Improving the usage of ribosomal RNA gene in microbiology and microbial ecology

Importance of standardization and biocuration

by

Pelin Yilmaz, M.Sc.

A thesis submitted in partial fulfillment

of requirements for the degree of

DOCTOR OF PHILOSOPHY

in Bioinformatics

Approved, Thesis Committee:

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

Prof. Dr. Matthias Ullrich (member)
Jacobs University

Dr. Wolfgang Ludwig (member)
Technical University of Munich

Date of Defense: December, 9 2011

School of Engineering and Science

Statement of Sources

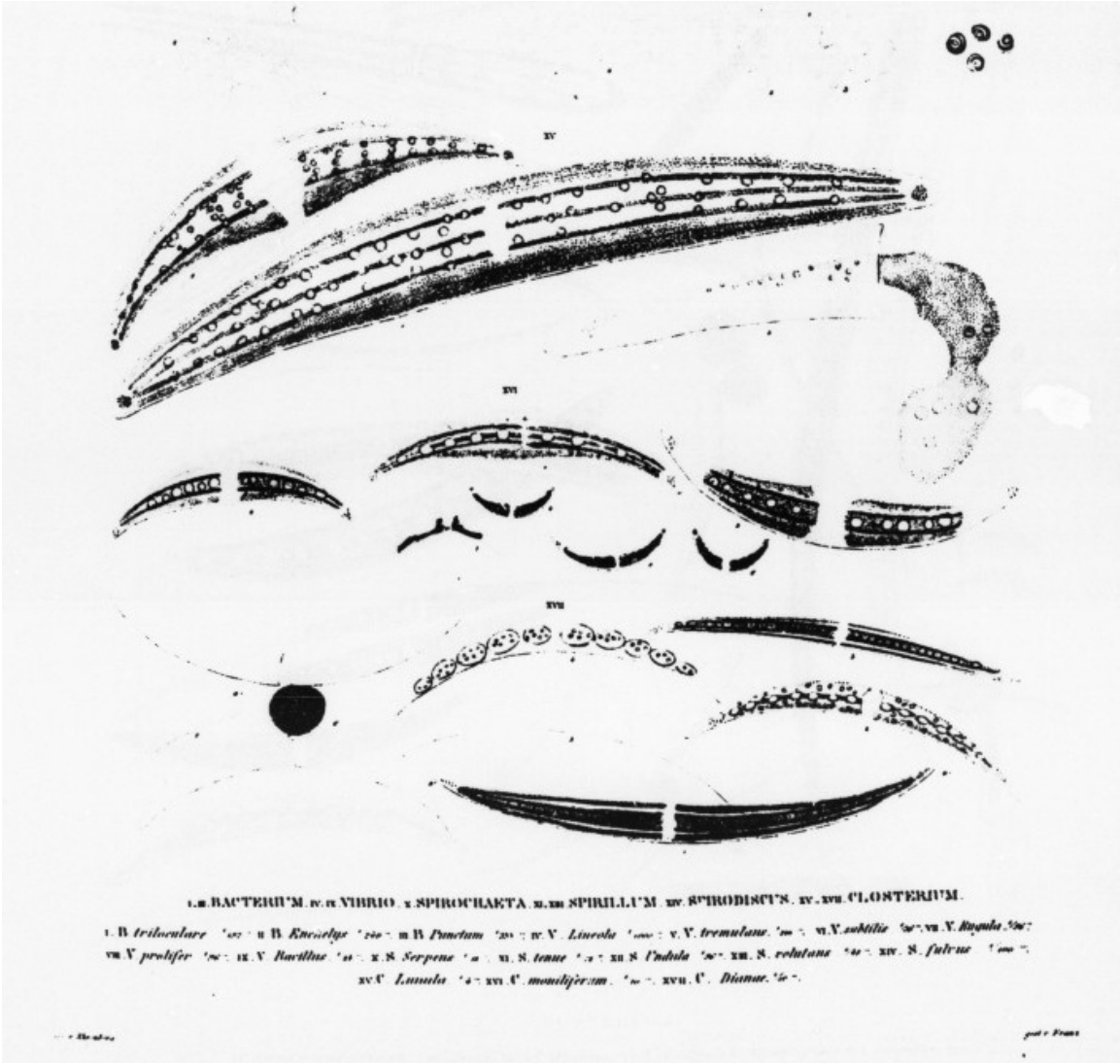
DECLARATION

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 Information derived from published or unpublished scientific work has been cited in the text and listed in the references.

Signature **Date**

First classification of Bacteria:

1. Bacterium, inflexible rod forms;
2. Vibrio, flexible rod forms;
3. Spirillum, spiral, inflexible;
4. Spirochaeta, spiral, flexible
5. Spirodiscus, flattened spirals.



* Adapted from; *Die Infusionsthierchen als vollkommene Organismen: ein Blick in das tiefere organische Leben der Natur,* Christian Gottfried Ehrenberg (1838).

Thesis Abstract

Environmental surveys of ribosomal RNA genes have become the standard approach to microbial ecology in the last twenty-five years. Although the ribosomal RNA approach has been competing with metagenomics during the past years, with high-throughput sequencing technologies presenting the ability to document the diversity and richness of the microbial biome at an astonishing rate, it is once again under spotlight.

High-throughput sequencing technologies bring with them a sequence data deluge, and present challenges to the traditional data management, analysis and retrieval practices. Biocuration, which leads to specialized ribosomal RNA datasets, has become an important part of microbial ecology. More recently, the development and use of metadata standards has emerged as an important aid for biocuration in microbiology. With the help of such standards, it will be possible to curate specialized datasets, which integrate data from even more diverse resources, and bring the organisms, its functions, and the environment together.

The work accomplished during this Ph.D. thesis aimed in improving the usage of ribosomal RNA gene in microbiology and microbial ecology by means of curating high-quality specialized datasets for ribosomal RNA gene sequences, and performing research that demonstrates the value of such datasets. More importantly, a metadata standard for marker genes sequences, such as ribosomal RNA, was developed and disseminated to a variety of bioinformatics service providers for implementation. This standard will greatly aid biocuration work in microbiology, and will improve the quantity and quality of integrative resources for microbiology, ultimately providing researchers the ability to understand the diversity and functioning of the microbial biosphere.

Table of Contents

INTRODUCTION

Early Microbiology and Microbial Ecology

The story of *Bacteria* started as a curiosity of nature with Leeuwenhoek's microscopy observations in 1670s. Until mid-1800s, *Bacteria* were largely ignored; in fact it was not even sure that they were living organisms. The actual realization of the significance of *Bacteria*, without doubt, came with Robert Koch's postulation of the "Germ Theory of Disease". Having found an "applied" perspective, microbiology (specifically the study of *Bacteria* and *Archaea*) made a kick-start, and numerous findings ranging from biogenesis to development of vaccines dotted the early timeline of this field.

Due to their implications in human health, other aspects of microbiology lagged behind, but did attain pace with the seminal works of Sergei Winogradsky and Martinus Beijerinck. Perhaps, those of the medical microbiology unfairly dominated their achievements. In fact, microbial ecology's existence would not have been, if it were not for their discoveries of nitrogen fixation or chemoautotrophy, which are cornerstones in today's understanding of the biogeochemical cycling of elements in the environment.

Medical microbiology not only overshadowed Winogradsky and Beijericks's achievements, but also made it difficult for microbiology to be realized as a "basic science". However, Beijerinck had another vision for microbiology, rather than confining it to the applications of medicine. In the statement he made during his speech at the Dutch Royal Academy of Sciences in 1905, he outlines the best practices to studying microorganisms [1]:

"[My] approach can be concisely stated as the study of microbial ecology, i.e., of the relation between environmental conditions and the special forms of life corresponding to them. It is my conviction that ... this is the most necessary and fruitful direction to guide us in organizing our knowledge of that part of nature which deals with the lowest limits

of the organic world, and which constantly keeps before our minds the profound problem of the origin of life itself."

Woese and Goldenfeld interpreted this statement as Beijerinck's vision of a three-tier microbiology; where the organism, ecology and evolution are studied together, not in a reductionist fashion, but as an integrative approach to understanding the world of microorganisms [2].

However, almost for two centuries, Beijerinck's vision was not to be realized. The methodological difficulties for studying *Bacteria* were vast; observing *Bacteria* was an obvious problem. Reasonable illumination methods were invented as early as 1893, but better methods such as phase contrast microscopy were not developed until 1950s. It was realized as early as 1850s that *Bacteria* were notoriously hard to culture, and cultivation methods were painstakingly time-consuming. This inability to observe and analyze *Bacteria* led to suspicions about their significance in the organismal world. Linnaeus could not classify *Bacteria*, therefore he termed the microscopic life "chaos" [3]. For practical necessities, classification was not left under chaos, and numerous different classification schemes were produced until 1970s [4]. These classification schemes were all non-phylogenetic, since it was believed that there are no species of microorganisms, and that morphological or physiological characteristics do not provide enough information [5]. Essentially, *Bacteria* as living organisms were ignored for about two centuries; fuelled by the overwhelming practical applications, medical and biochemical, and dragged on by lack of technology necessary to study them.

The Molecular Dawn for Bacteria

The necessary tools and methods to actually understand and study the world of microorganisms were provided by molecular biology. Firstly, *Bacteria* were used as a media to study molecular biology and genetics, but then scientists started realizing that molecular biology held the key to secrets of *Bacteria* by giving means to understand the characteristics of the organisms, the evolution and finally the ecology.

Understanding the organism

Oswald Avery was the first to utilize *Bacteria* in his studies of hereditary material, and incidentally the first to discover that DNA was the carrier of hereditary material [6]. Although Avery did not have *Bacteria* in mind while performing these studies, this discovery is considered an important step to understanding organismal characteristics of *Bacteria*.

The first group to focus on genetics of *Bacteria* was Joshua Lederberg, Edward L. Tatum, and George Beadle. In the late 1940s, they discovered that *Bacteria* can mate and exchange genes [7]. This discovery not only brought a Nobel Prize [1], but also marked the start of bacterial genetics field. The study of exchange of genetic material in *Bacteria* is considered an important cornerstone in understanding the biology of *Bacteria*, because, prior to the discovery that they can exchange genetic material, i.e recombine, it was believed that *Bacteria* only reproduced asexually, and that all cells in a lineage were just clones of a parent cell. The fact that *Bacteria* can evolve in a Darwinian fashion put *Bacteria* under a new light; they were biologized just like “higher organisms”.

Revisiting evolution

On the contrary to general knowledge, it was not Pauling and Zuckerkandl who first voiced the idea of molecular phylogenies. Just a few years before the structure of DNA was resolved, Francis Crick stated [8]:

“*Biologists should realize that before long we shall have a subject which might be called “protein taxonomy”—the study of amino acid sequences of proteins of an organism and*

[1] http://www.nobelprize.org/nobel_prizes/medicine/laureates/1958

the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away within them."

The story from this point on is well known; Pauling and Zuckerkandl articulated the idea of molecular clocks, which states that changes in nucleotide or amino acid sequences of DNA and proteins are constant over time and can be used to estimate times of lineage-splitting events, or simply track evolution [9]. This idea intrigued Carl Woese into searching for a universal molecular clock. He saw that molecular evidence would revolutionize how bacterial phylogeny and taxonomy was perceived, since it held the potential of eliminating uninformative comparative anatomy and physiology approaches. Not surprisingly, Woese decided on small subunit ribosomal (SSU rRNA; 16S and 18S) RNAs to build the universal tree of life [10]. The prime feature of rRNA gene is ubiquity in all domains of life, as they are the functional part of ribosomes; the protein translation machinery. Additionally, the sequence properties make rRNA gene a suitable molecular clock; stretches of extensive sequence and structure conservation spanned by other stretches of hypervariable regions. First comparisons that Woese performed were based on direct sequencing of bacterial 16S and eukaryotic 18S rRNAs. In 1976, with the inclusion of a methanogen in these analyses, unexpected results surfaced. The methanogenic "bacterium" did not belong to *Bacteria*, in fact it was more closely related to eukaryotic organisms. Careful consideration of the data led to the proposal of the three-domain life comprising *Bacteria*, *Archaea* and *Eukaryota* [11, 10, 12]. Although the proposal for three-domain life was not well received at the time, there is very little dispute about the molecular evidence today [13, 14], and the rRNA gene sequence data is being used at an unprecedented rate to document the tree of life.

In addition to being able to study the evolutionary relationships among all organisms, rRNA based bacterial and archaeal phylogeny also led the way to a unified taxonomy. Earlier volumes of "Bergey's Manual of Systematic Bacteriology" used a mixture of physiological, morphological and molecular data [15, 16], however from volume 3 on, the higher order taxonomic hierarchy of *Bacteria* and *Archaea* was based solely on the phylogenetic framework provided by rRNA gene sequence data [17, 18], bringing order to centuries long chaos.

Accessing the microbial biosphere

Only a decade after Woese's proposal for three domains of life, scientists started using rRNA sequences as indicators of phylogeny in microbial ecology studies. The first community composition studies involved sequencing 5S rRNA genes, which were short with 120 base pairs and were easily handled with the available sequencing capabilities [19]. However, the resolution provided by 5S rRNA gene sequences was not sufficient for complex community analyses, due to its short size; hence, the researchers turned their interest towards 16S rRNA genes. The size of the molecule, around 1500 base pairs, posed a problem until the sequencing technologies advanced in the second half of 1980s. After that, the 16S rRNA gene became the molecule of choice for microbial diversity studies [20-22]. In combination with polymerase chain reaction and cloning, sequencing of the 16S rRNA genes led to groundbreaking discoveries in microbial ecology. The rRNA gene-based molecular approach to characterizing natural communities of organisms provided, for the first time, culture-independent access to the diversity and distribution of microorganisms *in situ*, which resulted in the astonishing discovery that a vast majority (90-99%) of microorganisms have evaded existing cultivation methods [23, 24]. With the use of rRNA sequence information, it also became possible to identify uncultured microorganisms in their habitat. Information retrieved from sequencing rRNA genes was used to design fluorescent probes, which were then hybridized with environmental samples, revealing the structure of microbial communities *in situ* [25].

Currently, the rRNA approach to microbial ecology is well established, and is broadly known as the full-cycle rRNA approach (Figure 1). Total DNA is extracted from an environmental sample by applying an appropriate method. The DNA is used to construct a clone library or subjected to selective PCR amplification of rRNA genes. The clones containing rRNA genes are selected, and sequenced in a second step. Retrieved sequences are used for comparative sequence analysis, and design of nucleic acid probes to be used in fluorescence *in situ* hybridization (FISH). This automation of the microbial ecology studies brought along the necessity of indexing and storing the rRNA sequence data produced at an ever-increasing rate. Specialized rRNA databases were built for this purpose, which are currently the backbone of the rRNA approach.

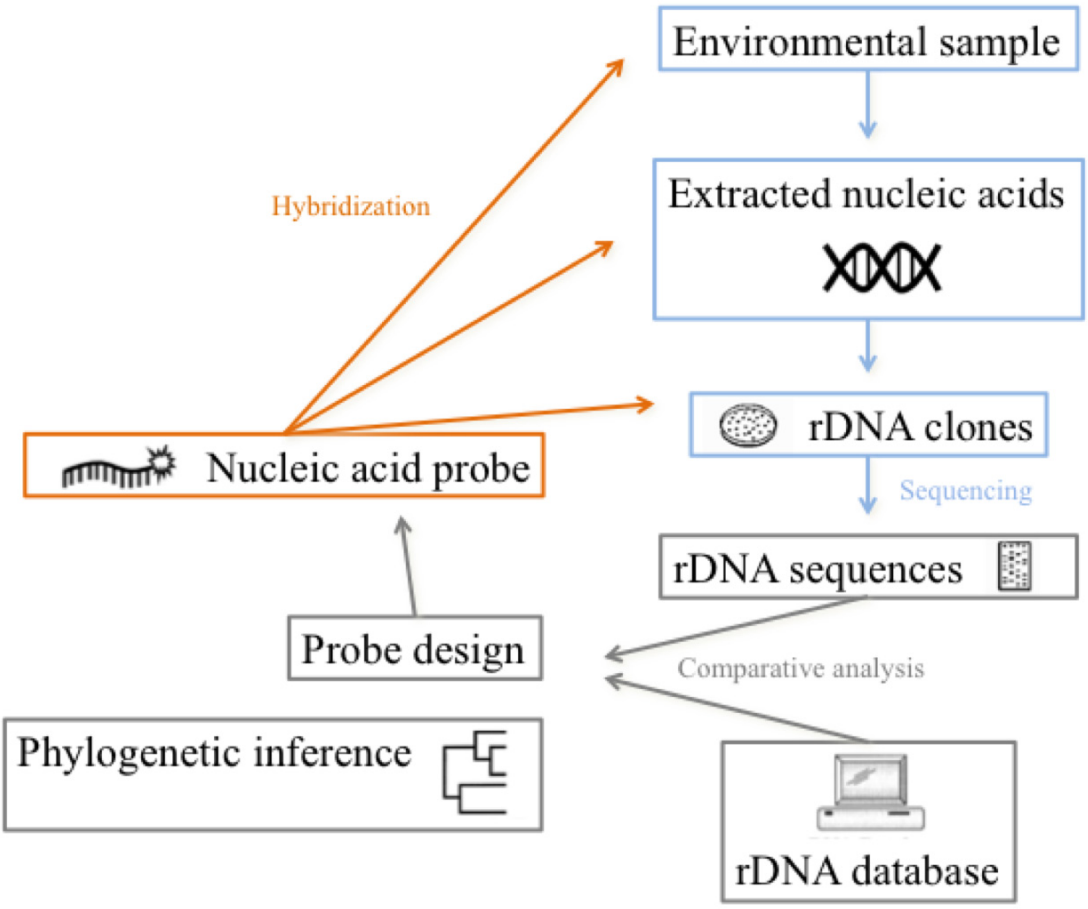


Figure 1. Generalized workflow for the full cycle rRNA approach. The approach uses the sequence information of cloned rRNA-encoding genes from environmental samples to develop phylogenetically specific oligonucleotide probes. Figure modified from http://www.arb-silva.de and [25].

Cataloging Bacterial and Archaeal Diversity

The first public nucleic acid sequence databases were started by manual curation of sequences from publications. Now, deposition of sequences into public databases is a prerequisite in the scientific publication pathway. Certainly, these databases are one of the major bioinformatics success stories of our time, and they have become an indispensable tool for scientists working on all aspects of natural sciences. As microbiology transformed into a "sequence-intensive" field, development of nucleic acid sequence databases became vital to the field. Our current understanding of microbial diversity is represented as rRNA sequences in public databases. In addition to public nucleic acid databases, there are additional databases specializing in rRNA sequences, which greatly aid microbial diversity and phylogeny studies.

Primary databases

The first biological molecules to be sequenced and collected in databases were proteins. In the 1960s, Margaret Dayhoff and coworkers at the National Biomedical Research Foundation assembled databases of protein sequences into a protein sequence atlas, which eventually became known as the Protein Information Resource. Nucleic acid databases appeared in the late 1970s [26]. DNA sequence databases were first assembled at the Los Alamos National Laboratory as a prototype of the current GenBank database. In 1982, the National Institutes of Health, the National Science Foundation, the Department of Energy and the Department of Defense of the United States of America started the funding of the public GenBank project. Currently, GenBank is an annotated collection of all publicly available DNA sequences [27]. It contains 130,671,233,801 bases in 142,284,608 sequence records in the traditional GenBank divisions and 208,315,831,132 bases in 64,997,137 sequence records in the whole genome shotgun sequencing (WGS) division as of August 2011[2].

The European counterpart of GenBank, the European Nucleotide Archive (ENA) was launched as the European Molecular Biology Laboratory (EMBL) Data Library at the EMBL's Heidelberg headquarters in 1980 [28]. Asia followed by establishing the DNA

[2] ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

Data Bank of Japan in 1986 at the National Institute of Genetics with the support of the Ministry of Education, Science, Sport and Culture of Japan [29].

Collectively, these three nucleic acid databases function as the International Nucleotide Sequence Database Collaboration (INSDC) [30]. New and updated data on nucleotide sequences contributed by researchers to each of the three databases have been synchronized on a daily basis through continuous interaction for over 18 years. The INSDC databases can be referred to as primary databases; or repositories, where the data is coming directly from the submitter, and from which one or more secondary databases can be created and maintained.

Secondary databases

Secondary databases incorporate knowledge, in addition to the basic data in the primary databases. rRNA databases constitute a good example for secondary databases. These databases aim to provide the scientific community with small/large subunit (SSU/LSU; 18S and 28S) rRNA sequences, annotations, alignments, secondary structures, and software for alignment and tree reconstruction. There are three widely recognized databases serving for this purpose. SILVA, developed and maintained by the Microbial Genomics Group at the Max Planck Institute (MPI) for Marine Microbiology in Bremen, Germany, in cooperation with the Department of Microbiology at the Technical University Munich and the company Ribocon, provides comprehensive, quality-checked and regularly updated databases of aligned SSU and LSU rRNA sequences for all three domains of life [31]. Release 108 (August 2011) of SILVA contains 2,492,653 sequences in the SSU Parc dataset (all sequences above 300 bases), 269,440 sequences in LSU Parc dataset, and 618,442 and 23,206 sequences in its high-quality SSU Ref (sequences above 1,200 bases) and LSU Ref (sequences above 1,900 bases) datasets.

The Ribosomal Database Project (RDP-II) is maintained by the Center for Microbial Ecology at the Michigan State University. The objectives of the project is to provide rRNA and related data and services to the scientific community, including online data analysis, aligned, and annotated bacterial and archaeal 16S rRNA sequences [32]. With the release 10.27 in August 2011, RDP-II hosts 1,921,179 16S rRNAs.

Greengenes of the Center for Environmental Biotechnology at the Lawrence Berkeley National Laboratory provides access to the current and comprehensive bacterial and archaeal 16S rRNA gene sequences and alignments for browsing, blasting, probing, and downloading [33]. The latest release of Greengenes, on July 2011, contains 1,021,768 aligned 16S rDNA records.

During the past three years, the SSU rRNA sequences have tripled, while the LSU rRNA sequences doubled in amount (Figure 2). It is important to point out that none of these specialized rRNA databases account for the majority of rRNA sequences produced by next-generation sequencing (NGS) technologies [34-36]. In just one round, a study from 2008 has produced nearly as many 16S rRNA genes as have been sequenced to date by Sanger sequencing [37]. Currently, prolific research groups produce even more data, with even faster and cheaper sequencing technologies. In light of these technological developments, it is important to ask, "What is the future holding for rRNA-based approaches?"

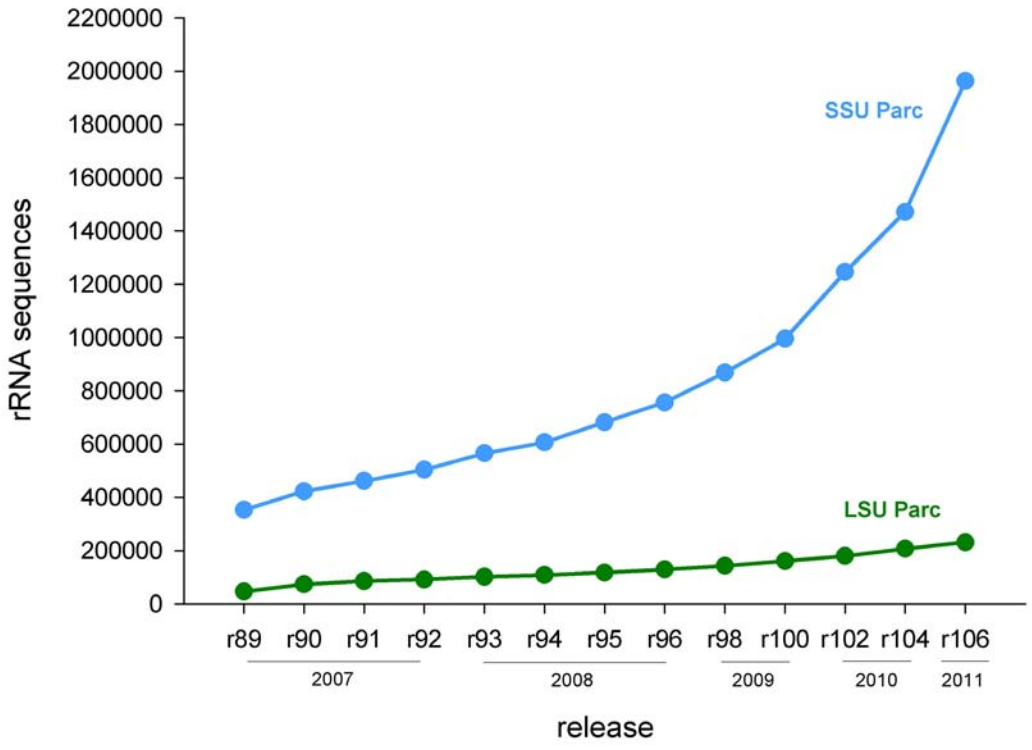


Figure 2. Growth of SSU and LSU databases according to SILVA releases. While the SSU rRNA sequences are growing exponentially, the LSU rRNA sequences are following a linear growth.

rRNA Approach in the Age of "Omics"

Since its introduction, the rRNA approach has been disparaged as much as it has been praised. There are inherent methodological pitfalls of the full-cycle rRNA approach, but a lot of the criticism also stems from more philosophical concerns about the usefulness of rRNA genes in areas such as ecology, evolution and taxonomy. Although such concerns have been raised repeatedly since the 1990s, around the time when rRNA gained momentum as a proxy for bacterial and archaeal species, the effects on the broader community have been minimal, as the quantity of the current rRNA gene based bacterial studies elaborately demonstrate.

Is the rRNA approach still relevant?

In order to represent both sides of the argument, I will try to cover a list of limitations and disadvantages that are often listed as reasons being against the rRNA approach, as well as arguments why rRNA gene is still useful despite those reasons.

Experimental problems range from sample collection, sample preservation and DNA extraction methods, to PCR amplification steps [38]. For example, one study reported changing bacterial structures with changing sample storage techniques [39], while others reported the effects of fragmented nucleic acid molecules on the subsequent PCR step [40, 41]. While such experimental errors are always a concern, they can be overcome with the use of standardized procedures and applying the correct method for the analysis. More severe problems can occur during PCR amplification of extracted nucleic acids [38]. A general assumption is that amplification efficiencies are the same for all molecules and the products reflect the quantitative abundance of species in the environment, but differential amplification is observed frequently. This can be due to a number of reasons, such as, differences in genome size and rRNA operon copy number [42], sub-optimal hybridization of primers with templates [43], G+C composition of rRNA genes [44], or template concentration in the PCR mixture [45]. A final problem associated with PCR amplification is the development of PCR artifacts, which also come in various different forms. Chimeric molecules [46], deletion mutants [47], and point mutants [48] can all lead to inaccurate conclusions on microbial diversity. Finally, sequencing technologies

are imperfect, and can introduce sequencing errors [49]. It is challenging to account for genome sizes or operon copy numbers in studies involving unknown organisms, however any biological experiment is expected to have this kind of background error rate. As for problems with primers, artifacts and sequencing errors, again “the correct method for the analysis” should prove useful; numerous software exists to predict and correct problems [50-52].

Another set of considerations about the rRNA approach originates from inherent properties of rRNA gene sequences. The *rrn* operon contains heterogeneities, which lead to intraspecies or intrastrain sequence variations. This implies that, after analysis, it may not be clear that one rRNA sequence or pattern is representing a distinct organism or just a variant of the *rrn* operon genes [38]. While it is true that *Bacteria* can contain up to 9 16S rRNA genes [53], and the sequence variations can be up to 6.4% [54], these are isolated cases and the average diversity between different 16S rRNA genes in a genome is 1-1.3% [55], which is below or equal to species detection limit of 98.7% [56].

The most significant disadvantage is the limited information content of the 16S rRNA gene sequences. Since rRNA genes are essential for the organism, the sequences are quite well conserved during evolution. This constrained evolution leads to questionable information content of rRNA sequences, and decreases the extent of phylogenetic conclusions that can be derived [57]. rRNA gene has limited resolution power in lower taxonomic levels, apparent from the threshold of 98.7% sequence identity by which two different species can be defined [56].

