255 nmol N l^{-1} d^{-1} from combined denitrification, anammox and DNRA (see section below).

Although our combined results suggest that H_2S was detoxified by both microaerobic activity and the sulfur-driven autotrophic denitrification well below the oxic zone (>1µM), the flux calculations indicate that the larger part of the H_2S was probably oxidized anaerobically with NO_x.

Nitrogen cycling

To shed light on the nitrogen cycling carried out by the microbial community, we measured potential rates of various nitrogen transformation processes using $^{15}NO_3^-$, $^{15}NO_2^-$, $^{15}N_2O$ or $^{15}NH_4^+$ incubations and compared them to corresponding abundances of functional genes and transcripts involved in their turnover (BLAST-hits and EC number-assignments; Figure 7). Throughout the anoxic and sulfidic zones we could measure active NO_3^- reduction to NO_2^-, with highest potential rates at 40 m (2500 nmol N l^{-1} d^{-1}; Figure 7B). The in comparison much lower rate at 30 m (150 nmol N l^{-1} d^{-1}) might actually be caused by limitations in reductants, given the much lower H_2S concentrations. The transcript abundance for respiratory nitrate reductase (EC 1.7.99.4) peaked in the anoxic zone at 20 m (0.35% of all protein-coding sequences) and then dropped within the sulfidic waters following the decrease in NO_3^- concentrations. Gene abundances however, increased with depth (~0.5%) and it may well be that a fast response of the microbial community to release the actual potential for NO_3^- reduction to NO_2^- was stimulated by the addition of NO_3^-. The transcripts at 5 and 20 m were mostly similar to *K. stuttgartiensis*, while transcripts in sulfidic waters belonged to diverse groups of α-, β-, γ-, δ- and ε-proteobacteria, indicating a clear taxonomic separation of the microorganisms expressing this gene with depth.

Potential rates for the subsequent steps in denitrification, the reduction of NO_2^- to N_2, were highest close to bottom waters at 80 m (900 nmol N l^{-1} d^{-1}, Figure 7C), whereas the transcript abundance for NO-forming cd-cytochrome nitrite reductase (EC 1.7.2.1) was highest in the oxic and anoxic zones at 5 and 20 m (0.4% and 0.2%), mirroring the availability of NO_2^-. The majority of the transcripts at 5 and 20 m were similar to *N. maritimus*, while those recovered from sulfidic depths were affiliated with diverse groups of proteobacteria. We also found high expression of ammonia monooxygenase transcripts related to *N. maritimus* (1.5%) in the oxic surface (5 m), which may be partly responsible for the high NO_2^- concentrations in the surface waters. However in all sulfidic depths, DNA and RNA sequences related to *N. maritimus* were very scarce, suggesting that *N. maritimus* played only a minor role in sulfidic waters.

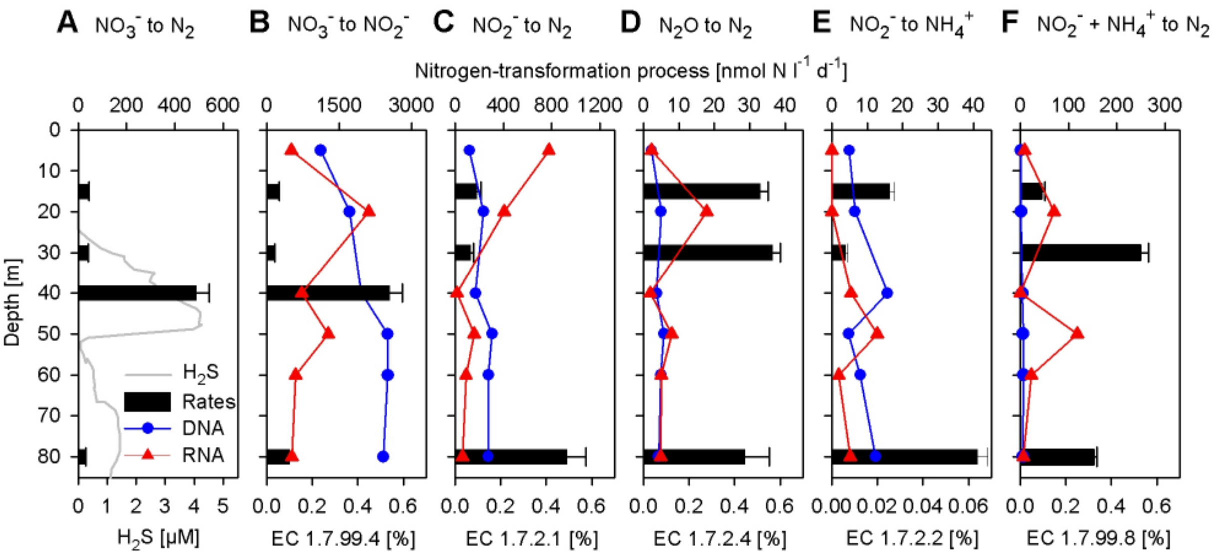


Figure 7: Vertical distribution of nitrogen transformation process rates and abundance of sequences encoding for involved enzymes. Shown in percent of all protein-coding sequences for the DNA and RNA datasets, respectively. (A) NO_3^- reduction to N_2 (denitrification). (B) NO_3^- reduction to NO_2^-, respiratory nitrate reductase (1.7.99.4). (C) NO_2^- reduction to N_2, (NO forming) nitrite reductase (1.7.2.1). (D) N_2O reduction to N_2, Nitrous-oxide reductase (1.7.2.4). (E) NO_2^- reduction to NH_4^+ (DNRA), (NH_4^+ forming) nitrite reductase (1.7.2.2). (F) NO_{2-} + NH_4^+ to N_2 (anammox, based on the sole addition of NO_2^-), hydrazine oxidoreductase (EC 1.7.99.8).

Some of the NO_2^- reduction (N_2 production) could be attributed to anammox activity, where the highest rates were found at 30 m (250 nmol N l^{-1} d^{-1} based on $^{15}NO_2^-$ addition, Figure 7B and 96 nmol N l^{-1} d^{-1} based on $^{15}NH_4^+$ addition, data not shown). At 80 m anammox activity could only be detected by $^{15}NO_2^-$ addition (152 nmol N l^{-1} d^{-1}), an incubation with added $^{15}NH_4$ did not stimulate any N_2 production (data not shown), most likely due to limitations in NO_x. The abundance of genes encoding for hydrazine oxidoreductase (EC 1.7.99.8) was generally very low at all depths (less than 0.015%, Figure 7F) while transcripts, mostly annotated as similar to *K. stuttgartiensis*, peaked in abundance at 20 and at 50 m (0.15 and

0.25%, respectively). The 50 m transcript maximum of hydrazine oxidoreductase is in good agreement with a small dip in NH_4^+ concentrations (Figure 2C).

N_2O is an intermediate in denitrification (NO_x to N_2) and we measured the reduction of N_2O to N_2, which turned out to be of smaller magnitude (30 nmol N l^{-1} d^{-1}) than the NO_2^- reduction to N_2 (Figure 7D). The N_2O concentrations, which ranged around 20-40 nM from surface to ~80 m, dropped below detection limit at 80 m (Figure 2B), indicating either a complete reduction of N_2O or a lack of production due to the limitation in NO_x. Gene and transcript abundance for nitrous oxide reductase (EC 1.7.2.4) were highest in the anoxic waters (20 m) reaching 0.3% and thus was of similar magnitude when compared to nitrate and nitrite reductases from the same depths.

We also conducted potential rate measurements of complete denitrification (NO_3^- to N_2; Figure 7A). Here, the highest potential rate (490 nmol N l^{-1} d^{-1}) was found at 40 m within the first H_2S maximum. Much lower rates were observed for the other depths (26-41 nmol N l^{-1} d^{-1}), which might be attributed to incomplete denitrification (e.g. NO_3^- reduction to NO_2^-, N_2O or NO, Figure 7B).

Potential rates for the dissimilatory nitrate reduction to ammonia (DNRA) were the lowest of the transformations we measured, not exceeding 40 nmol N l^{-1} d^{-1} (Figure 7E). Gene abundance for cytochrome c nitrite reductase (EC 1.7.2.2) was also smaller than reported for genes implicated in the other nitrogen transformation processes and except for 50 m depth, the transcripts were even rarer than the gene abundances (less than 0.01%). This suggests an only minor importance of DNRA in the microbial community at the time of sampling.

Although the rate measurements do not provide information on the organisms carrying out the nitrogen-transformations, the predominance of sulfur-oxidizing proteobacteria found in our sequence data and the corroboration of their dominance in FISH-counts suggest that the sulfur-driven autotrophic denitrification was most likely the dominant pathway for N-loss during our sampling campaign.

The reduction of NO_x in the sulfidic waters seemed to have been rapidly stimulated by the high NO_x concentrations in our experiments. This could indicate that the low NO_x concentrations at the time point of sampling were restricting the activity of the microbial community, or that the measured NO_x was available only to certain microorganisms. Mechanisms to overcome limitations, e.g. for NO_3^-, were shown for the γ-proteobacteria *Beggiatoa spp.*, *Thioploca spp.* and *Thiomargarita spp.* These organisms are thought to be able to accumulate NO_3^- intracellular to concentrations 3,000 to 20,000 times higher than ambient [75-77]. Similar mechanisms are so far not known for SUP05 or *V. okutanii*, but it was hypothesized that the integration of a toxin-antitoxin module in the denitrification regulon of SUP05 might enable this organism to withstand extreme NO_3^- limitation [33]. This regulon was indeed highly expressed in our samples (data not shown), and it would be difficult to explain this in the presence of upto 0.1 µM NO_x in the sulfidic waters down to 80 m. Thus, an intracellularly storage of NO_x would allow H_2S oxidation and growth of SUP05 deep within the sulfidic waters. Moreover, NO_x stored intracellular could be released during the sample fixation for NO_3^- and NO_2^- measurements, and potentially explain the low, but consistent concentrations we measured down to 80 m water depth.

Carbon assimilation

Eastern Boundary Upwelling Systems are known for high primary productivity due to photoautotrophic growth in surface waters, which stimulates heterotrophic respiration processes in underlying waters. However, also autotrophic lifestyles within these underlying waters, significantly contributing to NO_x and PO_4^{3-} depletion, are found, e.g. by organisms responsible for anammox [18-21] or the sulfur-driven autotrophic denitrification [26,34,73]. To investigate the magnitude of inorganic carbon assimilation of the microbial community, we conducted rate measurements at selected depths with ^{13}C-bicarbonate (Figure 8A) and compared them to the relative abundance of transcripts encoding for CO_2-fixing enzymes (Figure 8B).

The metatranscriptomic datasets indicated a great abundance of transcripts for ribulose-bisphosphate carboxylase/oxygenase (RuBisCo, EC 4.1.1.39) in all depths, but especially in the sulfidic waters at 60 and 80 m (1.9 and 1.3% of all protein-coding sequences, respectively; Figure 8B and Figure S6). The phylogenetic diversity of the RuBisCo transcripts varied throughout the water column, with photosynthetic organisms such as eukaryotic algae and cyanobacteria dominating the 5 and 20 m samples. In the sulfidic zone (40-80 m samples), β- and especially γ-proteobacterial transcripts were most abundant (Figure 8C). Approximately 25% of all RuBisCo-transcripts were most similar to a bacterial artificial chromosome-clone of unknown bacterium 560, which also appears to belong to a SUP05 genome [78]. Thus, adding up counts for all γ-proteobacteria, they contributed to about 70% of all RuBisCo-encoding transcripts at our particular study site in sulfidic waters. While the high proportions of RuBisCo transcripts were indicative of an active Calvin-Benson-Bassham Cycle, other CO_2 fixation pathways were represented by the presence of transcripts for key enzymes of the Arnon-Buchanan cycle (ATP citrate lyase, EC 2.3.3.8) and the Wood-Ljungdahl pathway (CO-dehydrogenase, EC 1.2.99.2).

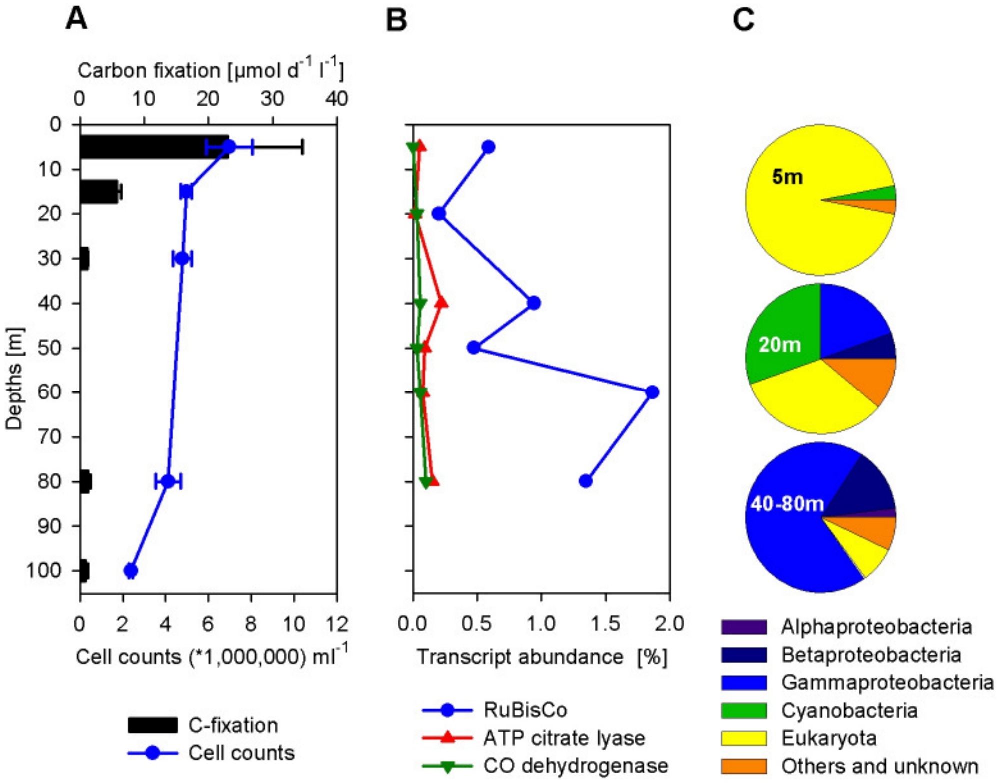


Figure 8: Vertical distribution of CO_2-fixation rates, flow-cytometry cell counts and transcript abundance of sequences encoding for key CO_2-fixing enzymes. Shown in percent of all protein-coding RNA sequences. (A) CO_2-fixation rates and total bacterial cell counts with flow-cytometry. (B) Abundance of transcripts encoding for CO_2-fixing enzymes: Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo, 4.1.1.39, Calvin-Benson-Bassham Cycle), ATP citrate lyase (2.3.3.8, Arnon-Buchanan cycle) and CO-dehydrogenase (1.2.99.2, Wood-Ljungdahl pathway). (C) Phylogenetic affiliation of the transcripts encoding for RuBisCo.

Transcripts for these enzymes were also recruited onto the *Sulfurovum* and *D. autotrophicum* genomes, respectively (Figure S5). Although less abundant than RuBisCo transcripts, RNA-sequences encoding for both enzymes accounted for approximately 0.3% (Figure 8B).

CO_2-fixation rates measured with ^{13}C-bicarbonate incubations (Figure 8A) were highest in the nutrient rich oxic surface waters (23 µmol C l^{-1} d^{-1}), reflecting the dominance of large-sized eukaryotic phytoplankton (RuBisCo transcripts were similar to e.g. the red tide flagellate *Heterosigma akashiwo* and to the diatoms *Odontella sinensis* and *Thalassiosira spp.*). However, dark incubations with ^{13}C-bicarbonate yielded CO_2-fixation rates ranging from 0.9 to 1.4 µmol C l^{-1} d^{-1} at depths of 30, 80 and 100 m, demonstrating the high chemolithoautotrophic activity. These rates are comparable to chemoautotrophic activity found in the redoxclines and the sulfidic zones of the Baltic and the Black Sea [79-83]. Integrating the rates over the predicted predominantly photic (0-20 m) and aphotic (20-100 m) zones, we calculated ~288 mmol C m^{-2} d^{-1} for the photic, and ~96 mmol C m^{-2} d^{-1} for the light-independent CO_2-fixation in sulfidic waters. Thus, at our particular study site, about ~25% of the total CO_2-fixation was carried out by chemolithoautotrophic microorganisms. This is similar to the CO_2-fixation in the redoxcline of the more permanently sulfidic Baltic Sea, which was shown to account for ~30% of the surface productivity [84].

Furthermore, the chemolithoautotrophic CO_2-fixation at our study site would represent as much as 33-53% of the estimated average CO_2-fixation rate per square meter for the Humboldt Current system (182-290 mmol C m^{-2} d^{-1}) [2,4]. On average, each cell fixed about ~0.3 fmol C d^{-1}, which is in the range of chemolithoautotrophic CO_2-fixation found in the sulfidic zone of the Baltic Sea [81]. In comparison, the carbon-assimilation rates of 0.3 µmol C l^{-1} d^{-1} by heterotrophic bacteria, after the addition of ^{13}C-glucose (data not shown), was only one third when compared to the carbon-assimilation of chemoautotrophs induced by the addition of ^{13}C-bicarbonate, further underlining the dominance of autotrophy.

Integrating the calculated aphotic ~96 mmol C m^{-2} d^{-1} CO_2-fixation over the entire extent of the sulfidic plume (~8000 km^2), the CO_2 fixed through chemolithoautotrophy would have been ~7.7 x 10^8 mol d^{-1}, equalling 9.2 kilotons C per day. In comparison, in this area an estimate of the total the primary production by remote sensing is 550-880 kilotons C per day [2]. Consequently, the chemoautotrophic activity during this sulfidic event would have contributed to approximately 1.7% of total CO_2 fixed in the Humboldt Current system off the Peruvian coast at the given time. This is intriguing, since the Humboldt Current system is one of the most productive marine systems world-wide and supports the production of more fish per unit area than anywhere else in the world [2-4,85]. Considering that the sulfidic plume may have been considerably larger than the extent we recorded and was either recurrent or prevailed for several months, the chemoautotrophic growth is significant in terms of carbon retention. The chemoautotrophic growth may act as an important, but until now neglected factor promoting SO_4^{2-} reduction and thus stabilizing sulfidic conditions in OMZ waters.

Conclusions

We provide here detailed insight into the phylogenetic diversity and the functional potential as well as the activity of microbial communities within sulfidic waters, sampled in one of the worlds largest and most productive oceanic upwelling regions [2,4]. To our knowledge, this is the first time H_2S has been measured within the OMZ waters off Peru and represented with >8000 km^2 the largest sulfidic plume ever observed in ocean waters with an estimated ~3.5 x 10^4 tons of toxic H_2S.