The limited information content, combined with the horizontal gene transfer (HGT) phenomenon in *Bacteria* and *Archaea* [58, 59], has led to heated debates about the phylogeny of *Bacteria* and *Archaea*, and the use of a single gene to study evolution, ecology and classification/taxonomy. Opponents of 16S rRNA gene usage in systematics are of the opinion that basing classification and taxonomy on molecular phylogeny is a misconception, and that species or taxa definitions should be based on phenotypic or functional characteristics [60]. This statement is completely ignoring the fact that the proposal of new species should be done by the polyphasic approach [4], which combines molecular and functional evidence. Rearrangements of taxa are indeed done based on 16S

rRNA gene evidence, which has proven useful and productive in systematics by bringing a standard approach to delineation of taxa. At the very least, the biggest suppliers of rRNA gene sequence data, SILVA, RDP-II and Greengenes, more or less conform to the outlines proposed by the Bergey's Manual of Systematic Bacteriology, and promote the use of these guidelines by researchers who use their services. It is also incorrect that systematics is solely based on 16S rRNA gene; other conserved core genes (23S rRNA, elongation factors, RNA polymerases and heat-shock proteins), as well as multi-locus sequence typing approaches are used frequently, and validate the picture provided by 16S rRNA gene phylogenetic trees [61]. Whole genome approaches, such as average nucleotide or amino acid identity, should certainly be used more often in systematics [62], however the genomic datasets are still limited for providing an unbiased view of *Bacteria* and *Archaea* from a wide variety of habitats [63, 64]. Additionally, they do not provide any assistance in the case of uncultured organisms, which constitute the majority of known diversity.

The implications of HGT on systematics are more complicated, since these events can completely erase bacterial phylogenetic tracks. Here, having a practical approach is probably the most productive. Unless a researcher is interested deeply in evolutionary biology, where 16S rRNA gene is known to be of little use [61], HGT should not cause serious errors in taxonomy and classification, or studies of microbial ecology/environmental microbiology, provided that one keeps in mind that rRNA is a generalized "name-tag" for the identity and physiology of the organism, and a particular function performed by this organism may or may not have arisen due to millennia of promiscuous gene exchange.

The above examples constitute an overview of criticism against the usage of rRNA gene in phylogeny and taxonomy, and their corresponding counter-arguments. As a final point to the discussion, it is important to note that there will always be a need for a viable phylogenetic marker for *Bacteria* and *Archaea*, and despite the drawbacks, rRNA gene is still a valid marker for these studies. After all, 16S rRNA gene easily complies with the requirements of being a phylogenetic marker (ubiquity, sequence and structure conservation), and gives richer information compared to any other indirect or direct nucleic acid-based systematics and identification method, due to the presence of

extensive rRNA sequence resources, databases, and tools to analyze the sequences. Additionally, both full-length rRNA gene and parts of it (hypervariable regions) are informative markers for studies of ecology topics such as species diversity, evenness richness, distribution, or biogeography [65, 66]. With these advantages, instead of the rRNA approach phasing out, as predicted about a decade ago [60], an even stronger presence is observed at the moment; current rRNA surveys are larger, and encompass multiple habitats, documenting bacterial and archaeal diversity at an astonishing rate [66]. Unfortunately, this rRNA revolution is not without drawbacks; while a decade ago it was challenging to find enough reference sequences, currently it is a problem to navigate in the sequence space to obtain relevant data and convert this data into knowledge

High-throughput microbial diversity

The quality and quantity of rRNA gene sequence data used to make phylogenetic assignments, to unravel succession as a function of the environment, and to assess biogeographic distributions continues to increase rapidly due to the availability of NGS technologies. Discoveries of microbial dynamics in relation to the environment and geography were achieved for cultured microorganisms long before the advent of high-throughput technologies [67-69]. However, with the new powerful technologies at our service, it is possible to unravel the diversity of the uncultured majority and to study increasingly complex and/or divergent ecosystems.

There are numerous examples of such research from recent years, ranging from correlations between phylogeny and living conditions [70], to latitudinal effects on microbial biogeography [71, 72]. Thorough within and cross-habitat studies have provided insights into variables shaping the microbial communities, both in our bodies and the environment [73-75]. These examples show that when placed in an environmental context, the sequence data could provide an unparalleled opportunity to produce a more comprehensive understanding of diversity, biogeography and ecology. Ideally, such studies should not be limited to the “first-hand” analysis of sequence data, but secondary analyses (meta or re-analysis) should also be possible. However, most researchers mistakenly think that secondary analyses will be possible through the simple increase of the amount of data available. The increase in sequence data however, is and will mostly

be due to the deposition of environmental sequences, implying that the sequences were obtained from uncultured and unidentified organisms, which means that an important fraction of known biodiversity is represented by organisms with otherwise unknown properties and ecological roles. For example, there are only 33,842 SSU rRNA sequences from cultivated organisms in the latest SILVA release, while the number of sequences from uncultured organisms is 2,458,811. Therefore, it is vital that deposited sequence data contains information other than just bare sequence data; a simple example being the geographical coordinates of the original sample which enable other researchers to retrieve this sequence data based on location, and combine it with other sequence data from the same location. Easy retrieval of sequence data from the right environment, location, or conditions enables re-analysis which in turn serves some basic principles of scientific progress such as verification of previous results, presentation of novel methods and new interpretations to existing data, comparative studies of existing datasets, and optimization the use of resources [76].

Missing ingredient; contextual data

The exponential growth of rRNA sequences by environmental sequences is enlightening in terms of microbial diversity, but it also raises the possibility that, at one point in time there will be an overwhelming amount of "abandoned" sequence information. The simple BLASTN search of a query sequence will just return the uninformative "uncultured bacterium", leaving researchers in the tedious position of sieving through thousands of such hits. INSDC databases do provide opportunities to enrich sequence entries with additional data, in the form of sequence entry "Feature Table"[3]. The overall goal of the feature table fields is to provide an extensive vocabulary for describing features of a sequence in a flexible framework. This field can contain valuable information regarding the isolation source of the sequence, physical and chemical properties of isolation source, habitat type, as well as information about experimental procedures that led to the sequence in question, which can be collectively referred to as the "contextual (meta) data" of the sequence. The term contextual is self-explanatory; it provides a context to the primary data. "Meta" derives from the Greek word denoting a nature of a higher order or

[3] http://www.ncbi.nlm.nih.gov/collab/FT/#3.2

more fundamental kind. Contextual (meta) data is necessary for the interpretation and utilization of a dataset. For microbial diversity studies, the contextual data associated with the sequence can be roughly divided into the following groups:

1. Sample identifiers
2. Sample source, material, location and information regarding sample collection and processing
3. Physical environmental factors of the sampling location, which may be temperature, measured irradiation, or weather conditions
4. Chemical environmental factors of the sampling location, which may include any concentration measurement or presence of pollutants
5. Biotic environmental factors of the sampling location, such as the surrounding vegetation, bacterial number counts or important facts about the ecosystems biotic components
6. Description of the habitat
7. Experimental procedures used during nucleic acid extraction and sequencing
8. Taxonomy

This valuable contextual data in the feature table is often not obtainable. It is possible to track down the contextual data from the original publication, however the required information is missing most of the time. In addition, tracking down each sequence for relevant additional information is a time consuming process, not to mention that the use of automated processes in order to extract information is not feasible.

While there is a clear need to educate and encourage the community with respect to enriching sequences with contextual data, it will be naïve to just assume that this will be enough. Reporting contextual data is only one part of the problem; there is also a lack of a unified standard on how to report the contextual data. A very simple example would be the case of reporting the collection date. A qualifier in the feature table exists, but it is reported in at least six different formats, rendering any kind of comparison impossible. For the future of nucleic acid sequence databases, it is important both to increase the

amount of rRNA sequence contextual data, and to find a common syntax to report the sequence contextual data.

Development and implementation of standards in biological sciences is a relatively new paradigm, but a sizeable number of such standards have appeared during the past ten years, dealing with a range of fields from microarray experiments to clinical trials [77]. Some of these standards are quite narrow-ranged [78], and the field of microbiology certainly does not fall into this category, hence making the development of a standard non-trivial. However, other standards cover a wide range of research fields, such as the Minimum Information About a Microarray Experiment [79], and have been successfully implemented and adopted by the community, proving that a similar minimum information guideline for marker gene sequences, such as rRNA, contextual data is an achievable task, provided that the guidelines are simple and conform to the community needs.

The goal of enhancing rRNA sequence contextual data therefore requires a well-defined roadmap. As noted above, the first step will certainly be to raise awareness for the issue in the scientific community. The Genomic Standards Consortium (GSC)[4], established in 2005, a group of international scientists with different backgrounds, has so far been very active in making the call for contextual data heard [80-84]. This community has the goal of developing standardized procedures for the description of genomes, in addition to facilitating the exchange and integration of genomic data. Although the GSC is focused primarily on genomic and metagenomic studies, a workgroup has been established within the body of the GSC, which handles determination of the major requirements for rRNA sequence reporting standards. This workgroup has provided the next steps for the roadmap to enhance rRNA sequence metadata. Firstly, there is a need to determine critical contextual data that needs to be reported along with an rRNA sequence. Secondly, these new contextual data fields can be combined with the Minimum Information about a Genome Sequence (MIGS/MIMS) checklist [85], already published by the GSC, and be used to generate an rRNA sequence contextual data submission checklist. Finally, it is necessary to provide the scientific community with tools to assist in effective sequence

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

and contextual data submission to nucleic acid sequence databases.

Development and adoption of such standards now, is very timely, given the amount and speed of sequence data is generated. The opportunity to produce a comprehensive and mechanistic understanding of microbiology, integrated across many systems and scales will be lost if such standards are not develop and complied with, and an unusable mess of data will be left to future researchers.

Motivation and Research Aims

The molecular approach to microbiology, with a focus on the rRNA approach, is flourishing, but requires a slight course correction in data management and integration practices, in order to continue producing groundbreaking research and deal with the sequence data deluge.

Sophisticated algorithms, faster software and advanced analysis methods appear to be the answer to the sequence data deluge. However, this is a somewhat misguided view; ultimately any software or analysis method requires the "right kind" of data. With growing database sizes and varieties, only accessing of the right data is currently up to the end user; location and integration of the data is now up to the "biocurator".

Typical tasks for a biocurator include managing and extracting raw biological data from primary databases, extracting information from publications, integrating biological data with the publication data, developing ways to present this integrated data in a structured and standardized format, and finally making this enriched data available to public.

Automation of the biocurator's tasks is only possible to a certain extent; manual intervention is almost always a necessity. Thus, biocuration is time-consuming and the production of results is seemingly slow. Nevertheless, these results are vital to biological sciences; basic research is possible since researchers have access to gold-standard curated datasets to be used in downstream analysis.

The overall aim of this thesis work is centered on biocuration in order to improve the utilization of the rRNA approach in microbiology and microbial ecology fields. Three main aims are, facilitating the curation of gold-standard datasets, curating such datasets, and using these datasets in microbial ecology research (Figure 3).

The facilitation of curation will be accomplished by development, dissemination and implementation of contextual data standards for marker gene sequences (Papers I-IV). Such a standard will enable environment-centric acquisition of sequence data, and ease the burden of integrating sequence organismal and environmental data.

Curation will be performed on existing database services provided by the Microbial Genomics and Bioinformatics Group (Papers V and VI). Specifically, taxonomic

classification of full-length bacterial, archaeal and eukaryotic rRNA sequence datasets will be updated based on authoritative resources such as Bergey's Manual of Systematic Bacteriology or List of Prokaryotic Species with Standing in Nomenclature. Geographical and environmental data of rRNA sequences will be used to integrate rRNA sequence data into the existing framework of environmental genomics and metagenomics platform.

Finally, these integrated, high-quality datasets will be used to address questions in microbial ecology, with an emphasis on marine-origin datasets (Papers VII and VIII). The common theme in both of these projects is the use of rRNA gene fragments from metagenomic datasets to deduce microbial diversity and community structure. One study is an exploratory study on the use of 23S rRNA gene sequences; the other study is focusing on the ecological distribution of bacterial and archaeal taxa at higher taxon ranks.

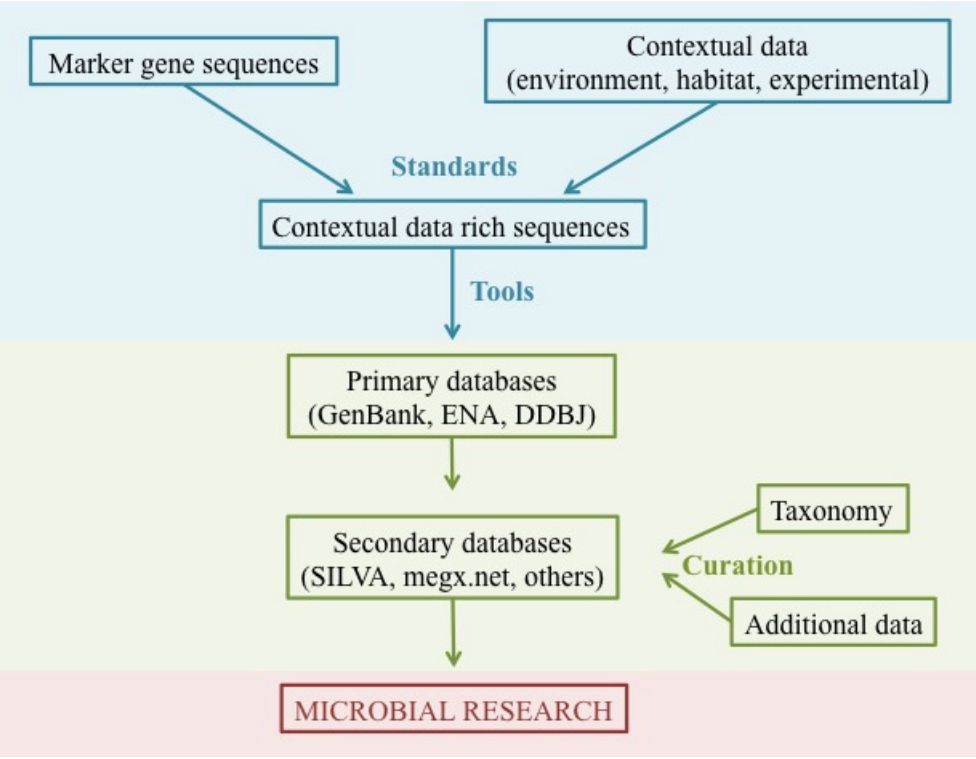


Figure 3. Schematic representation of research aims in three levels; standards and their implementation, curation, and finally research.

RESULTS AND DISCUSSION

Overview

This chapter presents eight research articles that best illustrate the achievements of this thesis with regard to the research aims. The papers are grouped under three research aims; contextual data standards for biocuration, curation of rRNA gene datasets, and usage of curated datasets in microbial ecology.

Contextual data standards for biocuration

I. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Authors: Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi et al.

Published in: Nature Biotechnology. 2011; 29 (5): 415-420

Contribution: designed the standard, held community sessions for agreement on the standard, wrote the manuscript

II. The genomic standards consortium: bringing standards to life for microbial ecology

Authors: Pelin Yilmaz, Jack A Gilbert, Rob Knight, Linda Amaral-Zettler, Ilene Karsch-Mizrachi, Guy Cochrane, Yasukazu Nakamura, Susanna-Assunta Sansone, Frank Oliver Glöckner, Dawn Field

Published in: ISME Journal. 2011; 5: 1565-1567

Contribution: conceived the design of the manuscript, wrote the manuscript

III. MetaBar - a tool for consistent contextual data acquisition and standards compliant submission

Authors: Wolfgang Hankeln, Pier Luigi Buttigieg, Dennis Fink, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner

Published in: BMC Bioinformatics. 2010; 11 (1): 358

Contribution: correct implementation of MIxS standards components

IV. CDinFusion – Submission-ready, on-line Integration of Sequence and Contextual Data

Authors: Wolfgang Hankeln, Norma Wendel, Jan Gerken, Jost Waldmann, Pier Luigi Buttigieg, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner

Published in: PLoS ONE. 2011; 6 (9): e24797

Contribution: correct implementation of MIxS standards components

Curation of rRNA datasets

V. SILVA: Comprehensive Databases for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB

Authors: Elmar Prüsse, Christian Quast, Pelin Yilmaz, Wolfgang Ludwig, Jörg Peplies, Frank Oliver Glöckner

Published in: Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches (ed. Frans J. de Bruijn). 2011; advance online publication 3 May 2011 doi: 10.1002/9781118010518.ch45

Contribution: curated bacterial and archaeal taxonomy for small and large subunit ribosomal RNA reference datasets, wrote the section on taxonomy

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

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

Published in: Nucleic Acids Research. 2010; 38: D391-D395.

Contribution: integration of 16S/18S and 23S/28S rRNA sequence data by linking SILVA with megx.net

Usage of curated datasets in microbial ecology

VII. Analysis of 23S rRNA genes in metagenomes – A case study from the Global Ocean Sampling Expedition

Authors: Pelin Yilmaz, Renzo Kottmann, Elmar Pruesse, Christian Quast and Frank Oliver Glöckner

Published in: Systematic and Applied Microbiology. 2011; 38 (6): 462-469

Contribution: designed and performed the research, analyzed the data and wrote the manuscript

VIII. Ecological structuring of bacterial and archaeal taxa in ocean surface waters.

Authors: Pelin Yilmaz, Wolfgang Hankeln, Renzo Kottmann, Christian Quast and Frank Oliver Glöckner

Awaiting revised manuscript after receipt of peer reviews at FEMS Microbiology Ecology on 05-09-2011

Contribution: designed and performed the research, analyzed the data and wrote the manuscript

Contextual Data Standards for Biocuration

I. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Authors: Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Philippe Rocca-Serra, Peter Sterk et al.

Published in: Nature Biotechnology. 2011; 29 (5): 415-420

Contribution: designed the standard, held community sessions for agreement on the standard, wrote the manuscript

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Pelin Yilmaz[1,2], Renzo Kottmann[1], Dawn Field[3], Rob Knight[4,5], James R Cole[6,7], Linda Amaral-Zettler[8], Jack A Gilbert[9,10,11], Ilene Karsch-Mizrachi[12], Anjanette Johnston[12], Guy Cochrane[13], Robert Vaughan[13], Christopher Hunter[13], Joonhong Park[14], Norman Morrison[3,15], Philippe Rocca-Serra[16], Peter Sterk[3], Manimozhiyan Arumugam[17], Mark Bailey[3], Laura Baumgartner[18], Bruce W Birren[19], Martin J Blaser[20], Vivien Bonazzi[21], Tim Booth[3], Peer Bork[17], Frederic D Bushman[22], Pier Luigi Buttigieg[1,2], Patrick S G Chain[7,23,24], Emily Charlson[22], Elizabeth K Costello[4], Heather Huot-Creasy[25], Peter Dawyndt[26], Todd DeSantis[27], Noah Fierer[28], Jed A Fuhrman[29], Rachel E Gallery[30], Dirk Gevers[19], Richard A Gibbs[31,32], Inigo San Gil[33], Antonio Gonzalez[34], Jeffrey I Gordon[35], Robert Guralnick[28,36], Wolfgang Hankeln[1,2], Sarah Highlander[31,37], Philip Hugenholtz[38], Janet Jansson[23,39], Andrew L Kau[35], Scott T Kelley[40], Jerry Kennedy[4], Dan Knights[34], Omry Koren[41], Justin Kuczynski[18], Nikos Kyrpides[23], Robert Larsen[4], Christian L Lauber[42], Teresa Legg[28], Ruth E Ley[41], Catherine A Lozupone[4], Wolfgang Ludwig[43], Donna Lyons[42], Eamonn Maguire[16], Barbara A Methé[44], Folker Meyer[10], Brian Muegge[35], Sara Nakielny[4], Karen E Nelson[44], Diana Nemergut[45], Josh D Neufeld[46], Lindsay K Newbold[3], Anna E Oliver[3], Norman R Pace[18], Giriprakash Palanisamy[47], Jörg Peplies[48], Joseph Petrosino[31,37], Lita Proctor[21], Elmar Pruesse[1,2], Christian Quast[1], Jeroen Raes[49], Sujeevan Ratnasingham[50], Jacques Ravel[25], David A Relman[51,52], Susanna Assunta-Sansone[16], Patrick D Schloss[53], Lynn Schriml[25], Rohini Sinha[22], Michelle I Smith[35], Erica Sodergren[54], Aymé Spor[41], Jesse Stombaugh[4], James M Tiedje[7], Doyle V Ward[19], George M Weinstock[54], Doug Wendel[4], Owen White[25], Andrew Whiteley[3], Andreas Wilke[10], Jennifer R Wortman[25], Tanya Yatsunenko[35], Frank Oliver Glöckner[1,2]

[1]Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany. [2]Jacobs University Bremen gGmbH, Bremen, Germany. [3]Natural Environment Research Council Environmental Bioinformatics Centre, Wallington CEH, Oxford, UK. [4]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. [5]Howard Hughes Medical Institute, San Francisco, California, USA. [6]Ribosomal Database Project, Michigan State University, East Lansing, Michigan, USA. [7]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, USA. [8]The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA. [9]Plymouth Marine Laboratory, Plymouth, UK. [10]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. [11]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA. [12]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA. [13]European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [14]School of Civil and Environmental Engineering, Yonsei University, Seoul, Republic of Korea. [15]School of Computer Science, University of Manchester, Manchester, UK. [16]Oxford e-Research Centre, University of Oxford,

Oxford, UK. [17]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [18]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. [19]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [20]Department of Medicine and the Department of Microbiology, New York University Langone Medical Center, New York, USA. [21]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. [22]Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. [23]DOE Joint Genome Institute, Walnut Creek, California, USA. [24]Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA. [25]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA. [26]Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium. [27]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [28]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA.[29]Department of Biological Sciences, University of Southern California, Los Angeles, California, USA. [30]National Ecological Observatory Network, Boulder, Colorado, USA. [31]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [32]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.[33]Department of Biology, University of New Mexico, LTER Network Office, Albuquerque, New Mexico, USA. [34]Department of Computer Science, University of Colorado, Boulder, Colorado, USA. [35]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA. [36]University of Colorado Museum of Natural History, University of Colorado, Boulder, Colorado, USA. [37]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA. [38]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia. [39]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [40]Department of Biology, San Diego State University, San Diego, California, USA. [41]Department of Microbiology, Cornell University, Ithaca, New York, USA. [42]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA. [43]Lehrstuhl für Mikrobiologie, Technische Universität München, Freising, Germany. [44]J. Craig Venter Institute, Rockville, Maryland, USA. [45]Department of Environmental Sciences, University of Colorado, Boulder, Colorado, USA. [46]Department of Biology, University of Waterloo, Ontario, Canada. [47]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. [48]Ribocon GmbH, Bremen, Germany. [49]VIB - Vrije Universiteit Brussel, Brussels, Belgium. [50]Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada. [51]Departments of Microbiology and Immunology and of Medicine, Stanford University School of Medicine, Stanford, California, USA. [52]Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA. [53]Department of Microbiology and Immunology, Ann Arbor, Michigan, USA. [54]The Genome Center, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA

ABSTRACT

Here we present a standard developed by the Genomic Standards Consortium (GSC) for reporting marker gene sequences—the minimum information about a marker gene sequence (MIMARKS). We also introduce a system for describing the environment from which a biological sample originates. The 'environmental packages' apply to any genome sequence of known origin and can be used in combination with MIMARKS and other GSC checklists. Finally, to establish a unified standard for describing sequence data and to provide a single point of entry for the scientific community to access and learn about GSC checklists, we present the minimum information about any (x) sequence (MIxS). Adoption of MIxS will enhance our ability to analyze natural genetic diversity documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

Introduction

Without specific guidelines, most genomic, metagenomic and marker gene sequences in databases are sparsely annotated with the information required to guide data integration, comparative studies and knowledge generation. Even with complex keyword searches, it is currently impossible to reliably retrieve sequences that have originated from certain environments or particular locations on Earth—for example, all sequences from 'soil' or 'freshwater lakes' in a certain region of the world. Because public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprising DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) and GenBank (http://www.insdc.org/)) depend on author-submitted information to enrich the value of sequence data sets, we argue that the only way to change the current practice is to establish a standard of reporting that requires contextual data to be deposited at the time of sequence submission. The adoption of such a standard would elevate the quality, accessibility and utility of information that can be collected from INSDC or any other data repository.

The GSC has previously proposed standards for describing genomic sequences— the “minimum information about a genome sequence” (MIGS)—and metagenomic sequences—the “minimum information about a metagenome sequence” (MIMS)[1]. Here we introduce an extension of these standards for capturing information about marker genes. Additionally, we introduce 'environmental packages' that standardize sets of

measurements and observations describing particular habitats that are applicable across all GSC checklists and beyond[2]. We define 'environment' as any location in which a sample or organism is found, e.g., soil, air, water, human-associated, plant-associated or laboratory. The original MIGS/MIMS checklists included contextual data about the location from which a sample was isolated and how the sequence data were produced. However, standard descriptions for a more comprehensive range of environmental parameters, which would help to better contextualize a sample, were not included. The environmental packages presented here are relevant to any genome sequence of known origin and are designed to be used in combination with MIGS, MIMS and MIMARKS checklists.

To create a single entry point to all minimum information checklists from the GSC and to the environmental packages, we propose an overarching framework, the MIxS standard (http://gensc.org/gc_wiki/index.php/MIxS). MIxS includes the technology-specific checklists from the previous MIGS and MIMS standards, provides a way of introducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages. A schematic overview of MIxS along with the MIxS environmental packages is shown in Figure 1.

Development of MIMARKS and the environmental packages

Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed spacer gene sequences (ITS) from *Bacteria*, *Archaea* and microbial *Eukaryotes* have provided deep insights into the topology of the tree of life[3, 4] and the composition of communities of organisms that live in diverse environments, ranging from deep sea hydrothermal vents to ice sheets in the Arctic[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Numerous other phylogenetic marker genes have proven useful, including RNA polymerase subunits (*rpoB*), DNA gyrases (*gyrB*), DNA recombination and repair proteins (*recA*) and heat shock proteins (*HSP70*)[3]. Marker genes can also reveal key metabolic functions rather than phylogeny; examples include nitrogen cycling (*amoA*, *nifH*, *ntcA*)[17, 18], sulfate reduction (*dsrAB*)[19] or phosphorus metabolism (*phnA*, *phnI*, *phnJ*)[20, 21]. In this paper we define all phylogenetic and functional genes (or gene fragments) used to profile natural genetic diversity as

'marker genes'. MIMARKS (Table 1) complements the MIGS/MIMS checklists for genomes and metagenomes by adding two new checklists, a MIMARKS survey, for uncultured diversity marker gene surveys, and a MIMARKS specimen, for marker gene sequences obtained from any material identifiable by means of specimens. The MIMARKS extension adopts and incorporates the standards being developed by the Consortium for the Barcode of Life (CBOL)[22]. Therefore, the checklist can be universally applied to any marker gene, from small subunit rRNA to cytochrome oxidase I (COI), to all taxa, and to studies ranging from single individuals to complex communities.

Both MIMARKS and the environmental packages were developed by collating information from several sources and evaluating it in the framework of the existing MIGS/MIMS checklists. These include four independent community-led surveys, examination of the parameters reported in published studies and examination of compliance with optional features in INSDC documents. The overall goal of these activities was to design the backbone of the MIMARKS checklist, which describes the most important aspects of marker gene contextual data.

Results of community-led surveys

Four online surveys about descriptors for marker genes have been conducted to determine researcher preferences for core descriptors. The Department of Energy Joint Genome Institute and SILVA[23] surveys focused on general descriptor contextual data for a marker gene, whereas the Ribosomal Database Project (RDP)[24] focused on prevalent habitats for rRNA gene surveys, and the Terragenome Consortium[25] focused on soil metagenome project contextual data (Supplementary Results 1). The above recommendations were combined with an extensive set of contextual data items suggested by an International Census of Marine Microbes (ICoMM) working group that met in 2005. These collective resources provided valuable insights into community requests for contextual data items to be included in the MIMARKS checklist and the main habitats constituting the environmental packages.