We demonstrated, contrasting previous studies carried out in the (non-sulfidic) OMZ off northern Chile [12,32], that some of the 'classical' OMZ-inhabitants (similar to *N. maritimus*, *K. stuttgartiensis* and *Pelagibacter spp.*) were much less abundant in sulfidic waters. Instead we found several distinct γ- and ε-proteobacteria related to SUP05, *R. magnifica*, *V. okutanii* and *Sulfurovum* to be dominant at the sampling site and to express a broad range of genes involved in sulfur (H_2S, $S_2O_3^{2-}$, S^0 and SO_3^{2-}) oxidation. Group-specific FISH counts further identified considerable populations of α- and ε-proteobacteria (e.g. *Roseobacter* spp. and *Arcobacter* spp.), which were of lower abundance in the sequence datasets, but have previously been reported within sulfidic waters [26].

Our data further suggested that these proteobacteria probably utilized several different oxidants ranging from O_2, NO_3^-, NO_2^-, N_2O to NO to oxidize the H_2S well below the oxic surface. While the sequences related to SUP05 indicated that NO_x and NO reduction genes were being expressed, *R. magnifica*-, *V. okutanii*- and *Sulfurovum*-like transcripts related to a microaerophilic cbb3-type cytochrome c oxidase also pointed to the use of O_2 for H_2S detoxification.

Furthermore, transcripts matching SO_4^{2-} reduction genes from δ-proteobacterial *D. autotrophicum* suggested that SO_4^{2-} reduction may have occurred in the water column, but based on the higher abundance of sulfur oxidizers and from our flux calculation we concluded that the main source for H_2S in the water column was sedimentary SO_4^{2-} reduction. Moreover, given the fact that many sulfur cycling enzymes can both oxidize and reduce sulfur species [61,62] and the presence of large elemental sulfur plumes within the sampling area, the disproportionation of sulfur seems a likely way of energy acquisition for the proposed SO_4^{2-} reducers [63].

Most identified microorganisms, both sulfur oxidizers and SO_4^{2-} reducers were expressing transcripts encoding for CO_2 fixing enzymes, which is concurrent with high CO_2 fixation we measured deep with in the dark, sulfidic waters. This light-independent CO_2 fixation at our study site indicated considerable chemoautotrophic growth, similar to observations in permanently stratified systems like the Baltic Sea and the Black Sea [79-83]. In addition to the postulated increased occurrence of sulfidic events due to anthropogenic activity [22,26,30], the carbon retention due to the chemoautotrophic activity presented in this study may enhance SO_4^{2-} reduction and could thus act as an important mechanism to stabilize sulfidic conditions in OMZ waters, and potentially reduce the liveable habitat of many higher organisms.

Material and Methods

Sample Collection

All waters samples were collected in January 2009 in the course of the RV Meteor cruise M77/3 on the Peruvian shelf. The sampling site (station 19; 12° 21,88' S, 77° 0,00' W) for the detailed depths profile (January 9th, 2009) was located approximately 15 km off the coast of Lima with a bottom depths of 100 m. During upcast, water was pumped from depth directly on board using a pump-conductivity-temperature-depth water sampler. We monitored density as well as O_2 concentrations to account for minor changes of the depths due to internal waves (and moved the pump CTD accordingly).

Samples for nucleic acid extraction were filled (oxygen-free) in 4.5 litre polycarbonate bottles. For each sample 1,5-2 litres of water were prefiltered through 10 µm pore size filters (Millipore/Durapore Membrane filters) and then collected upon 0,22 µm pore size filters (Millipore/Durapore Membrane filters) using a vacuum pump (Sartorius eJet). From the time point the water was pumped on board, less than 18 minutes elapsed until the filters were put in microcentrifuge reaction tubes and flash frozen in liquid nitrogen. Samples for incubation experiments and nutrient analysis were taken from same waters (see description below).

RNA-extraction and cDNA-synthesis

DNA and RNA were extracted using the DNA/RNA-Allprep kit (Qiagen) with minor modifications in the protocol for the lyses step: The frozen filters were crushed using a disposable pestle and incubated with 200 µl lysozyme (10 µg/µl) and 1mM EDTA at ambient temperatures for 5 minutes. 40 µl of Proteinase K (10 µg/µl) were added and followed by another incubation of 5 minutes at ambient temperatures. After adding 500 µl buffer RLT-Plus (containing 10 µl/ml β-mercaptoethanol) the manufacture's instructions were followed. Total RNA was eluted in 50 µl nuclease-free water, followed by a subsequent step of DNA digestion (Turbo DNA-free kit, Ambion). rRNA was removed with Epicentre mRNA only prokaryotic. Further depletion of bacterial rRNA was achieved with Ambion Microb*Express*. Cleaned mRNA was subjected to an *in vitro*-amplification step using Ambion Message*Amp*. Finally, cDNA was created by using Invitrogen superscript III cDNA synthesis kit with random hexameric primers (Qiagen). Throughout the procedure all DNA and RNA-samples were subsequently quantified using nano-litre spectrophotometry (NanoDrop) and checked for degradation with BioRad Experion (RNA Standard & High Sense). Leftover reactants and reagents were removed using the PCR Mini Elute Kit (Qiagen). The cDNA was immediately stored at -80°C until pyrosequencing. Throughout the whole procedure nuclease-free plastic consumables and nuclease-free water and reagents were used to hinder any possible degradation of RNA or cDNA.

Sequencing

50 µl of the DNA/cDNA-samples were sequenced with the GS-FLX (Roche) pyrosequencer at the Institute of Clinical Molecular Biology (IKMB, Kiel), each sample was loaded on one quarter of a PicoTiter plate (except the 5 m cDNA-sample, which was loaded on two quarters of a PicoTiter plate). This resulted in 1.888.768 (DNA) and 1.560.959 (cDNA) raw reads, accounting for 757.439.211 and 599.103.110 base pairs of sequence information, respectively.

Sequence annotation pipeline

The sequences were organized and analyzed with an in-house meta-omic annotation pipeline. The raw reads were clustered using Cd-hit [86] with a sequence identity threshold of 98% and word length of 8, delivering about 1,581,637 / 592,711 cluster representative sequences in total. The ribosomal RNA sequences in these cluster representatives were identified by a BLASTn [87] search against the SILVA database (http://www.arb-silva.de) [88] including both prokaryotic and eukaryotic small and large subunit sequences with a bit score cut off of 86. The bit score cut off of 86 described earlier [89] was validated using a simulation exercise. All sequences deposited in SILVA database (477,749) were randomly fragmented into one million sequences to simulate a 454-sequencing profile. The generated fragments had a normal length distribution with the mean and standard deviation value equal to those from our own dataset (mean: 420 base pairs and standard deviation: 150 base pairs). This simulated dataset was CDHIT-clustered as above and BLASTed against the SILVA database itself. All queries which hit themselves in the database (e.g. the sequence of origin of the query) or hit a sequence belonging to the same taxonomic lineage (allowing a mismatch of up to 2 taxonomic levels) were considered true positives, while those fragments that hit sequences from other taxonomic lineages were false positives. The bit score distributions for the true and false positives were binned and a threshold sweep was used to calculate the sensitivity and specificity value at each bit score threshold. A bit score value of 86 in the resulting plot of the sensitivity and specificity distributions for this threshold sweep gives a specificity of 99.35% and a sensitivity of 99.85% respectively. This cut off was used for all further analysis of the sequences against the SILVA database. This cut off was also used in MEGAN to make taxonomic assignments for the sequences using a minimum support of 5 and a 10% score range for its LC Algorithm [90]. The cluster representative sequences without a hit in the

SILVA rRNA database (non rRNA-sequences) were compared against the non-redundant database from NCBI using BLASTx with a bit score cut off of 35. For the functional assignment of the cluster representatives the top hit of each BLAST-search was used. The non-rRNA sequences were also scanned with profile hidden Markov models of the ModEnzA EC groups [58] and the Pfam protein families [59]. The Pfam hits were converted to EC numbers and along with the ModEnzA hits, mapped to the KEGG reference pathways using an in-house java-based pathway mapping and visualization tool. The sequences, clustering information (cluster size, cluster ID) from Cd-hit, results from the BLAST, Pfam and EC searches and the taxonomic assignment from MEGAN for each cluster representative sequence were added to a MySQL database for ease of analysis.

Sequencing statistics

For all sampled depths, we sequenced DNA as well as cDNA, synthesized from rRNA-depleted amplified mRNA (water sample fraction 0.2-10 µm) using Roche 454 FLX technology. A total of 757,539,211 base pairs (1,888,768 sequences) for DNA and 599,103,110 base pairs (1,560,959 sequences) for RNA, respectively, were recovered with an average sequence length of 392 base pairs (Table S1).

Combining all DNA sequences, 0.3% were of ribosomal-gene origin. For the RNA-sequences, the percentage of rRNA reads varied between 58 and 76% with the exception of the sample collected at 5m which was dominated by eukaryotic microorganisms, and contained a high percentage of eukaryotic rRNA sequences as a result (>80%). The high count of eukaryotic rRNA reads indicates that a further depletion of eukaryotic rRNAs supplementing the depletion of bacterial rRNAs would have been worthwhile, whereas the number of archaeal rRNAs was mostly negligible. The rDNA and rRNA-sequences were excluded from further analysis, leaving 1.882.842 DNA and 421.528 RNA non-ribosomal sequences.

Of all non-ribosomal reads, 54% (DNA) and 53% (RNA) of the sequences could be identified as protein-coding; the remainder could not be assigned. Of the protein-coding sequences, 82% of the DNA and 99% of the RNA sequences could be taxonomically assigned (matched the non-redundant database using BLASTx). The remaining protein-coding sequences (18% for DNA and 1% for RNA) matched profile hidden Markov models of either ModEnzA EC groups [58] or Pfam protein families [59].

Hierarchical clustering of samples using the taxonomic profiles

The taxonomic profiles of the metagenome and metatranscriptome samples (the occurrence of the microbial taxa in the samples as a percentage of the total number of sequences having a BLASTx hit) were used for hierarchical clustering using the PRIMER 6 program [91]. The samples were grouped into 6 categories according to the depths and a multivariate statistical test (ANOSIM) was used to determine if the groupings were distinct from each other in terms of the microbial communities. The relationships between the depth groups were visualized in non-parametric Multidimensional Scaling plot using PRIMER 6.

Metabolic and taxonomic diversity measures

The EC activity matrix (with sample sizes equalized to the smallest sample, randomly selecting the same amount from the other samples) was exported from our in-house pathway mapping tool and the EC counts for each sample were used to calculate the inverse of Simpson's index (1/D) where $D = \Sigma P_i^2$ and P_i representing the proportional abundance of species i, and the Evenness E = (1/D)/S with S being the number of unique species.

Similar calculations were also performed for the taxonomic assignments at the phylum level from the BLASTx searches normalized to total number of sequences having a BLASTx hit.

Sequence recruitment to (meta-) genomes of the most abundant organisms

The sequence data was recruited onto the reference (meta-) genomes of the five most abundant organisms SUP05, *R. magnifica*, *V. okutanii*, *Sulfurovum* and *D. autotrophicum* using the MUMmer program [92]. For SUP05, the ordered assembly of contigs from Walsh and colleagues was used as the reference [33]. The recruited reads were re-assessed using a BLAST search against the reference genomes. Sequences hitting more than one genome with a bit score difference of less than 5% between the first and second hits were discarded giving rise to a non-overlapping set of reads for each genome. These reads were then BLASTed again to genome to calculate the average coverage for each base over non-overlapping windows of size 300 base pairs from the reference genomes. The coverage of the reads for each metatranscriptome sample for each reference genome window was normalized by the total number of BLASTx hits for that sample and divided by coverage of the same window from the corresponding metagenome reads (which had also been similarly normalized). This value, the expression ratio, was then corrected for the differences in sizes of the metatranscriptome and metagenome and plotted for selected regions of the reference genomes using customized R and PERL scripts.

Flow-cytometry cell counts

Samples for flow-cytometry were fixed with a final concentration of 1% paraformaldehyde and stored at -80°C until analysis. Cell counts were performed on a FACSCalibur. After 20 minutes staining of samples with SybrGreen (double stranded DNA stain) at 4°C, cells were counted for two minutes or until a total count of 50000 was reached. Sample flow rate was calibrated with standard beads (Trucount, BD Biosciences) and cell numbers calculated via the time of measurement. Events that did not pass a certain threshold of SybrGreen fluorescence (viruses, particles) were excluded from cell counts.

Fluorescence *in situ* hybridization cell counts

Water samples were fixed with paraformaldehyde (final concentration 2%) for DAPI-staining and CARD-FISH analyses and were filtered through 0.22-µm polycarbonate filters. Filter sections were treated for cell permeabilization using lysozyme as described elsewhere [93]. Endogenous peroxidases were inhibited by the addition of 3% H_2O_2. Afterwards filter sections were hybridized with horseradish peroxidase-labelled oligonucleotide probes [93]. For all probes hybridizations were conducted at 46°C, expect for probe GSO, which was hybridized at 35°C as described before [26]. The following probes were applied: Bacteria, EUBI-III [94]; γ-proteobacteria, Gam42a [95]; GSO-cluster (γ-proteobacterial sulfur oxidizer), GSO477 [26]; Roseobacter clade, Ros537 [96], Arcobacter, Arc94 [97], some ε-proteobacteria, Epsy682 [98]. Washing, signal amplification and epifluorescence microscopy were conducted as described previously [93].

Nutrient analysis and oxygen measurements

Our pump CTD system was equipped with a conventional amperometric O_2 microsensor to obtain vertical profiles of dissolved O_2. In addition, the recently developed STOX (Switchable Trace amount OXygen) sensor, which allows high-accuracy O_2 measurements in near anoxic environments (detection limit: 50-100 nmol L^{-1} during our measurements), was deployed [11,14]. At least five measuring cycles after a minimum of ten minutes sensor equilibration at a given sampling depth were used to calculate O_2 concentrations.

Water samples for nutrient analysis were taken with a depth resolution of 1 to 2 m. NH_4^+ was measured fluorometrically [99] and NO_2^- was analyzed spectrophotometrically [100] on board. Water samples for NO_3^- and PO_4^{3-} were stored frozen until spectrophotometric determination [100] with an autoanalyzer (TRAACS 800, Bran & Lubbe). Detection limits for NH_4^+, NO_2^-, NO_3^- and PO_4^{3-} were 10, 10, 100 and 100 nmol L^{-1}, respectively. Dissolved N_2O

concentrations were determined onboard in triplicates measurements using the GC headspace equilibration method as described elsewhere [101]. H_2S concentrations were measured on discrete samples and with a calibrated microsensor [102].

The calculation of the ~8000 km^2 area covering sulfidic plume is based on the detection of H_2S in water column along ~200 km of the Peruvian coast. On average, the sulfidic zone was ~40 km broad and contained mean H_2S-maxima close to the bottom of the water column of 3.4 µM. H_2S was extending vertically over the water column with an averaged depths of ~80 m, yielding ~640 km^3 of H_2S-containing waters. Based on an average H_2S concentration of 1.5 µM, we estimated a total content of ~3.5 x 10^4 tons H_2S within the plume.

Flux calculations

Density was calculated using the data processing program SeaSoft (Sea-Bird Electronics). The stability of the water column was expressed using the Brunt-Väisälä frequency N, defined as:

$$N^2 = \frac{g}{\rho}\frac{\partial\rho}{\partial z}$$

Where ρ is the water density, g is the gravitational acceleration and z is the water depth. The density gradient was calculated over 4 m bins. The turbulent diffusivity Ez was calculated according to [103] from the Brunt-Väisälä-Frequency and the dissipation rate of turbulent kinetic energy ε:

$$Ez = \frac{\gamma\varepsilon}{N^2}$$

with the mixing coefficient γ=0.2. We applied a mean ε of 1.85 x 10^{-9} W kg^{-1} [104]. This value was measured by Gregg et al., 1986 for the open-ocean thermocline and was applied in several rate diffusion models [105,106]. Vertical concentration gradients for O_2, H_2S, and NO_3^- were calculated over 4 m bins. Fluxes of O_2, H_2S, and NO_3^- were calculated according to Fick's law $\mathrm{J_i = Ez\ \partial C/\partial z}$ at respective depths.

Satellite images

MODIS (Moderate Resolution Imaging Spectroradiometer) and MERIS (Medium Resolution Imaging Spectrometer) were implemented to study milky turquoise discolouration in waters off the Peruvian coast. Cloudy weather north of Pisco during most of our cruise made remote sensing difficult, but turquoise discolorations were also observed off Lima on January 20-21st and 27-28th (Figure S1B). The estimation of the extent of the plumes requires data of higher spatial resolution; the algorithm for the identification of elemental sulfur plumes is based on the high spectral resolution of MERIS data [47]. Because full resolution MERIS data are not available for this region we present a high resolution MODIS true colour image for optical reasons in our main figure (Figure 1E).

However, the reflectance spectra derived from reduced resolution MERIS data (Figure S1C), reveal that the turquoise plume southwest of Pisco conform the criteria for elemental sulfur in the identification algorithm distinguishable from optically similar coccolithophore blooms [107]. Differences in the shape and appearance of the sulfur cloud on the MERIS image (Figure S1C) and the MODIS images (Figure 1E) both from May 7th are mainly attributed to the different acquisition time depending on the overpasses of the satellite. The temporal development is also clearly seen in the large changes in cloud structure between the two images.

Rate measurements of nitrogen transformation processes

^{15}N-labeling incubation experiments were carried out at three or four depths, respectively. N-loss by either anammox or denitrification was measured as the production of ^{15}N-labeled N_2 in $^{15}NH_4^+$ + ($^{14}NO_2^-$), $^{15}NO_2^-$ (+ $^{14}NH_4^+$), $^{15}NO_3^-$ + ($^{14}NO_2^-$) and $^{15}N_2O$ (isotopes: Campro scientific, concentrations: 5 µmol NH_4^+, 10 µmol NO2$^-$, 20 µmol NO_3^- 1 µmol N_2O) time-series incubations carried out in 12-ml Exetainers (Labco, UK). At each time interval (about 0, 6, 12, 24 and 48h) production in one replicate Exetainer was terminated by the addition of saturated mercuric chloride to stop biological activity. The N-isotopic composition of N_2 gas produced in these experiments was determined by gas chromatography isotope-ratio mass spectrometry (GC/IRMS, Fisons VG Optima) [108]. The N-isotopic composition of NO_2^- was determined by GC/IRMS after conversion to either N_2O by sodium azide or to N_2 by sulfamic acid [109]. The production rates of N_2 isotopes were calculated from the increase of ^{15}N-concentration over time. Only significant and linear production of ^{15}N-species without an initial lag-phase were considered (*t*-tests, $p < 0.05$; $R^2 > 0.8$). The net production rates presented here have been corrected for the mole fractions of ^{15}N in the original substrate pools but not for isotope dilution due to any other concurrent N-consumption or production processes in the course of the incubation.

Rate measurements of carbon fixation

Triplicate incubations of 4.5 l seawater were set up with water from 5, 15, 30, 80 and 100 m depth. To each bottle 4.5 ml of ^{13}C bicarbonate solution (1 g bicarbonate in 50 ml water) was added and bottles were incubated in on deck incubators shaded with blue lagoon light foil to 25% (5 and 15 m) surface irradiance and cooled with surface seawater or incubated at 12°C in the dark (30, 80 and 100 m) for 24 hours. At the end of the incubation 1-2 l were filtered on precombusted (450°C, 6 hours) GF/F filters. Filters were oven dried (50°C, 24 hours) and stored for later analysis. After smoking with 37% HCl over night, filters were dried 2 hours at 50°C and analyzed in a CHN analyser coupled to an isotope ratio monitoring mass spectrometer. The carbon fixation rate was calculated according to the enrichment of ^{13}C in the samples relative to unlabeled background values:

C_{fix} = (At% sample – At% background) / (At% label – At% background) x POC/time

with At% sample, the ratio of $^{13}C/^{12}C$ times 100 in the particulate organic carbon pool (POC), At% background the same ratio in unlabeled POC and At% label the final ratio of $^{13}C/^{12}C$ in the incubation bottle after label addition. The resulting ratio is multiplied with the concentration of POC and divided by the incubation time in days.

Averaging the CO_2-fixation rates over the predicted photic (0-20 m, ~14.4 µmol C l^{-1} d^{-1}) and aphotic (20-100 m, ~1.2 µmol C l^{-1} d^{-1}, dark incubations) zones and multiplying by the respective water depth, a photic zone CO_2-fixation of ~288 mmol C m^{-2} d^{-1} and a light-independent CO_2-fixation of ~96 mmol C m^{-2} d^{-1} was estimated.

Using the overall mean CO_2-fixation rate for the Humboldt Current system off Peru (2.18 g C m^{-2} d-1 or 182 mmol C m^{-2} d^{-1} [2] and 3.5 g C m^{-2} d-1 or 292 mmol C m^{-2} d^{-1} (Montecino and Lange, 2009)), our measured dark CO_2-fixation over an 80 m deep aphotic zone (~96 mmol C m^{-2} d^{-1}) contributed to 33-53% of the total CO_2-fixation. Extrapolating the 80 m deep dark CO_2-fixation (~96 mmol C m^{-2} d^{-1}) over the entire sulfidic plume (~8000 km^2, see calculations above), we calculated a total CO_2-fixation of ~7.7 x 10^8 mol C d^{-1} or ~9.2 kilotons C d^{-1}. Using this CO_2-fixation estimate, it would contribute 1.7% to the total primary production of the Peru Humboldt Current system as presented in Carr, 2002 (~0.2 Gtons C y^{-1} or ~550 kilotons C d^{-1}). The average CO_2-fixation rates per cell were calculated from total bacterial cell numbers as obtained from flow-cytometry.

Acknowledgments

We thank the Government of Peru for permitting research in their territorial waters, the chief scientist Martin Frank (GEOMAR) and the captain and crew of RV Meteor for support onboard ship. We acknowledge Gabriele Klockgether, Andreas Ellrott (MPIMM), Violeta Leon (IMARPE) and Hermann Bange (GEOMAR) for technical and analytical assistance. Additionally, we would like to thank Mary Ann Moran, James T Hollibaugh, Nasreen Bano

(University of Georgia, Athens) and Rachel S Poretsky (Georgia Institute of Technology, Atlanta) for methodological help with the nucleic acid sample preparation.

Financial Disclosure

Funding was provided by the WGL-pact project “REAL” (Leibniz Association), the Max Planck Society and the Helmholtz Association. The Cluster of Excellence “The Future Ocean” provided structural support. This work is a contribution of the Collaborative Research Centre 754 “climate - biogeochemistry interactions in the tropical oceans” (www.sfb754.de), which is supported by the German Research Association.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interest

The authors declare no conflict of interest.

Abbreviations

anammox, anaerobic ammonia oxidation; *apr*, adenylylsulfate reductase; DNRA, dissimilatory nitrate reduction to ammonia; *dsr*, intracellular dissimilatory sulfite reduction; EC, Enzyme Commission; FISH, fluorescence *in situ* hybridization; GSO, γ-proteobacterial sulfur oxidizer; OMZ, oxygen minimum zone; RuBisCo, Ribulose-bisphosphate carboxylase/oxygenase; *sat*, sulfate adenylyltransferase; *sox*, periplasmic sulfur oxidation; *sqr*, sulfide:quinone oxidoreductase; STOX, Switchable Trace amount Oxygen; SUP05, uncultured SUP05 cluster bacterium;

References

1. Friederich GE, Codispoti LA (1987) An Analysis of Continuous Vertical Nutrient Profiles Taken during a Cold-Anomaly Off Peru. Deep-Sea Res 34: 1049-1065.

2. Carr ME (2002) Estimation of potential productivity in Eastern Boundary Currents using remote sensing. Deep-Sea Res Pt II 49: 59-80.

3. Chavez FP, Messie M (2009) A comparison of Eastern Boundary Upwelling Ecosystems. Prog Oceanogr 83: 80-96.

4. Montecino V, Lange CB (2009) The Humboldt Current System: Ecosystem components and processes, fisheries, and sediment studies. Prog Oceanogr 83: 65-79.

5. Ryther JH (1969) Photosynthesis and fish production in the sea. Science 166: 72-76.

6. Pauly D, Christensen V (1995) Primary Production Required to Sustain Global Fisheries. Nature 374: 255-257.

7. Wyrtki K (1962) The Oxygen Minima in Relation to Ocean Circulation. Deep-Sea Res 9: 11-23.

8. Helly JJ, Levin LA (2004) Global distribution of naturally occurring marine hypoxia on continental margins. Deep-Sea Res Pt I 51: 1159-1168.

9. Stramma L, Johnson GC, Sprintall J, Mohrholz V (2008) Expanding oxygen-minimum zones in the tropical oceans. Science 320: 655-658.

10. Paulmier A, Ruiz-Pino D (2009) Oxygen minimum zones (OMZs) in the modern ocean. Prog Oceanogr 80: 113-128.

11. Revsbech NP, Larsen LH, Gundersen J, Dalsgaard T, Ulloa O, et al. (2009) Determination of ultra-low oxygen concentrations in oxygen minimum zones by the STOX sensor. Limnol Oceanogr Meth 7: 371-381.

12. Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, et al. (2010) A Cryptic Sulfur Cycle in Oxygen-Minimum-Zone Waters off the Chilean Coast. Science 330: 1375-1378.

13. Jensen MM, Lam P, Revsbech NP, Nagel B, Gaye B, et al. (2011) Intensive nitrogen loss over the Omani Shelf due to anammox coupled with dissimilatory nitrite reduction to ammonium. ISME J 5: 1660-1670.

14. Kalvelage T, Jensen MM, Contreras S, Revsbech NP, Lam P, et al. (2011) Oxygen Sensitivity of Anammox and Coupled N-Cycle Processes in Oxygen Minimum Zones. PLoS One 6: e29299. doi: 10.1371/journal.pone.0029299.

15. Emery KO, Orr WL, Rittenberg SC (1955) Nutrient budgets in the ocean. In: Essays in Natural Sciences in Honor of Captain Allan Handcock. Los Angeles: University of Southern California Press. pp. 299–310.

16. Codispoti LA, Brandes JA, Christensen JP, Devol AH, Naqvi SWA, et al. (2001) The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? Sci Mar 65: 85-105.

17. Gruber N (2004) The dynamics of the marine nitrogen cycle and atmospheric CO2. In: Oguz T, Follows M, editor. Carbon Climate Interactions. Dordrecht: Kluwer Academic, NATO ASI Series, pp. 97–148.

18. Kuypers MMM, Lavik G, Woebken D, Schmid M, Fuchs BM, et al. (2005) Massive nitrogen loss from the Benguela upwelling system through anaerobic ammonium oxidation. Proc Natl Acad Sci U S A 102: 6478-6483.

19. Hamersley MR, Lavik G, Woebken D, Rattray JE, Lam P, et al. (2007) Anaerobic ammonium oxidation in the Peruvian oxygen minimum zone. Limnol Oceanogr 52: 923-933.

20. Lam P, Lavik G, Jensen MM, van de Vossenberg J, Schmid M, et al. (2009) Revising the nitrogen cycle in the Peruvian oxygen minimum zone. Proc Natl Acad Sci U S A 106: 4752-4757.

21. Thamdrup B, Dalsgaard T, Jensen MM, Ulloa O, Farias L, et al. (2006) Anaerobic ammonium oxidation in the oxygen-deficient waters off northern Chile. Limnol Oceanogr 51: 2145-2156.

22. Diaz RJ, Rosenberg R (2008) Spreading dead zones and consequences for marine ecosystems. Science 321: 926-929.

23. Ward BB, Glover HE, Lipschultz F (1989) Chemoautotrophic activity and nitrification in the oxygen minimum zone off Peru. Deep-Sea Res 36: 1031-1051.

24. Jorgensen BB (1982) Ecology of the bacteria of the sulphur cycle with special reference to anoxic-oxic interface environments. Philos Trans R Soc Lond B 298: 543-561.

25. Bruchert V, Jorgensen BB, Neumann K, Riechmann D, Schlosser M, et al. (2003) Regulation of bacterial sulfate reduction and hydrogen sulfide fluxes in the central Namibian coastal upwelling zone. Geochim Cosmochim Ac 67: 4505-4518.

26. Lavik G, Stuhrmann T, Bruchert V, Van der Plas A, Mohrholz V, et al. (2008) Detoxification of sulphidic African shelf waters by blooming chemolithotrophs. Nature 457: 581-584.

27. Shao MF, Zhang T, Fang HHP (2010) Sulfur-driven autotrophic denitrification: diversity, biochemistry, and engineering applications. Appl Microbiol Biotechnol 88: 1027-1042.

28. Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ (2011) Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. Microbiol Mol Biol Rev 75: 361-422.

29. Dugdale RC, Goering JJ, Barber RT, Smith RL, Packard TT (1977) Denitrification and Hydrogen-Sulfide in Peru Upwelling Region during 1976. Deep-Sea Res 24: 601-608.

30. Naqvi SW, Jayakumar DA, Narvekar PV, Naik H, Sarma VV, et al. (2000) Increased marine production of N2O due to intensifying anoxia on the Indian continental shelf. Nature 408: 346-349.

31. Hannig M, Lavik G, Kuypers MMM, Woebken D, Martens-Habbena W, et al. (2007) Shift from denitrification to anammox after inflow events in the central Baltic Sea. Limnol Oceanogr 52: 1336-1345.

32. Stewart FJ, Ulloa O, DeLong EF (2011) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. Environ Microbiol 14: 23-40.

33. Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, et al. (2009) Metagenome of a Versatile Chemolithoautotroph from Expanding Oceanic Dead Zones. Science 326: 578-582.

34. Brettar I, Rheinheimer G (1991) Denitrification in the Central Baltic - Evidence for H2S-Oxidation as Motor of Denitrification at the Oxic-Anoxic Interface. Mar Ecol-Prog Ser 77: 157-169.

35. Brettar I, Labrenz M, Flavier S, Botel J, Kuosa H, et al. (2006) Identification of a Thiomicrospira denitrificans-like epsilonproteobacterium as a catalyst for autotrophic denitrification in the central Baltic Sea. Appl Environ Microbiol 72: 1364-1372.

36. Glaubitz S, Lueders T, Abraham WR, Jost G, Jurgens K, et al. (2009) (13)C-isotope analyses reveal that chemolithoautotrophic Gamma- and Epsilonproteobacteria feed a microbial food web in a pelagic redoxcline of the central Baltic Sea. Environ Microbiol 11: 326-337.

37. Jorgensen BB, Fossing H, Wirsen CO, Jannasch HW (1991) Sulfide Oxidation in the Anoxic Black-Sea Chemocline. Deep-Sea Res 38: S1083-S1103.

38. Luther GW, Church TM, Powell D (1991) Sulfur Speciation and Sulfide Oxidation in the Water Column of the Black-Sea. Deep-Sea Res 38: S1121-S1137.

39. Sorokin YI, Sorokin PY, Avdeev VA, Sorokin DY, Ilchenko SV (1995) Biomass, Production and Activity of Bacteria in the Black-Sea, with Special Reference to Chemosynthesis and the Sulfur Cycle. Hydrobiologia 308: 61-76.

40. Zhang JZ, Millero FJ (1993) The Chemistry of the Anoxic Waters in the Cariaco Trench. Deep-Sea Res Pt I 40: 1023-1041.

41. Hayes MK, Taylor GT, Astor Y, Scranton MI (2006) Vertical distributions of thiosulfate and sulfite in the Cariaco Basin. Limnol Oceanogr 51: 280-287.

42. Tebo BM, Emerson S (1986) Microbial manganese(II) oxidation in the marine environment - a quantitative study. Biogeochemistry 2: 149-161.

43. Hamukuaya H, O'Toole MJ, Woodhead PMJ (1998) Observations of severe hypoxia and offshore displacement of Cape hake over the Namibian shelf in 1994. S Afr J Mar Sci 19: 57-59.

44. Hart TJ, Currie RI (1960) The Benguela Current. In: Discovery Reports. Cambridge: Cambridge University Press. pp. 123-297.

45. Copenhagen WJ (1953) The periodic mortality of fish in the Walvis region - A phenomenon within the Benguela current. S Afr Div Sea Fish Invest Rep 14: 1-35.

46. Weeks SJ, Currie B, Bakun A (2002) Massive emissions of toxic gas in the Atlantic. Nature 415: 493-494.

47. Ohde T, Siegel H, Reissmann J, Gerth M (2007) Identification and investigation of sulphur plumes along the Namibian coast using the MERIS sensor. Cont Shelf Res 27: 744-756.

48. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105: 3805-3810.

49. Shi YM, Tyson GW, DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature 459: 266-U154.

50. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2010) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME J 5: 461-472.

51. Stevens H, Ulloa O (2008) Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. Environ Microbiol 10: 1244-1259.

52. Lavin P, Gonzalez B, Santibanez JF, Scanlan DJ, Ulloa O (2010) Novel lineages of Prochlorococcus thrive within the oxygen minimum zone of the eastern tropical South Pacific. Environ Microbiol Rep 2: 728-738.

53. Sunamura M, Higashi Y, Miyako C, Ishibashi J, Maruyama A (2004) Two bacteria phylotypes are predominant in the Suiyo seamount hydrothermal plume. Appl Environ Microbiol 70: 1190-1198.

54. Kuwahara H, Yoshida T, Takaki Y, Shimamura S, Nishi S, et al. (2007) Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, Calyptogena okutanii. Curr Biol 17: 881-886.

55. Newton ILG, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, et al. (2007) The Calyptogena magnifica chemoautotrophic symbiont genome. Science 315: 998-1000.

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

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

58. Desai DK, Nandi S, Srivastava PK, Lynn AM (2011) ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities. Adv Bioinformatics: 743782. doi: 10.1155/2011/743782.

59. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-222.

60. Albert DB, Taylor C, Martens CS (1995) Sulfate Reduction Rates and Low-Molecular-Weight Fatty-Acid Concentrations in the Water Column and Surficial Sediments of the Black-Sea. Deep-Sea Res Pt I 42: 1239-1260.

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

62. Meyer B, Kuever J (2007) Phylogeny of the alpha and beta subunits of the dissimilatory adenosine-5'-phosphosulfate (APS) reductase from sulfate-reducing prokaryotes - origin and evolution of the dissimilatory sulfate-reduction pathway. Microbiology+ 153: 2026-2044.

63. Finster K, Liesack W, Thamdrup B (1998) Elemental sulfur and thiosulfate disproportionation by Desulfocapsa sulfoexigens sp nov, a new anaerobic bacterium isolated from marine surface sediment. Appl Environ Microbiol 64: 119-125.

64. Fossing H (1990) Sulfate Reduction in Shelf Sediments in the Upwelling Region Off Central Peru. Cont Shelf Res 10: 355-367.

65. Niggemann J (2005) Composition and degradation of organic matter in sediments from the Peru-Chile upwelling region. Bremen: University of Bremen Press. 200 p.

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

67. Yamamoto M, Nakagawa S, Shimamura S, Takai K, Horikoshi K (2010) Molecular characterization of inorganic sulfur-compound metabolism in the deep-sea epsilonproteobacterium Sulfurovum sp NBC37-1. Environ Microbiol 12: 1144-1152.

68. Pitcher RS, Watmough NJ (2004) The bacterial cytochrome cbb3 oxidases. Biochim Biophys Acta 1655: 388-399.

69. Preisig O, Zufferey R, Thony-Meyer L, Appleby CA, Hennecke H (1996) A high-affinity cbb3-type cytochrome oxidase terminates the symbiosis-specific respiratory chain of Bradyrhizobium japonicum. J Bacteriol 178: 1532-1538.

70. Stolper DA, Revsbech NP, Canfield DE (2010) Aerobic growth at nanomolar oxygen concentrations. Proc Natl Acad Sci U S A 107: 18755-18760.

71. Forte E, Urbani A, Saraste M, Sarti P, Brunori M, et al. (2001) The cytochrome cbb3 from Pseudomonas stutzeri displays nitric oxide reductase activity. Eur J Biochem 268: 6486-6491.

72. Vanderoost J, Deboer APN, Degier JWL, Zumft WG, Stouthamer AH, et al. (1994) The Heme-Copper Oxidase Family Consists of 3 Distinct Types of Terminal Oxidases and is Related to Nitric-Oxide Reductase. FEMS Microbiol Lett 121: 1-9.

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

74. Wirsen CO, Sievert SM, Cavanaugh CM, Molyneaux SJ, Ahmad A, et al. (2002) Characterization of an Autotrophic Sulfide-Oxidizing Marine Arcobacter sp. That Produces Filamentous Sulfur. Appl Environ Microbiol 68: 316-325.

75. Schulz HN, Brinkhoff T, Ferdelman TG, Marine MH, Teske A, et al. (1999) Dense populations of a giant sulfur bacterium in Namibian shelf sediments. Science 284: 493-495.

76. Fossing H, Gallardo VA, Jorgensen BB, Huttel M, Nielsen LP, et al. (1995) Concentration and Transport of Nitrate by the Mat-Forming Sulfur Bacterium Thioploca. Nature 374: 713-715.

77. McHatton SC, Barry JP, Jannasch HW, Nelson DC (1996) High nitrate concentrations in vacuolate, autotrophic marine Beggiatoa spp. Appl Environ Microbiol 62: 954-958.

78. Stewart FJ (2011) Dissimilatory sulfur cycling in oxygen minimum zones: an emerging metagenomics perspective. Biochem Soc Trans 39: 1859-1863.

79. Grote J, Jost G, Labrenz M, Herndl GJ, Juergens K (2008) Epsilonproteobacteria Represent the Major Portion of Chemoautotrophic Bacteria in Sulfidic Waters of Pelagic Redoxclines of the Baltic and Black Seas. Appl Environ Microbiol 74: 7546-7551.

80. Grote J, Labrenz M, Pfeiffer B, Jost G, Jurgens M (2007) Quantitative distributions of Eppsilonproteobacteria and a Sulfulimonas subgroup in pelagic redoxclines of the central Baltic sea. Appl Environ Microbiol 73: 7155-7161.

81. Jost G, Zubkov MV, Yakushev E, Labrenz M, Jurgens K (2008) High abundance and dark CO2 fixation of chemolithoautotrophic prokaryotes in anoxic waters of the Baltic Sea. Limnol Oceanogr 53: 14-22.