Survey of published parameters

We reviewed published rRNA gene studies, retrieved from SILVA and the ICoMM database MICROBIS (The Microbial Oceanic Biogeographic Information System, http://icomm.mbl.edu/microbis/) to further supplement contextual data items that are included in the respective environmental packages. In total, 39 publications from SILVA and >40 ICoMM projects were scanned for contextual data items to constitute the core of the environmental package subtables (Supplementary Results 1).

In a final analysis step, we surveyed usage statistics of INSDC source feature key qualifier values of rRNA gene sequences contained in SILVA (Supplementary Results 1). Notably, <10% of the 1.2 million 16S rRNA gene sequences (SILVA release 100) were associated with even basic information such as latitude and longitude, collection date or PCR primers.

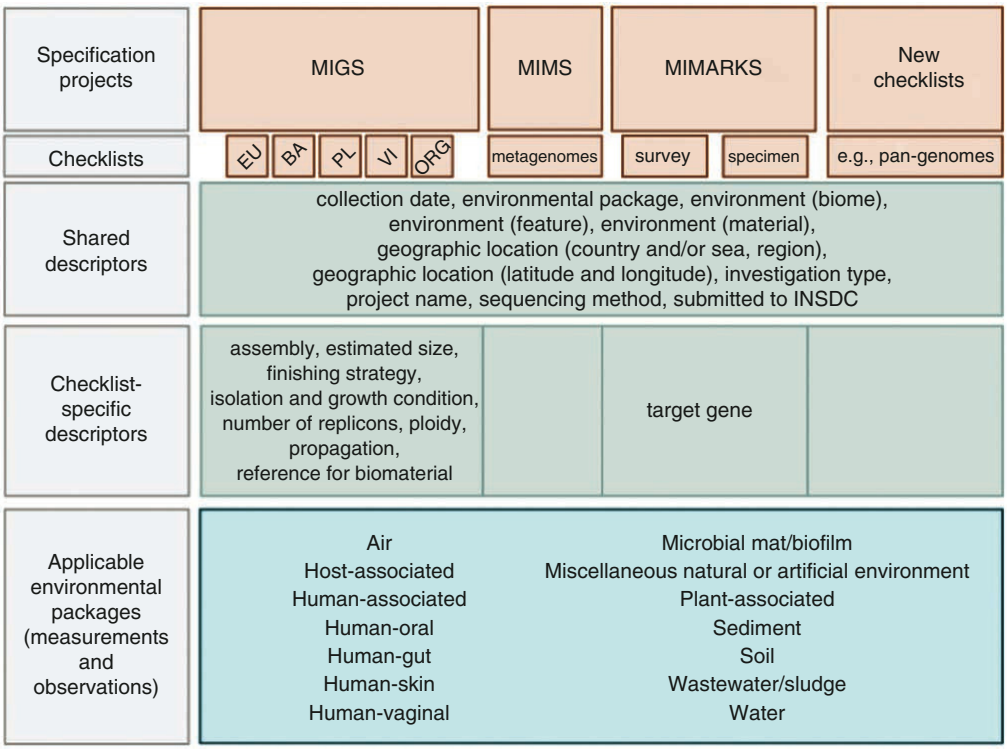


Figure 1: Schematic overview about the GSC MIxS standard (brown), including combination with specific environmental packages (blue). Shared descriptors apply to all MIxS checklists; however, each checklist has its own specific descriptors as well. Environmental packages can be applied to any of the checklists. EU, eukarya; BA, bacteria/archaea; PL, plasmid; VI, virus; ORG, organelle.

The MIMARKS checklist

The MIMARKS checklist provides users with an 'electronic laboratory notebook' containing core contextual data items required for consistent reporting of marker gene investigations. MIMARKS uses the MIGS/MIMS checklists with respect to the nucleic acid sequence source and sequencing contextual data, but extends them with further experimental contextual data such as PCR primers and conditions, or target gene name.

For clarity and ease of use, all items within the MIMARKS checklist are presented with a value syntax description, as well as a clear definition of the item. Whenever terms from a specific ontology are required as the value of an item, these terms can be readily found in the respective ontology browsers linked by URLs in the item definition. Although this version of the MIMARKS checklist does not contain unit specifications, we recommend all units to be chosen from and follow the International System of Units (SI) recommendations. In addition, we strongly urge the community to provide feedback regarding the best unit recommendations for given parameters. Unit standardization across data sets will be vital to facilitate comparative studies in future. An Excel version of the MIMARKS checklist is provided on the GSC web site (http://gensc.org/gc_wiki/index.php/MIMARKS).

The MIxS environmental packages

Fourteen environmental packages provide a wealth of environmental and epidemiological contextual data fields for a complete description of sampling environments. The environmental packages can be combined with any of the GSC checklists (Fig. 1 and Supplementary Results 2). Researchers within The Human Microbiome Project[26] contributed the host-associated and all human packages. The Terragenome Consortium contributed sediment and soil packages. Finally, ICoMM, Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites and the Max Planck Institute for Marine Microbiology contributed the water package. The MIMARKS working group developed the remaining packages (air, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated and wastewater/sludge). The package names describe high-level habitat terms in order to be exhaustive. The

miscellaneous natural or artificial environment package contains a generic set of parameters, and is included for any other habitat that does not fall into the other thirteen categories. Whenever needed, multiple packages may be used for the description of the environment.

Table 1: The core items of the MIMARKS checklists, along with the value types, descriptions and requirement status. Items for the MIMARKS specification and their mandatory (M), status for both MIMARKS-survey and MIMARKS-specimen checklists.

		Report type	
Item	Description	MIMARKS survey	MIMARKS specimen
Investigation			
Submitted to INSDC [boolean]	Depending on the study (large-scale, e.g., done with next-generation sequencing technology, or small-scale) sequences have to be submitted to SRA (Sequence Read Archives), DRA (DDBJ Sequence Read Archive) or through the classical Webin/Sequin systems to GenBank, ENA and DDBJ	M	M
Investigation type [mimarks-survey or mimarks-specimen]	Nucleic Acid Sequence Report is the root element of all MIMARKS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIMARKS survey or MIMARKS specimen	M	M
Project name	Name of the project within which the sequencing was organized	M	M
Environment			
Geographic location (latitude and longitude [float, point, transect and region])	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system	M	M
Geographic location (depth [integer, point, interval, unit])	Please refer to the definitions of depth in the environmental packages	E	E
Geographic location (elevation of site [integer, unit]; altitude of sample [integer, unit])	Please refer to the definitions of either altitude or elevation in the environmental packages	E	E
Geographic location (country and/or sea [INSDC or GAZ]; region [GAZ])	The geographical origin of the sample as defined by the country or sea name. Country, sea or region names should be chosen from the INSDC list (http://insdc.org/country.html), or the GAZ (Gazetteer, v1.446) ontology (http://bioportal.bioontology.org/visualize/40651)	M	M
Collection date [ISO8601]	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated, that is, all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; except for 2008-01 and 2008, all are ISO6801 compliant	M	M
Environment (biome [EnvO])	In environmental biome level are the major classes of ecologically similar communities of plants, animals and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing and other factors like climate. Examples include desert, taiga, deciduous woodland or coral reef. Environment Ontology (EnvO) (v1.53) terms listed under environmental biome can be found at http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00000428	M	M
Environment (feature [EnvO])	Environmental feature level includes geographic environmental features. Examples include harbor, cliff or lake. EnvO (v1.53) terms listed under environmental feature can be found at http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297	M	M
Environment (material [EnvO])	The environmental material level refers to the matter that was displaced by the sample, before the sampling event. Environmental matter terms are generally mass nouns. Examples include: air, soil or water. EnvO (v1.53) terms listed under environmental matter can be found at http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483	M	M
MIGS/MIMS/MIMARKS extension			
Environmental package [air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water]	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported	M	M
Nucleic acid sequence source			
Isolation and growth conditions [PMID, DOI or URL]	Publication reference in the form of PubMed ID (PMID), digital object identifier (DOI) or URL for isolation and growth condition specifications of the organism/material	–	M
Sequencing			
Target gene or locus (e.g., 16S rRNA, 18S rRNA, nif, amoA, rpo)	Targeted gene or locus name for marker gene study	M	M
Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)	Sequencing method used, e.g., Sanger, pyrosequencing, ABI-solid	M	M

Furthermore, "–" denotes that an item is not applicable for a given checklist. E denotes that a field has environment-specific requirements. For example, whereas "depth" is mandatory for the environments water, sediment or soil, it is optional for human-associated environments. MIMARKS-survey is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. MIMARKS-specimen, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIMARKS-survey and specimen checklists can be used for any type of marker gene sequence data, ranging from 16S, 18S, 23S, 28S rRNA to COI, hence the checklists are universal for all three domains of life. Item names are followed by a short description of the value of the item in parentheses and/or value type in brackets as a superscript. Whenever applicable, value types are chosen from a controlled vocabulary (CV) or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (http://www.obofoundry.org/). This table only presents the very core of MIMARKS checklists, that is, only mandatory items for each checklist. Supplementary Results 2 contains all MIMARKS items, the tables for environmental packages in the MIGS/MIMS/MIMARKS extension and GenBank structured comment name that should be used for submitting MIMARKS data to GenBank. In case of submitting to EBI-ENA, the full names can be used.

Examples of MIMARKS-compliant data sets

Several MIMARKS-compliant reports are included in Supplementary Results 3. These include a 16S rRNA gene survey from samples obtained in the North Atlantic, an 18S pyrosequencing tag study of anaerobic protists in a permanently anoxic basin of the North Sea, a *pmoA* survey from Negev Desert soils, a *dsrAB*survey of Gulf of Mexico sediments and a 16S pyrosequencing tag study of bacterial diversity in the western English Channel (SRA accession no. SRP001108).

Adoption by major database and informatics resources

Support for adoption of MIMARKS and the MIxS standard has spread rapidly. Authors of this paper include representatives from genome sequencing centers, maintainers of major resources, principal investigators of large- and small-scale sequencing projects, and individual investigators who have provided compliant data sets, showing the breadth of support for the standard within the community.

In the past, the INSDC has issued a reserved 'barcode' keyword for the CBOL[7]. Following this model, the INSDC has recently recognized the GSC as an authority for the

MIxS standard and issued the standard with official keywords within INSDC nucleotide sequence records[27]. This greatly facilitates automatic validation of the submitted contextual data and provides support for data sets compliant with previous versions by including the checklist version as a keyword.

GenBank accepts MIxS metadata in tabular format using the sequin and tbl2asn submission tools, validates MIxS compliance and reports the fields in the structured comment block. The EBI-ENA Webin submission system provides prepared web forms for the submission of MIxS compliant data; it presents all of the appropriate fields with descriptions, explanations and examples, and validates the data entered. One tool that can aid submitting contextual data is MetaBar[28], a spreadsheet and web-based software, designed to assist users in the consistent acquisition, electronic storage and submission of contextual data associated with their samples in compliance with the MIxS standard. The online tool CDinFusion (http://www.megx.net/CDinFusion) was created to facilitate the combination of contextual data with sequence data, and generation of submission-ready files.

The next-generation Sequence Read Archive (SRA) collects and displays MIxS-compliant metadata in sample and experiment objects. There are several tools that are already available or under development to assist users in SRA submissions. The myRDP SRA PrepKit allows users to prepare and edit their submissions of reads generated from ultra-high-throughput sequencing technologies. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to checklists such as MIMARKS. The Quantitative Insights Into Microbial Ecology (QIIME) web application (http://www.microbio.me/qiime) allows users to generate and validate MIMARKS-compliant templates. These templates can be viewed and completed in the users' spreadsheet editor of choice (e.g., Microsoft Excel). The QIIME web-platform also offers an ontology lookup and geo-referencing tool to aid users when completing the MIMARKS templates. The Investigation/Study/Assay (ISA) is a software suite that assists in the curation, reporting and local management of experimental metadata from studies using one or a combination of technologies, including high-throughput sequencing[29]. Specific ISA configurations (http://isa-tools.org/tools.html) have been developed to ensure MIxS compliance by providing templates and validation capability.

Another tool, ISAconverter, produces SRA.xml documents, facilitating submission to the SRA repository. MIxS checklists are also registered with the BioSharing catalog of standards (http://biosharing.org/), set to progressively link minimal information specifications to the respective exchange formats, ontologies and compliant tools.

Further detailed guidance for submission processes can be found under the respective wiki pages (http://gensc.org/gc_wiki/index.php/MIxS) of the standard.

Maintenance of the MIxS standard

To allow further developments, extensions and enhancements of MIxS, we set up a public issue tracking system to track changes and accomplish feature requests (http://mixs.gensc.org/). New versions will be released annually. Technically, the MIxS standard, including MIMARKS and the environmental packages, is maintained in a relational database system at the Max Planck Institute for Marine Microbiology Bremen on behalf of the GSC. This provides a secure and stable mechanism for updating the checklist suite and versioning. In the future, we plan to develop programmatic access to this database to allow automatic retrieval of the latest version of each checklist for INSDC databases and for GSC community resources. Moreover, the Genomic Contextual Data Markup Language is a reference implementation of the GSC checklists by the GSC and now implements the full range of MIxS standards. It is based on XML Schema technology and thus serves as an interoperable data exchange format for infrastructures based on web services[30].

Conclusions and call for action

The GSC is an international body with a stated mission of working towards richer descriptions of the complete collection of genomes and metagenomes through the MIxS standard. The present report extends the scope of GSC guidelines to marker gene sequences and environmental packages and establishes a single portal where experimentalists can gain access to and learn how to use GSC guidelines. The GSC is an

open initiative that welcomes the participation of the wider community. This includes an open call to contribute to refinements of the MIxS standards and their implementations.

The adoption of the GSC standards by major data providers and organizations, as well as the INSDC, supports efforts to contextually enrich sequence data and complements recent efforts to enrich other (meta) 'omics data. The MIxS standard, including MIMARKS, has been developed to the point that it is ready for use in the publication of sequences. A defined procedure for requesting new features and stable release cycles will facilitate implementation of the standard across the community. Compliance among authors, adoption by journals and use by informatics resources will vastly improve our collective ability to mine and integrate invaluable sequence data collections for knowledge- and application-driven research. In particular, the ability to combine microbial community samples collected from any source, using the universal tree of life as a measure to compare even the most diverse communities, should provide new insights into the dynamic spatiotemporal distribution of microbial life on our planet and on the human body.

Note: Supplementary information is available on the Nature Biotechnology website.

References

1. Field, D. et al. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 26, 541–547 (2008).

2. Taylor, C.F. et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat. Biotechnol. 26, 889–896 (2008).

3. Ludwig, W. & Schleifer, K.H. in Microbial Phylogeny and Evolution, Concepts and Controversies. (ed. Sapp, J.) 70–98 (Oxford University Press, New York, 2005).

4. Ludwig, W. et al. Bacterial phylogeny based on comparative sequence analysis. Electrophoresis 19, 554–568 (1998).

5. Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. Genetic diversity in Sargasso Sea bacterioplankton. Nature 345, 60–63 (1990).

6. Stahl, D.A. Analysis of hydrothermal vent associated symbionts by ribosomal RNA sequences. Science 224, 409–411 (1984).

7. Ward, D.M., Weller, R. & Bateson, M.M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345, 63–65 (1990).

8. DeLong, E.F. Archaea in coastal marine environments. Proc. Nat. Acad. Sci. USA 89,

5685–5689 (1992).

9. Diez, B., Pedros-Alio, C. & Massana, R. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Appl. Environ. Microbiol. 67, 2932–2941 (2001).

10. Fuhrman, J.A., McCallum, K. & Davis, A.A. Novel major archaebacterial group from marine plankton. Nature 356, 148–149 (1992).

11. Hewson, I. & Fuhrman, J.A. Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. Appl. Environ. Microbiol. 70, 3425–3433 (2004).

12. Huber, J.A., Butterfield, D.A. & Baross, J.A. Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. Appl. Environ. Microbiol. 68,

1585–1594 (2002).

13. Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature 409, 603–607 (2001).

14. Moon-van der Staay, S.Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature 409, 607–610 (2001).

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

16. Rappe, M.S. & Giovannoni, S.J. The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394 (2003).

17. Francis, C.A., Beman, J.M. & Kuypers, M.M.M. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. ISME J. 1, 19–27 (2007).

18. Zehr, J.P., Mellon, M.T. & Zani, S. New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. Appl. Environ. Microbiol. 64, 3444–3450 (1998).

19. Minz, D. et al. Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. Appl. Environ. Microbiol. 65, 4666–4671 (1999).

20. Gilbert, J.A. et al. The seasonal structure of microbial communities in the Western English Channel. Environ. Microbiol. 11, 3132–3139 (2009).

21. Martinez, A.W., Tyson, G. & DeLong, E.F. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. Environ. Microbiol. 12, 222–238 (2009).

22. Hanner, R. Data Standards for BARCODE Records in INSDC (BRIs) (Database Working Group, Consortium for the Barcode of Life, 2009). <http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG_data_standards-Final.pdf>.

23. Pruesse, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35,

7188–7196 (2007).

24. Cole, J.R. et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141–D145 (2009).

25. Vogel, T.M. et al. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat. Rev. Microbiol. 7, 252 (2009).

26. Turnbaugh, P.J. et al. The Human Microbiome Project. Nature 449, 804–810 (2007).

27. Benson, D.A. et al. GenBank. Nucleic Acids Res. 36, D25–D30 (2008).

28. Hankeln, W. et al. MetaBar—a tool for consistent contextual data acquisition and standards compliant submission. BMC Bioinformatics 11, 358 (2010).

29. Rocca-Serra, P. et al. ISA infrastructure: supporting standards-compliant experimental reporting and enabling curation at the community level. Bioinformatics 26, 2354–2356 (2010).

30. Kottmann, R. et al. A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS 12, 115–121 (2008).

II. The genomic standards consortium: bringing standards to life for microbial ecology

Authors: Pelin Yilmaz, Jack A Gilbert, Rob Knight, Linda Amaral-Zettler, Ilene Karsch-Mizrachi, Guy Cochrane, Yasukazu Nakamura, Susanna-Assunta Sansone, Frank Oliver Glöckner, Dawn Field

Published in: ISME Journal. 2011; 5: 1565-1567

Contribution: conceived the design of the manuscript, wrote the manuscript

The genomic standards consortium: bringing standards to life for microbial ecology

Pelin Yilmaz [1,2], Jack A. Gilbert[3,4], Rob Knight[5], Linda Amaral-Zettler[6], Ilene Karsch-Mizrachi[7], Guy Cochrane[8], Yasukazu Nakamura[9], Susanna-Assunta Sansone[10], Frank Oliver Glöckner[1,2,*] and Dawn Field[11]

1 Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

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

3 Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

4 Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

5 Howard Hughes Medical Institute and Department of Chemistry & Biochemistry, University of Colorado at Boulder, Boulder, CO 80309, USA

6 Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

7 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

8 EMBL Outstation. The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

9 Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima 411-8510, Japan

10 Oxford e-Research Centre, University of Oxford, Oxford OX1 3QG, UK

11 NERC Centre for Ecology and Hydrology, Oxford, OX1 3SR, UK.

Correspondence: Frank Oliver Glöckner, E-mail: fog@mpi-bremen.de

Adoption of easy-to-follow standards will vastly improve our ability to interpret data from genomes, metagenomes and marker studies

Interest in sampling of diverse environments, combined with advances in high-throughput sequencing, vastly accelerates the pace at which new genomes and metagenomes are generated. For example, as of January 2011, 12 500 user-generated metagenomes have been submitted to the public MG-RAST Annotation server (http://metagenomics.nmpdr.org; Meyer *et al.*, 2008), >90% of which were produced using high-throughput sequencing methodologies. We have entered into an era of 'mega-sequencing projects' that include the Genomic Encyclopaedia of Bacteria and Archaea project (http://www.jgi.doe.gov/programs/GEBA), the Microbial Earth Project (http://genome.jgi-psf.org/programs/bacteria-archaea/MEP/index.jsf), the Human Microbiome Project (http://nihroadmap.nih.gov/hmp), the Metagenomics of the Human Intestinal Tract consortium (http://www.metahit.eu), the Terragenome Initiative (http://www.terragenome.org), the Tara Oceans Expedition (http://oceans.taraexpeditions.org), the National Ecological Observatory Network (NEON-http://www.neoninc.org), the International Census of Marine Microbes (ICoMM-http://icomm.mbl.edu), Microbial Inventory Research Across Diverse Aquatic Long-Term Ecological Research Sites (http://amarallab.mbl.edu/mirada/mirada.html), the Earth Microbiome Project (http://www.earthmicrobiome.org) and other funded and unfunded projects, with many more visionary projects on the horizon.

Additionally, studies of emerging metatranscriptomes (community transcript profiles), metaproteomes (community protein profiles) and metametabolomes (community metabolite profiles) now complement genomes and metagenomes. Comparative studies of multi-omic data sets from the same community hold the promise of unparalleled insights into fundamental questions across a range of fields including evolution, ecology, environmental science, physiology and medicine. Advances stem from improvements in the annotation and quantification of genes, pathways, organisms and consortia within these communities. We are just starting to exploit these technologies to understand the microbial world, and have only scratched the surface in terms of sampling microbial diversity across temporal and spatial scales (Delmotte *et al.*, 2009;Gilbert *et al.*, 2010a). To fully exploit the promise of these data, we need both scientific innovation and community agreement on how to provide appropriate stewardship of these resources for the benefit of all.

Although we have collected billions of nucleic-acid sequences from thousands of ecosystems, illuminating uncharacterized microbial lifestyles remains far from trivial. For example, in each analysed genome or metagenome, about 40% of the putative protein-coding genes cannot be assigned to any known function or taxon. Only 42% of the 61 known bacterial phyla have even a single cultured representative (Hugenholtz and Kyrpides, 2009), with the remainder being known only from 16S rRNA gene environmental surveys. Surprisingly, only 14% of cultured bacterial taxa have a single complete genome sequenced. Holistic approaches that will centralize (meta) omics data are needed, which will allow investigators to analyze these data within the context of space, time, habitat and characteristics of the environment. Networks of information arising from these studies will allow us to describe and predict ecological patterns of organisms, genes, transcripts and proteins.

One key insight into the function of a gene or organism is the environment where it occurs. Collection of contextual (meta) data, which delineates the source of a sequence in terms of the space, time, habitat and characteristics of the environment, is thus essential in interpreting these unknown genes and species, as well as gaining new insights into the known fraction. Although early comparative studies of metagenomes (Tringe *et al.*, 2005) relied on a few, deeply sequenced samples, the experience from 16S rRNA gene surveys suggests that additional insight is gained from observing spatial and temporal variation across hundreds of samples, whether examining the distribution of bacteria in soils across a continent (Lauber *et al.*, 2009) or various skin sites from many subjects (Grice *et al.*, 2009).

At present, the valuable contextual data halo is often missing for sequences deposited in the International Nucleotide Sequence Database Collaboration (INSDC; GenBank, European Nucleotide Archive (ENA, including EMBL-Bank) and the DNA Databank of Japan (DDBJ)). This leaves researchers in the position of searching in electronic resources, literature or contacting the authors for even the most basic contextual data, such as geographic location, date and time of sampling or the habitat where the sample was obtained. Molecular ecologists should immediately recognize the inherent value of these data to the community, because without them their own sequence data sets will

have extremely limited comparability with the wealth of other data available. Sequences without contextual data are like unlabeled cans in a supermarket—you do not know what you are purchasing until you open it and examine the contents. The present inability to automatically retrieve rich contextual data hampers comparative research, and constitutes a considerable misuse of the vast global resources currently being applied to microbial ecology. Just as food-safety laws emphasize clear and accurate labeling based on the product, process and producer, so should sequence data be properly annotated.

Standardization of the required information will greatly facilitate the annotation of sequence data. To achieve this, we must first have community collaboration and participation. Second, as a result of this collaboration, a contextual data set must be standardized in terms of content, syntax and terminology to which the community can adhere. In 2005, members of the community came together to form the Genomic Standards Consortium (GSC), an open-membership working body with the stated mission of working towards better descriptions of our genomes, metagenomes and related data (http://www.gensc.org). Supported by the expertize of the members involved in many of the aforementioned mega-sequencing projects, the GSC has formalized contextual data requirements for genomes and metagenomes as the Minimum Information about a Genome/Metagenome Sequence checklist (MIGS/MIMS) (Field *et al.*, 2008). Furthermore, to cover the description of phylogenetic and functional marker genes an extended standard, the Minimum Information about a MARKer gene Sequence (MIMARKS) checklist (http://gensc.org/gc_wiki/index.php/MIMARKS) has been developed (Yilmaz *et al.*, 2011). This family of minimum information checklists provides researchers with a condensed set of contextual data requirements, which range from description of the environment to sampling and sequencing procedures. The GSC is also driving the evolution of omics data sharing in a broader context through participation in the BioSharing (http://biosharing.org) portal. This forum aims to enable a broader dialog among funders, journals, standards and technology developers, and researchers on the critical issue of data sharing within the metagenomics community and beyond (Field *et al.*, 2009). It provides an example of what an infrastructure to support standards-compliant reporting of contextual data might look like; as well as encouraging and

enabling curation at community level (Rocca-Serra *et al.*, 2010; http://isatab.sourceforge.net).

The primary sequence databases' adoption of these standards is integral to their success. The INSDC partners have recognized this support for submission of compliant data sets with the adoption of an official keyword for the family of minimum standards reserved for compliant INSDC sequence records. Additionally, the development of a number of tools and formats to aid in data exchange (Kottmann *et al.*, 2008) and compliance during sequence submissions with these standards are ongoing within specialized genomics and metagenomics resources.

The application of high-throughput sequencing technologies has transformed the way microbial ecologists approach questions in their field (Gilbert *et al.*, 2010b). The shift of sequencing capacity to individual labs is creating a data bonanza. With appropriate contextual information, these data sets could herald a new era of discovery for microbial ecology. This will only be possible, if each study, from each environment, and from each lab maintains, at the very least, a minimum contextual data standard to facilitate cross-comparison and meta-analysis of global microbial communities. Inadequate implementation of these standards threatens progress in our field of research, as we will lose the best opportunity to produce a complete mechanistic understanding of microbial life. Every investigator will benefit immensely by being able to obtain a rapid, comprehensive answer to the question 'Have my microbes been seen before, and, if so, where, with whom, and what were they doing? Only by accepting the relatively small responsibility of entering their own contextual data into a global system will they realize this dream. Just as standardized deposition of sequence data contributed an immensely valuable resource, standardization of contextual data will allow us to reap vast dividends for decades to come and enable us to finally escape the burden of 'my sequence matches 1500 uncultured environmental isolates—now what'?

To provide a better understanding of the requirements, we included three examples for MIGS, MIMS and MIMARKS compliant data sets in the Supplementary Table 1 and Supplementary File 2 provides links to detailed submission and compliance guidelines.

With this open letter to the ISME community, we not only hope to advertise the existence of the GSC and invite more microbial ecologists investigating marker genes and doing 'omics' work to join us, but also make a call for compliance with current and future GSC standards. To learn how to describe your data according to MIGS/MIMS/MIMARKS (MIxS) standards, please visit the GSC website for details and options for submitting compliant data sets into public domain databases (http://gensc.org/gc_wiki/index.php/MIGS/MIMS/MIMARKS).

References

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R et al. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proceedings of the National Academy of Sciences 106: 16428-16433.

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: 541-7.

Field D, Sansone S-A, Collis A, Booth T, Dukes P, Gregurick SK et al. (2009). 'Omics data sharing. Science 326: 234-236.

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B et al. (2010a). The taxonomic and functional diversity of microbes at a temperate coastal site: A 'multi-omic' study of seasonal and diel temporal variation. PLoS ONE 5: e15545.