82. Jost G, Martens-Habbena W, Pollehne F, Schnetger B, Labrenz M (2009) Anaerobic sulfur oxidation in the absence of nitrate dominates microbial chemoautotrophy beneath the pelagic chemocline of the eastern Gotland Basin, Baltic Sea. FEMS Microbiol Ecol 71: 226-236.

83. Glaubitz S, Labrenz M, Jost G, Jurgens K (2010) Diversity of active chemolithoautotrophic prokaryotes in the sulfidic zone of a Black Sea pelagic redoxcline as determined by rRNA-based stable isotope probing. FEMS Microbiol Ecol 74: 32-41.

84. Detmer AE, Giesenhagen HC, Trenkel VM, Venne HAD, Jochem FJ (1993) Phototrophic and Heterotrophic Pico-Plankton and Nanoplankton in Anoxic Depths of the Central Baltic Sea. Mar Ecol-Prog Ser 99: 197-203.

85. Chavez FP, Bertrand A, Guevara-Carrasco R, Soler P, Csirke J (2008) The northern Humboldt Current System: Brief history, present status and a view towards the future. Prog Oceanogr 79: 95-105.

86. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215: 403-410.

88. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188-7196.

89. Urich T, Lanzen A, Qi J, Huson DH, Schleper C, et al. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One 3: e2527. doi: 10.1371/journal.pone.0002527.

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

91. Clarke KR (1993) Nonparametric Multivariate Analyses of Changes in Community Structure. Aust J Ecol 18: 117-143.

92. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5(2): R12. doi: 10.1186/gb-2004-5-2-r12.

93. Pernthaler A, Pernthaler J (2007) Fluorescence in situ hybridization for the identification of environmental microbes. Methods Mol Biol 353: 153-164.

94. Daims H, Bruhl A, Amann R, Schleifer KH, Wagner M (1999) The domain-specific probe EUB338 is insufficient for the detection of all Bacteria: Development and evaluation of a more comprehensive probe set. Syst Appl Microbiol 22: 434-444.

95. Manz W, Amann R, Ludwig W, Wagner M, Schleifer KH (1992) Phylogenetic Oligodeoxynucleotide Probes for the Major Subclasses of Proteobacteria - Problems and Solutions. Syst Appl Microbiol 15: 593-600.

96. Eilers H, Pernthaler J, Peplies J, Glockner FO, Gerdts G, et al. (2001) Isolation of novel pelagic bacteria from the German bight and their seasonal contributions to surface picoplankton. Appl Environ Microbiol 67: 5134-5142.

97. Snaidr J, Amann R, Huber I, Ludwig W, Schleifer KH (1997) Phylogenetic analysis and in situ identification of bacteria in activated sludge. Appl Environ Microbiol 63: 2884-2896.

98. Moussard H, Corre E, Cambon-Bonavita MA, Fouquet Y, Jeanthon C (2006) Novel uncultured Epsilonproteobacteria dominate a filamentous sulphur mat from the 13 degrees N hydrothermal vent field, East Pacific Rise. FEMS Microbiol Ecol 58: 449-463.

99. Holmes RM, Aminot A, Kerouel R, Hooker BA, Peterson BJ (1999) A simple and precise method for measuring ammonium in marine and freshwater ecosystems. Can J Fish Aquat Sci 56: 1801-1808.

100. Grasshoff K, Ehrhardt M, Kremling K (1999) Methods of seawater analysis. Weinheim: Wiley. 600 p.

101. Walter S, Bange HW, Breitenbach U, Wallace DWR (2006) Nitrous oxide in the North Atlantic Ocean. Biogeosciences 3: 607-619.

102. Cline JD (1969) Spectrophotometric Determination of Hydrogen Sulfide in Natural Waters. Limnol Oceanogr 14: 454-458.

103. Osborn TR (1980) Estimates of the Local-Rate of Vertical Diffusion from Dissipation Measurements. J Phys Oceanogr 10: 83-89.

104. Gregg MC, Dasaro EA, Shay TJ, Larson N (1986) Observations of Persistent Mixing and near-Inertial Internal Waves. J Phys Oceanogr 16: 856-885.

105. Lam P, Jensen MM, Lavik G, McGinnis DF, Muller B, et al. (2007) Linking crenarchaeal and bacterial nitrification to anammox in the Black Sea. Proc Natl Acad Sci U S A 104: 7104-7109.

106. Fennel W (1995) A Model of the Yearly Cycle of Nutrients and Plankton in the Baltic Sea. J Marine Syst 6: 313-329.

107. Siegel H, Ohde T, Gerth M, Lavik G, Leipe T (2007) Identification of coccolithophore blooms in the SE Atlantic Ocean off Namibia by satellites and in-situ methods. Cont Shelf Res 27: 258-274.

108. Holtappels M, Lavik G, Jensen MM, Kuypers MMM (2011) (15)N-Labeling Experiments to Dissect the Contributions of Heterotrophic Denitrification and Anammox to Nitrogen Removal in the OMZ Waters of the Ocean. Method Enzymol 486: 223-251.

109. Fussel J, Lam P, Lavik G, Jensen MM, Holtappels M, et al. (2011) Nitrite oxidation in the Namibian oxygen minimum zone. ISME J 6: 1200-1209.

Supplementary information:

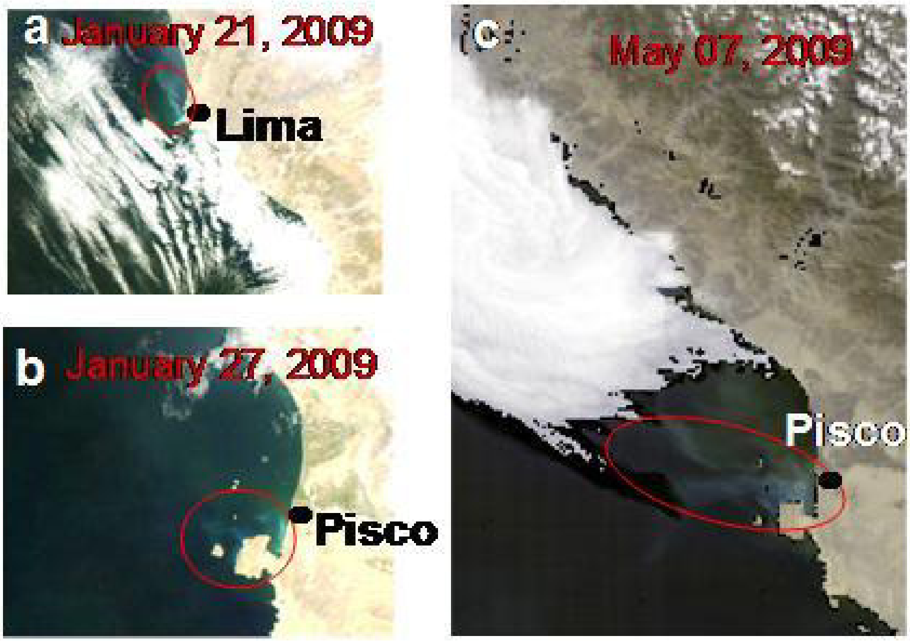


Figure S1: Satellite images of the Peruvian coast. The elemental sulfur paths are visible within the red circles. (A) Satellite image (MODIS) of the area around Lima on January, 29th, 2009. (B) Satellite image (MODIS) of the area around Pisco on January, 27th, 2009. (C) Satellite image (MERIS) of the area around Pisco on May 7th, 2009.

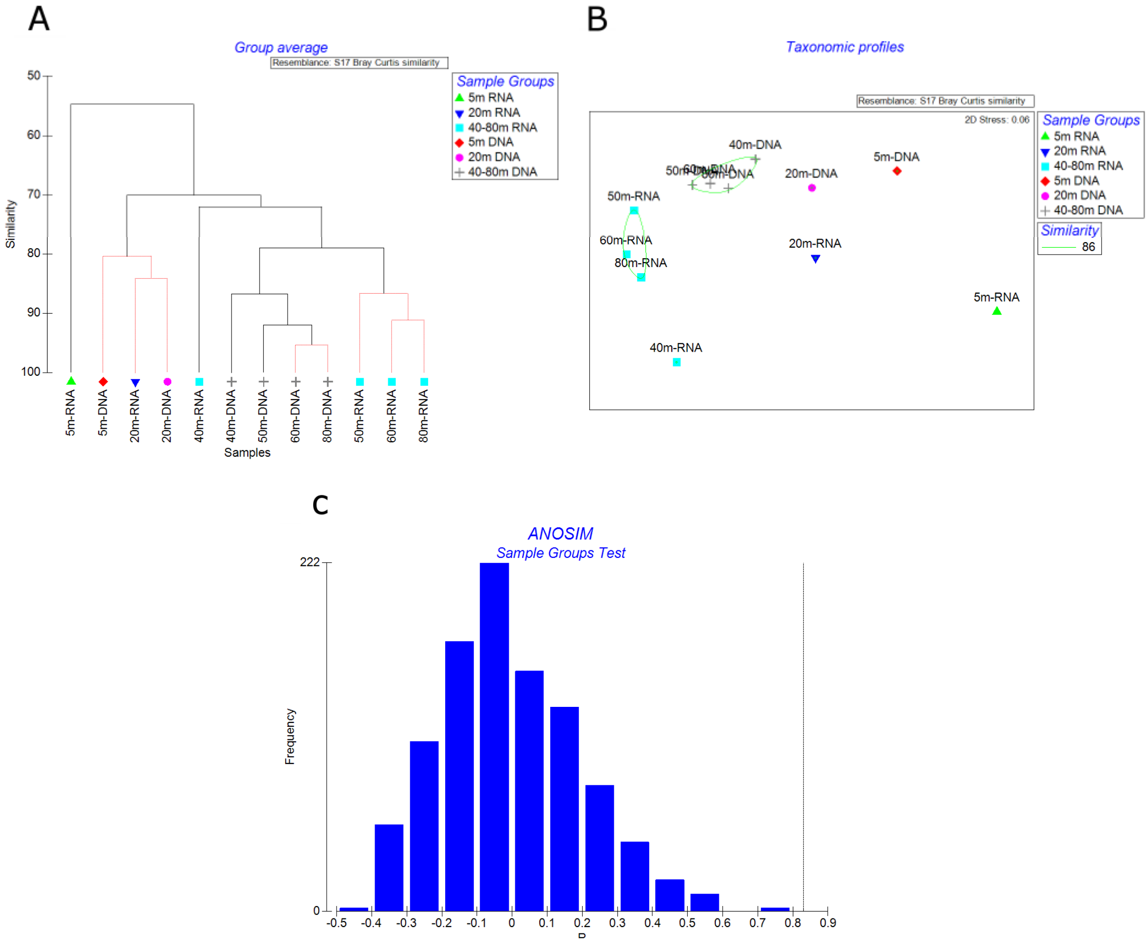


Figure S2: Multivariate statistical analysis and clustering of all protein-coding sequences based on shared taxonomic categories. Categories are chosen according to Figure 3. (A) Hierarchical clustering. (B) Non-metric Multi dimensional Scaling. Plot is labelled by prior groupings of the samples. The solid green circles mark the hierarchical clusters obtained using a similarity cut off of 86%. (C) ANOSIM test for significance of difference between the prior groupings.

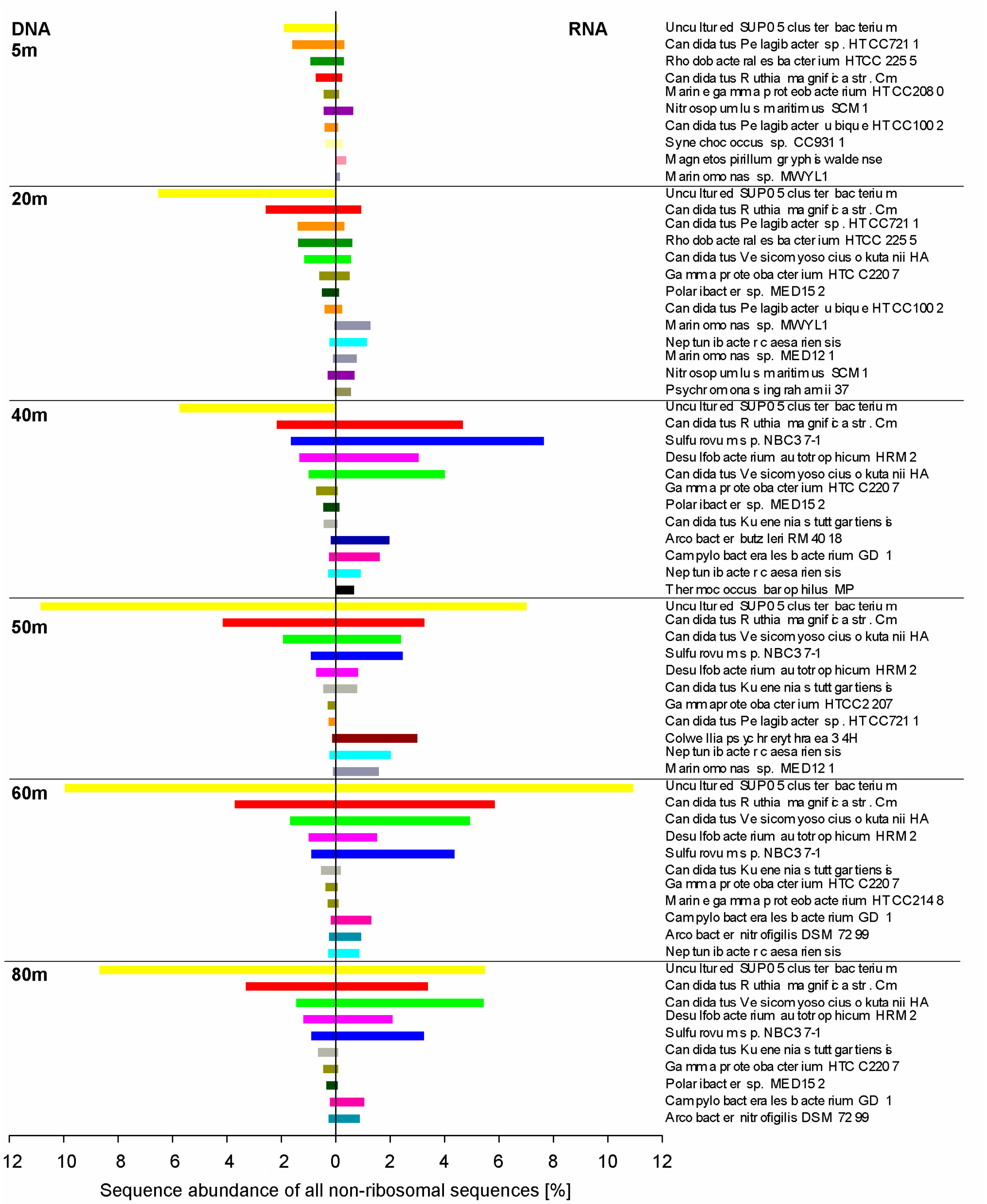


Figure S3: Vertical distribution of taxonomic affiliations. Shown are the eight most abundant organisms (on species-level) in percent of all non-ribosomal sequences in the DNA and RNA datasets; ordered descending according to DNA-counts and supplemented with the remainder of the top eight organisms from the RNA-dataset if not already present in the DNA-dataset. Note: no RNA-sequences were identified for the SUP05 cluster bacterium at 20 and 40m using BLASTx-searches against the non-redundant database of NCBI.

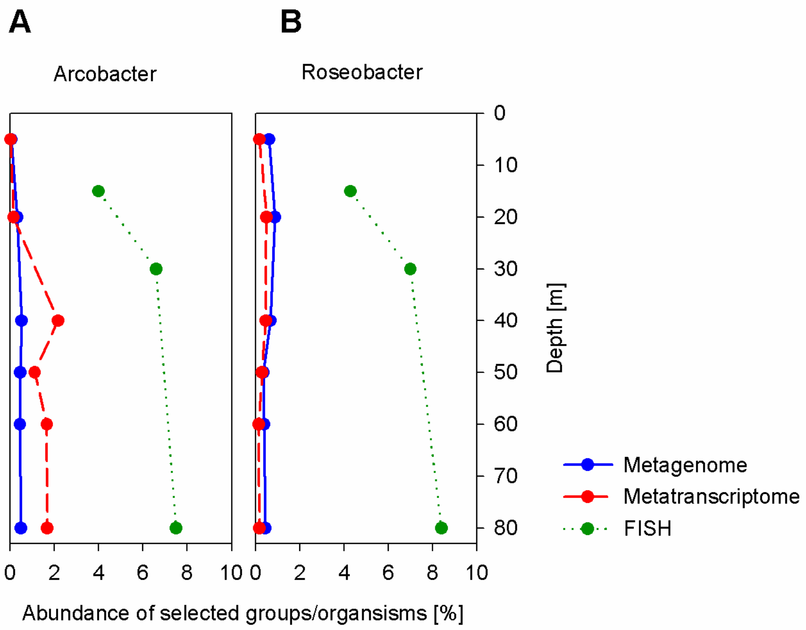
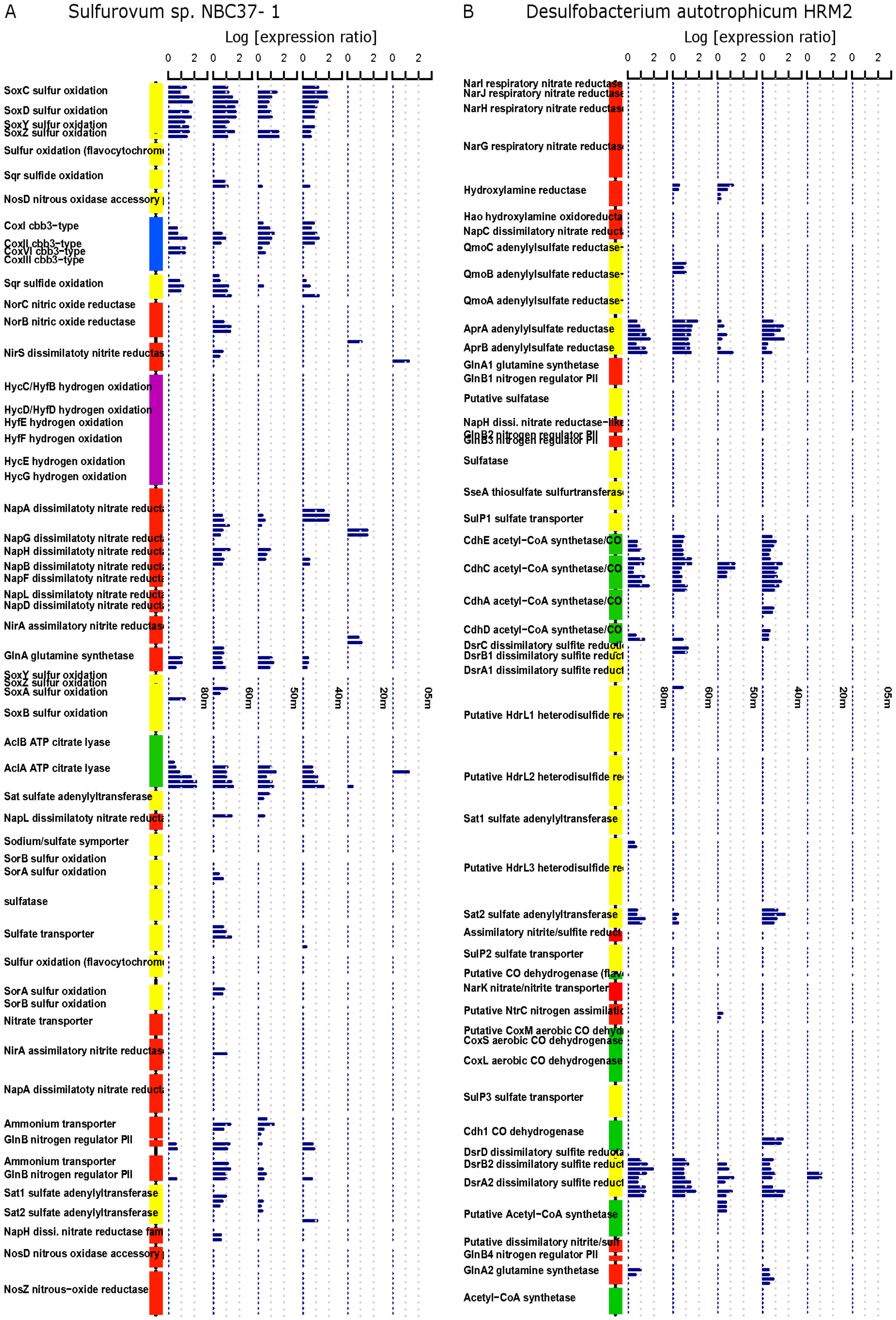


Figure S4: Vertical distribution of *Arcobacter*-group and *Roseobacter*-group FISH counts. Sequence data shown in percent of all non-ribosomal sequences and FISH counts in percent of all DAPI-stained cells, respectively.

Figure S5 (next page): Vertical distribution of sequences recruited onto the genomes of (A) *Sulfurovum* sp. NBC37- 1 and (B) *Desulfobacterium autotrophicum* HRM2. Shown are selected genes (in the corresponding order of the genomes), mostly involved in oxygen- (blue), sulfur- (yellow), nitrogen- (red), carbon- (green) and hydrogen-metabolism (purple). The y-axis depicts the log of the expression-ratio, a measure for the selective enrichment of transcripts over the corresponding gene, normalized to the total pool of protein-coding sequences. A list of the start and end position of each gene and the full name of the corresponding enzyme are shown in Table S3.


A
Sulfurovum sp. NBC37- 1
Log [expression ratio]
0 2 0 2 0 2 0 2 0 2 0 2
SoxC sulfur oxidation
SoxD sulfur oxidation
SoxY sulfur oxidation
SoxZ sulfur oxidation
Sulfur oxidation (flavocytochrom
Sqr sulfide oxidation
NosD nitrous oxidase accessory
CoxI cbb3−type
CoxII cbb3−type
CoxVI cbb3−type
CoxIII cbb3−type
Sqr sulfide oxidation
NorC nitric oxide reductase
NorB nitric oxide reductase
NirS dissimilatoty nitrite reductas
HycC/HyfB hydrogen oxidation
HycD/HyfD hydrogen oxidation
HyfE hydrogen oxidation
HyfF hydrogen oxidation
HycE hydrogen oxidation
HycG hydrogen oxidation
NapA dissimilatoty nitrate reduct
NapG dissimilatoty nitrate reduct
NapH dissimilatoty nitrate reduct
NapB dissimilatoty nitrate reduct
NapF dissimilatoty nitrate reduct
NapL dissimilatoty nitrate reduct
NapD dissimilatoty nitrate reduct
NirA assimilatory nitrite reductas
GlnA glutamine synthetase
SoxY sulfur oxidation
SoxZ sulfur oxidation
SoxA sulfur oxidation
SoxB sulfur oxidation
AclB ATP citrate lyase
AclA ATP citrate lyase
Sat sulfate adenylyltransferase
NapL dissimilatoty nitrate reduct
Sodium/sulfate symporter
SorB sulfur oxidation
SorA sulfur oxidation
sulfatase
Sulfate transporter
Sulfur oxidation (flavocytochrom
SorA sulfur oxidation
SorB sulfur oxidation
Nitrate transporter
NirA assimilatory nitrite reductas
NapA dissimilatoty nitrate reduct
Ammonium transporter
GlnB nitrogen regulator PII
Ammonium transporter
GlnB nitrogen regulator PII
Sat1 sulfate adenylyltransferase
Sat2 sulfate adenylyltransferase
NapH dissi. nitrate reductase fam
NosD nitrous oxidase accessory
NosZ nitrous−oxide reductase
80m
60m
50m
40m
20m
05m
B
Desulfobacterium autotrophicum HRM2
Log [expression ratio]
0 2 0 2 0 2 0 2 0 2 0 2
NarI respiratory nitrate reductase
NarJ respiratory nitrate reductase
NarH respiratory nitrate reductase
NarG respiratory nitrate reductase
Hydroxylamine reductase
Hao hydroxylamine oxidoreducta
NapC dissimilatory nitrate reduct
QmoC adenylylsulfate reductase-
QmoB adenylylsulfate reductase-
QmoA adenylylsulfate reductase-
AprA adenylylsulfate reductase
AprB adenylylsulfate reductase
GlnA1 glutamine synthetase
GlnB1 nitrogen regulator PII
Putative sulfatase
NapH dissi. nitrate reductase−like
GlnB2 nitrogen regulator PII
GlnB3 nitrogen regulator PII
Sulfatase
SseA thiosulfate sulfurtransferas
SulP1 sulfate transporter
CdhE acetyl−CoA synthetase/CO
CdhC acetyl−CoA synthetase/CO
CdhA acetyl−CoA synthetase/CO
CdhD acetyl−CoA synthetase/CO
DsrC dissimilatory sulfite reducti
DsrB1 dissimilatory sulfite reduct
DsrA1 dissimilatory sulfite reduct
Putative HdrL1 heterodisulfide re
Putative HdrL2 heterodisulfide re
Sat1 sulfate adenylyltransferase
Putative HdrL3 heterodisulfide re
Sat2 sulfate adenylyltransferase
Assimilatory nitrite/sulfite reduct
SulP2 sulfate transporter
Putative CO dehydrogenase (flav
NarK nitrate/nitrite transporter
Putative NtrC nitrogen assimilatio
Putative CoxM aerobic CO dehydr
CoxS aerobic CO dehydrogenase
CoxL aerobic CO dehydrogenase
SulP3 sulfate transporter
Cdh1 CO dehydrogenase
DsrD dissimilatory sulfite reducta
DsrB2 dissimilatory sulfite reduct
DsrA2 dissimilatory sulfite reduct
Putative Acetyl−CoA synthetase
Putative dissimilatory nitrite/sulf
GlnB4 nitrogen regulator PII
GlnA2 glutamine synthetase
Acetyl−CoA synthetase
80m
60m
50m
40m
20m
05m

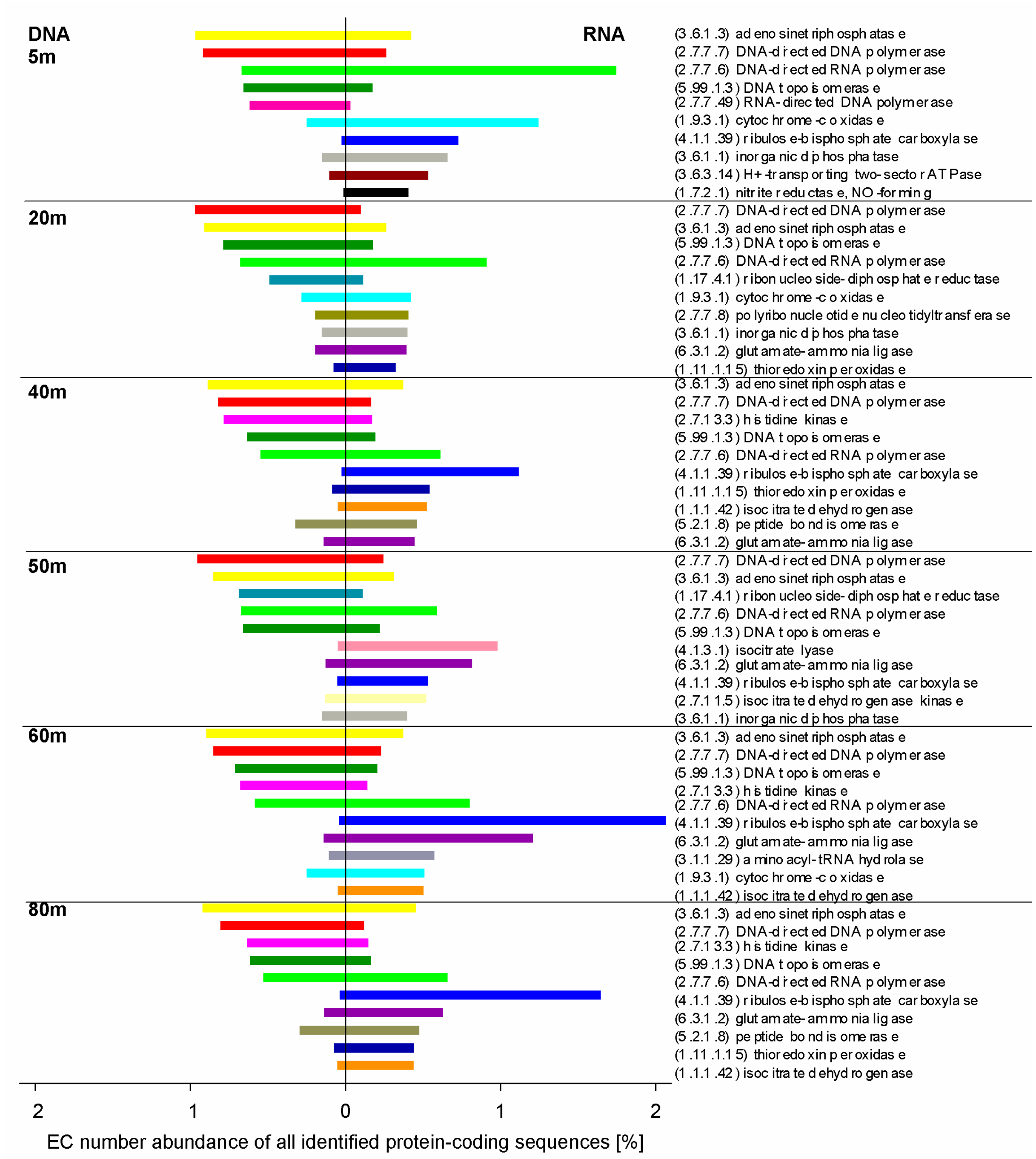


Figure S6: Vertical distribution of EC number assignments. Shown are the top five most abundant EC numbers (including enzyme names as denoted in KEGG) in percent of all protein-coding sequences in the DNA and RNA-datasets (as identified by profile hidden Markov model scans of the Pfam protein families and the ModEnzA EC groups); ordered descending according to the DNA-counts and supplemented with the remainder of the top five EC numbers from the RNA-dataset if not already present in the DNA-dataset.

Table S1: Read numbers and sequencing statistics.

[a]as obtained with Cd-hit

[b]as obtained by BLASTn-searches against the SILVA database

[c]as obtained by BLASTx-searches against the non-redundant database of NCBI and by scans with profile hidden Markov models of the ModEnzA EC groups and of the Pfam protein families

[d]average

	5m - DNA	20m - DNA	40m - DNA	50m - DNA	60m - DNA	80m - DNA	Total - DNA
Total base pairs	122,220,795	100,015,865	118,370,710	124,119,866	138,696,504	154,015,471	757,439,211
Total sequences	315,414	265,343	296,291	312,453	347,136	352,131	1,888,768
Average sequence (bp)	387	380	402	397	399	437	400
Unique sequences[a]	277,246	213,668	240,667	249,195	287,253	313,608	263,606[d]
Ribosomal sequences[b]	1,036	644	983	913	1,189	1,158	5,923
Protein-coding sequnces[c]	105,802	138,433	175,639	178,679	208,405	218,562	1,025,520
Not identified sequences	208,576	126,266	119,669	132,861	137,542	132,411	857,325
	5m - RNA	20m - RNA	40m - RNA	50m - RNA	60m - RNA	80m - RNA	Total - RNA
Total base pairs	127,903,752	71,241,493	100,799,133	101,156,732	100,252,938	97,749,062	599,103,110
Total sequences	354,281	204,205	250,292	259,169	245,161	247,851	1,560,959
Average sequence (bp)	364	349	403	390	408	394	385
Unique sequences[a]	99,476	83,474	99,238	106,424	103,656	100,443	98,785[d]
Ribosomal sequences[b]	320,329	155,959	188,480	153,495	142,620	177,068	1,137,951
Protein-coding sequnces[c]	5,599	17,605	35,500	65,181	63,450	36,809	224,144
Not identified sequences	28,353	30,641	26,312	40,493	39,091	33,974	198,864

Table S2: Metabolic and taxonomic evenness and diversity in all protein-coding sequences. For the metabolic diversity a collection of all identified EC numbers was used, while for the taxonomic diversity and all assignments from BLASTx-searches were considered. Shown are the evenness and the diversity (inverse of the Simpson's index) for both DNA and RNA datasets.

	Metabolism				Taxonomy			
	DNA		RNA		DNA		RNA	
Depths	Evenness	Diversity	Evenness	Diversity	Evenness	Diversity	Evenness	Diversity
5m	0.37	130.80	0.15	36.25	0.15	5.28	0.18	5.71
20m	0.46	163.98	0.24	66.72	0.12	4.21	0.15	5.15
40m	0.42	157.81	0.26	66.51	0.17	6.00	0.17	5.92
50m	0.38	132.64	0.25	64.09	0.10	3.67	0.10	3.53
60m	0.48	172.15	0.20	46.41	0.12	4.25	0.10	3.53
80m	0.36	128.77	0.18	42.95	0.14	4.97	0.12	4.35

Table S3: Genomic regions of the sequences recruited onto the (meta-) genomes as plotted in Figures 5 and S5. Shown are the start and end position of each gene and the corresponding enzyme name for the SUP05 cluster bacterium, *Candidatus* Ruthia magnifica str. Cm, *Candidatus* Vesicomyosocius okutanii HA, *Sulfurovum* sp. NBC37- 1 and *Desulfobacterium autotrophicum* HRM2.

***This DsrM-like protein has also high similarity to a *narG* respiratory nitrate reductase.**

SUP05 cluster bacterium

23505, 25394, SoxB sulfur oxidation
107746, 109647, AprA adenylylsulfate reductase
109864, 110197, AprB adenylylsulfate reductase
111266, 112474, Sat sulfate adenylyltransferase
173044, 173866, NirK dissi. nitrite reductase
210707, 211918, Ammonia transport
211929, 212276, GlnB nitrogen regulator PII
332050, 333357, FccB sulfur oxidation
333370, 333957, FccA sulfur oxidation
335190, 335672, NapF dissi. nitrate reductase
335722, 336267, NapB dissi. nitrate reductase
336492, 338819, NapA dissi. nitrate reductase
338925, 339794, NapH dissi. nitrate reductase
339791, 340633, NapG dissi. nitrate reductase
340643, 340897, NapD dissi. nitrate reductase
349545, 351011, NarK nitrate/nitrite transport
351026, 353407, NirB assi. nitrite reductase
359192, 359527, DsrC-like protein
359743, 360435, NarI dissi. nitrate reductase
360447, 360962, NarJ dissi. nitrate reductase
360972, 362522, NarH dissi. nitrate reductase
362538, 366281, NarG dissi. nitrate reductase
366479, 368092, NarK nitrate/nitrite transport
368124, 369533, NarK2 nitrate/nitrite transport
370535, 372781, CbbO RuBisCo (ribulose-bisphosphate carboxylase/oxygenase) regulator
372797, 373447, CbbQ RuBisCo (ribulose-bisphosphate carboxylase/oxygenase) regulator
373711, 375093, RuBisCo (ribulose-bisphosphate carboxylase/oxygenase)
448830, 449651, SoxA sulfur oxidation
449675, 449977, SoxZ sulfur oxidation
450011, 450454, SoxY sulfur oxidation
450465, 450812, SoxX sulfur oxidation
493067, 494722, Sulfate permease family protein
514544, 515821, Ammonia permease
515832, 516170, GlnB nitrogen regulator PII
588232, 589524, Sqr (sulfide:quinone oxidoreductase) sulfide oxidation

624824, 625366, DsrA sulfur oxidation
625368, 626129, DsrA sulfur oxidation
626165, 627238, DsrB sulfur oxidation
627253, 627654, DsrE sulfur oxidation
627656, 628060, DsrF sulfur oxidation
628073, 628369, DsrH-like protein
628396, 628719, DsrC sulfur oxidation
628795, 629577, DsrM sulfur oxidation
629579, 630946, DsrK sulfur oxidation
774165, 775056, Sulfate permease
868785, 870200, GlnA glutamine synthetase
1173500, 1174903, NorB nitric oxide reductase
1174900, 1175574, NorC nitric oxide reductase

Candidatus **Ruthia magnifica str. Cm**

39283, 40350, CoxII (cytochrome c oxidase)
40373, 41989, CoxI (cytochrome c oxidase)
42080, 42613, Cox (cytochrome c oxidase) assembly protein
42631, 43515, CoxIII (cytochrome c oxidase)
44618, 46036, CoxI (cytochrome c oxidase), cbb3-type
46050, 46781, CoxII (cytochrome c oxidase), cbb3-type
46945, 47847, CoxIII (cytochrome c oxidase), cbb3-type
101492, 102700, Sat sulfate adenylyltransferase
103798, 104277, AprB adenylylsulfate reductase
104277, 106160, AprA adenylylsulfate reductase
182253, 184142, SoxB sulfur oxidation
230535, 230885, GlnB nitrogen regulator PII
301598, 301936, GlnB nitrogen regulator PII
301983, 303221, Ammonium transport
307708, 308055, GlnB nitrogen regulator PII
344806, 345219, Thiosulfate sulfurtransferase
543750, 545165, GlnA glutamine synthetase
676385, 676711, DsrC family protein
739720, 741186, Nitrate/nitrite transport
741202, 743577, NirB assi. nitrite reductase
752989, 753798, CbbQ RuBisCo (ribulose-bisphosphate carboxylase/oxygenase) regulator
753897, 755279, RuBisCo (ribulose-bisphosphate carboxylase/oxygenase)
858411, 859226, SoxA sulfur oxidation protein
859254, 859556, SoxZ sulfur oxidation protein
859589, 860032, SoxY sulfur oxidation protein
860043, 860390, SoxX sulfur oxidation protein
916759, 917961, NrfD polysulfide reductase
922717, 923490, DsrM-like protein*
923564, 923887, DsrC family protein
923914, 924210, DsrH family protein