Gilbert JA, Meyer F, Bailey MJ. The Future of microbial metagenomics (or is ignorance bliss?). ISME J 2010b; e-pub ahead of print 25 November 2010, doi:10.1038/ismej.2010.178

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC et al. (2009). Topographical and temporal diversity of the human skin microbiome. Science 324: 1190-1192.

Hugenholtz P, Kyrpides NC (2009). A changing of the guard. Environ Microbiol 11: 551-553.

Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T et al. (2008). A standard MIGS/MIMS compliant XML schema: Toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS 12: 115-121.

Lauber CL, Hamady M, Knight R, Fierer N (2009). Soil pH as a predictor of soil bacterial community structure at the continental scale: A pyrosequencing-based assessment. Appl Environ Microbiol 75: 5111-5120.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, 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.

Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K et al. (2010). ISA infrastructure: supporting standards-compliant experimental reporting and enabling curation at the community level. Bioinformatics 26: 2354-2356.

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

Yilmaz P, Kottman R, Field D, Knight R, Cole JR, Amaral-Zettler L et al. The “Minimum Information about a MARKer gene Sequence” (MIMARKS) specification. Nat Biotechnol 2011; accepted 18 February 2011

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

III. MetaBar - a tool for consistent contextual data acquisition and standards compliant submission

Authors: Wolfgang Hankeln, Pier Luigi Buttigieg, Dennis Fink, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner

Published in: BMC Bioinformatics. 2010; 11 (1): 358

Contribution: correct implementation of MIxS standards components

MetaBar - a tool for consistent contextual data acquisition and standards compliant submission

Wolfgang Hankeln[1,2] , Pier Luigi Buttigieg[1,2] , Dennis Fink[3] , Renzo Kottmann[1] , Pelin Yilmaz[1,2] and Frank Oliver Glöckner[1,2]

[1] Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

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

[3] Symbiosis Group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

ABSTRACT

Background

Environmental sequence datasets are increasing at an exponential rate; however, the vast majority of them lack appropriate descriptors like sampling location, time and depth/altitude: generally referred to as metadata or contextual data. The consistent capture and structured submission of these data is crucial for integrated data analysis and ecosystems modeling. The application MetaBar has been developed, to support consistent contextual data acquisition.

Results

MetaBar is a spreadsheet and web-based software tool designed to assist users in the consistent acquisition, electronic storage, and submission of contextual data associated to their samples. A preconfigured Microsoft® Excel® spreadsheet is used to initiate structured contextual data storage in the field or laboratory. Each sample is given a unique identifier and at any stage the sheets can be uploaded to the MetaBar database server. To label samples, identifiers can be printed as barcodes. An intuitive web interface provides quick access to the contextual data in the MetaBar database as well as user and project management capabilities. Export functions facilitate contextual and sequence data submission to the International Nucleotide Sequence Database Collaboration (INSDC), comprising of the DNA DataBase of Japan (DDBJ), the European Molecular Biology Laboratory database (EMBL) and GenBank. MetaBar requests and stores contextual data in compliance to the Genomic Standards Consortium specifications. The MetaBar open source code base for local installation is available under the GNU General Public License version 3 (GNU GPL3).

Conclusion

The MetaBar software supports the typical workflow from data acquisition and field-sampling to contextual data enriched sequence submission to an INSDC database. The integration with the megx.net marine Ecological Genomics database and portal facilitates georeferenced data integration and metadata-based comparisons of sampling sites as well as interactive data visualization. The ample export functionalities and the INSDC submission support enable exchange of data across disciplines and safeguarding contextual data.

Background

The technological advancement in molecular biology facilitates investigations of biodiversity and functions on a temporal and geospatial scale. Improved sampling and laboratory methods, together with fast and affordable sequencing technologies [1], provide the framework to create a network of data points capable to answer basic ecological questions such as: 'Who is out there?' and 'What are these organisms doing?' To shed light on the complex interplay, adaptation and survival mechanisms of organisms in times of global change, contextual data describing the surrounding environment of sampling locations are of crucial importance [2]. At the very least, the latitude and longitude (x, y), the depth/altitude (z) in relation to sea level, and the sampling date and time (t) must be provided to allow anchoring molecular sequence data to their environmental context. If every sequence entry in the INSDC databases, comprising of DDBJ, EMBL and GenBank, would be thus georeferenced, researchers would have the post factum opportunity to contextualize these sequences with environmental data [3]. The power of contextual data enriched sequence data sets for the environmental and medical field has been recently documented [4-12].

Unfortunately, a survey in the EMBL sequence repository has shown that only a minor set of sequences are accompanied by a relevant amount of contextual data. For example, latitude, longitude (INSDC: lat_lon), and time (INSDC: collection_date), elements of the key contextual data tuple (x,y,z,t), are only reported in 7.3% and 7.2% of all submissions [Guy Cochrane, personal communication, October 2009]. But even if these data are available, correctness is not guaranteed.

The paucity of sequence associated contextual data has been recognized by the primary database providers and biocuration efforts are currently underway for specific subsets. The National Center for Biotechnology Information (NCBI), for example, curates the Reference Sequence (RefSeq) database which aims to provide a comprehensive, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts and proteins http://www.ncbi.nlm.nih.gov/RefSeq/ The European Molecular Biology Laboratory (EMBL) provides the UniProt/Swiss-Prot Knowledgebase which focuses on high quality protein sequence annotations http://www.ebi.ac.uk/uniprot/ [13]. However, the common aim of these efforts is to enhance the quality of the sequence or protein data and annotations rather than to provide more information on the data processing or the environment where the sample or organism has been taken.

To improve the quantity and quality of contextual data describing the environment of a sample is currently addressed by several projects which systematically collect georeferenced sequence data, environmental parameters, and further curated metadata [14]. SILVA [15] or RDP II [16] are examples for specialized databases that offer users curated and quality checked ribosomal RNA sequences that are often enriched with more reliable contextual and taxonomic information than originally annotated by the sequence submitters.

Furthermore, there are projects which curate the contextual data associated to the primary sequence data to facilitate specific analysis purposes. For example the Genomes OnLine Database (GOLD) collects metadata for ongoing and completed genome sequencing projects [17]. The Visualization and Analysis of Microbial Population Structures (VAMPS) project, with its integrated collection of tools for researchers, aims to visualize and analyze data for microbial population structures and distributions. All the contextual data in VAMPS comes from the MICROBIS database management system of the International Census of Marine Microbes (ICoMM: http://icomm.mbl.edu/microbis/). The megx.net portal http://www.megx.net [18] systematically integrates environmental parameters and sequence data of marine microbial genomes and metagenomes using georeferencing as an anchor.

In 2005 the international Genomic Standards Consortium (GSC) introduced checklists to promote standardized contextual data acquisition and storage. So far the Minimum Information about a Genome (Metagenome) Sequence (MIGS/MIMS) has been published [2] and the Minimum Information about an Environmental Sequence (MIENS) is in development http://gensc.org/gc_wiki/index.php/MIENS. For data exchange, the Genomic Contextual Data Markup Language (GCDML) [19] has been developed. A corollary of these ongoing efforts is the need to support field scientists in the consistent capture, storage and submission of both contextual and sequence data. Handlebar, a lightweight Laboratory Information Management System (LIMS) for the management of barcoded samples, in part addresses this issue by supporting the acquisition and processing of contextual data compliant with GSC standards [20]. The Barcode of Life Database (BOLD) initiative, which aims to identify and classify all eukaryotic life on Earth [21], also includes an advanced data acquisition and submission system. Unfortunately the system only supports phylogenetic markers which serve the eukaryotic domain e.g. the cytochrome c oxidase I (COI), which is only present in Eukarya and absent in the other domains of life, and so far exclude Archaea and Bacteria. Furthermore, it does not support the printing of database identifiers as barcodes to label collected samples.

Even though initiatives and tools exist to enhance the quantity and quality of contextual data subsequent to sequence submission, the amount of contextual data in the INSDC databases remains an issue. In summary, the most likely reasons for the persisting scarcity of consistent contextual data are: (1) Contextual data that are recorded in the field are often not stored electronically in structured databases. Consequently contextual data get rapidly unlinked from the sequence data and finally 'forgotten' in the sequence submission process. (2) There is a lack of automatic quality checking mechanisms active before data submission. Unfortunately, the flood of data entering the public databases prevents any manual curation process. (3) The sheer amount of potential contextual data with respect to the different fields of research ranging from textual data to images or even videos would rapidly exceed the capacities of the INSDC databases. Consequently, only a commonly agreed and standardized subset of data can be stored and made available.

Here the user-centric, web-based tool MetaBar is presented. MetaBar offers all the required features for sample identification and barcode labeling allowing robust sample tracking and inventorying. MetaBar is focused on the acquisition of contextual data recorded during sampling in the field 'offline' using spreadsheets. All recorded contextual data can be subsequently uploaded and consistently stored in an underlying database. The web Graphical User Interface (GUI) provides advanced user management and access to data and barcodes. Vitally, the tool captures GSC standards compliant data and it is integrated into a set of tools to facilitate further data usage such as integration, visualization and analysis available from the Marine Ecological Genomics database and portal, megx.net. Finally, MetaBar supports contextual data enriched sequence submission to the INSDC databases. The tool is not restricted to any given research field or domain of life, but can universally be applied to capture the contextual data of any biological sample. It is designed to support the complete workflow from the sampling event up to the sequence submission to an INSDC database.

Implementation

Programming languages, tools and frameworks

MetaBar is programmed in the object-oriented, platform-independent programming language, Java 1.5 http://www.java.com/en/. MetaBar is a multiuser web application using Apache Tomcat http://tomcat.apache.org/, the open source Spring frameworkhttp://www.springsource.org/about, jasig CAS http://www.jasig.org/cas, which is used as a central authentication service to implement the user management, and Apache POI http://poi.apache.org/ to parse the Microsoft® Excel® spreadsheets used for data input. Any Java objects generated are stored in a PostgreSQL database using the iBATIS persistence framework http://ibatis.apache.org/. The input fields in the Microsoft® Excel® spreadsheet are validated using Visual Basic (VBA) macros. The web interface has been continuously tested during development using Selenium IDE http://seleniumhq.org/projects/ide/ and JUnit http://www.junit.org/. The source code of the MetaBar application is available under the GNU GPL3

http://www.megx.net/metabar/data/metabar-1.0.tar.gz and as additional file 1 to this publication.

Core software components

The MetaBar application consists of (1) the Microsoft® Excel® acquisition spreadsheet which is used to capture and auto-correct the contextual data, (2) the MetaBar server which generates and receives the acquisition spreadsheets, parses the data and stores them in (3) MegDB, a PostgreSQL database which is the central database of the megx.net portal [18].

External software components

MetaBar is integrated into a set of external tools directly accessible from the web interface. The interpolation of environmental physical and chemical parameters of the oceans can be initiated via the WOA05 data extractor of the megx.net portal. On the fly visualization of sampling sites on a world map can be performed using the Genes Mapserver http://www.megx.net/gms and in Google Earth® via the KML export function. The data can be exported prior to sequence submission as a structured comment block for submission to INSDC by the Sequin toolhttp://www.ncbi.nlm.nih.gov/Sequin/index.html. MetaBar also includes a data export to GCDML [19] for report creation and data exchange.

Results

The core application can be best explained by describing the workflow across the different MetaBar components (Figure 1). First, users log on to the MetaBar web server (Figure 1, step 1). Upon entry, users can allocate a certain range of sample identifiers before, during or after a sampling campaign. The identifier consists of a six digit [sample-id] that is incremented with every new sample, a six digit [project-id] that is incremented with every new project and a two digit [institute-id] that is fixed and identifies a certain institute. The combination of these three parts in one identifier assures the unique identification of each sample. The identifiers can be printed (Figure 1, step 2) as barcodes onto labels (Figure 2) that can be placed on sample containers and pasted into laboratory

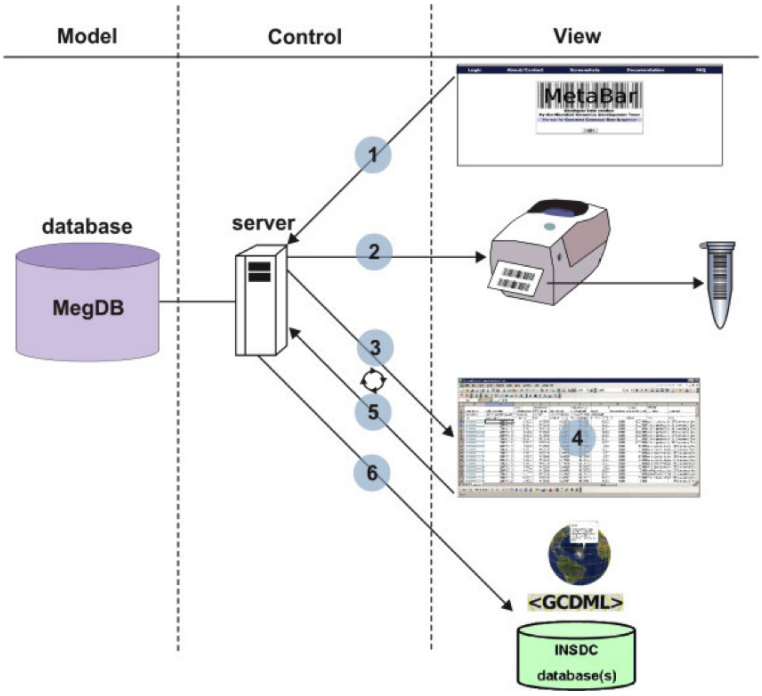


Figure 1. Scheme of the MetaBar workflow. Users are allowed to create barcodes, print them onto physical labels, and to capture, upload, and update contextual data for the barcodes using Excel® spreadsheets. The contextual data can be exported to various formats and submitted to the INSDC databases.

notebooks for consistency. Users can download (Figure 1, step 3) the acquisition spreadsheet containing the allocated identifiers and the empty contextual data fields in the first worksheet. As the user fills these fields, VBA validation macros check the inputs and users are prompted to use, for example, correct formats in the correct numerical range, where applicable. New worksheets can be added to the spreadsheet. Thus, any additional data outside of the MetaBar model can be added to the same file. Once the worksheets are filled (Figure 1, step 4) they can be uploaded (Figure 1, step 5) to the MetaBar web server.

Part:	[institute-ID]	[project-ID]	[sample-ID]
Length:	2 digits	6 digits	6 digits
Example:	01	000001	000001

Figure 2. The barcode identifier. Barcodes consist of three parts ([institute-ID], [project-ID] and [sample-ID]) which in combination uniquely identify a sample. The barcodes are printed onto physical labels that can be placed on sample containers.

After the upload is finished, the file is parsed and the values in the first worksheet are stored in the respective relational fields of the central database. The additional worksheets in the file are not lost, but stored as binary data in the database. The latter three steps can be repeated whenever it is necessary to edit and update the data. Users can log in to the system at any time to search and browse their data via the web GUI (Figure 3).

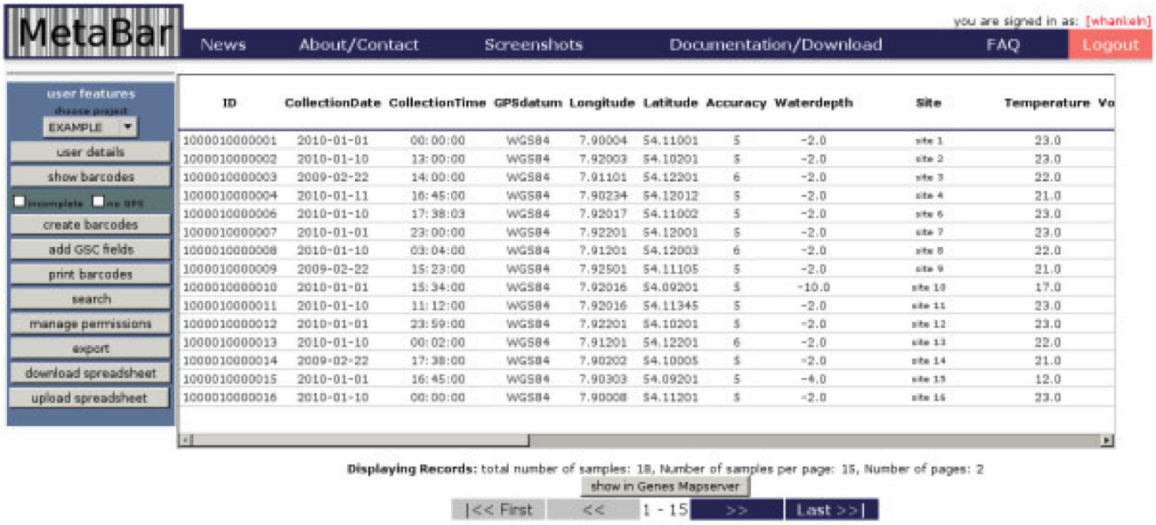


Figure 3. Screenshot of the graphical user interface I. Uploaded contextual data can be browsed and queried online.

Additionally, the MetaBar core set of contextual data fields can be extended for each sample with further GSC compliant parameters. These additional fields are organized into different types of report and environmental packages, each containing further parameters. The parameters can be directly selected and updated via the web interface (Figure 4).

MetaBar can also be used as an inventory e.g. for freezer contents. The database may be queried using sample identifiers by scanning their barcodes with an appropriate device, by manually entering their corresponding numeric code, or by text search on a metadata field. The query then retrieves all corresponding contextual data stored in the system.

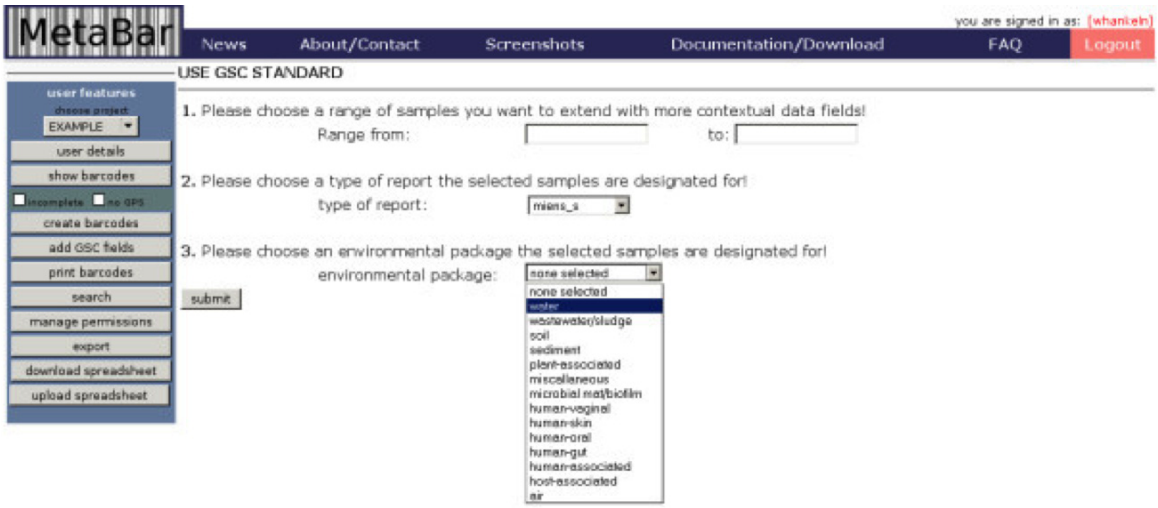


Figure 4. Screenshot of the graphical user interface II. Contextual data entries can be extended with GSC parameters. All contextual data can be exported and submitted to the INSDC databases.

MetaBar is integrated into a set of external tools with direct access from the web interface. The interpolation of physical and chemical parameters such as temperature, nitrate, phosphate, salinity, silicate, dissolved oxygen, oxygen saturation, apparent oxygen utilization and chlorophyll of a marine sampling site can be initiated via the WOA05 data extractor of the megx.net portal. On the fly visualization of sampling sites on a world map can be performed using the Genes Mapserver http://www.megx.net/gms. Furthermore, four export functions (Figure 1, step 6) are currently supported: (1) an export to KML to visualize sampling sites including their contextual data in Google Earth®, (2) an export to GCDML[19] for report creation and for data exchange, (3) an export to a GSC compliant MIGS/MIMS/MIENS spreadsheet, and (4) an export as a structured comment for sequence data submission to the INSDC databases using the Sequin tool http://www.ncbi.nlm.nih.gov/Sequin/index.html.

Role, ownership and permission concept

MetaBar's user management provides a "MetaBar admin", a "project admin" and a "MetaBar user" mode. These modes depend on the role that is assigned to a certain user account. The web GUI possesses a cascading menu on the left which contains the

"MetaBar admin features", the "project admin features", and the "MetaBar user features", respectively. Furthermore, a sophisticated ownership and permission concept offers the users to share their data with other users in the same project giving them read or write permissions, or to prevent access for others. Project admins have the possibility to transfer the ownership of a set of samples to another user, create projects, assign users to a project and to remove users from a project. For a given MetaBar installation there is only one MetaBar admin who can create users, assign or dismiss project admins and delete samples or whole projects.

A quick reference guide describing the general workflow of MetaBar from the acquisition to the submission of data is available on the website. Examples for metadata enriched INSDC database entries, created with MetaBar, are available through the accession numbers: [GenBank:GU949561 and GenBank:GU949562].

Discussion

MetaBar can be used whenever it is necessary to capture contextual data that describe the environmental origin of a sample. The system has been tested in several studies in close collaboration with biologists taking samples in the field. By integrating their feedback MetaBar should qualify as user-friendly, scientist-centric software tool.

Case study

The above mentioned studies allow the typical MetaBar workflow to be generalized as follows. A scientist acting as the project administrator (PA) plans a sampling campaign together with two members in his research team (PM1, PM2). The project EXAMPLE is created in MetaBar and the users PM1 and PM2 are added to EXAMPLE. It is anticipated that PM1 will collect five push core samples of sediment while PM2 will collect five water samples. Thus, PM1 and PM2 create five barcodes each and download their acquisition spreadsheets. These barcodes are printed in multiple copies to label sample containers, appear in field notebooks and for contingency.

During sample collection, PM1 and PM2 log contextual data such as latitude, longitude, depth and sampling time next to the barcode labels pasted in their field notebooks. The

sample containers are labeled with the corresponding barcodes and transported back to the laboratory. The link between sample and environment is thus established. At the end of the sampling campaign the contextual data gathered in the field are transferred to PM1 and PM2's acquisition spreadsheets. Barcodes pasted on the sample, the field records and present in the spreadsheet ensure fidelity and the data are then uploaded to the MetaBar server over the internet. PM1 and PM2 may enter further contextual data specific to their sampling environments by selecting the relevant GSC-compliant metadata packages (e.g. "sediment" and "water", respectively) through the web GUI. The PA and both members of the project can now review the consolidated contextual data for errors or missing values. Corrective action at this stage improves the quality of the data prior to submission.

During laboratory processing, every new subsample is labeled with a copy of the original sample's barcode, preserving the link to the *in situ* sampling event. Native laboratory protocols and practices are otherwise unaffected and are documented in laboratory books. PM1 sequences the genomes of several sediment sample isolates and PM2 sequences microbial metagenomes from the community in the water sample. Congruent to the environmental extensions, GSC packages corresponding to various study types are available. PM1 and PM2 may use the "MIENS culture (miens_c)" and "metagenome (me)" packages, respectively, to record data specific to their study type (Figure 4). PM1 and PM2 receive their genome and metagenome sequences as FASTA files with automatically generated sequence identifiers in the header. The researchers enter these identifiers into the "seqID" field in the acquisition spreadsheet and export the data to a format for submission to INSDC. With this mapping, these contextual datasets can easily be combined with one or more FASTA sequences using a suitable submission tool. The researchers then submit their metadata-enriched sequences from the EXAMPLE project to an INSDC database.

MetaBar implements a neat trade-off between universality and specificity. The export functions assure that the collected data can be publicly stored and shared with the scientific community.

Comparison of MetaBar and Handlebar

The idea of uniquely identifying samples and storing data about these samples in databases is not new and is widely used in many applications and disciplines. However, tools able to capture the contextual data of environmental samples combined with barcode labeling are rare. To our knowledge with the exception of MetaBar, the only open source tool using barcoding to identify georeferenced samples from the environment is Handlebar [20]. A tabular comparison of the programs' general features can be found in Table 1.

Table 1. Features of Handlebar and MetaBar

	Handlebar	**MetaBar**
Focus	Web-based lightweight LIMS for handling barcoded samples	Web-based tool for consistent contextual data acquisition with barcoded samples
System requirements	Operating system: Windows® or GNU Linux Apache, Perl, PostgreSQL, OpenOffice or Microsoft® Excel®	Without local MetaBar server installation: Operating system: Windows® Internet connection, web browser (e.g. Firefox), Microsoft® Excel® 2003 or higher Optional: EPL barcode printer (e.g. a Zebra® TLP 2824) With local MetaBar server installation: Operating system: Windows® or GNU Linux Apache, Java, Spring, jasig CAS, PostgreSQL, Microsoft® Excel® 2003 or higher
Coverage	Metadata that emerges during sampling events and subsequent processing step data	Contextual data that emerges during sampling events (other data optional)
Sample type templates	Various	One generic and extensible template
Input validation	Done by the server	Done by VBA® macros in the acquisition spreadsheet and on the server
Integration into data analysis tool set	GenQuery	http://www.megx.net
Export functions	-	GCDML, KML (for Google Earth)
Contextual data enriched sequence submission support	-	Export to MIGS/MIMS/MIENS and structured comment

HandleBar, as a lightweight LIMS, not only covers contextual data that are recorded during sampling, but also aims to document subsequent sample processing steps in the laboratory. In this respect, MetaBar is a simplification focusing only on the capture of

contextual data in the field. MetaBar does not seek to replace well established laboratory bookkeeping or professional LIM systems, but rather aims to complement this process to ensure that contextual data are electronically accessible. Nevertheless, users may choose to use the tool as a storage inventory manager or to store intermediate results of sample processing because it is possible to store additional data in the spreadsheets. It is important to note that coupling contextual data with sequence data before submission to the INSDC databases is a unique feature of MetaBar.

The barcodes are, by concept, solely used to link environmental samples to contextual and, if available, sequence and species data derived from a labeled sample, thus, no hierarchy or processing method is encoded in the identifiers. Also, sample hierarchies and complex identifier schemes are avoided. This concept does not interfere with native laboratory sample tracking methods, yet ensures consistency in environmental contextual data capture.

It is important that users have the flexibility to cover different sample types. MetaBar offers a single template in which a restricted part is parsed to the database and an unrestricted part of the spreadsheet can be changed to contain sample specific additional data. HandleBar offers a set of non-constrained sample templates depending on the sample type and also individual templates can be created. In MetaBar each sample can be extended with further parameters organized into types of report and environmental packages suggested by the GSC.

In contrast to HandleBar, data entered into MetaBar's acquisition spreadsheet is validated on input, ensuring correct format before upload to the MetaBar server. This avoids frequent rejection of the acquisition sheet. In HandleBar the validation is done by the web server and erroneous sheets have to be corrected retrospectively by the uploading user. The variety of export features are currently unique to MetaBar.

MetaBar is integrated into the megx.net tool set and connected to MegDB. This offers opportunity to work with the data and to analyze them alone, or in the context of other research project data stored in the megx.net database. This level of integration necessitated a user authentication and authorization management system and SSL encryption. Consequently, the local installation of MetaBar requires modification of the

open source code base. The software and a detailed installation manual are available at http://www.megx.net/metabar. However, accounts on the MetaBar installation hosted at the MPI for Marine Microbiology in Bremen can easily be given to interested users and an "anonymous" project exists where data of external users can be stored anonymously. It is the intention of the Microbial Genomics and Bioinformatics Group at the MPI-Bremen to support this tool as open source in the future.

Applicability

MetaBar has been developed at the Max Planck Institute for Marine Microbiology; however, the tool may be readily applied to a wide range of research fields outside the marine sciences. Contextual data fields relevant to air, host associated, human associated, sediment, soil, wastewater sludge or water samples are available via the "add GSC fields" function. The parameters in each of these environmental packages have been selected based on community usage and consensus http://gensc.org/gc_wiki/index.php/MIENS. For example, fields requesting data on barometric pressure, carbon dioxide, carbon monoxide, chemical administration, humidity, methane, organism count, oxygen, oxygenation status of sample, perturbation, pollutants, respirable particulate matter, sample salinity, sample storage duration, sample storage location, sample storage temperature, solar irradiance, temperature, ventilation rate, ventilation type, volatile organic compounds, wind direction, and wind speed would be presented to users using the air environmental package. Users may easily add new, custom fields as columns using standard Microsoft® Excel® operations. Combined, the GSC extensions and freedom for customization generalize MetaBar's applicability to any scenario necessitating the capture of contextual data describing a sample's environmental origin.

Conclusion

MetaBar offers an integrated contextual data acquisition, storage, and submission solution to the INSDC system. The impact of better contextual data availability and correctness in the primary sequence databases will greatly improve the possibilities to reach a higher level of data integration and interpretation to address basic ecological questions. MetaBar's integration into the megx.net tool set and its export mechanisms

offer extended analysis possibilities via comparison to other scientific studies and with complementary interpolated environmental data. The visualization of the sampling sites on the Genes Mapserver and in Google Earth® offers the users a simple way to show sampling events on the globe and to relate them to other publicly available scientific studies.

Statistical analysis of phylogenetic and functional biodiversity in their environmental context will reveal new insights into the biogeography and habitat adaptation of organisms. In the medical field, for example, it will be possible to create detailed disease maps which reveal mutation patterns of a certain pathogenic organism over time [5,11]. Such maps might help to predict the dispersal of epidemics and pandemics around the globe. For marine microbiology, Ed DeLong and coworkers have successfully shown that there is a stratification of genomic variability along the depth continuum in the water column at a specific sampling location [4]. It has also been demonstrated that specific diversity patterns are annually recurring [7]. A dense network of data points, enriched with contextual data, will lead to new insights into the complex interplay of organisms by comparing different sampling sites around the globe and over time. The denser this network of data points, the more will be revealed about the influence of the biotic factor in the elementary nutrient cycles that profoundly affect Earth's climate.

Availability and Requirements

Project name: MetaBar

Software

Project homepage: http://www.megx.net/metabar

Operating systems: Linux and Windows

Programming language: Java JRE 1.5 or higher

Other requirements: Microsoft® Excel® 2003 or higher, Google Earth® (optional)

License: GNU General Public License version 3 (GNU GPL3)

Hardware

At least 1024 Mb of RAM

EPL barcode printer (e.g. a Zebra TLP 2824) (optional)

Barcode handscanner (optional)

The software can be tested anonymously using the login: "anonymous" with the password: "testmetabar".

Authors' contributions

WH developed and implemented MetaBar and wrote the manuscript. RK advised programming design and helped with the integration of MetaBar with MegDB and megx.net. DF tested the software on cruises and provided feedback for design improvements. PY assured the MIGS/MIMS/MIENS standard compliance in MetaBar. PLB critically revised the manuscript and took care of EnvO integration. PY, PLB and RK tested the tool in the field. FOG supervised the work and helped with writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Thanks to Jens Harder, Christine Klockow, Mirja Meiners and all the testers. Thanks to Dawn Field and Tim Booth at NERC, UK for feedback and useful input during the design phase of MetaBar. Thanks to Norma Wendel for her help in finalizing the MetaBar source code. The study was supported by the Max Planck Society.

References

1. Hall N: Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* 2007, 210:1518-1525.

2. Field D, *et al*.: The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008, 26:541-547

3. Wieczorek J: The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International journal of geographical information science* 2004, 18:745-767

4. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N.-U, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM:Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 2006, 311:496-503

5. Janies D, Hill AW, Guralnick R, Habib F, Waltari E, Wheeler WC: Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Syst Biol* 2007, 56:321-329.

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

7. Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S: Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* 2006, 103:13104-13109.

8. Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, Tunlid A, Hagström A: Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* 2007, 16:867-880.

9. Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko R: GenGIS: A geospatial information system for genomic data. *Genome Res* 2009, 10:1896-1904.

10. Green JL, Bohannan BJM, Whitaker RJ: Microbial biogeography: from taxonomy to traits. *Science* 2008, 320:1039-1043.

11. Schriml LM, Arze C, Nadendla S, Ganapathy A, Felix V, Mahurkar A, Phillippy K, Gussman A, Angiuoli S, Ghedin E, White O, Hall N:GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. *Nucleic Acids Res* 2009, 38:D754-D764.

12. Field D: Working together to put molecules on the map. *Nature* 2008, 453:978.

13. Consortium, U: The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 2010, 38:D142-D148.

14. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill D, Kania PR, Schaeffer M, Pierre SS, Twigger S, White O, Rhee SY: Big data: The future of biocuration. *Nature* 2008, 455:47-50.

15. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO: SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007, 35:7188-7196.

16. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM: The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* 2005, 33:D294-D296.

17. Liolios K, Chen IMA, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2009, 38:D346-D354.

18. Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, Waldmann J, Glöckner FO: Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* 2009, 38:D391-D395.

19. Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T, Field D, Glöckner FO, Consortium GS: A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 2008, 12:115-121.

20. Booth T, Gilbert J, Neufeld JD, Ball J, Thurston M, Chipman K, Joint I, Field D: Handlebar: a flexible, web-based inventory manager for handling barcoded samples. *Biotechniques* 2007, 42:300-302.

21. Ratnasingham S, Hebert PDN: bold: The Barcode of Life Data System. [http://www.barcodinglife.org] *Mol Ecol Notes* 2007, 7:355-364.

IV. CDinFusion – Submission-ready, on-line Integration of Sequence and Contextual Data

Authors: Wolfgang Hankeln, Norma Wendel, Jan Gerken, Jost Waldmann, Pier Luigi Buttigieg, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner

Published in: PLoS ONE. 2011; 6 (9): e24797

Contribution: correct implementation of MIxS standards components

CDinFusion – Submission-ready, on-line Integration of Sequence and Contextual Data

Wolfgang Hankeln[1,2], Norma Johanna Wendel[1,3], Jan Gerken[1,2], Jost Waldmann[1], Pier Luigi Buttigieg[1,2], Ivaylo Kostadinov[1,2], Renzo Kottmann[1], Pelin Yilmaz[1,2], Frank Oliver Glöckner[1,2*]

1 Max Planck Institute for Marine Microbiology, Bremen, Germany

2 Jacobs University gGmbH, Bremen, Germany

3 Fachhochschule Bingen, Bingen am Rhein, Germany

ABSTRACT

State of the art (DNA) sequencing methods applied in "Omics" studies grant insight into the 'blueprints' of organisms from all domains of life. Sequencing is carried out around the globe and the data is submitted to the public repositories of the International Nucleotide Sequence Database Collaboration. However, the context in which these studies are conducted often gets lost, because experimental data, as well as information about the environment are rarely submitted along with the sequence data. If these contextual or metadata are missing, key opportunities of comparison and analysis across studies and habitats are hampered or even impossible. To address this problem, the Genomic Standards Consortium (GSC) promotes checklists and standards to better describe our sequence data collection and to promote the capturing, exchange and integration of sequence data with contextual data. In a recent community effort the GSC has developed a series of recommendations for contextual data that should be submitted along with sequence data. To support the scientific community to significantly enhance the quality and quantity of contextual data in the public sequence data repositories, specialized software tools are needed. In this work we present CDinFusion, a web-based tool to integrate contextual and sequence data in (Multi)FASTA format prior to submission. The tool is open source and available under the Lesser GNU Public License 3. A public installation is hosted and maintained at the Max Planck Institute for Marine Microbiology at http://www.megx.net/cdinfusion. The tool may also be installed locally using the open source code available at http://code.google.com/p/cdinfusion.

Introduction

The introduction of the first deoxyribonucleic acid (DNA) sequencing methods in 1977 marked a major breakthrough in life science [1], [2]. Subsequently, developments in these technologies allow the routine sequencing of organismal genomes, metagenomes and

marker genes from all domains of life. Genomic information can be seen as the ‘blueprint’ of life and being able to decode and to interpret it, grants insight into life's fundamental mechanisms [3], [4]. However, microbes pose a challenge to genomic description as the vast majority of microbial life cannot readily be isolated in pure cultures [5], [6]. The rise of cultivation independent approaches like metagenomic and sequencing of marker genes addresses this limitation [7]. In these approaches, bulk DNA is extracted from an environmental sample and either specific genes are amplified and sequenced or random sequencing is performed. Thus, a fragmented, but cultivation-independent, overview of an environment's biological diversity and functional potential is provided [8], [9].

Early on, scientists recognized the necessity to share sequence data to facilitate reuse, reproducibility and comparisons. This has become an integral part of the research and publication process. In the ‘Bermuda Principles’, on the first international strategy meeting on human genome sequencing in 1996, it was agreed upon, that all human genomic sequence information, generated by centers funded for large-scale human sequencing, should be freely available in the public domain to encourage research and to maximize its benefits to society (http://www.ornl.gov/sci/techresources/HumanGenome/research/bermuda.shtml, accessed: 11.03.2011). In the Fort Lauderdale meeting in 2003 organized by the Wellcome Trust, it was finally agreed to deposit all kinds of sequencing data that are analyzed in scientific publications in public databases. Over the past two decades, the amount of sequence data submitted to the world's largest public nucleotide sequence data repository INSDC (International Nucleotide Sequence Database Collaboration, comprising of DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive), and GenBank) has grown exponentially [10]. Recently, Next Generation Sequencing (NGS) technologies [11] allow even faster and more economical sequence generation, resulting in an unprecedented sequence accumulation.

Despite the impressive magnitude of sequence data generation, numerous life science studies have shown that contextual (meta)data (CD) are crucial for their interpretation [12]–[14]. CD are metadata about features such as the environmental origin

and the processing steps that were applied to obtain the sequences. These range from data about the geographic location (latitude, longitude), sampling time, habitat, to experimental procedures used to obtain the sequences up to video data recorded during sampling. The fact however that e.g. latitude, longitude (INSDC: lat_lon), and time (INSDC: collection_date), which can be submitted to the public repositories for years, have so far only been reported in 7.3% and 7.2% of all submissions [15], strongly implies that the procedure to deposit these data is hampered. Common reasons are: 1) no clear descriptors exist to guide the submitters which metadata should be deposited and 2) no appropriate tools exist that support the combined submission of sequence data and CD.

These concerns have recently prompted the Genomic Standards Consortium (GSC), an international consortium, which promotes mechanisms to standardize the description of genomes and the exchange of genomic data, to create a series of checklists defining the minimal set of CD that should accompany a sequence submission. The Minimum Information About a (Meta)Genome Sequence (MIGS/MIMS) checklist[16] outlines a conceptual structure for extending the core information that has been traditionally captured by the INSDC (DDBJ/EMBL/GenBank) to describe genomic and metagenomic sequences. The Minimum Information about a MARKer gene Sequence (MIMARKS) standard complements the MIGS/MIMS specification by adding two new “report types”, a “MIMARKS-survey” and a “MIMARKS-specimen”, the former being the checklist for uncultured diversity marker gene surveys, the latter is designed for marker gene sequences obtained from any material identifiable via specimens. The standards also cover sets of measurements and observations describing particular habitats, termed “environmental packages”. Collectively the MIGS/MIMS/MIMARKS standards are now called MIxS (Minimum Information about any (x) Sequence) [17], [18]. Through collaboration with the GSC, the INSDC now offers the structures to store the data items specified in the GSC checklists. This facilitates an early integration of sequence data and CD. However, specialized tools to allow this integration for different user scenarios are needed.

The European Nucleotide Archive (ENA) provides an on-line submission system called Webin which contains prepared web forms for the submission of GSC compliant data. It

shows all fields with descriptions, explanations and examples and does data validation in the forms (https://www.ebi.ac.uk/embl/genomes/submission/login.jsf, accessed: 16.03.2011). The Investigation Study Assay (ISA) Infrastructure offers a software suite that produces documents that can be submitted to the Sequence Read Archive (SRA) repository [19]. With the Quantitative Insights Into Microbial Ecology (QIIME) web application [20] users can generate and validate MIMARKS-compliant templates. Finally, MetaBar is a spreadsheet and web-based software tool which assists users in the consistent acquisition, electronic storage and submission of CD associated to their samples [15]. However, a tool that integrates CD and sequence data by directly enriching FASTA files for submission does not exist yet.

Here we present CDinFusion (Contextual Data and FASTA in fusion). CDinFusion has been designed to submit sequence data together with CD to INSDC. CDinFusion intends to facilitate the integration of CD and sequence data prior to submission by directly enriching sequence data using the FASTA format. It generates submission-ready outputs for INSDC by implementing the MIxS standard defined by the GSC. CDinFusion processes single as well as MultiFASTA files, containing up to millions of sequences. It was successfully applied to several use cases. Example submissions to the INSDC can be accessed with the following accession numbers: JF681370, JF268327–JF268425 and Genome Project ID 63253. A public installation is hosted and maintained at the Max Planck Institute for Marine Microbiology, Bremen, Germany: http://www.megx.net/cdinfusion. The tool is easy to install and released under the LGPL 3 open source license to promote distribution in aid of increasing the quantity and quality of CD in the public repositories.

Results and Discussion

CDinFusion has been designed as a web-based tool, which enables users to upload single or MultiFASTA files from single sequence to high-throughput analysis and enrich them with CD. After uploading the sequences, the user is requested to select the appropriate

GSC checklist and environmental package. CD can be entered in the web forms or CSV templates can be downloaded, filled with CD off-line and uploaded. The CSV files help to store and share the data. The merged sequence and CD can be downloaded for subsequent submission to INSDC.

The implemented workflow covers the three typical scenarios of sequence submission to an INSDC database namely: 1) Enriching a single sequence with one CD set, 2) Enriching many sequences in a MultiFASTA file with one CD set, and 3) Enriching subsets of sequences in a MultiFASTA file with several CD sets (Figure 1).

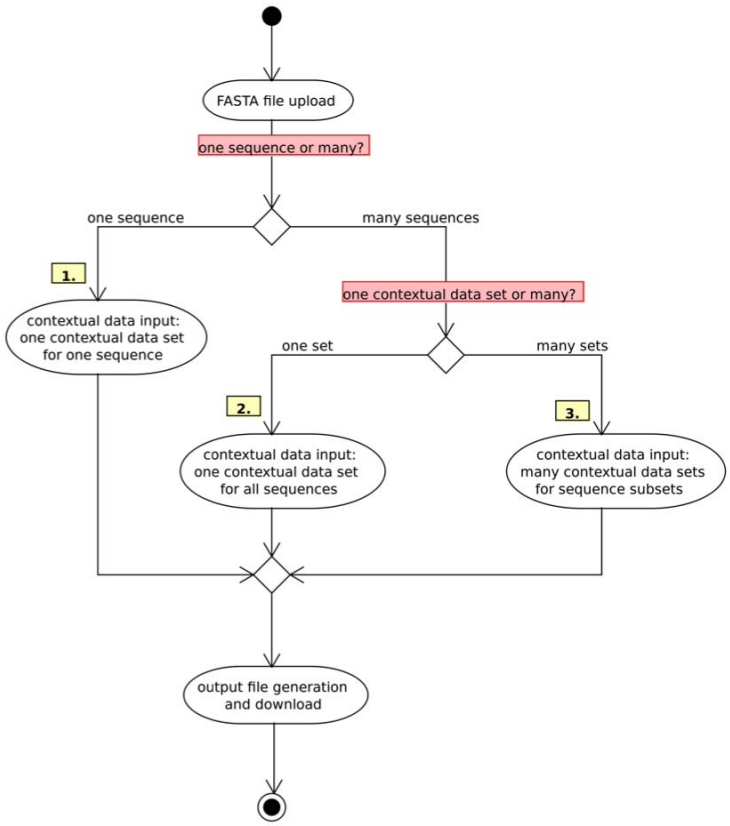


Figure 1. Overview of submission scenarios. Three primary scenarios of sequence data submission to INSDC can be distinguished and are all covered by the CDinFusion workflow:1) The submission of a single FASTA sequence file along with one CD set, 2) The submission of a MultiFASTA file along with one CD set for all sequences in the file and 3) The submission of a MultiFASTA file annotated with several CD sets.

The functionality of each of these different scenarios has been tested in dedicated use cases. The first use case was conducted with a single 16S rRNA sequence obtained from a bacterium isolated from a coastal water sample taken off the coast of the Wadden Sea island Sylt. After uploading the FASTA file the tool directly proceeded to the CD package selection for one CD set, as the file contained only a single sequence. The MIMARKS survey (mimarks_s) package and the water package were selected to provide suitable CD fields for this environmental survey sequence obtained from seawater. Subsequently the web forms were filled with all the CD available for this particular sequence (example Figure 2). After generating and downloading the output file, the CD enriched FASTA was imported into Sequin version 11.00. CDinFusion inserted qualifiers specified by GenBank into the header line of the FASTA file. The tool placed the rest of the CD into a tab delimited structured comment file. This file was loaded into Sequin with the "Advanced Table Readers" option in the "Annotate" menu. The CD appeared in the metadata section between the header and the feature table section. By selecting "Done", the Sequin file was saved and the complete submission was prepared. The INSDC database entry for this submission can be accessed at [Accession number: JF681370].

This use case exemplifies submission scenarios, where a single sequence and its CD are to be submitted to the INSDC databases. Single sequences can, for example, be marker genes or genomes that consist of a single sequence or contig.

In the second use case, a permanent draft genome from a *Rhodopirellula baltica* strain along with its associated CD was prepared for submission. After the 6.9 Mb MultiFASTA file was uploaded, the user was offered the option to annotate all sequences in this file with one CD set or to enter many CD sets for sequence subsets. As all sequence fragments were parts of the same bacterial genome, isolated from a sediment sample, one CD set for all sequences was selected using the MIGS bacterial genome (ba) checklist and the sediment package. The user filled in all CD fields available and the CD enriched files were generated, downloaded and imported into Sequin. The data of this genome project can be accessed by ID 63253 and with the accession number: AFAR00000000.

The genome will be analyzed in a separate study in preparation (Richter et al., Permanent draft genome sequence of Rhodopirellula baltica WH47).

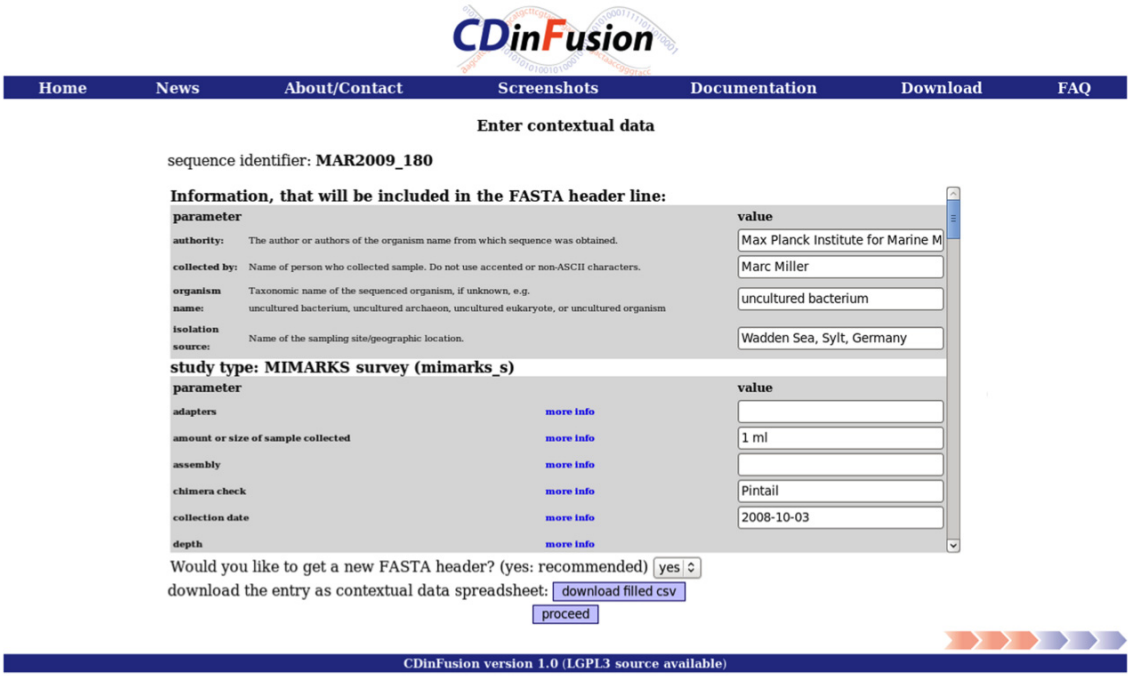


Figure 2. CDinFusion web user interface. The CD are entered into the auto-generated web forms. Details about each parameter are accessible with the “more info” link. These details are retrieved using a web service accessing the GSC database and are therefore always up to date.

This use case describes a procedure that may also be applied to metagenomic MultiFASTA files originating from one sampling site, which should be annotated with the same CD.

In the third use case a MultiFASTA file containing 99 16S rRNA sequences, obtained from a clone library, was enriched with CD. This file comprised four sequence subgroups, each with distinct CD. After the MultiFASTA file was uploaded, the CD for each of the groups was entered sequentially until all sequence subgroups were annotated.

After the user selected the MIMARKS (mimarks_s) and the “environmental package” sediment the CD were entered in the web forms.

The output files created were a CD enriched MultiFASTA file and a compressed ZIP archive containing four structured comment files, one for each of the subgroups. After the FASTA file had been imported to Sequin, the structured comment files were loaded one by one with the “Advanced Table Readers” function. The file was then saved and submitted. This clone library and its CD [21] will be analyzed in a separate study in preparation (Ruff et al., Microbial Communities of Submarine Methane Seeps at Hikurangi Margin, New Zealand). The INSDC database entries for this submission will be available under the accession numbers: JF268327–JF268425.

The same procedure has been applied to ten 16S rRNA sequences of an environmental culturability study conducted by the M.Sc. Marine Microbiology (MarMic) class of 2014 at the island of Sylt. The sequences of that study will be analyzed in a separate study in preparation (Hahnke et al., Flavobacteria of the North Sea: Diversity of Culturability) available under the accession numbers: JF710778–JF710788.

This use cases apply, whenever batches of sequences have to be submitted and subgroups of these sequences have to be annotated with individual CD sets. These MultiFASTA files can for example contain batches of marker genes or a pooled metagenome.

To test if high-throughput data can be processed with CDinFusion, metagenomic FASTA files from the Global Ocean Survey (GOS, http://jcvi.org/cms/research/projects/gos/overview/, accessed: 16.03.2011), and metagenome data from the Microbial Interactions in Marine Systems project (MIMAS, http://www.mimas-projekt.de/mimas/, accessed: 16.03.2011) were loaded into CDinFusion. FASTA files containing over two million sequences with file sizes of two GigaBytes (GB) could be processed in less than three minutes in an AMD™ 64Bit, 2 GHz and 4 GB RAM environment.

All described test cases were recorded with the Selenium IDE (http://seleniumhq.org/) test case recorder. The test cases along with the test data, except for the metagenomic

datasets, are deposited at http://code.google.com/p/cdinfusion. Descriptions how to run the tests, can be found in the documentation section of the public CDinFusion installation at http://www.megx.net/cdinfusion.

Materials and Methods

Languages, Tools and detailed Workflow

CDinFusion has been designed to allow users to add CD to single and MultiFASTA files that may comprise one to several million sequences. The CD enriched output can readily be submitted to the INSDC archives. The tool is programmed in the object-oriented, platform-independent programming language Java SE 5.0 (http://www.oracle.com/technetwork/java/index.html) using the Eclipse IDE (http://www.eclipse.org/). The open source Spring framework (http://www.spring-source.org/about/) was used, which supports the Model-View-Controller (MVC) design pattern. The functionality of the tool was continuously tested using the Selenium IDE (http://seleniumhq.org/). It runs on an Apache Tomcat 5.5.25 web server (http://tomcat.apache.org/). The project has been built using Apache Ant 1.7.1 (http://ant.apache.org/) and has been deployed on a web server with 2 GHz AMD Opteron™ processor 246, with 4 GB main memory and Debian GNU/Linux 5.0.3 (lenny).

Figures 3a and 3b show the implementation details of the software's workflow. FASTA files are parsed and validated, when uploaded by the FastaReader class. It implements the FastaValidatorCallback interface of the FastaValidator package (http://www.megx.net/FastaValidator), which has been developed within the frame of this project. This event-driven parser is designed to quickly parse and validate arbitrarily large FASTA files with minimal time and memory requirements. It facilitates the processing of gigabases of FASTA files containing millions of sequences on common desktop PC architectures. The parser is available separately and is also released under the GNU LGPL 3 license. It may also be used for other projects.

If only one sequence is detected in the FASTA upload, the control flow will be directed towards the 2a_GSC_SELECTED_1to1 JSP (use case 1 in the Results section), shown in Figure 3a. If the user opts to annotate all sequences of a MultiFASTA file (Figure 3b) with either one CD set or many CD sets, the control flow will be directed either to the 3b_CD_INPUT_1tom JSP (use case 2 in the Results section) or to the 3b1_CD_INPUT_ntom JSP (use case 3 in the Results section), respectively.

After the CD have been entered into the web forms, these data may be downloaded as comma separated value (CSV) files. The CSV files may serve as local backups and can be edited off-line and uploaded to CDinFusion to re-populate the web forms. Each session concludes with a confirmation step, where users can revisit any previous step and correct CD input if necessary. This holds true for all three branches of the workflow (Figure 3a and 3b). If the user chooses to proceed to the file download, a CD FASTA file and a structured comment file are generated and can, depending on their size, either be imported to Sequin or merged on the command line using tbl2asn (http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html, accessed: 30.03.2011) before submission.

Implementation of the GSC checklists in CDinFusion

Once the user has uploaded a MultiFASTA file and its contents have been validated, the data is processed along the data model (Figure S1). For each CD set a CDElement is created that contains an object for a “type of report” and an object for an “environmental package”. The GSC MIxS standard, including all “type of reports” and the “environmental packages”, is maintained in a relational database system called the GSC database at the Max Planck Institute for Marine Microbiology Bremen on behalf of the GSC. A non- authoritative version of the database can be downloaded at http://gcdml.gensc.org/wiki/GscDb [17]. Java classes were auto-generated from the relations in the GSC database using the Ibator tool from the iBatis project

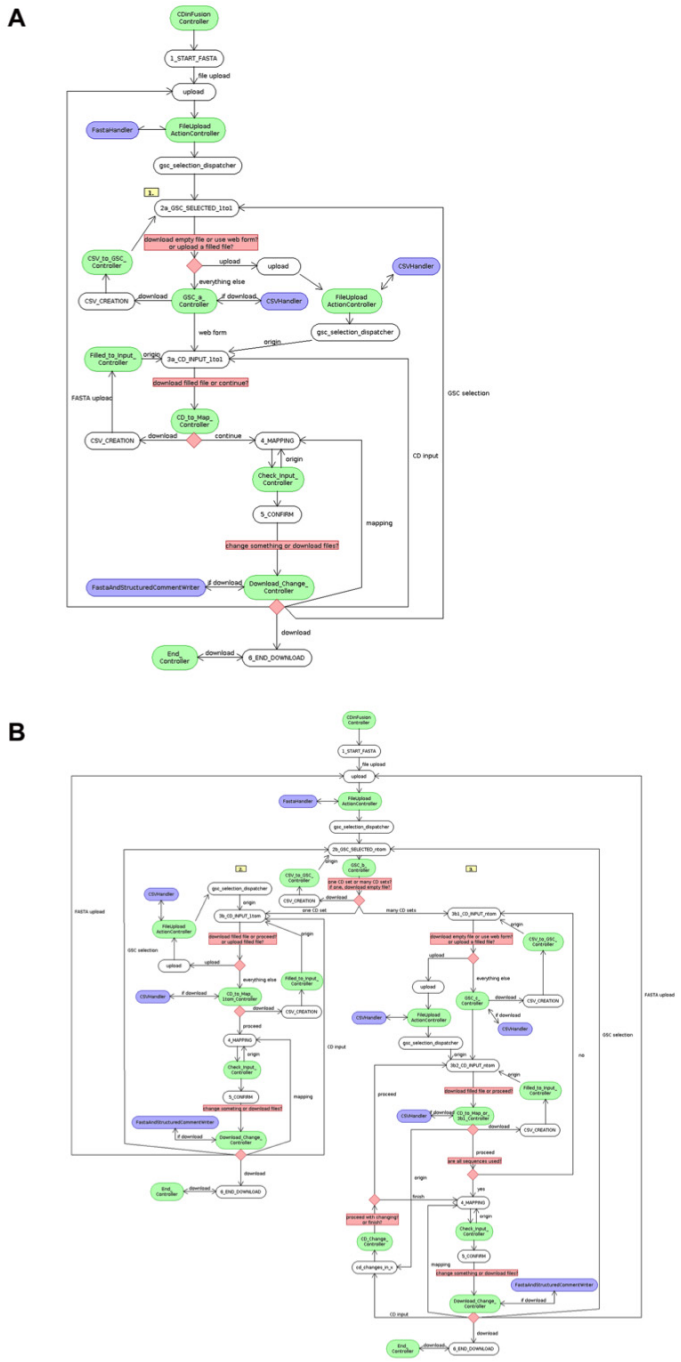


Figure 3. CDinFusion implementation details. The implementation details along the workflows 1–3 covering the primary scenarios of sequence data submission to the INSDC are shown. CDinFusion implements the Model-View-Controller design pattern. Classes implementing the data model and its manipulation methods are shown in blue, components belonging to the web user interface (view) are shown in white and components directing the workflow (control) are shown in green.