924223, 924621, DsrE family protein
924623, 925024, DsrE family protein
925037, 926110, DsrB sulfur oxidation
926183, 927484, DsrA sulfur oxidation
939512, 939802, DsrC-related protein
1130921, 1132216, Sqr (sulfide:quinone oxidoreductase) sulfide oxidation

Candidatus Vesicomyosocius okutanii HA

34428, 35537, CoxII (cytochrome c oxidase)
35563, 37164, CoxI (cytochrome c oxidase)
37255, 37788, Cox (cytochrome c oxidase) assembly protein
37806, 38690, CoxIII (cytochrome c oxidase)
39798, 41216, CoxI (cytochrome c oxidase), cbb3-type
41231, 41962, CoxII (cytochrome c oxidase), cbb3-type
41962, 42102, CoxIV (cytochrome c oxidase), cbb3-type
42121, 43023, CoxIII (cytochrome c oxidase), cbb3-type
98093, 99301, Sat sulfate adenylyltransferase
100417, 100896, AprB adenylylsulfate reductase
100896, 102779, AprA adenylylsulfate reductase
172596, 174485, SoxB sulfur oxidation
211500, 211850, GlnB nitrogen regulator PII
264163, 264501, GlnB nitrogen regulator PII
264491, 265786, Ammonium transport
270118, 270465, GlnB nitrogen regulator PII
473999, 475414, GlnA glutamine synthetase
600392, 600733, DsrC-like protein
660239, 661705, NarK nitrate transport
661708, 664077, NirB assi. nitrite reductase
669765, 670445, NarI resp. nitrate reductase
670461, 671117, NarJ resp. nitrate reductase
671114, 672655, NarH resp. nitrate reductase
672655, 676386, NarG resp. nitrate reductase
678095, 680341, CbbO RuBisCo (ribulose-bisphosphate carboxylase/oxygenase) regulator
680357, 681166, CbbQ RuBisCo (ribulose-bisphosphate carboxylase/oxygenase) regulator
681254, 682636, RuBisCo (ribulose-bisphosphate carboxylase/oxygenase)
770792, 771607, SoxA sulfur oxidation
771635, 771937, SoxZ sulfur oxidation
771971, 772414, SoxY sulfur oxidation
772425, 772772, SoxX sulfur oxidation
790563, 791690, Sqr (sulfide:quinone oxidoreductase) sulfide oxidaton
817196, 817537, DsrR sulfur oxidation
817534, 818910, DsrN sulfur oxidation
818938, 820140, DsrP sulfur oxidation
820166, 820897, DsrO sulfur oxidation
820894, 821277, DsrJ sulfur oxidation

821307, 823271, DsrL sulfur oxidation
823327, 824892, DsrK sulfur oxidation
824894, 825667, DsrM sulfur oxidation
825744, 826067, DsrC sulfur oxidation
826094, 826390, DsrH sulfur oxidation
826402, 826806, DsrF sulfur oxidation
826810, 827211, DsrE sulfur oxidation
827224, 828297, DsrB sulfur oxidation
828373, 829674, DsrA sulfur oxidation
839956, 840246, DsrC sulfur oxidation
995954, 997240, Sqr (sulfide:quinone oxidoreductase) sulfide oxidation

***Sulfurovum* sp. NBC37- 1**

53157, 54473, SoxC sulfur oxidation
54454, 55626, SoxD sulfur oxidation
55666, 56178, SoxY sulfur oxidation
56258, 56560, SoxZ sulfur oxidation
60805, 62172, Sulfur oxidation (flavocytochrome c)
73399, 74571, Sqr (sulfide:quinone oxidoreductase) sulfide oxidation
159880, 161109, NosD nitrous oxidase accessory protein
181741, 183207, CoxI (cytochrome c oxidase), cbb3-type
183219, 183908, CoxII (cytochrome c oxidase), cbb3-type
183914, 184132, CoxVI (cytochrome c oxidase), cbb3-type
184129, 185019, CoxIII (cytochrome c oxidase), cbb3-type
194939, 196390, Sqr (sulfide:quinone oxidoreductase) sulfide oxidation
242443, 243114, NorC nitric oxide reductase
243104, 244549, NorB nitric oxide reductase
244879, 246594, NirS dissimilatoty nitrite reductase
259164, 261047, HycC/HyfB hydrogen oxidation
261044, 261964, HycD/HyfD hydrogen oxidation
261968, 262561, HyfE hydrogen oxidation
262558, 263973, HyfF hydrogen oxidation
263970, 265337, HycE hydrogen oxidation
265334, 265840, HycG hydrogen oxidation
300766, 303603, NapA dissimilatoty nitrate reductase
303608, 304441, NapG dissimilatoty nitrate reductase
304441, 305244, NapH dissimilatoty nitrate reductase
305263, 306231, NapB dissimilatoty nitrate reductase
306231, 306743, NapF dissimilatoty nitrate reductase
307192, 308145, NapL dissimilatoty nitrate reductase
308156, 308515, NapD dissimilatoty nitrate reductase
313115, 314764, NirA assimilatory nitrite reductase
366880, 368310, GlnA glutamine synthetase
503764, 504243, SoxY sulfur oxidation
504299, 504637, SoxZ sulfur oxidation

504641, 505402, SoxA sulfur oxidation
505412, 507187, SoxB sulfur oxidation
551581, 552891, AclB ATP citrate lyase
552907, 554727, AclA ATP citrate lyase
526620, 527798, Sat sulfate adenylyltransferase
696198, 697148, NapL dissimilatoty nitrate reductase
993735, 995081, Sodium/sulfate symporter
1126255, 1126572, SorB sulfur oxidation
1126584, 1127789, SorA sulfur oxidation
1184827, 1186743, sulfatase
1201373, 1202956, Sulfate transporter
1378920, 1380239, Sulfur oxidation (flavocytochrome c)
1521996, 1523204, SorA sulfur oxidation
1523201, 1523530, SorB sulfur oxidation
1544689, 1545969, Nitrate transporter
1546196, 1548115, NirA assimilatory nitrite reductase
1554543, 1556867, NapA dissimilatoty nitrate reductase
1695756, 1697063, Ammonium transporter
1697198, 1697539, GlnB nitrogen regulator PII
1698115, 1699287, Ammonium transporter
1699300, 1699641, GlnB nitrogen regulator PII
1774429, 1775868, Sat1 sulfate adenylyltransferase
1775870, 1776781, Sat2 sulfate adenylyltransferase
2265182, 2266093, NapH dissi. nitrate reductase family protein
2268064, 2269287, NosD nitrous oxidase accessory protein
2270271, 2272871, NosZ nitrous-oxide reductase

***Desulfobacterium autotrophicum* HRM2**

79712, 80398, NarI respiratory nitrate reductase
80428, 80994, NarJ respiratory nitrate reductase
81017, 82438, NarH respiratory nitrate reductase
82428, 86126, NarG respiratory nitrate reductase
107440, 109101, Hydroxylamine reductase
111904, 113301, Hao hydroxylamine oxidoreductase
113343, 113867, NapC dissimilatory nitrate reductase
513380, 514555, QmoC adenylylsulfate reductase-like protein
514591, 516927, QmoB adenylylsulfate reductase-like protein
516935, 518209, QmoA adenylylsulfate reductase-like protein
518490, 520451, AprA adenylylsulfate reductase
520508, 520945, AprB adenylylsulfate reductase
765848, 767260, GlnA1 glutamine synthetase
767296, 767634, GlnB1 nitrogen regulator PII
1089653, 1091512, Putative sulfatase
1092844, 1093680, NapH dissi. nitrate reductase-like protein
1107061, 1107411, GlnB2 nitrogen regulator PII

1107408, 1107788, GlnB3 nitrogen regulator PII
1190379, 1192217, Sulfatase
1450969, 1452849, SseA thiosulfate sulfurtransferase-like protein
1515688, 1516887, SulP1 sulfate transporter
1907545, 1908885, CdhE acetyl-CoA synthetase/CO dehydrogenase
1908976, 1911189, CdhC acetyl-CoA synthetase/CO dehydrogenase
1911259, 1913295, CdhA acetyl-CoA synthetase/CO dehydrogenase
1913836, 1915137, CdhD acetyl-CoA synthetase/CO dehydrogenase
2487823, 2488140, DsrC dissimilatory sulfite reduction
2488276, 2489121, DsrB1 dissimilatory sulfite reduction
2489140, 2490195, DsrA1 dissimilatory sulfite reduction
2491674, 2496122, Putative HdrL1 heterodisulfide reductase
2577948, 2581370, Putative HdrL2 heterodisulfide reductase
2872705, 2874399, Sat1 sulfate adenylyltransferase
3505532, 3509995, Putative HdrL3 heterodisulfide reductase
3538111, 3539388, Sat2 sulfate adenylyltransferase
3795726, 3796382, Assimilatory nitrite/sulfite reductase
3797720, 3799492, SulP2 sulfate transporter
3847454, 3847750, Putative CO dehydrogenase (flavoprotein)
3913360, 3914553, NarK nitrate/nitrite transporter
4187991, 4189358, Putative NtrC nitrogen assimilation regulator
4266961, 4267788, Putative CoxM aerobic CO dehydrogenase
4267785, 4268255, CoxS aerobic CO dehydrogenase
4268246, 4270555, CoxL aerobic CO dehydrogenase
4540182, 4542308, SulP3 sulfate transporter
4606186, 4608150, Cdh1 CO dehydrogenase
4749506, 4749763, DsrD dissimilatory sulfite reductase
4749859, 4751007, DsrB2 dissimilatory sulfite reductase
4751024, 4752349, DsrA2 dissimilatory sulfite reductase
4872760, 4874640, Cdh2 CO dehydrogenase
5045877, 5048279, Putative Acetyl-CoA synthetase
5061740, 5062543, Putative dissimilatory nitrite/sulfite reductase
5130721, 5131059, GlnB4 nitrogen regulator PII
5142825, 5144153, GlnA2 glutamine synthetase
5322777, 5324543, Acetyl-CoA synthetase

Discussion

4. Discussion

The major effort of the work presented here was the analysis of microbial communities collected in oxygen-depleted and sulfidic OMZ waters. In order to achieve this, the basis for nucleic acid sample preparation and the analysis of the sequence data via different bioinformatic approaches had to be laid.

In general, high-throughput sequencing technologies have already proven to be powerful tools to describe the structure and function of microbial communities (e.g. Leininger et al., 2006; Bailly et al., 2007; Frias-Lopez et al., 2008; Shi et al., 2009; Canfield et al., 2010; Stewart et al., 2011; Toulza et al., 2012), however, metagenomic and metatranscriptomic approaches are far from being standardized or perfected. Thus, in this thesis an approved method of nucleic acid sample preparation was developed, biases resulting from the sampling process were assessed and an integrated bioinformatic analysis pipeline was established. These technical and methodological improvements are discussed first, and then followed by the major findings of the metagenomic and metatranscriptomic characterization of microbial communities from OMZ waters.

4.1. Technical improvements

Molecular biology – standardizing sample collection and processing

The initial protocol for RNA sample preparation used in this thesis was based on the work of Rachel Poretsky (Poretsky et al., 2005; Poretsky et al., 2009). Poretsky and co-workers published the first study targeting explicitly mRNA isolated from a natural environment, which is therefore the first (Sanger-sequencing-based) metatranscriptomic study focussing on environmental microbial functions (Poretsky et al., 2005). The quality of the final sequence data relies largely on the efficiency of rRNA removal. Although the molecular biological methods employed in the thesis presented here were mainly based on products commercially available, the enrichment for mRNAs is neither trivial nor free from biases. The removal of bacterial rRNA has proven to be successful in this thesis, but depending on the environment, an additional removal of eukaryotic and archaeal rRNA might be necessary in the future (manuscript C). In environments of low microbial cell density, an amplification of mRNA could be needed additionally, to receive the minimum amount of nucleic acids required for sequencing (manuscript C). In this case, a trade-off has to be considered: Either more water needs to be sampled and filtered, which would increase the sampling time and lead to biases due to changes within the microbial community (manuscript A), or an amplification of mRNA has to be carried out. Although a linear amplification presumably does not alter the relative abundance of the sequences, every manipulation of nucleic acids harbours the potential of introducing artefacts (Stewart et al., 2010). However, so far there is no optimal method of nucleic acid sample preparation and the decision on which protocol to use has to be carefully taken within the environmental context of the sampling site.

Among other parameters, the time that elapses during sample collection and processing is of pivotal importance to capture a realistic picture of the communities *in situ*. For fragile RNA samples, a maximum of 18 minutes from initial water collection to flash freezing of samples was kept in this thesis. Experiments with several prolonged sampling times showed that changes in the microbial community structure and function occurred already within a few minutes (manuscript A). The changes were not only relatively large, but most important, also selective. The most abundant taxonomic group to be detected in the first time point, the β-proteobacteria, decreased in relative abundance by 85% after only 20 minutes. After an incubation of five hours, the detectable number of β-proteobacterial sequences had decreased by 97%. In contrast, most other taxa stayed constant or showed only minor changes in abundance (manuscript A). Therefore, the sampling time had great impact on the detected microbial community and should be as short as possible.

A method for sampling seawater with a minimum of biases has been presented recently (Feike et al., 2012). The authors show the use of an *in situ* fixation system, which instantly stops all biological activity through the introduction of a fixation agent directly at the sampling site. However, for large-scale sampling campaigns over several sampling sites and water depths, this would require the parallel use of several of these fixation systems, yet self-made and not commercially available. Furthermore, the fixation agent is toxic, and thus a second set of sampling and filtration devices would be required, if also unfixed water is needed for other measurements or incubation experiments. Nevertheless, the use of in situ fixation is the only way to capture a realistic picture of the microbial community. A solution could eventually be using disposable bags, in which the fixation agent is introduced. This would minimize the contamination of sampling and filtration devices with the toxic agent.

Bioinformatics – supplementing BLAST-searches

Given the immense size of high-throughput sequencing datasets, there is an urgent need to avoid time-consuming manual corrections and to streamline and automate the analysis of unknown sequences. Up to now, the overwhelming majority of studies rely mainly on

BLAST-searches against the non-redundant database of NCBI. However, there are problems associated with BLAST and the commonly used NCBI databases. In this thesis, a set of tools was developed or acquired, which can be used to supplement BLAST-searches.

Using BLAST, the standard procedure is collecting and exporting all obtained hits to a database, in which the analysis is conducted manually in most cases. Then the hits are searched, grouped and categorized according to the scientific question and eventually counted to assess their abundance. The information of a certain BLAST-hit though is often either redundant, incomplete or even wrong. If an unknown sequence hits an entry in the database, it is considered as identified. However, if the entry reads ‘unknown protein of unknown bacteria’ there is not necessarily a gain of information. In addition, a search for a clearly identified sequence, e.g. the ‘cytochrome c oxidase’, which is an enzyme used for oxic respiration, might also deliver false negative results, for sequences annotated as ‘oxic respiration enzyme’ or ‘c*ox*’ (a commonly used abbreviation for the cytochrome c oxidase). On the other hand, sequence annotations which include the terminology ‘cytochrome c oxidase’ like ‘cytochrome c oxidase binding factor’ or ‘maturation protein of cytochrome c oxidase’ could be misinterpreted during data analysis and thus harbour the potential of being recognized as false positives. Thus, the quality of the annotations in a given database largely determines the quality of the information to be drawn out of a search.

A simulation with ribosomal RNA fragments was carried out in this study to assess both the sensitivity and the specificity of BLAST-searches. It was shown that the most commonly used thresholds and cut offs for BLAST-searches (e.g. Urich et al., 2008) result in about 20% false negatives (80% sensitivity) and 20% false positives (80% specificity, manuscript C). Therefore, BLAST-searches alone are not sufficient to characterize microbial communities realistically. There are indeed several different methods for sequence annotation available and new ones are currently being developed (e.g. Huson et al., 2007; Meyer et al., 2008; Pond et al., 2009; Finn et al., 2010; Desai et al., 2011; Kanehisa et al., 2012). Those can surpass BLAST-searches in either the sensitivity or the specificity. As a consequence, in this thesis three different methods were applied to investigate sequence data. In addition to the BLAST-searches, profile HMM scans of the ModEnzA EC groups (Desai et al., 2011) and of the Pfam protein families (Finn et al., 2010), as well as the recruitment of sequences onto reference genomes were used.

The great advantage of using EC numbers or Pfam assignments, the second approach of sequence analysis applied in this thesis is the distinct character of the gathered information. A period-separated four digit number (EC number) or an alphanumeric code (Pfam assignment) that is assigned to unknown sequences allows a clear functional identification and opens possibilities for an automated analysis with a minimum of manual correction. The sequence data can thus be categorized and analyzed according to functional groupings. FROMP, the java-based analysis tool that was developed during this thesis, applies this concept (manuscript B). However, EC numbers still lack phylogenetic information and rely on a smaller database when compared to BLAST and NCBI, and therefore these methods should be used to complement each other. This allows to combine sequence-sequence alignments of high sensitivity with profile-sequence alignments of high specificity, delivering a maximum of information from the sequence datasets (manuscript C).

The third method for sequence identification employed in this thesis, the recruitment of the raw sequence data onto prominent reference genomes, can eventually visualize expression patterns and reveal the full activity of the microorganisms of interest (manuscript C). However, this approach is limited as the initial identification of the abundant organisms is based on BLAST-searches and as it is only applicable, if reference genomes are available.

In combination though, all three techniques are powerful tools to investigate sequence data and probably deliver a most realistic view of microbial communities.

As part of the developed analysis pipeline, the data obtained (sequences, BLAST-hits, EC numbers and Pfam assignments) are stored subsequently in a user-friendly MySQL database, together with all relevant metadata (manuscript B). This database can be accessed via a phpMyAdmin tool through a common web browser, even by persons with only little background in bioinformatics. This is a serious possibility for laymen to analyze high-throughput sequence data. All scripts and tools needed for the use of this pipeline are freely available online (manuscript B).

4.2. Sulfidic ocean waters – a rare occurrence or increasingly common?

All techniques and bioinformatic methods developed during this thesis were applied in manuscript C to provide a detailed and quantitative description of microbial communities and their functional role in the sulfidic OMZ waters off the Peruvian coast. The dataset obtained, represents the first coupled metagenomic-metatranscriptomic analysis of sulfidic ocean waters. It was combined with cell counts, rate measurements, water column profiles, flux calculations and data from satellite remote sensing to fully characterize the H_2S-containing waters. This allowed a first insight into the lifestyles of the endemic microbial community and into potential mechanisms of H_2S formation and detoxification.

Generally, the microbial communities were dominated by several different chemolitho-autotrophic α-, γ-, δ- and ε-proteobacterial groups involved in either sulfur oxidation or SO_4^{2-} reduction (manuscript C). The recruitment of the sequence data on reference genomes allowed the identification and partial reconstruction of three different lineages of sulfur oxidizers (related to the SUP05 cluster bacterium, *Candidatus* Ruthia magnifica str. Cm, *Candidatus* Vesicomyosocius okutanii HA) and two of SO_4^{2-} reducers (similar to *Sulfurovum* sp. NBC37-1 and *Desulfobacterium autotrophicum* HRM2), potentially using different metabolic strategies. These organisms, representing the five most abundant in the sulfidic waters, contributed only a minor proportion to the total bacterioplankton in non-sulfidic OMZ waters (Canfield et al., 2010; Stewart et al., 2011). On the other hand, 'classical' inhabitants of OMZ waters were very scarce in sulfidic waters.