(http://ibatis.apache.org). The Java classes cover the MIGS, MIMS and MIMARKS (MIxS) specifications. The GSC plans to refine these standards annually. With every new version of the standards the Java classes can easily be updated using the Ibator tool.

The short names of the parameters are resolved using a web service that was developed within the frame of this project. The web service offers details about all GSC parameters stored in the GSC database. Web forms (see Figure 2) are dynamically rendered during runtime and therefore always contain the latest information including all definitions and descriptions of the GSC checklists parameters. If a user wants to know how a certain GSC parameter is specified, the “more info” link opens a window with information about the full name of the parameter, its definition, the expected value, the syntax and an example. This information is directly retrieved from the GSC database. For CDinFusion to be fully functional, there needs to be Internet access to the web service. If a certain type of report and environmental package has been selected, these parameters are cached. The next time these packages are selected the web forms are rebuild from cache without re-using the web service.

Two Strings “first SequenceID” and “lastSequenceID” in the CDElement object store the range of the associated sequence identifiers for each CD set. The CDFastaHeader object contains those parameters that are covered by the web forms in addition to the GSC parameters that are later used to extend the FASTA header lines.

Installation details

There are two ways to install CDinFusion: 1) CDinFusion can be installed by downloading and deploying the pre-compiled web archive file (war) on an Apache Tomcat (version >5.5.25). In this case the war file only has to be uploaded in the Tomcat manager. Afterwards the application can be accessed under http://<local_tomcat_installation>/CDinFusion.This method is preferable if users do not want to compile the program from its source code. 2) CDinFusion can also be installed by downloading and compiling the source code and subsequently deploying the software on an Apache Tomcat web server (version >5.5.25). To compile the code, the generic

build.xml and build.properties files can be adjusted to local settings. If the standard settings in these files are not changed, the war file will be compiled into the CDinFusion root folder. The project can be compiled by executing the Apache ant build tasks, “deploy” or “deploywar”, respectively. The build.xml can additionally be configured to directly deploy the tool on an Apache Tomcat web server or to create the war file and upload it with the Tomcat manager. Further installation details can be found in the README.txt file that is included in the source bundle and that is also available in the documentation section of the CDinFusion web page. On some platforms the CATALINA_HOME environment variable needs to be set, in order for CDinFusion to write and read files. Relative to the path specified, CDinFusion will create a “data” folder, where temporary files will be saved. The application has been tested on Debian GNU Linux installations, but should be platform-independent and run on all platforms that support Java and Apache Tomcat installation such as Windows™ or MAC OS™.

Availability and Future Directions

The public installation of CDinFusion is hosted and maintained at the Microbial Genomics and Bioinformatics Group (MGG) of the Max Planck Institute of Marine Microbiology Bremen and accessible under: http://www.megx.net/cdinfusion. The source code is available under GNU LGPL 3 and deposited in a public repository: http://code.google.com/p/cdinfusion.

As open source software it is the intention of the MGG to support this software well into the future. Currently CDinFusion supports submission of CD enriched sequence data to the INSDC using Sequin and tbl2asn for large data sets. Support for installations outside the MPI cannot be granted. The direct submission to EMBL/ENA and DDBJ is planned. Furthermore the integration of GCDML [22] as an exchange format would be advantageous. The GSC and life science community is encouraged to download the source code and to modify and extend the software to make it even more useful.

Supporting Information

Figure S1. In the CDinFusion data model the central Java class is the CDElement class, which is a composition of the classes “report type” and “environmental package”. These classes implement the MIGS, MIMS and MIMARKS (MIxS) checklists specified by the GSC. The two strings “firstSequenceID” and “lastSequenceID” define if the CDElement contains CD for a single or a range of sequences. Instances of the CDFastaHeader class contain the data that is generated into the FASTA headers in the FASTA file.

Acknowledgements

Thanks to S. Emil Ruff and the Geotechnologien project COMET/MUMM II (03G0608A, BMBF) for providing unpublished data and for the beta-testing. Thanks to Michael Richter for genome data and beta-testing. Thanks to Rudolf Amann, Bernhard Fuchs and Hanno Teeling for the MIMAS data. Thanks to Jens Harder, Richard Hahnke and the M.Sc. Marine Microbiology (MarMic) class of 2014 for data and beta-testing.

Author Contributions

Conceived and designed the experiments: WH NJW PY PLB RK FOG. Performed the experiments: WH NJW. Analyzed the data: WH NJW JG JW PLB IK RK PY FOG. Contributed reagents/materials/analysis tools: JW IK RK. Wrote the paper: WH FOG. Helped to optimize and finalize the source code: JG. Developed the FastaValidator library: JW. Helped to document the tool: PLB. Developed the GSC web service: IK RK. Wrote the GSC descriptions: PY.

References

1. Gilbert W, Maxam A (1973) The nucleotide sequence of the lac operator. Proc Natl Acad Sci U S A 70: 3581–3584.
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463–5467.
3. Moxon ER, Higgins CF (1997) E. coli genome sequence. A blueprint for life. Nature 389: 120–121
4. Henry C, Overbeek R, Stevens RL (2010) Building the blueprint of life. Biotechnol J 5: 695–704.
5. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59: 143–169.
6. Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A 99: 10494–10499.
7. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669–685.
8. 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.
9. Ratnasingham S, Hebert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes 7: 355–364.
10. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719–724.
11. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387–402.
12. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496–503.

13. Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. Proc Natl Acad Sci U S A 103: 13104–13109.

14. Schriml LM, Arze C, Nadendla S, Ganapathy A, Felix V, et al. (2009) GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. Nucleic Acids Res.

15. Hankeln W, Buttigieg PL, Fink D, Kottmann R, Yilmaz P, et al. (2010) MetaBar - a tool for consistent contextual data acquisition and standards compliant submission. BMC Bioinformatics 11: 358.

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

17. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, et al. (2011) The Minimum information about a maker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications, in press.

18. Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, et al. (2011) The genomic standards consortium: bringing standards to life for microbial ecology. ISME Journal 1751–7370.

19. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, et al. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 26: 2354–2356.

20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335–336.

21. Bialas J, Greinert J, Linke P, Pfannkuche O (2007) 190 p.

22. Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, et al. (2008) A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS 12: 115–121.

Curation of rRNA Datasets

V. SILVA: Comprehensive Databases for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB

Authors: Elmar Prüsse, Christian Quast, Pelin Yilmaz, Wolfgang Ludwig, Jörg Peplies, Frank Oliver Glöckner

Published in: Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches (ed. Frans J. de Bruijn). 2011; advance online publication 3 May 2011 doi: 10.1002/9781118010518.ch45

Contribution: curated bacterial and archaeal taxonomy for small and large subunit ribosomal RNA reference datasets, wrote the section on taxonomy

SILVA: comprehensive databases for quality checked and aligned ribosomal RNA sequence data compatible with ARB

Elmar Prüsse[1,2#], Christian Quast[1#], Pelin Yilmaz[1,2], Wolfgang Ludwig[3], Jörg Peplies[4] and Frank Oliver Glöckner[1,2]

1 Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

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

3 Department for Microbiology, Technical University Munich, D-85354 Freising, Germany

4 Ribocon GmbH, D-28359 Bremen

[#]both authors contributed equally to the paper

Correspondence should be addressed to F.O.G (fog@mpi-bremen.de)

ABSTRACT

SILVA (from Latin *silva*, forest), is a comprehensive web resource for up to date, quality controlled databases of aligned ribosomal RNA sequences from the *Bacteria*, *Archaea* and *Eukarya* domains. All sequences are checked for anomalies, carry a rich set of sequence associated contextual information, have multiple taxonomic classifications, and the latest validly described nomenclature. Four precompiled sequence datasets, fully compatible with the widely used ARB (from Latin *arb*, tree) software suite, are regularly offered for download on the SILVA website: (1) the SSU and LSU reference (Ref) datasets, comprising only high quality, nearly full length sequences suitable for in-depth phylogenetic analysis and probe design and (2) the comprehensive SSU and LSU Parc datasets with all publicly available rRNA sequences longer than 300 nucleotides suitable for biodiversity analyses. As of December 2009 the SILVA databases include more than 1.2 million small and large subunit rRNA gene sequences.

Introduction

Initiated by the pioneering studies of Fox and Woese [Fox et al. 1977] 30 years ago and later on pursued by Pace, Olsen, Giovannoni, and Ward [Pace et al. 1985; Olsen et al. 1986; Giovannoni et al. 1988; Ward et al. 1990], the ribosomal RNA (rRNA) molecule has been established as the “gold-standard” for the investigation of the phylogeny and ecology of microorganisms [Amann et al. 1995; Pace 2009] (see Chapter 17, Vol. I). Today, more than 1,200,000 publicly available small and large subunit (SSU and LSU)

rRNA sequences demand for appropriate software tools and specialized quality controlled databases. In anticipation of this impending deluge of rRNA data, the development of the ARB software suite and the curation of its associated databases began more than 15 years ago [Ludwig et al. 2004] (see Chapter 53, Vol. I). ARB offers a graphical user interface and a wide variety of interacting software tools built around a common database. It is estimated that ARB is currently employed by several thousand users worldwide, coming from both academia and industry. Since 2007, the corresponding SILVA database project provides structured, integrative knowledge datasets for SSU and LSU rRNAs fully compatible with ARB [Pruesse et al. 2007]. Besides the SILVA project, there are currently two projects offering access to a set of curated rRNA sequences and alignments: the Ribosomal Database Project II at Michigan State University in East Lansing, MI [Cole et al. 2009] (see Chapter 41, Vol I), and the greengenes project maintained by the Lawrence Berkeley National Laboratory in Berkeley, CA [DeSantis et al. 2006]. All projects offer at least one 16S rRNA dataset, but vary in the amount of sequences, quality checks, alignments, taxonomies and update procedures. However, the SILVA project is the only platform that actively incorporates homologous SSU as well as LSU sequences from all three domains of life, the *Bacteria*, *Archaea* (16S/23S) and *Eukarya* (18S/28S). To compensate for the limited phylogenetic resolution of the SSU rRNA [Peplies et al. 2004; Ludwig et al. 2005] the two fold larger LSU rRNA should now also be included in the rRNA approach [Amann et al. 1995] (see Chapter 3, Vol. I). Especially for Eukaryotes, the highly variable regions in the LSU rRNA are already commonly used for species discrimination [Wuyts et al. 2001].

The recent introduction of accelerated and less expensive sequencing technologies, such as pyrosequencing [Margulies et al. 2006] and their application in microbial ecology [Tringe et al. 2008; Reeder et al. 2009] (see Chapter ‘Consortia and Databases’, Vol. I), further substantiates the need for comprehensive quality controlled datasets for comparisons. The SILVA website was officially launched in January 2007 and this book chapter is an updated version of on the corresponding publication by Pruesse et al. [Pruesse et al. 2007].

Materials And Methods

Sequence data retrieval and rRNA extraction

The SILVA release cycle and numbering corresponds to that of the EBI-EMBL database, a member of the International Nucleotide Sequence Database Collaboration. A complex combination of keywords including all permutations of 16/18S, 23/28S, SSU, LSU, ribosomal and RNA is used to retrieve a comprehensive subset of all available SSU and LSU rRNA sequences. Additionally, the complete EBI-EMBL database is searched for rRNAs using Hidden Markov Models provided by RNAmmer [Lagesen et al. 2007]. The internal reference database providing the seed alignment for the automatic alignment of the SSU sequences includes a representative set of 56,354 aligned rRNA sequences from *Bacteria*, *Archaea* and *Eukarya* with 50,000 alignment positions. The database providing the LSU reference alignment contains 2,868 sequences with 150,000 alignment positions. Both datasets were iteratively cross-checked by expert curators during database build-up.

Quality checks

Every imported SSU and LSU sequence has to pass a multi-stage quality inspection. Sequences are rejected if they are shorter than 300 unaligned nucleotides, if they are composed of more than 2% of ambiguous bases, if homopolymers longer than four bases comprise more than 2% of the sequence, or if they have more than 5% identity to vector sequences. The identity is checked by querying a database of commonly used vector sequences, based on the EMVEC and UniVec databases using the blastn tool [Korf et al. 2003]. All thresholds to reject sequences were defined based on statistical analysis of the retrieved SSU and LSU sequences. Each sequence in the SILVA databases carries the percentages of ambiguities, homopolymers, and vector contamination. A summary "sequence quality" score is calculated according to the following formula (with Sq = sequence quality, A = % ambiguities, H = % homopolymers and V = % vector identity):

$$Sq = 1 - \left(\frac{\frac{A}{A\max} + \frac{H}{H\max} + \frac{V}{V\max}}{3} \right) * 100$$

This score represents the mean of the three individual parameters, such that 100 is the best possible value.

Aligner

To guaranty the specificity of the SILVA databases and a high quality alignment of the rRNAs the fast and accurate sequence aligner SINA (SILVA INcremental Aligner) was developed. In the first step the aligner uses the suffix tree index of ARB [Ludwig et al. 2004] to find up to 40 closely related sequences within the reference-alignment. These reference sequences are then transferred into a partial order graph as used in [Lee et al. 2002], but preserving the positional identity from the reference alignment. The graph concept allows “jumping” between the different references to find an optimal alignment for different sequence regions. To further improve the alignment quality a variability statistic is applied to give more weight to conserved positions. Results of each step of the aligner are reported to the database and shown in the corresponding fields of the exported ARB file. The “alignment quality” score is a measure of the similarity with the reference sequences that are taken into account for the alignment process. High values (>90) indicate that very similar sequences have been found within the seed alignment, resulting in a high likeliness for the alignment to be accurate. Due to the size of the seed alignment, low values are rather rare and suggest manual inspection of the particular sequences. The “basepair” score is calculated from the number of bases involved in helix binding according to the secondary structure model of Gutell et al. [Gutell et al. 1994]. To fit our unified scoring scheme, the alignment quality and the base pair score were normalized to values between 0 and 100, such that 100 represents the maximum score. After alignment, the constraint on the sequence length is tightened to at least 300 aligned bases within the rRNA gene boundaries.

Anomaly check

To check for sequence anomalies, a customized version of the Pintail software [Ashelford et al. 2005] is used. The software was initially adapted for batch processing by the RDP II team (see Chapter 41, Vol. I). Pintail checks whether a pair of sequences is mutually anomalous by computing a distance profile and comparing it to a predicted distance profile. The result is 'yes', 'likely', or 'no', depending on the amount of measured deviation

from expectation. From this operation, the SILVA pintail score is constructed by running each sequence against the ten most similar sequences within a cleaned reference set. Sequences that have passed all tests with 'no' (not anomalous) get a score of '100%', whereas all tests returning 'likely' would yield a 50% score. Only SSU sequences are checked for anomalies because the Pintail software does not contain profiles for sequences other than 16S rRNA.

Taxonomy and type strain information

Every sequence in the SILVA databases carries the EBI-EMBL taxonomy assignment. Where available, the greengenes and RDP taxonomies are added for comparison. The EMBL taxonomy is retrieved simultaneously with the sequences, whereas the other taxonomies are assigned to the sequences based on accession numbers. For LSU rRNA sequences no additional up to date datasets are available. A substantial revision of the classification of all sequences in the Ref datasets was first published with SILVA release 100. Based on the guide trees, all phylogenetic assignments are manually curated, taking into account taxonomic information provided by Bergey's Taxonomic Outline of the Prokaryotes [Garrity et al. 2004], the taxonomic outlines for Volumes 3, 4 and 5 of *Bergey's Manual* and the List of Prokaryotic names with Standing in Nomenclature [Euzeby 1997]. Furthermore, extensive effort is spent to represent prominent uncultured, and not-validly published environmental clades, groups, and taxa, respectively. The majority of these clades and groups are annotated in the guide tree for the SSU Ref dataset based on literature surveys and personal communications. Taxonomic groups consisting only of sequences from uncultured organisms are named after the clone sequence submitted earliest. Due to this exhaustive manual approach SILVA currently contains the most up to date and detailed bacterial and archaeal taxonomic classification.

Type strain information for *Bacteria* and *Archaea* is added to the field 'strain' and indicated by '[T]'. Mapping is based on the 'All-Species Living Tree' project [Yarza et al. 2008], the Straininfo.net database [Dawyndt et al. 2005] and RDP II [Cole et al. 2009].

Nomenclature and rDNAs from genome projects

With every release all organism names are synchronized with the 'Nomenclature up to date' website of the 'Deutsche Sammlung für Mikroorganismen und Zellkulturen'

(DSMZ, http://www.dsmz.de/download/bactnom/names.txt) and the ‘All-Species Living Tree' project [Yarza et al. 2008]. All rRNA sequences marked by EBI-EMBL as genome projects are labeled by ‘[G]’ in the ‘strain’ field. Manually curated information about the isolation environment (habitat) of the rRNAs of genome sequences is added based on the EnvO-Lite annotations in the megx.net database [Kottmann et al. 2009].

SSU and LSU rRNA databases for ARB

Two types of pre-compiled databases for both SSU and LSU rRNA sequences are available in ARB format: the high-quality Ref databases and the comprehensive Parc databases. Each Ref database is based on a subset of its Parc database comprising only full length or nearly full length 16/18S and 23/28S rRNA sequences. A SSU sequence is considered ‘full length’ if it contains at least 1200 aligned bases within the gene boundaries. This constrained is loosened to 900 bases for sequences belonging to the domain *Archaea* as applying a strict cut-off at 1200 bases would result in the loss of the majority of these sequences. LSU sequences are considered full length if they are at least 1900 bases long. For quality assurance, sequences that could not be unambiguously aligned (alignment quality score <50 for SSU or <30 for LSU) are removed from the Ref databases. Both Ref databases are supplemented with a fully classified guide tree. The trees are incrementally built using the ARB parsimony tool with filters to remove highly variable positions.

The rRNA Parc databases are a collection of all quality checked and automatically aligned rRNA sequences longer than 300 bases of the aligned rRNA gene. All sequences in the SILVA databases are associated with a rich set of sequence and process parameters, including information taken directly from the EBI-EMBL sequence record, as well as information from the initial quality checks of the alignment process. Using the search and query features of the web site or of ARB, one can quickly locate problematic sequences and generate individual high or low quality sequence subsets.

Availability/Website

The SILVA databases are available via a web-based interface. Downloads of the complete Parc and Ref datasets in the ARB file format are available in the download section. Subsets of aligned sequences from the Parc dataset can be retrieved from the

taxonomic browser and with the advanced search functions. After selecting a database and the desired taxonomy in the browser, the user can navigate through the taxonomy by clicking on the respective nodes. A cart system is used to easily select and download subsets of single sequences, complete groups or even whole phyla.

The advanced search functionality offered on the SILVA website allows the user to easily compile custom subsets of sequences. In addition to simple searches e.g. for accession numbers, organism names, taxonomic entities, or publication DOI/PubMed IDs, complex queries over several database fields using constraints such as sequence length or quality values are possible. All sequences or subsets can be added to the cart for subsequent export in the ARB and multi-FASTA file formats.

The colored bars on the search page and in the short and detailed sequence views of the browser give a fast overview of the different quality aspects assigned to each sequence. A wealth of additional information about the current status of the databases, as well as FAQs are available in the background section of the website. Furthermore, the SILVA website hosts a set of projects like 'The All-Species Living Tree' project [Yarza et al. 2008], the ‘Standard Operating Procedure for Phylogenetic Inference (SOPPI)’ [Peplies et al. 2008] and is part of the international Genomic Standards Consortium [Field et al. 2008] currently developing the Minimum Information about an ENvironmental Sequence (MIENS) checklist and standard (see Chapter 40, Vol. 1).

Results And Discussion

Data retrieval and processing

Cross checks with RDP II and greengenes indicate a sensitivity of the SILVA rDNA sequence retrieval procedure of >99%. A comparison of the length distribution immediately after importing the SSU sequences with the length distribution of sequences after the specific alignment for SILVA releases 89 to 100 shows that partial sequences between 300 and 800 bases were more frequently rejected than longer ones (>900 bases) (Fig. 1). The short ‘problematic’ sequences may be generated in diversity studies based on single strand sequencing. The high number of rejected sequences with less than 300

bases is an indicator for the increased number of projects employing tag sequencing based on next generation sequencing technologies.

As expected, the peaks of the SSU sequence length distribution follow the prominent primer sets used to sequence specific conserved regions on the 16S/18S rRNA gene [Marchesi et al. 1998] (Fig 1). The large number of sequences with 300 and 600 bases is typical for diversity studies that use single reads or fingerprinting techniques. It is interesting to note that up to SILVA release 94, the 500 base peak clearly dominated over the full length sequences. Recent releases show a trend towards the submission of higher quality, nearly full length rRNA sequences.

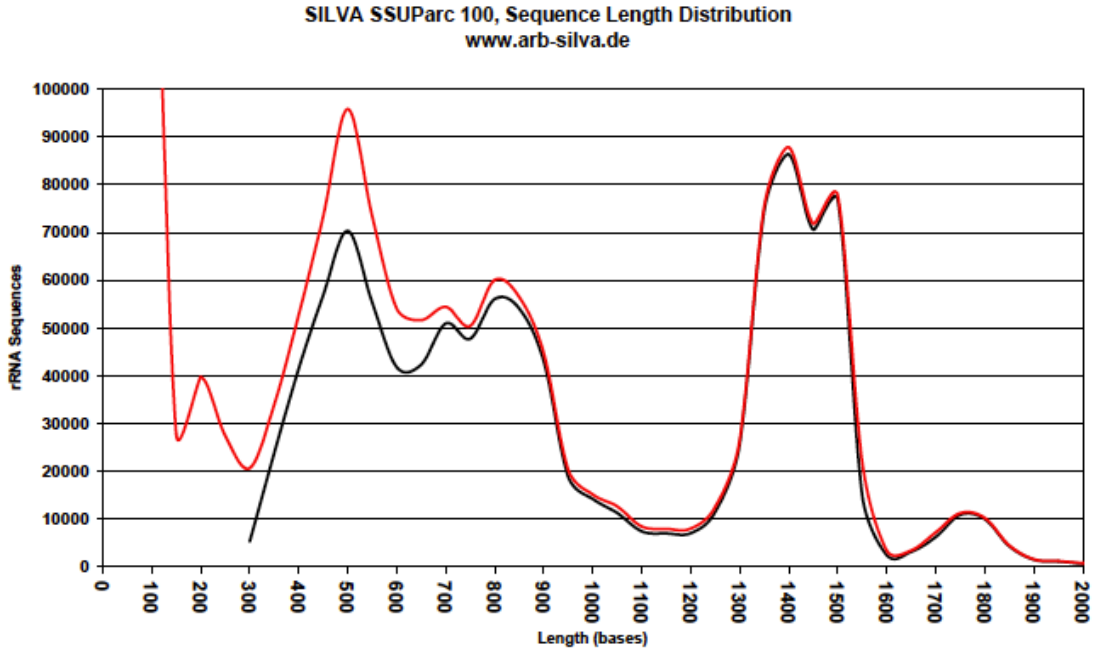


Figure 1 Sequence length distribution of rRNA genes in the SILVA 100 SSU database. The red line represents the sequence distribution directly after importing, the black line after quality checks and alignment. The huge amount of sequences up to 200 bases reflects the impact of tag sequencing approaches.

It has to be emphasized that the primary intention of the SILVA project is to provide reliable rRNA datasets with an informative set of processing and quality values assigned to each sequence. Such quality values enable users to easily evaluate sequences in order to create subsets of sequences for specific applications, or to identify sequences that need

further attention with respect to sequence and/or alignment quality or anomalies. The alternative taxonomies and type strain information, as well as the latest nomenclature will facilitate the daily workflow of diversity analysis using classical clone based and high throughput sequencing approaches. Additionally, SILVA provides two LSU databases to support the increasing use of molecular markers with a higher resolution than the SSU rRNA [Ludwig et al. 2005]. A taxonomic breakdown of the LSU Parc database contents shows that 91% of the sequences are of eukaryotic origin. A closer look indicates that the LSU rRNA is becoming more and more attractive for the molecular identification of e.g. Fungi.

Alignment

The current SILVA alignment is based on 50,000 and 150,000 alignment positions for the small and large subunit rRNA, respectively. The reasons for the large amount of alignment positions are: (1) large insertions often present in *Eukarya* and (2) sequencing errors, such as additional artificial bases often found in homopolymeric sequence stretches. Such errors are common and require placement to be filtered before phylogenetic tree reconstruction, without corrupting the rest of the alignment.

To further improve the quality of the SSU and LSU seed databases a manual curation process is performed and over time additional curated sequences are added to under-represented sections of the seed. The SSU seed currently includes over 1000 unpublished sequences that primarily cover the domain *Archaea*. The SILVA team highly appreciates the return of manually inspected and corrected alignments of sequence subsets for inclusion in the SILVA seed. This will allow to further increase the quality of future alignments.

Conclusions

The SILVA system provides comprehensive, quality controlled, richly annotated and aligned, reference rRNA datasets to support the molecular assessment of biodiversity, as well as investigations of the evolution of organisms. Applications of the datasets range from basic research in microbiology and molecular ecology to the detection of

contaminants and pathogens in biotechnology and medicine. Molecular taxonomy and diagnostics have already revolutionized our view on microbial diversity on Earth [Hong et al. 2006; Pedros-Alio 2006; Tringe et al. 2008]}, and the added value of molecular techniques for the determination of eukaryotic diversity has recently been documented by Tautz et al. [Tautz et al. 2002]. The SILVA databases combined with the ARB software suite provide a stable and easy to use workbench for researchers worldwide to perform in depth sequence analysis and phylogenetic reconstructions. They are designed as specialist databases to assist in the daily effort to keep pace with the increasing amount of data flooding the general-purpose primary databases.

Internet Resources

- The SILVA project (www.arb-silva.de)
- The Ribosomal Database Project II (http://rdp.cme.msu.edu/)
- The greengenes project (http://greengenes.lbl.gov/)
- The International Nucleotide Sequence Database Collaboration (http://www.insdc.org).
- The EMVEC database (http://www.ebi.ac.uk/blastall/vectors.html)
- The UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html)
- Documentation of the SILVA database fields in ARB (http://www.arb-silva.de/documentation/faqs/).
- Bergey's Manual (http://www.bergeys.org/outlines.html)
- List of Prokaryotic names with Standing in Nomenclature (http://www.bacterio.net/)
- The Megx.net database (www.megx.net)
- The Minimum Information about an Environmental Sequence (MIENS) checklist and standard (http://gensc.org/gc_wiki/index.php/MIENS)

Acknowledgments

We would like to thank Ralf Westram for expert assistance with the ARB software suite and all colleagues and students who helped with the manual curation of the databases. We would also like to thank James Cole, George Garrity and the RDP II team for help with Pintail and fruitful discussions. We are grateful for funding from the Max Planck Society.

References

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

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. 71:7724-7736

Cole JR et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acid Res. 37:D141-D145

Dawyndt P, Vancanneyt M, De Meyer H, Swings J (2005) Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. IEEE Transactions on Knowledge and Data Engineering 17:1111-1126

DeSantis TZ et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069-5072

Euzeby JP (1997) List of bacterial names with standing in nomenclature: A folder available on the Internet. Int. J. Syst. Bacteriol. 47:590-592

Field D et al. (2008) The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 26:541-547

Fox GE, Pechman KR, Woese CR (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. International Journal of Bacteriology 27:44-57

Garrity GM, Bell JA, Lilburn TG (2004) Taxonomic outline of the prokaryotes. In, release 5.0 edn. Springer-Verlag, New York

Giovannoni SJ, DeLong EF, Olsen GJ, Pace NR (1988) Phylogenetic groupspecific oligodeoxynucleotide probes for identification of single microbial cells. J. Bacteriol. 170:720-726

Gutell RR, Larsen N, Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol. Rev. 58:10-26

Hong SH, Bunge J, Jeon SO, Epstein SS (2006) Predicting microbial species richness. Proc. Natl. Acad. Sci. USA 103:117-122

Korf I, Yandell M, Bedell J (2003) BLAST. O'Reilly & Associates, Beijing, Cambridge, Farnham, Köln, Paris, Sebastopol, Taipei, Tokyo

Kottmann R et al. (2009) Megx.net: integrated database resource for marine ecological genomics. Nucleic Acids Res online

Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acid Res. 35:3100-3108

Lee C, Grasso C, Sharlow MF (2002) Multiple sequence alignment using partial order graphs. Bioinformatics 18:452-464

Ludwig W, Schleifer KH (2005) Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes. In: Sapp J (ed) Microbial phylogeny and evolution, concepts and controversies. Oxford university press, New York, pp 7098

Ludwig W et al. (2004) ARB: a software environment for sequence data. Nucleic Acid Res. 32:1363-1371

Marchesi JR, Sato T, Weightman AJ, Martin TA, Fry JC, Hiom SJ, Wade WG (1998) Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. Appl. Environ. Microbiol. 64:795-799

Margulies M et al. (2006) Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005). Nature 441:120-120

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu. Rev. Microbiol. 40:337-365

Pace NR (2009) Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev 73:565-576

Pace NR, Stahl DA, Olsen GJ, Lane DJ (1985) Analyzing natural microbial populations by rRNA sequences. ASM News 51:4-12

Pedros-Alio C (2006) Marine microbial diversity: can it be determined? Trends Microbiol. 14:257-263

Peplies J, Glöckner FO, Amann R, Ludwig W (2004) Comparative sequence analysis and oligonucleotide probe design based on 23S rRNA genes of Alphaproteobacteria from North Sea bacterioplankton. Syst. Appl. Microbiol. 27:573-580

Peplies J, Kottmann R, Ludwig W, Glöckner FO (2008) A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. Syst Appl Microbiol 31:251-257

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acid Res. 35:7188-7196

Reeder J, Knight R (2009) The 'rare biosphere': a reality check. Nat. Methods 6:636-637

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2002) DNA points the way ahead of taxonomy - In assessing new approaches, it's time for DNA's unique contribution to take a central role. Nature 418:479479

Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol. 11:442-446

Ward DM, Weller R, Bateson MM (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345:63-65

Wuyts J, De Rijk P, Van de Peer Y, Winkelmans T, De Wachter R (2001) The European Large Subunit Ribosomal RNA Database. Nucleic Acid Res. 29:175-177

Yarza P et al. (2008) The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 31:241-250

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

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

Published in: Nucleic Acids Research. 2010; 38: D391-D395.

Contribution: integration of 16S/18S and 23S/28S rRNA sequence data by linking SILVA with megx.net

Megx.net: integrated database resource for marine ecological genomics

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

1 Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen

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

*To whom correspondence should be addressed. Tel: +49 421 2028974; Fax: +49 421

ABSTRACT

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

Introduction

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

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

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

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

Megx.net is the first database to integrate curated contextual data with their respective genes, genomes and metagenomes in the marine environment (10). Now, the extended megx.net database resource allows post factum retrieval of interpolated environmental parameters, such as temperature, nitrate, phosphate, etc. for any location in the ocean waters based on profile and remote sensing data. Furthermore, the content has been significantly updated to include prokaryote and marine phage genomes, metagenomes from the GOS project (2) and all georeferenced small and large subunit ribosomal RNA (rRNA) sequences from the SILVA database project (11).

The extended megx.net portal is the first resource of its kind to offer access to this unique combination of data, including manually curated habitat descriptors for genomes,

metagenomes and marker genes, their respective contextual data and additionally integrated environmental data. See the megx.net online video tutorial for a guided introduction and overview at http://www.megx.net/portal/tutorial.html (Supplementary Data).

New Database Structure And Content

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

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

Contextual and sequence data content

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

MegDB, the first step towards a community call for integration of viral genomic and biogeochemical data (12).

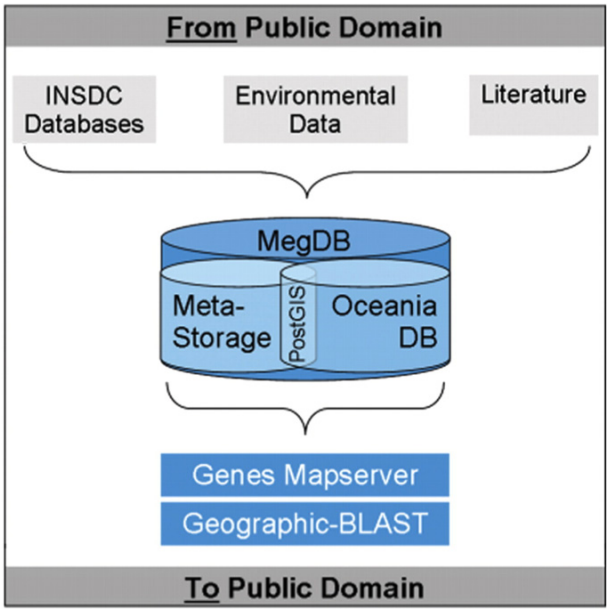


Figure 1. General architecture of megx.net: DNA sequence data (from INSDC) is integrated with contextual data from diverse resources (i.e. manual literature mining and the GOLD database) and interpolated environmental data. MegDB integrates the data conforming to OGC standards and MIGS/MIMS specification.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

In an effort towards integrating microbial diversity with specific sampling sites, megx.net has been extended to include georeferenced small and large subunit rRNA sequences from the SILVA rRNA databases project (11). Currently, only 9% (16S/18S) and 2% (23S/28S) of over 1 million sequences in SILVA SSUParc (16S/18S) and LSUParc (23S/28S) databases are georeferenced. With the implementation of the Minimal Information about an Environmental Sequence (MIENS) standard for marker gene sequences (http://gensc.org/gc_wiki/index.php/MIENS), efforts are ongoing to significantly improve this situation.

All genomic sequences in megx.net are supplemented by contextual data from GOLD (13) and NCBI Genome Projects

(http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). The database is designed to store all contextual data recommended by the Genomics Standards Consortium, and is thus compliant with the Minimum Information about a Genome Sequence (MIGS) standard and its extension, Minimum Information about a Metagenome Sequence (MIMS) (7,9).

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

Environmental data content

OceaniaDB was added to MegDB to supplement the georeferenced molecular data of MetaStorage with interpolated environmental parameters. When sufficient date, depth and location measurements are provided, any ‘on site’ contextual data taken at a sampling site can be supplemented by environmental data describing physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll.

The environmental data is retrieved from three sources:

(1) World Ocean Atlas: a set of objectively analysed (one decimal degree spatial resolution) climatological fields of in situ measurements (http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html);

(2) World Ocean Database: a collection of scientific, quality-controlled ocean profiles (http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html); and

(3) SeaWIFS chlorophyll a data (http://seawifs.gsfc.nasa.gov).

These data are described at 33 standard depths for annual, seasonal and monthly intervals. Together, the location and time data (x, y, z and t) serve as a universal anchor, and link environmental data to the sequence and contextual data in MetaStorage (Figure 1). As such, megx.net integrates biologist-supplied sequence and contextual data (measured at

the time of sampling) with oceanographic data provided by third-party databases. All environmental data are compatible with OGC standards (http://www.opengeospatial.org/standards) and are described with exhaustive meta-information consistent with the ISO 19115 standard.

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

User Access

Genes Mapserver

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

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

Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome, metagenome, marine phages and rRNA sequence data using the BLAST algorithm (15). The results are reported according to the sample locations (when provided) of the database hits. With the updated Geographic-BLAST, results are plotted on the Genes Mapserver world map, where they

are labeled by number of hits per site (Figure 2). Standard BLAST results are shown in a table, which also provides direct access to the associated contextual data of the hits.

Software extensions to the portal

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

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

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

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

User test case

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

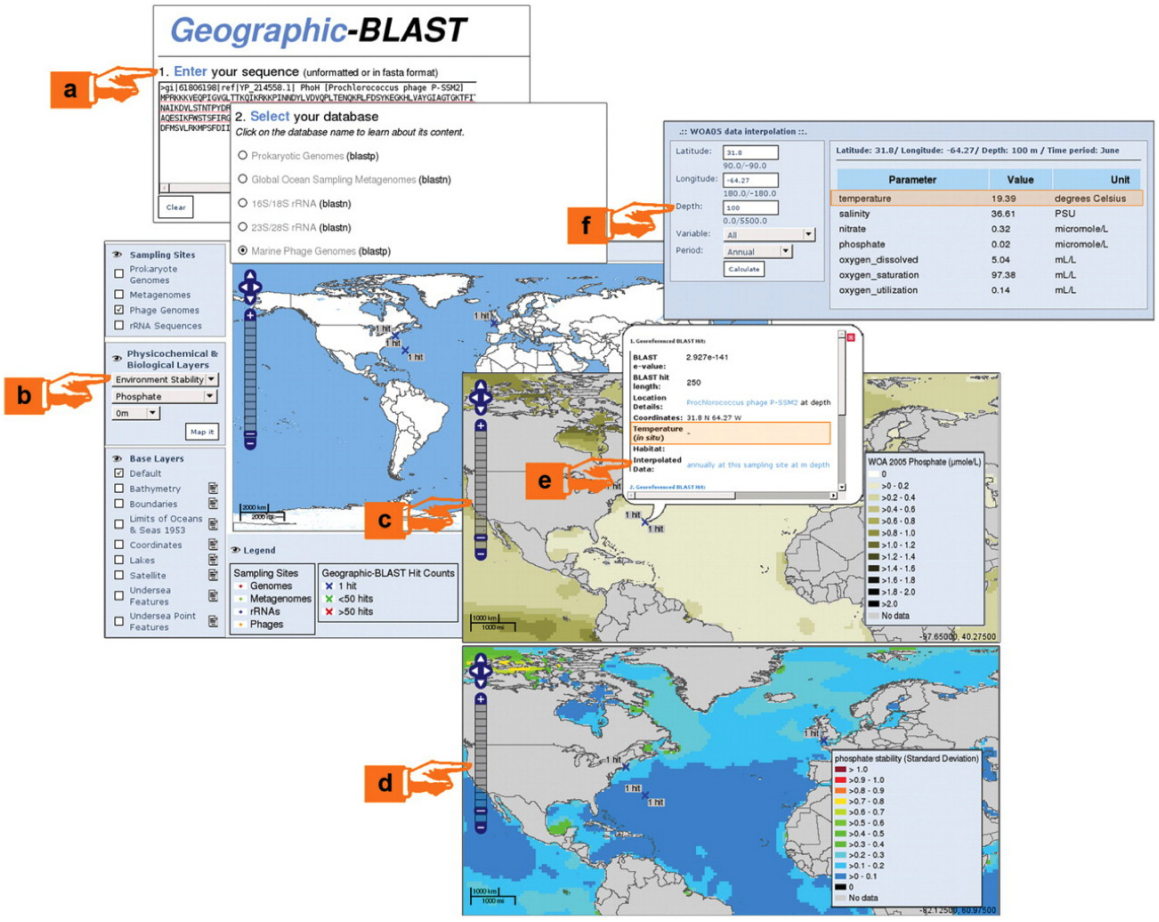


Figure 2. User test case: (a) BLAST sequence against the marine phage genomes to see the results on the Genes Mapserver. (b) View the BLAST hits with underlying environmental data, such as (c) average annual phosphate values, or (d) stability of phosphate concentrations in terms of monthly standard deviations.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

Web Services

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

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

Other changes

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

Summary

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

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

Database access

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

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

Supplementary Data

Supplementary Data are available at NAR Online.

Funding

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

Conflict of interest statement. None declared.

Acknowledgements

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

References

1. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 2004;304:66-74.
2. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, et al. The Sorcerer II Global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007;5:e77.
3. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 2007;5:e16.
4. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. PLoS Biol. 2007;5:e75.
5. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IMA, Grechkin Y, Dubchak I, Anderson I, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acid Res. 2008;36:D534-D538.
6. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. The Metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 2008;9:386.
7. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 2008;26:541-547.
8. Field D, Morrison N, Glöckner FO, Kottmann R, Cochrane G, Vaughan R, Garrity G, Cole J, Hirschman L, Schriml L, et al. Working together to put molecules on the map. Nature 2008;453:978.
9. Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T, Field D, Glöckner FO, Genomic Standards Consortium. A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS 2008;12:115-121.

10. Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glöckner FO. Megx.net–database resource for marine ecological genomics. Nucleic Acid Res. 2006;34:D390-D393.
11. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, Glöckner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acid Res. 2007;35:7188-7196.
12. Brussaard CPD, Wilhelm SW, Thingstad F, Weinbauer MG, Bratbak G, Heldal M, Kimmance SA, Middelboe M, Nagasaki K, Paul JH, et al. Global-scale processes with a nanoscale drive: the role of marine viruses. ISME J. 2008;2:575-578.
13. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acid Res. 2008;36:D475-D479.
14. Hirschman L, Clark C, Cohen KB, Mardis S, Luciano J, Kottmann R, Cole J, Markowitz V, Kyrpides N, Morrison N, et al. Habitat-Lite: a GSC case study based on free text terms for environmental metadata. OMICS 2008;12:129-136.
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990;215:403-410.
16. Richter M, Lombardot T, Kostadinov I, Kottmann R, Duhaime MB, Peplies J, Glöckner FO. JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta) genomes. BMC Bioinformatics 2008;9:177.
17. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, et al. GenDB–an open source genome annotation system for prokaryote genomes. Nucleic Acid Res. 2003;31:2187-2195.
18. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol. 2005;3:790-806.
19. Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, Andrews-Pfannkoch C, Fadrosh D, Miller CS, Sutton G, et al. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 2008;3:e1456.

20. Metagenomics versus Moore’s law. Nat. Methods 2009;6:623.

Usage of Curated Datasets in Microbial Ecology

VII. Analysis of 23S rRNA genes in metagenomes – A case study from the Global Ocean Sampling Expedition

Authors: Pelin Yilmaz, Renzo Kottmann, Elmar Pruesse, Christian Quast and Frank Oliver Glöckner

Published in: Systematic and Applied Microbiology. 2011; 38 (6): 462-469

Contribution: designed and performed the research, analyzed the data and wrote the manuscript

Analysis of 23S rRNA genes in metagenomes – A case study from the Global Ocean Sampling Expedition

Pelin Yilmaz[a, b], Renzo Kottmann[a], Elmar Pruesse[a, b], Christian Quast[a], Frank Oliver Glöckner[a, b, *]

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

b Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany

* Corresponding author at: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany.

Tel.: +49 0 421 2028 970; fax: +49 0 421 2028 580. E-mail address: fog@mpi-bremen.de (F.O. Glöckner)

ABSTRACT

As an evolutionary marker, 23S ribosomal RNA (rRNA) offers more diagnostic sequence stretches and greater sequence variation than 16S rRNA. However, 23S rRNA is still not as widely used. Based on 80 metagenome samples from the Global Ocean Sampling (GOS) Expedition, the usefulness and taxonomic resolution of 23S rRNA were compared to those of 16S rRNA. Since 23S rRNA is approximately twice as large as 16S rRNA, twice as many 23S rRNA gene fragments were retrieved from the GOS reads than 16S rRNA gene fragments, with 23S rRNA gene fragments being generally about 100 bp longer. Datasets for 16S and 23S rRNA sequences revealed similar relative abundances for major marine bacterial and archaeal taxa. However, 16S rRNA sequences had a better taxonomic resolution due to their significantly larger reference database. Reevaluation of the specificity of previously published PCR amplification primers and group specific fluorescence in situ hybridization probes on this metagenomic set of non-amplified 23S rRNA sequences revealed that out of 16 primers investigated, only two had more than 90% target group coverage. Evaluations of two probes, BET42a and GAM42a, were in accordance with previous evaluations, with a discrepancy in the target group coverage of the GAM42a probe when evaluated against the GOS metagenomic dataset.

Introduction

Metagenomics, the study of community genomes taken directly from the environment, allows the cultivation-independent access to the diversity and functional information of microbial communities in their natural habitats [12]. For marine habitats, at least 51 metagenome studies are currently available [18]. One of the largest and geographically most comprehensive is the Global Ocean Sampling (GOS) Expedition. The initial dataset

consisted of 6.3 billion bp of Sanger sequence reads obtained from 41 surface water samples. These 41 samples covered a region from the North Atlantic to the South Pacific [32]. Furthermore, the publicly available GOS dataset has recently been augmented by samples from the Atlantic, Pacific and Indian Oceans [44].

The taxonomic diversity of the GOS metagenomic dataset has been assessed previously based on 16S ribosomal RNA (rRNA) gene fragments [4,32]. The distribution of 23S rRNA gene sequences in the GOS and other metagenomes remains unexplored. Although the 16S rRNA gene has been established as the standard molecule for analyzing the taxonomic diversity in metagenomes [36,41], 23S rRNA offers advantages over 16S rRNA. With an average length of 2900 bases, it is almost twice as long as the 16S rRNA and, therefore, is theoretically a more informative phylogenetic marker than the 16S rRNA gene [19,20,22]. The 23S and 16S rRNA molecules share the same properties in terms of molecule-ubiquity, as well as sequence and structure conservation. Furthermore, phylogenetic trees based on 16S rRNA and on 23S rRNA genes have comparable topologies [21,31].

A disadvantage of the 23S rRNA gene is the relatively low number of sequences available in the public databases as compared to 16S rRNA genes. Currently (March 2011), only 231,356 23S/28S sequences are publicly available, compared to 1,962,952 16S/18S sequences [29]. Furthermore, the low number of 23S/28S rRNA sequences (20,959) longer than 1900 bases (full-length) limits the assessment of taxonomic diversity due to reduced resolution in taxonomic assignments. The lower number of available 23S rRNA gene sequences can historically be explained by the technical difficulty and higher cost of sequencing the larger molecule with Sanger sequencing technology. However, with new technologies and constantly decreasing sequencing costs, these difficulties are becoming less.

This study is a systematic analysis of 23S rRNA gene sequences in unassembled reads of 80 GOS samples, with the focus on the quantity of retrieved fragments, the fragment length distribution, and the high level taxonomic classification of the fragments.

In order to evaluate and validate the classification results obtained using 23S rRNA sequences, a comparison of the bacterial and archaeal diversity of the GOS sites was

undertaken based on 23S rRNA and 16S rRNA gene classifications. Additionally, previously reported 23S rRNA primers and probes have been evaluated based on the extended dataset.

Materials and methods

Retrieval, alignment and taxonomic classification of 23S/28S and 16S/18S rRNA fragments

Unassembled metagenomic reads for 80 GOS sample datasets were downloaded as a FASTA file from the CAMERA website [34] in September, 2009. A total of 10,085,737 reads, with an average read length of 822 bp, were processed with the SILVA pipeline [30] in order to retrieve 23S/28S and 16S/18S rRNA gene fragments. Aligned fragments were imported into the ARB software suite for further analysis [23]. The fragments were added to the guide trees of the large subunit (LSU (23S/28S)) and small subunit (SSU (16S/18S)) datasets of the SILVA Reference (Ref) release 102 using the ARB Parsimony tool. Fragments with 300–600 aligned bases within the 23S/28S rRNA gene boundaries, and 100–500 aligned bases within the 16S/18S rRNA gene boundaries were added to the guide tree using positional variability filters (an all domain filter for 23S/28S; individual *Bacteria*, *Archaea* and *Eukarya* filters for 16S/18S) excluding highly variable positions indicated by numbers between 1 and 7, which resulted in 2903 out of 3546 valid positions for 23S/28S rRNA sequences, and 1391 out of 1444 positions for 16S/18S rRNA sequences. Fragments with more than 600 aligned bases for 23S/28S rRNA, and 500 aligned bases for 16S/18S rRNA sequences were added with the same positional variability filters but excluding highly variable positions between 1 and 9, leaving 2345 and 1224 valid positions for 23S/28S and 16S/18S rRNA sequences, respectively. Taxonomic assignments are based on membership of the fragments to the existing clades of the SILVA taxonomy, as represented by the guide trees of the high quality SILVA Ref datasets [30]. Taxonomic path assignments were stored in the “tax slv” field of ARB files using the taxonomy(n) function of ARB Command Interpreter (ACI).

A “Best-BLASTN (Nucleotide BLAST) hit” approach of 23S rRNA fragments was also

performed for comparison with the ARB-parsimony approach [1]. Unaligned 23S/28S rRNA fragments retrieved by the SILVA pipeline were used to query the reference dataset of SILVA LSU release 102, using the Tera-BLASTN algorithm (Tera-BLASTTM, TimeLogic Inc., Carlsbad, CA, USA). The parameters used for the BLASTN algorithm were as follows: word size = 11, extension threshold = 20, nucleic match = 1, nucleic mismatch =−3, gap open penalty =−5, gap extension penalty =−2. Best-BLASTN hits were selected as the top-scoring hit from a group of hits having an expect value of less than 0.00001, and an identity to the query of more than or equal to 97%. The taxonomy of the best hit in the reference dataset was assigned to the query sequence. Further processing of data for taxa abundance counts and method comparisons was performed using MegDB [27].

Primer and probe matching

Sequence Associated Information (SAI) filters corresponding to binding sites of the primers and probes (Supplementary material 1) were manually constructed. These filters were used to count the number of bases within the primer/probe binding sites of all 23S rRNA sequences found in the GOS and SILVA LSU release 102. The target group sequences were chosen from all sequences having a full-length primer/probe-binding region according to these counts.

Table 1.

Percentage of 23S and 16S rRNA gene fragments that can be classified up to Domain, Phylum, Class, Order, Family and Genus levels. Total number of fragments classified are 20,036 and 12,491 for 23S and 16S rRNA, respectively, excluding *Eukarya* and fragments with less than 300 aligned bases for LSU and less than 100 aligned bases for SSU.

	23S rRNA gene fragments (%)	16S rRNA gene fragments (%)
Domain	99.9	100.0
Phylum	96.6	100.0
Class	94.4	99.1
Order	78.8	96.3
Family	35.4	80.0
Genus	16.6	31.2

The sizes of primer/probe target groups for GOS and LSU Parc, as well as sequences of the primers and probes are given inTable 2 and Supplementary material 1, respectively. Primer/probe matching was carried out manually using the PROBE MATCH module of the ARB software package with the “zero mismatches” and “no weighted mismatches” criteria. Results were parsed and the group coverage in each target group was calculated as the relative number of probe and primer hits to the total number of sequences in the respective target group.

Data access

23S/28S rRNA sequences retrieved from the GOS metagenomes that were analyzed in this study are publicly available from http://www.arbsilva.de/download/archive/GOS diversity/ in ARB format, as well as unaligned and aligned FASTA files. The ARB file contains fields created for the purpose of the primer/probe matching procedure; specifically, fields named with the primer or probe name (example, 129f) contain the PROBE MATCH results as ‘pos’ if the results reported were positive. The fields carrying the primer name and the suffix ‘ len’ (example, 129f len) contain the length of the primer/probe binding regions.

Results and discussion

Summary of rRNA gene fragment retrieval A total of 29,581 23S/28S rRNA (0.3% of total reads), and 142,783 16S/18S rRNA (1.4% of total reads) gene fragments were retrieved and aligned using SINA. Fragments with less than 100 aligned bases within the 23S/28S or 16S/18S rRNA gene boundaries were excluded from further analysis, which reduced the dataset to 22,575 23S/28S (76% of total 23S/28S) and 12,742 16S/18S (9% of total 16S/18S) rRNA fragments. For the majority of the excluded sequences (>98%) less than 50 bases could be aligned. Excluding these sequences from the analysis increased the reliability of taxonomic assignments, since sequences this short do not carry sufficient phylogenetic information. Ten GOS sample datasets (GS038–GS046, and GS050) had less than five rRNA gene fragments of sufficient length (Fig. 1A and B) and were excluded from further analysis. These sites contained, on average, only 700 total reads, explaining the low fragment retrieval. Furthermore, no rRNA fragments were

retrieved from the MOVE858 sample, which was obtained using 0.002–0.22 μm filters, representing the viral metagenome fraction.

The 23S/28S rRNA gene is twice the length of the 16S/18S rRNA gene, hence the probability of retrieving a 23S/28S rRNA gene fragment should be proportionately higher. This expectation was supported by the results of this study, since ratios of almost 2:1 were observed at sites GS000d (904 23S/28S vs. 438 16S/18S), GS029 (351 23S/28S vs. 162 16S/18S), or GS112a (227 23S/28S vs. 113 16S/18S) (Fig. 1A).

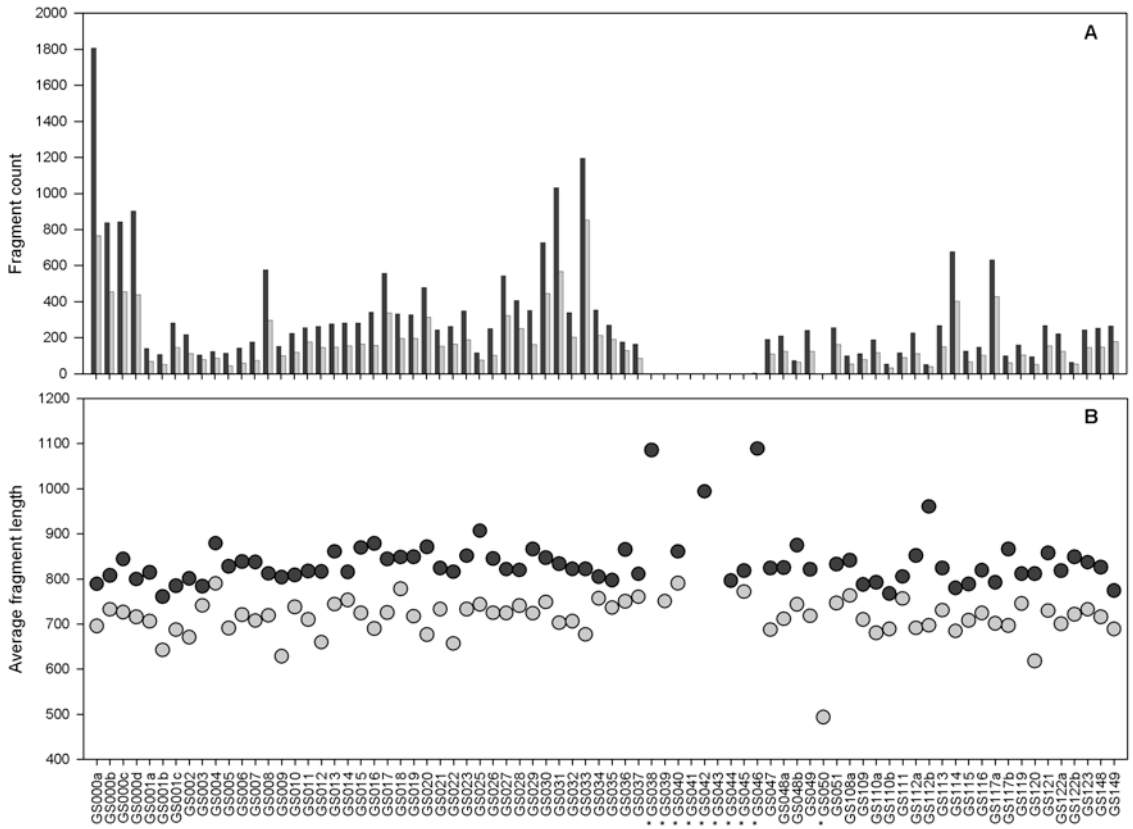


Fig. 1. (A) Comparison of number of 23S/28S (dark grey bars) and 16S/18S (light grey bars) rRNA fragments retrieved from each GOS sample dataset. (B) Average length of 23S/28S (dark grey circles) and 16S/18S (light grey circles) rRNA fragments from each GOS sample dataset in terms of number of aligned bases within rRNA gene boundaries, excluding any fragment (23S/28S or 16S/18S) that contained less than 100 aligned bases. Sites marked with a '*' indicate that less than five fragments were retrieved.