The combined data suggested that the detoxification of H_2S occurred not only in the expected way through the autotrophic reduction of NO_x, the sulfur-driven autotrophic denitrification (Lavik et al., 2009; Walsh et al., 2009), but also through aerobic respiration (reduction of O_2). Most likely, the microbial community was exploiting a wide range of oxidants (O_2, NO_3^-, NO_2^-, NO and N_2O) to detoxify the waters. Although the identified oxidants were mostly below or little above the detection limit, they probably played a greater role than expected. Thus, the detoxification mechanisms identified in this thesis relied mostly on substances that were actually limiting (manuscript C).

The sequence data further suggested that SO_4^{2-} reduction could have taken place. Since many sulfur cycling proteins can function in both the oxidation and reduction of sulfur species (Meyer and Kuever, 2007a, b), also the disproportionation of sulfur compounds could have occurred at the sampling site (Finster et al., 1998). The presence of large plumes of elemental sulfur in the study area, as observed in satellite data could have been used this chemoautotrophic reaction (manuscript C). Both processes, SO_4^{2-} reduction and disproportionation of sulfur compounds form H_2S and could therefore lead to a stabilization of sulfidic waters within the OMZ.

Additionally, the microbial community was dominated largely by chemoautotrophic microorganisms and as a consequence, light-independent carbon-fixation rates throughout the sulfidic zone were high (manuscript C). They reached about 30% of the photoautotrophic carbon fixation from the surface, which is somewhat unexpected, since the study area is known for high phototrophic production (Ryther, 1969; Pauly and Christensen, 1995; Chavez and Messie, 2009). High chemoautotrophic carbon fixation might be a major factor influencing the stability of sulfidic waters masses. The produced biomass could fuel further

SO_4^{2-} reduction and thereby contribute to additional H_2S formation, creating a negative feedback which stabilizes the sulfidic zones as a more permanent feature in OMZs.

The plume observed in January 2009 was the largest ever reported for oceanic waters and the first to be characterized in Peruvian waters. It might have persisted for several months, or reoccurred later that year, as the remote sensing of satellite data suggested. So far, initiation, duration and termination mechanisms of these toxic events are only poorly understood, but of major importance, since they limit strongly the habitat for most multicellular organisms and have already been attributed to massive fish kills (Copenhagen, 1954; Hart and Currie, 1960; Hamukuaya et al., 1998). Global climate change scenarios are thought to intensify oxygen-depletion (Stramma et al., 2008) and thus, as sulfidic events are extreme specification of OMZs, their frequency and extent are likely to increase in the future, too. Large-scale sulfidic waters could have dramatic consequences on fisheries and the quality of life on densely populated coastal areas, since the most productive fishing grounds of the world are found adjacent to upwelling-derived OMZs (Ryther, 1969; Pauly and Christensen, 1995; Carr, 2002; Montecino and Lange, 2009; Chavez and Messie, 2009).

4.3. Outlook

The data presented in this thesis raises several new questions. First of all, the observed highly abundant β-proteobacteria were rarely detected in oxygen-depleted waters so far and only in minor abundances (Molina et al., 2007; Lam et al., 2009; Stewart et al., 2011). Furthermore, the majority of β-proteobacterial sequences could not be affiliated to a known species. This might be due to the fact that a representative β-proteobacterial genome is currently not available, however, urgently needed to assess the functional role of this group. There have already been studies presenting genomes solely from metagenomic sequencing efforts (e.g. Tyson et al., 2004; Poinar et al., 2006; Zehr et al., 2008; Walsh et al., 2009). Promising DNA-samples from the Peruvian OMZ (the same station as analyzed in manuscript A) have been processed in this thesis and are actually available for a metagenomic reconstruction of a β-proteobacterial genome in the future.

Additionally, several other metagenomic samples, next to those that were presented above, have already been sequenced during the work for this thesis and the analysis is currently ongoing. These samples originated from an open-ocean site approximately 350 km offshore the Peruvian coast. Consequently, they are from a distinctly different oxygen-depleted habitat than those samples presented in manuscripts A and C, which were from coastal (15 km offshore) and shelf (100 km offshore) waters. The prospected way of analysis could be the use of functional categories based on EC numbers and metabolic pathways, which might eventually allow identifying those genes, which are distinctly different in abundance at the sampling sites. The ultimate aim would be the correlation of gene abundance with environmental parameters, using multivariate statistical analysis. This approach could potentially allow the prediction of the response of microbial communities to changing environmental parameters.

These capacious datasets from sampling sites with different O_2 and nutrient concentrations could help to further expand the knowledge of microbial dynamics in both sulfidic and non-sulfidic oxygen-depleted subsurface waters. However, more sampling campaigns, especially to sulfidic OMZ waters are necessary in the future.

Sequencing technologies have shown unexpectedly rapid advances within only a few years. The first sequencing efforts of the three billion base pairs of a human genome were international, collaborative projects, which lasted for about a decade (Lander et al., 2001; Venter et al., 2001). Only a few years later, it was claimed, that a human genome was sequenced within merely a few days (Hayden, 2009) or even from a single DNA molecule (Pushkarev et al., 2009).

If in the future technological enhancements will be of similar pace, automated sampling and analysis tools could be used routinely within a few years. It is imaginable that autonomous systems on-board of research vessels or moored instruments sample continuously and worldwide, automatically sequence and analyze the microbial communities, and eventually transferring the data to land-based research institutes. If bioinformatic capacities are sufficient to handle the generated data, they might after all allow the understanding of the microbial basis of such a complex ecosystem as the ocean.

References

5. References

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.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic Identification and in-Situ Detection of Individual Microbial-Cells without Cultivation. *Microbiol Rev* **59**: 143-169.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *J Exp Med* **79**: 137-158.

Bailly, J., Fraissinet-Tachet, L., Verner, M.C., Debaud, J.C., Lemaire, M., Wesolowski-Louvel, M., and Marmeisse, R. (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J* **1**: 632-642.

Barz, M., Beimgraben, C., Staller, T., Germer, F., Opitz, F., Marquardt, C. et al. (2010) Distribution Analysis of Hydrogenases in Surface Waters of Marine and Freshwater Environments. *PLoS One* **5**(11): e13846. doi: 10.1371/journal.pone.0013846.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S. et al. (2004) The Pfam protein families database. *Nucleic Acids Res* **32**: D138-D141.

Beman, M.J., Arrigo, K.R., and Matson, P.A. (2005) Agricultural runoff fuels large phytoplankton blooms in vulnerable areas of the ocean. *Nature* **434**: 211-214.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**: 365-370.

Bourne, P.E., Briedis, K., Dupont, C.L., Valas, R., and Yang, S. (2010) Chapter 8 - Genome Evolution Studied through Protein Structure. In: *Evolutionary Genomics and Systems Biology*. Chichester, UK: Wiley-Blackwell Inc.

Brettar, I., and Rheinheimer, G. (1991) Denitrification in the Central Baltic - Evidence for H2s-Oxidation as Motor of Denitrification at the Oxic-Anoxic Interface. *Mar Ecol-Prog Ser* **77**: 157-169.

Brettar, I., Labrenz, M., Flavier, S., Botel, J., Kuosa, H., Christen, R., and Hofle, M.G. (2006) Identification of a Thiomicrospira denitrificans-like epsilonproteobacterium as a catalyst for autotrophic denitrification in the central Baltic Sea. *Appl Environ Microbiol* **72**: 1364-1372.

Bruchert, V., Jorgensen, B.B., Neumann, K., Riechmann, D., Schlosser, M., and Schulz, H. (2003) Regulation of bacterial sulfate reduction and hydrogen sulfide fluxes in the central Namibian coastal upwelling zone. *Geochim Cosmochim Ac* **67**: 4505-4518.

Canfield, D.E., and Thamdrup, B. (2009) Towards a consistent classification scheme for geochemical environments, or, why we wish the term 'suboxic' would go away. *Geobiology* **7**: 385-392.

Canfield, D.E., Stewart, F.J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E.F. et al. (2010) A Cryptic Sulfur Cycle in Oxygen-Minimum-Zone Waters off the Chilean Coast. *Science* **330**: 1375-1378.

Carr, M.E. (2002) Estimation of potential productivity in Eastern Boundary Currents using remote sensing. *Deep-Sea Res PII* **49**: 59-80.

Castro-Gonzalez, M., and Farias, L. (2004) N(2)O cycling at the core of the oxygen minimum zone off northern Chile. *Mar Ecol-Prog Ser* **280**: 1-11.

Castro-Gonzalez, M., Braker, G., Farias, L., and Ulloa, O. (2005) Communities of nirS-type denitrifiers in the water column of the oxygen minimum zone in the eastern South Pacific. *Environ Microbiol* **7**: 1298-1306.

Chaudhary, A., Haack, S.K., Duris, J.W., and Marsh, T.L. (2009) Bacterial and Archaeal Phylogenetic Diversity of a Cold Sulfur-Rich Spring on the Shoreline of Lake Erie, Michigan. *Appl Environ Microbiol* **75**: 5025-5036.

Chavez, F.P., and Messie, M. (2009) A comparison of Eastern Boundary Upwelling Ecosystems. *Prog Oceanogr* **83**: 80-96.

Chen, K., and Pachter, L. (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* **1**(2): e24. doi: 10.1371/journal.pcbi.0010024.

Chen, Y., Wu, L.Q., Boden, R., Hillebrand, A., Kumaresan, D., Moussard, H. et al. (2009) Life without light: microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile Cave. *ISME J* **3**: 1093-1104.

Codispoti, L.A. (2007) An oceanic fixed nitrogen sink exceeding 400 Tg Na(-1) vs the concept of homeostasis in the fixed-nitrogen inventory. *Biogeosciences* **4**: 233-253.

Codispoti, L.A., Brandes, J.A., Christensen, J.P., Devol, A.H., Naqvi, S.W.A., Paerl, H.W., and Yoshinari, T. (2001) The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Sci Mar* **65**: 85-105.

Copenhagen, W.J. (1954) The periodic mortality of fish in the Walvis region - A phenomenon within the Benguela current. *S Afr Div Sea Fish Invest Rep* **14**: 1-35.

Correns, C. (1899) Untersuchungen über die Xenien bei Zea mays. *Ber Deut Bot Ges* **17**: 410-418.

Cypionka, H. (2010) Allgemeine Bioenergetik. In: *Grundlagen der Mikrobiologie*. Berlin, Heidelberg: Springer.

Danovaro, R., Dell'Anno, A., Pusceddu, A., Gambi, C., Heiner, I., and Kristensen, R.M. (2010) The first metazoa living in permanently anoxic conditions. *BMC Biol* **8**: 30. doi: 10.1186/1741-7007-8-30.

DeLong, E.F. (2009) The microbial ocean from genomes to biomes. *Nature* **459**: 200-206.

Desai, D.K., Nandi, S., Srivastava, P.K., and Lynn, A.M. (2011) ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities. *Adv Bioinformatics*: **743782**. doi: 10.1155/2011/743782.

Deutsch, C., Sarmiento, J.L., Sigman, D.M., Gruber, N., and Dunne, J.P. (2007) Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* **445**: 163-167.

Dugdale, R.C. (1972) Chemical oceanography and primary productivity in upwelling regions. *Geoforum* **3(3)**: 47–61.

Dugdale, R.C., Goering, J.J., Barber, R.T., Smith, R.L., and Packard, T.T. (1977) Denitrification and Hydrogen-Sulfide in Peru Upwelling Region during 1976. *Deep-Sea Res* **24**: 601-608.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863-14868.

Ekman, V.W. (1905) On the influence of the earth's rotation on ocean currents. *Arch. Math. Astron. Phys.* **2**: 11.

Eloe, E.A., Fadrosh, D.W., Novotny, M., Zeigler Allen, L., Kim, M., Lombardo, M.J. et al. (2011) Going deeper: metagenome of a hadopelagic microbial community. *PLoS One* **6**(5): e20388. doi: 10.1371/journal.pone.0020388.

Emery, K.O., Orr, W.L., and Rittenberg, S.C. (1955) Essays in Natural Sciences in Honor of Captain Allan Handcock. In: *Nutrient budgets in the ocean*. Los Angeles, USA: University of Southern California Press.

Engel, A.S. (2007) Observations on the biodiversity of sulfidic karst habitats. *J Cave Karst Stud* **69**: 187-206.

Engel, A.S., Porter, M.L., Stern, L.A., Quinlan, S., and Bennett, P.C. (2004) Bacterial diversity and ecosystem function of filamentous microbial mats from aphotic (cave) sulfidic springs dominated by chemolithoautotrophic "Epsilonproteobacteria". *FEMS Microbiol Ecol* **51**: 31-53.

Feike, J., Jurgens, K., Hollibaugh, J.T., Kruger, S., Jost, G., and Labrenz, M. (2012) Measuring unbiased metatranscriptomics in suboxic waters of the central Baltic Sea using a new in situ fixation system. *ISME J* **6**: 461-470.

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, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**: W29-W37.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E. et al. (2010) The Pfam protein families database. *Nucleic Acids Res* **38**: D211-222.

Finster, K., Liesack, W., and Thamdrup, B. (1998) Elemental sulfur and thiosulfate disproportionation by Desulfocapsa sulfoexigens sp nov, a new anaerobic bacterium isolated from marine surface sediment. *Appl Environ Microbiol* **64**: 119-125.

Finster, K.W., and Kjeldsen, K.U. (2010) Desulfovibrio oceani subsp oceani sp nov., subsp nov and Desulfovibrio oceani subsp galateae subsp nov., novel sulfate-reducing bacteria isolated from the oxygen minimum zone off the coast of Peru. *Anton Leeuw Int J G* **97**: 221-229.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R. et al. (1995) Whole-Genome Random Sequencing and Assembly of Haemophilus-Influenzae Rd. *Science* **269**: 496-512.

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. *Proc Natl Acad Sci U S A* **105**: 3805-3810.

Friederich, G.E., and Codispoti, L.A. (1987) An Analysis of Continuous Vertical Nutrient Profiles Taken during a Cold-Anomaly Off Peru. *Deep-Sea Res* **34**: 1049-1065.

Galloway, J.N., Dentener, F.J., Capone, D.G., Boyer, E.W., Howarth, R.W., Seitzinger, S.P. et al. (2004) Nitrogen cycles: past, present, and future. *Biogeochemistry* **70**: 153-226.

Gianoulis, T.A., Raes, J., Patel, P.V., Bjornson, R., Korbel, J.O., Letunic, I. et al. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106**: 1374-1379.

Gilbert, J.A., and Dupont, C.L. (2011) Microbial Metagenomics: Beyond the Genome. *Ann Rev Mar Sci* **3**: 347-371.

Gilbert, J.A., Field, D., Huang, Y., Edwards, R., Li, W.Z., Gilna, P., and Joint, I. (2008) Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *PLoS One* **3**(8): e3042. doi: 10.1371/journal.pone.0003042.

Glaubitz, S., Labrenz, M., Jost, G., and Jurgens, K. (2010) Diversity of active chemolithoautotrophic prokaryotes in the sulfidic zone of a Black Sea pelagic redoxcline as determined by rRNA-based stable isotope probing. *FEMS Microbiol Ecol* **74**: 32-41.

Glaubitz, S., Lueders, T., Abraham, W.R., Jost, G., Jurgens, K., and Labrenz, M. (2009) C-13-isotope analyses reveal that chemolithoautotrophic Gamma- and Epsilonproteobacteria feed a microbial food web in a pelagic redoxcline of the central Baltic Sea. *Environ Microbiol* **11**: 326-337.

Glockner, F.O., and Joint, I. (2010) Marine microbial genomics in Europe: current status and perspectives. *Microb Biotechnol* **3**: 523-530.

Grantham, B.A., Chan, F., Nielsen, K.J., Fox, D.S., Barth, J.A., Huyer, A. et al. (2004) Upwelling-driven nearshore hypoxia signals ecosystem and oceanographic changes in the northeast Pacific. *Nature* **429**: 749-754.

Gruber, N. (2004) The dynamics of the marine nitrogen cycle and atmospheric CO2. In: *Carbon Climate Interactions. NATO ASI Series*. Dordrecht, Netherlands: Kluwer Academic.

Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.

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

Hamersley, M.R., Lavik, G., Woebken, D., Rattray, J.E., Lam, P., Hopmans, E.C. et al. (2007) Anaerobic ammonium oxidation in the Peruvian oxygen minimum zone. *Limnol Oceanogr* **52**: 923-933.

Hamukuaya, H., O'Toole, M.J., and Woodhead, P.M.J. (1998) Observations of severe hypoxia and offshore displacement of Cape hake over the Namibian shelf in 1994. *S Afr J Mar Sci* **19**: 57-59.

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. *Chem Biol* **5**: R245-R249.

Hannig, M., Lavik, G., Kuypers, M.M.M., Woebken, D., Martens-Habbena, W., and Jurgens, K. (2007) Shift from denitrification to anammox after inflow events in the central Baltic Sea. *Limnol Oceanogr* **52**: 1336-1345.

Haque, M.M., Ghosh, T.S., Komanduri, D., and Mande, S.S. (2009) SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **25**: 1722-1730.

Hart, T.J., and Currie, R.I. (1960) The Benguela Current. In: *Discovery Reports*. Cambridge, UK: Cambridge University Press.

Hayden, E.C. (2009) Genome sequencing: the third generation. *Nature* **457**: 768-769.

Hayes, M.K., Taylor, G.T., Astor, Y., and Scranton, M.I. (2006) Vertical distributions of thiosulfate and sulfite in the Cariaco Basin. *Limnol Oceanogr* **51**: 280-287.

Helly, J.J., and Levin, L.A. (2004) Global distribution of naturally occurring marine hypoxia on continental margins. *Deep-Sea Res PI* **51**: 1159-1168.

Huang, X., and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res* **9**: 868-877.

Huber, J.A., Cantin, H.V., Huse, S.M., Welch, D.B.M., Sogin, M.L., and Butterfield, D.A. (2010) Isolated communities of Epsilonproteobacteria in hydrothermal vent fluids of the Mariana Arc seamounts. *FEMS Microbiol Ecol* **73**: 538-549.

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

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377-386.

Huson, D.H., Reinert, K., Kravitz, S.A., Remington, K.A., Delcher, A.L., Dew, I.M. et al. (2001) Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics* **17**: S132-139.

Inskeep, W.P., Rusch, D.B., Jay, Z.J., Herrgard, M.J., Kozubal, M.A., Richardson, T.H. et al. (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5**(3): e9773. doi: 10.1371/journal.pone.0009773.

Jensen, M.M., Lam, P., Revsbech, N.P., Nagel, B., Gaye, B., Jetten, M.S., and Kuypers, M.M. (2011) Intensive nitrogen loss over the Omani Shelf due to anammox coupled with dissimilatory nitrite reduction to ammonium. *ISME J* **5**: 1660-1670.

Jorgensen, B.B. (1982) Ecology of the bacteria of the sulphur cycle with special reference to anoxic-oxic interface environments. *Philos Trans R Soc Lond B Biol Sci* **298**: 543-561.

Jorgensen, B.B., Fossing, H., Wirsen, C.O., and Jannasch, H.W. (1991) Sulfide Oxidation in the Anoxic Black-Sea Chemocline. *Deep-Sea Res* **38**: S1083-S1103.

Kalvelage, T., Jensen, M.M., Contreras, S., Revsbech, N.P., Lam, P., Gunter, M. et al. (2011) Oxygen Sensitivity of Anammox and Coupled N-Cycle Processes in Oxygen Minimum Zones. *PLoS One* **6**(12): e29299. doi: 10.1371/journal.pone.0029299.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**: D109-114.

Karstensen, J., Stramma, L., and Visbeck, M. (2008) Oxygen minimum zones in the eastern tropical Atlantic and Pacific oceans. *Prog Oceanogr* **77**: 331-350.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**(2): R12. doi: 10.1186/gb-2004-5-2-r12.

Kuwahara, H., Yoshida, T., Takaki, Y., Shimamura, S., Nishi, S., Harada, M. et al. (2007) Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, Calyptogena okutanii. *Curr Biol* **17**: 881-886.

Kuypers, M.M.M., Lavik, G., Woebken, D., Schmid, M., Fuchs, B.M., Amann, R. et al. (2005) Massive nitrogen loss from the Benguela upwelling system through anaerobic ammonium oxidation. *Proc Natl Acad Sci U S A* **102**: 6478-6483.

Lam, P., and Kuypers, M.M. (2011) Microbial nitrogen cycling processes in oxygen minimum zones. *Ann Rev Mar Sci* **3**: 317-345.

Lam, P., Lavik, G., Jensen, M.M., van de Vossenberg, J., Schmid, M., Woebken, D. et al. (2009) Revising the nitrogen cycle in the Peruvian oxygen minimum zone. *Proc Natl Acad Sci U S A* **106**: 4752-4757.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Lavik, G., Stuhrmann, T., Bruchert, V., Van der Plas, A., Mohrholz, V., Lam, P. et al. (2009) Detoxification of sulphidic African shelf waters by blooming chemolithotrophs. *Nature* **457**: 581-584.

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G.W. et al. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**: 806-809.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

Lipman, D.J., and Pearson, W.R. (1985) Rapid and Sensitive Protein Similarity Searches. *Science* **227**: 1435-1441.

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

Luther, G.W., Church, T.M., and Powell, D. (1991) Sulfur Speciation and Sulfide Oxidation in the Water Column of the Black-Sea. *Deep-Sea Res* **38**: S1121-S1137.

Macalady, J.L., Dattagupta, S., Schaperdoth, I., Jones, D.S., Druschel, G.K., and Eastman, D. (2008) Niche differentiation among sulfur-oxidizing bacterial populations in cave waters. *ISME J* **2**: 590-601.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Maxam, A.M., and Gilbert, W. (1977) A New Method for Sequencing DNA. *Proc Natl Acad Sci U S A* **74**: 560-564.

McCarren, J., Becker, J.W., Repeta, D.J., Shi, Y.M., Young, C.R., Malmstrom, R.R. et al. (2010) Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci U S A* **107**: 16420-16427.

Mendel, G. (1866) Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereins in Brünn* **4**: 3-47.

Mendel, G. (1870) Über einige aus künstlicher Befruchtung gewonnene Hieracium-Bastarde. *Verhandlungen des naturforschenden Vereins in Brünn* **8**: 26-31.

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

Meyer, B., and Kuever, J. (2007b) Phylogeny of the alpha and beta subunits of the dissimilatory adenosine-5'-phosphosulfate (APS) reductase from sulfate-reducing prokaryotes - origin and evolution of the dissimilatory sulfate-reduction pathway. *Microbiology+* **153**: 2026-2044.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M. et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Bmc Bioinformatics* **9**: 386. doi: 10.1186/1471-2105-9-386.

Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W. (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**: 82-88.

Molina, V., Belmar, L., and Ulloa, O. (2010) High diversity of ammonia-oxidizing archaea in permanent and seasonal oxygen-deficient waters of the eastern South Pacific. *Environ Microbiol* **12**: 2450-2465.

Molina, V., Ulloa, O., Farias, L., Urrutia, H., Ramirez, S., Junier, P., and Witzel, K.P. (2007) Ammonia-oxidizing beta-Proteobacteria from the oxygen minimum zone off northern Chile. *Appl Environ Microbiol* **73**: 3547-3555.

Montecino, V., and Lange, C.B. (2009) The Humboldt Current System: Ecosystem components and processes, fisheries, and sediment studies. *Prog Oceanogr* **83**: 65-79.

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

Naqvi, S.W., Jayakumar, D.A., Narvekar, P.V., Naik, H., Sarma, V.V., D'Souza, W. et al. (2000) Increased marine production of N2O due to intensifying anoxia on the Indian continental shelf. *Nature* **408**: 346-349.

Newton, I.L.G., Woyke, T., Auchtung, T.A., Dilly, G.F., Dutton, R.J., Fisher, M.C. et al. (2007) The Calyptogena magnifica chemoautotrophic symbiont genome. *Science* **315**: 998-1000.

Niederberger, T.D., Perreault, N.N., Lawrence, J.R., Nadeau, J.L., Mielke, R.E., Greer, C.W. et al. (2009) Novel sulfur-oxidizing streamers thriving in perennial cold saline springs of the Canadian high Arctic. *Environ Microbiol* **11**: 616-629.

Ogata, H., Goto, S., Fujibuchi, W., and Kanehisa, M. (1998) Computation with the KEGG pathway database. *Biosystems* **47**: 119-128.

Orcutt, B.N., Sylvan, J.B., Knab, N.J., and Edwards, K.J. (2011) Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. *Microbiol Mol Biol Rev* **75**: 361-422.

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

Paulmier, A., and Ruiz-Pino, D. (2009) Oxygen minimum zones (OMZs) in the modern ocean. *Prog Oceanogr* **80**: 113-128.

Paulmier, A., Kriest, I., and Oschlies, A. (2009) Stoichiometries of remineralisation and denitrification in global biogeochemical ocean models. *Biogeosciences* **6**: 923-935.

Pauly, D., and Christensen, V. (1995) Primary Production Required to Sustain Global Fisheries. *Nature* **374**: 255-257.

Pearson, W.R., and Lipman, D.J. (1988) Improved Tools for Biological Sequence Comparison. *Proc Natl Acad Sci U S A* **85**: 2444-2448.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R.D.E., Buigues, B. et al. (2006) Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* **311**: 392-394.

Pond, S.K., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W.Y., Taylor, J. et al. (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Res* **19**: 2144-2153.

Poretsky, R.S., Gifford, S., Rinta-Kanto, J., Vila-Costa, M., and Moran, M.A. (2009) Analyzing gene expression from marine microbial communities using environmental transcriptomics. *J Vis Exp* **24**: e1086. doi: 10.3791/1086.

Poretsky, R.S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M. et al. (2005) Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* **71**: 4121-4126.

Porter, M.L., and Engel, A.S. (2008) Diversity of uncultured Epsilonproteobacteria from terrestrial sulfidic caves and springs. *Appl Environ Microbiol* **74**: 4973-4977.

Porter, M.L., Engel, A.S., Kane, T.C., and Kinkle, B.K. (2009) Productivity-Diversity Relationships from Chemolithoautotrophically Based Sulfidic Karst Systems. *Int J Speleol* **38**: 27-40.

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

Pushkarev, D., Neff, N.F., and Quake, S.R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**: 847-850.

Raes, J., Letunic, I., Yamada, T., Jensen, L.J., and Bork, P. (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* **7**: 473. doi: 10.1038/msb.2011.6.

Revsbech, N.P., Larsen, L.H., Gundersen, J., Dalsgaard, T., Ulloa, O., and Thamdrup, B. (2009) Determination of ultra-low oxygen concentrations in oxygen minimum zones by the STOX sensor. *Limnol Oceanogr Meth* **7**: 371-381.

Ronaghi, M., Uhlen, M., and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**: 363-365.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S. et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**(3): e77. doi: 10.1371/journal.pbio.0050077.

Ryther, J.H. (1969) Photosynthesis and fish production in the sea. *Science* **166**: 72-76.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N. et al. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**: 374-378.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA Sequencing with Chain-Terminating Inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463-5467.

Sarmiento, J.L., Hughes, T.M.C., Stouffer, R.J., and Manabe, S. (1998) Simulated response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393**: 245-249.

Seibel, B.A. (2011) Critical oxygen levels and metabolic suppression in oceanic oxygen minimum zones. *J Exp Biol* **214**: 326-336.

Shao, M.F., Zhang, T., and Fang, H.H.P. (2010) Sulfur-driven autotrophic denitrification: diversity, biochemistry, and engineering applications. *Appl Microbiol Biotechnol* **88**: 1027-1042.

Shi, Y.M., Tyson, G.W., and DeLong, E.F. (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**: 266-269.

Sorek, R., and Cossart, P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* **11**: 9-16.

Sorokin, Y.I., Sorokin, P.Y., Avdeev, V.A., Sorokin, D.Y., and Ilchenko, S.V. (1995) Biomass, Production and Activity of Bacteria in the Black-Sea, with Special Reference to Chemosynthesis and the Sulfur Cycle. *Hydrobiologia* **308**: 61-76.

Stevens, H., and Ulloa, O. (2008) Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environ Microbiol* **10**: 1244-1259.

Stewart, F.J., Ottesen, E.A., and DeLong, E.F. (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* **4**: 896-907.

Stewart, F.J., Ulloa, O., and Delong, E.F. (2011) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* **14**: 23-40.

Stramma, L., Johnson, G.C., Sprintall, J., and Mohrholz, V. (2008) Expanding oxygen-minimum zones in the tropical oceans. *Science* **320**: 655-658.

Streit, W.R., and Schmitz, R.A. (2004) Metagenomics - the key to the uncultured microbes. *Curr Opin Microbiol* **7**: 492-498.

Strous, M., Van Gerven, E., Kuenen, J.G., and Jetten, M. (1997) Effects of aerobic and microaerobic conditions on anaerobic ammonium-oxidizing (anammox) sludge. *Appl Environ Microbiol* **63**: 2446-2448.

Sunamura, M., Higashi, Y., Miyako, C., Ishibashi, J., and Maruyama, A. (2004) Two bacteria phylotypes are predominant in the Suiyo seamount hydrothermal plume. *Appl Environ Microbiol* **70**: 1190-1198.

Suzuki, Y., Sasaki, T., Suzuki, M., Nogi, Y., Miwa, T., Takai, K. et al. (2005) Novel chemoautotrophic endosymbiosis between a member of the Epsilonproteobacteria and the hydrothermal-vent gastropod Alviniconcha aff. hessleri (Gastropoda : Provannidae) from the Indian Ocean. *Appl Environ Microbiol* **71**: 5440-5450.

Taylor, G.T., Iabichella, M., Ho, T.Y., Scranton, M.I., Thunell, R.C., Muller-Karger, F., and Varela, R. (2001) Chemoautotrophy in the redox transition zone of the Cariaco Basin: A significant midwater source of organic carbon production. *Limnol Oceanogr* **46**: 148-163.

Tebo, B.M., and Emerson, S. (1986) Microbial Manganese(Ii) Oxidation in the Marine-Environment - a Quantitative Study. *Biogeochemistry* **2**: 149-161.

Toulza, E., Tagliabue, A., Blain, S., and Piganeau, G. (2012) Analysis of the global ocean sampling (GOS) project for trends in iron uptake by surface ocean microbes. *PLoS One* **7**(2): e30931. doi: 10.1371/journal.pone.0030931.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

Ulloa, O., and Pantoja, S. (2009) The oxygen minimum zone of the eastern South Pacific. *Deep-Sea Res PII* **56**: 987-991.

Urakawa, H., Dubilier, N., Fujiwara, Y., Cunningham, D.E., Kojima, S., and Stahl, D.A. (2005) Hydrothermal vent gastropods from the same family (Provannidae) harbour epsilon- and gamma-proteobacterial endosymbionts. *Environ Microbiol* **7**: 750-754.

Urich, T., Lanzen, A., Qi, J., Huson, D.H., Schleper, C., and Schuster, S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**(6): e2527. doi: 10.1371/journal.pone.0002527.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G. et al. (2001) The sequence of the human genome. *Science* **291**: 1304-1351.

von Tschermack, E. (1900) Über Künstliche Kreuzung bei Pisum sativum. *Ber Deut Bot Ges* **18**: 232-239.

Walsh, D.A., Zaikova, E., Howes, C.G., Song, Y.C., Wright, J.J., Tringe, S.G. et al. (2009) Metagenome of a Versatile Chemolithoautotroph from Expanding Oceanic Dead Zones. *Science* **326**: 578-582.

Ward, B.B., Glover, H.E., and Lipschultz, F. (1989) Chemoautotrophic Activity and Nitrification in the Oxygen Minimum Zone Off Peru. *Deep-Sea Res* **36**: 1031-1051.

Watson, J.D., and Crick, F.H.C. (1953) Molecular Structure of Nucleic Acids - A Structure for Deoxyribose Nucleic Acid. *Nature* **171**: 737-738.

Woebken, D., Lam, P., Kuypers, M.M.M., Naqvi, S.W.A., Kartal, B., Strous, M. et al. (2008) A microdiversity study of anammox bacteria reveals a novel Candidatus Scalindua phylotype in marine oxygen minimum zones. *Environ Microbiol* **10**: 3106-3119.

Wyrtki, K. (1962) The Oxygen Minima in Relation to Ocean Circulation. *Deep-Sea Res* **9**: 11-23.

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S.M., Meng, J. et al. (2010) Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J* **5**: 414-426.

Yamamoto, M., Nakagawa, S., Shimamura, S., Takai, K., and Horikoshi, K. (2010) Molecular characterization of inorganic sulfur-compound metabolism in the deep-sea epsilonproteobacterium Sulfurovum sp NBC37-1. *Environ Microbiol* **12**: 1144-1152.

Zaikova, E., Walsh, D.A., Stilwell, C.P., Mohn, W.W., Tortell, P.D., and Hallam, S.J. (2010) Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ Microbiol* **12**: 172-191.

Zehr, J.P., Bench, S.R., Carter, B.J., Hewson, I., Niazi, F., Shi, T. et al. (2008) Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**: 1110-1112.

Zhang, J.Z., and Millero, F.J. (1993) The Chemistry of the Anoxic Waters in the Cariaco Trench. *Deep-Sea Res PI* **40**: 1023-1041.

Zhang, Y., and Gladyshev, V.N. (2008) Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. *PLoS Genet* **4**(6): e1000095. doi: 10.1371/journal.pgen.1000095.

Danksagung

Danksagung

Ich möchte mich ganz herzlich bei Julie LaRoche bedanken. Dafür, dass Du mir die Arbeit an einem so spannenden Thema wie den Sauerstoffminimumzonen ermöglicht hast, und dafür dass ich diese während mehrerer Ausfahrten auch beproben konnte. Vielen Dank auch für die stetige und unermüdliche Hilfe, Licht ins Dunkel der Sequenzdatenanalyse und der biogeochemischen Kreisläufe zu bringen. Ich weiß außerdem sehr zu schätzen, wie geduldig Du mir das wissenschaftliche Arbeiten beigebracht hast (und vor allem auch, mit welcher Begeisterung).

Vielen Dank auch an Philip Rosenstiel, der mich während meiner Doktorarbeit stets mit neuen Ideen und konstruktiver Kritik begleitet und schließlich auch meine Doktorarbeit begutachtet hat.

Marcel Kuypers, Gaute Lavik und Ruth Schmitz-Streit danke ich für das Fachwissen, das Ihr mit mir geteilt habt, für Eure Hilfe und für die fruchtvolle Zusammenarbeit (die mir übrigens immer viel Spaß gemacht hat)!

Gefreut habe ich mich auch über die Zusammenarbeit mit den anderen Initiatoren des ‚REAL' Projektes: Klaus Jürgens, Matthias Labrenz und Hans-Peter Grossart.

Nicht zu vergessen, ich möchte auch noch ein Lob für Captain & Crew der Forschungsschiffe Meteor und Polarstern aussprechen, ebenso wie den Fahrtleitern Martin Frank und Andreas Macke.

Ganz klar, nun seit Ihr dran: Tania Klüver und Diana Gill. Ihr seid nicht nur die besten Techniker, sondern auch die besten Kollegen, die man sich vorstellen kann. Ohne Euch, Eure Hilfe, Eure Geduld und Eure stets gute Laune, wäre ich wahrscheinlich noch immer im Labor unterwegs…

Unersetzlich war für mich auch die Zusammenarbeit mit Dhwani. Was hätte ich nur ohne Dich gemacht? Ist gar nicht so einfach, mal eben ein paar Milliarden Basenpaare zu untersuchen. Ich hoffe, Du weißt, wie wichtig mir diese Zusammenarbeit war!

Es gibt es noch eine ganze Reihe anderer Leute, die ich im Laufe der Zeit in Kiel kennengelernt habe und mit denen ich viel Spaß hatte. Ob auf einem Schiff oder sonst wo. Um Einige zu nennen: Jassi Franz, Jan Czerny, Sophie Roy, Leni Hauss, Caro Löscher, Anna Dammshäuser, Flo Scholz, Isa Taubner, Christian Deusner, Jan Büdenbender, Tina Baustian, Holger Anlauf, Janine Böhdorn, Wiebke Mohr, Jockel Behnke, Steffi Stratil, Markus Lommer, Peter Fritsche, Rainer Kiko, Julia Goldhammer, Jogi Löser, Corinna Borchard und Sonja Endres.

Besonderer Dank auch an meinen Vorschoter Tobi, an Deckshand Kai und an unseren Smutje Paddy. Nicht nur, für die vielen Stunden gemeinsamen Segelns, sondern auch für die ein oder andere nervenaufreibende, aber nützliche Diskussion.

Für das Korrekturlesen der Arbeit bedanke ich mich außerdem bei Tim Haarmann, meinem Bruder und meinen Eltern!

Zuletzt möchte ich mich noch ganz besonders bei meiner Mutter bedanken. Dafür, dass Du mir immer vorgelebt hast, dass man manche Dinge einfach tun muss, trotz aller Bedenken oder Hindernisse… oder gerade deswegen… Und dass man bei allem was man tut, sich stets seine gute Laune behalten soll!

Eidesstattliche Erklärung

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass die von mir vorgelegte Dissertation, abgesehen von der Beratung durch meine Betreuer, nach Inhalt und Form meine eigene Arbeit ist und unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft angefertigt wurde. Genutzte Quellen, Hilfsmittel und die Zusammenarbeit mit anderen Wissenschaftlern wurden kenntlich gemacht.

Desweiteren versichere ich, dass die von mir vorgelegte Dissertation weder im Ganzen noch im Teil einer anderen Fakultät oder einer anderen Hochschule im Rahmen eines Prüfungsverfahrens vorgelegt wurde. Veröffentlichte oder zur Veröffentlichung eingereichte Manuskripte wurden kenntlich gemacht.

(Harald Schunck)