This two-fold difference was also reflected by the average number of fragments retrieved per site, which was 301 for 23S/28S rRNA and 177 for 16S/18S rRNA. Furthermore,

23S/28S rRNA gene fragments were considerably longer than 16S/18S gene fragments (Fig. 1B). Where an average 23S/28S rRNA fragment had 836 aligned bases within the rRNA gene boundaries, a 16S/18S rRNA fragment had 713 aligned bases. More abundant and larger rRNA gene fragments may provide additional information in assessing taxonomic diversity, both with phylogeny and operational taxonomic unit based methods, as well as increasing the chances to affiliate other gene fragments with specific lineages. Both 23S/28S and 16S/18S rRNA fragments were randomly distributed over the rRNA gene regions, meaning that no specific sequence region was over-or under-represented (Supplementary material 3).

Taxonomic diversity based on 23S and 16S rRNA genes

Few eukaryotic sequences (340 28S rRNA and 251 18S rRNA) were retrieved from samples obtained from 0.22 to 0.8 μm, 0.8 to 3 μm and 3 to 20 μm size fractions. These were excluded from further analyses due to the inconsistent taxonomic classification of eukaryotic sequences in databases and to allow greater focus on the bacterial and archaeal fraction. As a result, a total of 20,036 23S rRNA (>300 bases) and 12,491 16S rRNA (>100 bases) gene sequences were classified. Percentages of both 23S and 16S rRNA fragments associated with major marine bacterial and archaeal taxa showed good agreement with each other and with previous studies [8,9,28] (Fig. 2A and B). Specifically, based on 23S rRNA assignments, 43% of the retrieved rRNA fragments were found to be associated with *Alphaproteobacteria*, followed by 17% *Gammaproteobacteria*, 9% *Actinobacteria*, 8% *Cyanobacteria*, 8% *Bacteroidetes*, 3% *Betaproteobacteria*, 2% *Euryarchaeota*, and 0.4% *Crenarchaeota* (Fig. 2A). However, less agreement in the assignment of 23S rRNA and 16S rRNA fragments was observed with less abundant marine taxa. For example, *Chloroflexi* and *Deferribacteres* associated fragments were not observed in the 23S rRNA gene-based classification, which may be ascribed to the lack of annotated clades for these taxa. In such cases, 16S rRNA gene-based classifications appear to provide better estimations.

Similar trends were observed in sample-by-sample distribution of taxa at the “Class” level for both 23S and 16S rRNA-based assignments, as compared to the previous overall assessment (Fig. 3A and B, Supplementary material 2). *Alphaproteobacteria*, followed by

Gammaproteobacteria, *Actinobacteria*, *Cyanobacteria*, *Flavobacteria* and *Betaproteobacteria* were the most abundant taxa in the majority of sample datasets. However, differences were observed in the occurrence or relative abundance of minor groups, such as *Planctomycetacia* or *Aquificae*. For example, *Planctomycetacia* associated 16S rRNA fragments were found in 15 sample datasets, whereas only 13 sample datasets contained *Planctomycetacia* associated 23S rRNA fragments.

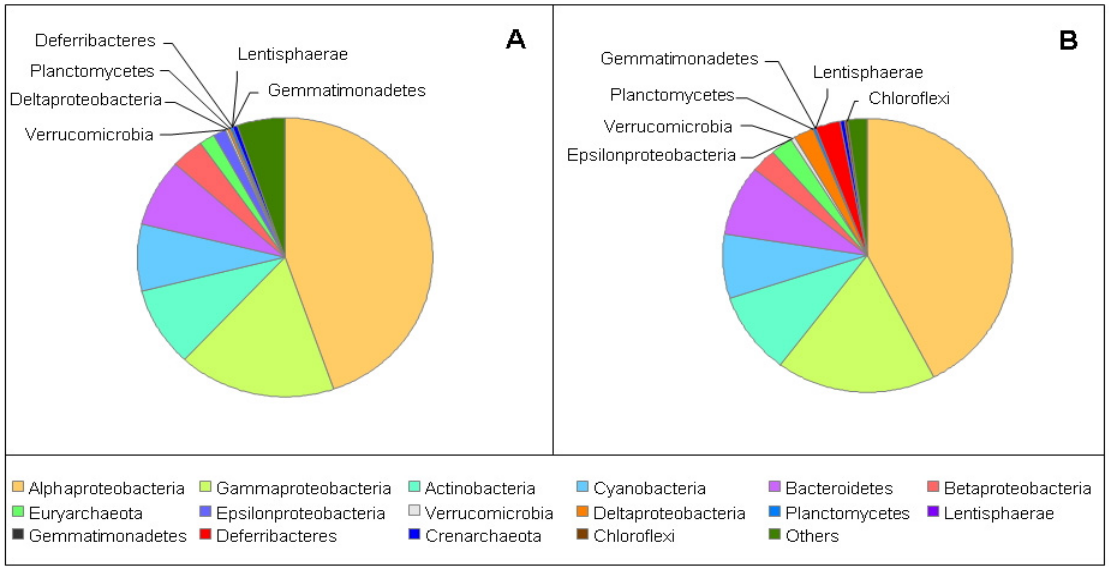


Fig. 2. Percentage of 23S (A) and 16S (B) rRNA fragments associated with major marine bacterial and archaeal taxa among all GOS sample datasets, except GS038–GS046 and GS050. Percentages were calculated based on absolute numbers of fragments associated with a given taxa.

The differences in relative abundance observed with 16S or 23S rRNA-based assignments in these sample datasets were up to six-fold (GS000a). Surprisingly, in certain cases, 23S rRNA-based assessments predicted higher relative abundances or occurrence in sample datasets for other taxa. Up to 12-fold more Epsilonproteobacteria associated 23S rRNA fragments were found in sample dataset GS000b compared to 16S rRNA fragments. Additionally, Lentisphaeria, which appeared to be present in ten sites according to 23S rRNA classifications, were observed only at two sites according to 16S rRNA gene classifications.

The former case, where 16S rRNA-based assignments estimated more taxa in more sample datasets, demonstrated the current drawback of 23S rRNA-based classification (i.e. its lack of resolution due to insufficient full-length reference sequences). On the other hand, the latter observations demonstrated that when reference sequences are present for a taxon, the higher number of 23S rRNA fragments retrieved can capture what is missed with 16S rRNA fragments.

An evaluation of the suitability of 23S rRNA-based diversity assessments can be obtained by comparing the community composition of contrasting habitats. Subtle differences in contrasting marine habitats are evident and comparable to each other and to general expectations for both 23S and 16S rRNA-based diversity assessments (Fig. 3A and B). For example, Gammaproteobacteria were less frequent in estuarine and freshwater habitats compared with coastal and open ocean habitats (GS000a vs. GS020). On the contrary, Actinobacteria and Betaproteobacteria were more abundant in estuarine and freshwater habitats than in coastal or open ocean habitats (GS011 vs. GS119), underlining previously reported trends [6,10,13]. Additionally, a distinct composition was evident in non-open ocean GOS habitats (GS033-hypersaline, GS030-mangrove).

Investigating relative abundances at lower taxonomic levels can shed light on more prominent habitat-specific diversity patterns. However, with the current size and content of LSU rRNA reference databases, the 23S rRNA has a distinct disadvantage in achieving this. As summarized in Table 1, the percentage of 23S rRNA gene fragments that can be classified to a certain taxa is comparable to that of the 16S rRNA gene-based classification at Domain, Phylum or Class levels. A decrease in percentage of classified 23S rRNA fragments was observed at lower levels, from 95% at the Class level, down to even 17% at the Genus level. This can be explained by the 23,197 sequences of taxonomically classified cultured organisms in the SILVA Ref release 102 SSU dataset vs. only 3602 sequences in the LSU Ref dataset.

In addition to the comparison of tree guided taxonomic classification methods, a comparison of the parsimony classification approach to a Best-BLAST hit approach was performed for 23S/28S rRNA gene fragments. BLAST, or modifications of this method, is increasingly popular in assessing the taxonomic diversity of highthroughput

metagenomic datasets and rRNA surveys. This is due to BLAST being faster than phylogenetic methods, such as ARB Parsimony, and it also provides a means of a multiple-alignment free taxonomic classification approach [15,16,25,35].

A total of 15,798 (excluding 86 Eukaryotic sequences) out of 29,581 unaligned GOS 23S/28S rRNA fragments could be classified using the Best-BLASTN hit approach. Sequences below 300 nucleotides were rejected, revealing a total of 14,656 classified sequences. The BLASTN approach was successful in classifying 5380 sequences, which were not classified by ARB Parsimony. However, the identity to the target sequence was below the chosen thresholds. The differences between the two methods could be settled by a sufficiently high bit score for the Best-BLASTN hit approach as the sole criterion for assigning taxonomy [5].

In the next step, the taxonomic assignments between Best-BLASTN hit and ARB parsimony were investigated. In summary, 97% of the 14,656 common sequences were assigned identical taxonomy by both methods. The remaining 3.4% (499) of the sequences, which had different taxonomic paths, fell into three different cases: (1) the taxonomic path assigned by the Best-BLASTN hit was at a lower rank compared to ARB Parsimony, (2) the taxonomic path assigned by ARB Parsimony was at a lower rank compared to the Best-BLASTN hit, and finally, (3) the assigned taxonomies were entirely different below a certain rank. For the majority of the sequences (408), the Best-BLASTN hit provided classification at a lower rank (case 1). This is an expected outcome because the taxonomic path is assigned directly from the next relative of the target sequence by the Best-BLASTN hit approach. On the contrary, in the ARB Parsimony approach the taxonomy is assigned based on a group membership, and a sequence can be placed close to, but outside, a group. At a lesser amount, with 28 sequences, a classification to a lower rank was achieved with the ARB Parsimony approach. Finally, 63 sequences had different taxonomic assignments, which could be broken down into 36 sequences assigned to different genera, 15 to different orders, and 12 to different classes. The relatively small differences in taxonomic assignments between the two methods were encouraging, especially regarding concerns about the suitability of large multiple alignments for taxonomic classification. In response to these concerns, it is important to point out that the SILVA alignment has been rated as having the ‘best-quality’ within

similar projects [33]. Furthermore, the SILVA alignment is based on a reference seed alignment, hence it is not subjected to the many drawbacks of large-scale multiple de novo alignments.

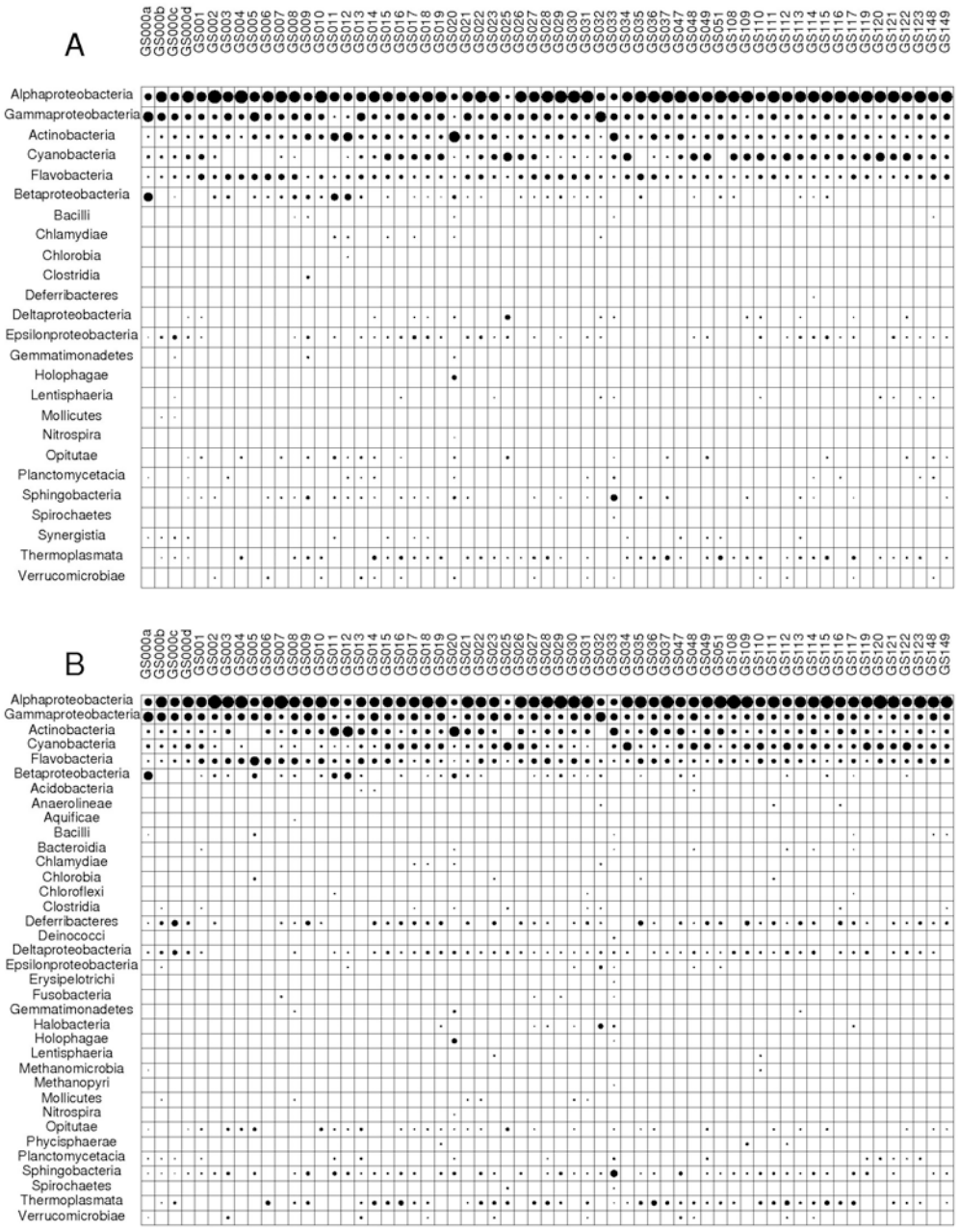


Fig. 3. The relative abundance of 23S (A) and 16S (B) rRNA fragments associated with different taxa (rows) at each GOS sample dataset (columns). Presence of a spot indicates the presence of fragments

associated with a given taxa, and the area of a spot represents the relative abundance. Relative abundances are based on absolute counts of all fragments from a given site associated with a certain taxa, which are then normalized according to the total fragment counts from that site. Abundances are not normalized with respect to single copy genes, and since rRNA operons can occur multiple times in a genome, the numbers do not represent cell abundances. The taxa shown here are on the 'Class' level, except *Cyanobacteria*, which is at the 'Phylum' level.

Finally, with this comparison, it was shown that both Best-BLAST hit and phylogenetic approaches, such as ARB Parsimony, can provide comparable and very similar results. This methodological comparison showed that if a congruent dataset for taxonomic classification is used, very similar results are obtained, regardless of the algorithms behind the taxonomic classifications.

Specificity of common 23S rRNA primers and probes

The addition of GOS 23S rRNA sequences increases the size of the current 23S/28S rRNA databases (based on SILVA 102 LSUParc) by 12%. Furthermore, they have not undergone PCR amplification, and hence provide a unique opportunity for testing the coverage of previously described universal amplification primers, as well as widely used class-specific probes.

The most recently developed primer sets (129f, 189f, 457r, 2490r) [14], as well as primer 2241r [17], showed reasonable group coverage in the GOS 23S dataset sequences with an average of 85% (Table 2), and the results were comparable to those obtained from matching the primers against the SILVA release 102 LSU Parc dataset with only a ±2% difference. The reference dataset used by Hunt et al. (2006) [14] was smaller with 2176 sequences than both the LSU Parc (average of 11,000 target group sequences) and the GOS 23S (average of 5400 target group sequences) datasets used in this study. However, the authors have included environmental shotgun sequences from the Sargasso Sea pilot study [39] in their dataset, which would account for the comprehensiveness of these primers also

Contrary to these results, the primers developed for the amplification of variable regions of bacterial 23S rRNA sequences (11a–97ar) [38] showed very poor group coverage in the GOS 23S dataset sequences, with generally less than 50% coverage of the target

group. A 90% group coverage was only observed for 69ar (Table 2). Although the primer binding sites were highly conserved, this was obviously counteracted by the very small dataset that these primers were based on [11]. Surprisingly, primers 53a to 97ar were observed to have higher group coverage within the GOS 23S rRNA sequences than within LSU Parc.

The two archaeal primers (LSU190-F and LSU2445a-R) [7] showed very low group coverage in the GOS 23S dataset (Table 2), with 14% and 5%, respectively. Nevertheless, while the percentages were higher in the LSU Parc, they did not exceed 50%.

Table 2.

Specificities of selected primers and probes, evaluated on the 23S/28S rRNA gene fragments retrieved from the GOS metagenomes having more than 300 aligned bases within the rRNA gene boundaries, and on the SILVA Parc release 102 LSU dataset. Outgroup hits are the sum of both *Archaea* and *Eukarya* in case of bacterial primers, both *Bacteria* and *Eukarya* in case of archaeal primers, only *Eukarya* in case of bacterial and archaeal primers, and non-*Betaproteobacteria* and non-*Gammaproteobacteria* for BET42a and GAM42A probes.

		GOS 23S/28S			LSU Parc		
Primer/probe	Target group	Size of target group	Group coverage (%)	Outgroup hits	Size of target group	Group coverage (%)	Outgroup hits
129f [14]	*Bacteria*	4853	74%	0	10640	82%	4
189f [14]	*Bacteria*	5285	87%	0	11508	87%	0
457r [14]	*Bacteria*	5551	86%	4	11177	83%	279
2241r [17]	*Bacteria*	5832	84%	10	11457	86%	3967
2490r [14]	*Bacteria*	5734	94%	0	10821	98%	0
11a [38]	*Bacteria*	5256	20%	0	11478	39%	0
23ar [38]	*Bacteria*	5619	23%	0	10526	49%	4
43a [38]	*Bacteria*	5633	6%	0	10999	44%	0
53a [38]	*Bacteria*	5320	3%	0	10594	1%	0
62ar [38]	*Bacteria*	5540	8%	0	11455	5%	0
69ar [38]	*Bacteria*	5731	90%	0	11443	87%	0
93a [38]	*Bacteria*	5737	62%	0	10322	55%	0
93ar [38]	*Bacteria*	5731	63%	0	10327	56%	2
97ar [38]	*Bacteria*	4969	55%	0	9165	29%	38
LSU190-F [7]	*Bacteria & Archaea*	5348	14%	0	11741	24% 28%	0
LSU2445a-R [7]	*Archaea*	142	5%	0	262	28%	0
BET42a [24]	*Betaproteobacteria*	209	79%	63	570	87%	348
GAM42a [24]	*Gammaproteobacteria*	980	42%	1	2877	78%	10

For the BET42a probe [24], 79% group coverage was found. This, as well as the number of outgroup hits within the GOS 23S dataset, was close to that reported by a previous evaluation [2] (Table 2). Group coverage within LSU Parc (87%) was in accordance with Amann and Fuchs [2] (Table 2), although considerably more outgroup hits, 348 in LSU Parc vs. 62, were observed.

The GAM42a probe coverage in the GOS 23S dataset (Table 2) was almost half (42%) of the value reported previously (76%) [2], and the corresponding evaluation of the LSU Parc (78%) dataset. Since the mismatches could result from sequencing errors, the alignments of sequences with mismatches to the probe GAM42a were manually inspected. A few cases were likely to be sequencing errors, and were mainly observed in fragments obtained from ends of sequencing reads. The majority of the mismatches revealed consistent, class-specific mismatches. These mismatches were up to four bases, and were found mainly between Escherichia coli positions 1030–1040. Although this evaluation of the GAM42a probe was based on a single environment, the surface ocean, limitations and anomalous results with the GAM42a probe have been reported previously for other environments as well, which were found to be mainly due to polymorphisms at E. coli position 1033 [3,43]. Our observation confirms these reports, by adding additional polymorphisms before and after this position. Consequently, the limitations of the GAM42a probe might be more severe than previously thought, and therefore we recommend the design and testing of novel Gammaproteobacteria probes.

Conclusions

This study exemplifies the possibility and power of using 23S rRNA genes for biodiversity surveys by providing a comparative overview of 16S and 23S rRNA fragments retrieved from the GOS metagenomes. High quality taxonomic classification for biodiversity analysis, as well as primer and probe design, depends on the size and extent of the reference dataset used. The advantage of using the larger 23S rRNA genes for biodiversity analysis, especially for the marine system, has been shown previously [27]. Additionally, a recent study assessing the diversity of paralogous 23S rRNA genes has shown that significant sequence diversification was observed in 184 species, further

supporting the suitability of this molecule for taxonomy [26]. Although an obvious limitation faced during this study was the small size of the 23S rRNA gene reference datasets, this is likely to be overcome in the near future with the contribution of (meta-) genomic sequences from mega-sequencing projects, such as the Human Microbiome Project [37], the TerraGenome [40], Tara Oceans (see http://oceans.taraexpeditions.org/) or the Genomic Encyclopedia of Bacteria and Archaea [42]. Moreover, studies assessing the characteristics and sequence diversity of 23S rRNA genes in bacterial and archaeal genomes, in combination with efforts to design, test and re-evaluate universal and group specific primers and probes [14], can renew the interest and utilization of this molecule. Application of continually advancing, cheaper sequencing technologies to the undiscovered fraction of the 23S rRNA gene sequences can result in a higher appreciation of this valuable phylogenetic marker.

Acknowledgements

We would like to thank Mar Fernández Méndez and Petra Pjevac for their assistance in phylogenetic and taxonomic analysis of the dataset. We would also like to thank Pier Luigi Buttigieg, Jörg Peplies and Hannah Marchant for their critical reading of the manuscript and helpful suggestions. This study was supported by the Max Planck Society.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.syapm.2011.04.005.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* **215** (1990), pp. 403-410.

[2] R. Amann, B. M. Fuchs, Single-cell identification in microbial communities by

improved fluorescence in situ hybridization techniques, *Nat. Rev. Microbiol.* **6** (2008), pp. 339-348.

[3] J. J. Barr, L. L. Blackall, P. Bond, Further limitations of phylogenetic group-specific probes used for detection of bacteria in environmental samples, *ISME. J.* **4** (2010), pp. 959-961.

[4] E. J. Biers, S. L. Sun, E. C. Howard, Prokaryotic genomes and diversity in surface ocean waters: interrogating the Global Ocean Sampling metagenome, *Appl. Environ. Microbiol.* **75** (2009), pp. 2221-2229.

[5] M. J. Claesson, O. O'Sullivan, Q. Wang, J. Nikkilä, J. R. Marchesi, H. Smidt, W. M. de Vos, R. P. Ross, P. W. O'Toole, Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine, *PLoS ONE* **4** (2009), pp. e6669.

[6] B. C. Crump, C. S. Hopkinson, M. L. Sogin, J. E. Hobbie, Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time, *Appl. Environ. Microbiol.* **70** (2004), pp. 1494-1505.

[7] E. DeLong, L. Taylor, T. Marsh, C. Preston, Visualization and enumeration of marine planktonic *Archaea* and *Bacteria* by using polyribonucleotide probes and fluorescent *in situ* hybridization, *Appl. Environ. Microbiol.* **65** (1999), pp. 5554-5563.

[8] J. Fuhrman, Å. Hagström. (2008) Bacterial and archaeal community structure and its patterns. In: Kirchman, D. L. (Eds.) Microbial ecology of the oceans. 2, Wiley-Blackwell, New York, pp. 45-90.

[9] S. J. Giovannoni, U. Stingl, Molecular diversity and ecology of microbial plankton, *Nature* **437** (2005), pp. 343-348.

[10] F. O. Glöckner, E. Zaichikov, N. Belkova, L. Denissova, J. Pernthaler, A. Pernthaler, R. Amann, Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of *Actinobacteria, Appl. Environ. Microbiol.* **66** (2000), pp. 5053-5065.

[11] R. R. Gutell, M. N. Schnare, M. W. Gray, A compilation of large subunit (23Sand 23S-like) ribosomal RNA structures, *Nucleic Acids Res.* **20** (1992), pp. 2095-2109.

[12] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, R. M. Goodman, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, *Chem. Biol.* **5** (1998), pp. R245-R249.

[13] I. Hewson, J. A. Fuhrman, Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia, *Appl. Environ. Microbiol.* **70** (2004), pp. 3425-3433.

[14] D. E. Hunt, V. Klepac-Ceraj, S. G. Acinas, C. Gautier, S. Bertilsson, M. F. Polz, Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity, *Appl. Environ. Microbiol.* **72** (2006), pp. 2221-2225.

[15] S. M. Huse, L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, M. L. Sogin, Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing, *PLoS Genet.* **4** (2008), pp. e1000255.

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

[17] D. J. Lane. (1991) 16S/23S rRNA sequencing. In: Stackebrandt, E. and Goodfellow, M. (Eds.) Nucleic acid techniques in bacterial systematics. J. Wiley & Sons, Chichester; New York, pp. 115-175.

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

[19] W. Ludwig, H. P. Klenk. (2001) A phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone, D. R. and Castenholz, R. W. (Eds.) The *Archaea* and the deeply branching and phototrophic *Bacteria*. Springer-Verlag, New York, pp. 49-65.

[20] W. Ludwig, R. Rossello-Mora, R. Aznar, S. Klugbauer, S. Spring, K. Reetz, C.

Beimfohr, E. Brockmann, G. Kirchhof, S. Dorn, M. Bachleitner, N. Klugbauer, N. Springer, D. Lane, R. Nietupsky, M. Weizenegger, K.-H. Schleifer, Comparative sequence analysis of 23S rRNA from *Proteobacteria, Syst. Appl. Microbiol.* **18** (1995), pp. 164-188.

[21] W. Ludwig, K. Schleifer, Phylogeny of *Bacteria* beyond the 16S rRNA standard, *ASM News* **65** (1999), pp. 752-757.

[22] W. Ludwig, K. H. Schleifer, Bacterial phylogeny based on 16S and 23S rRNA sequence analysis, *Fems Microbiol. Rev.* **15** (1994), pp. 155-173.

[23] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, K.-H. Schleifer, ARB: a software environment for sequence data, *Nucleic Acids Res.* **32** (2004), pp. 1363-1371.

[24] W. Manz, R. Amann, W. Ludwig, M. Wagner, K.-H. Schleifer, Phylogenetic oligodeoxynucleotide probes for the major subclasses of *Proteobacteria*: Problems and solutions, *Syst. Appl. Microbiol.* **15** (1992), pp. 593-600.

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

[26] A. Pei, C. W. Nossa, P. Chokshi, M. J. Blaser, L. Yang, D. M. Rosmarin, Z. Pei, Diversity of 23S rRNA genes within individual prokaryotic genomes, *PLoS ONE* **4** (2009), pp. e5437.

[27] J. Peplies, F. O. Glöckner, R. Amann, W. Ludwig, Comparative sequence analysis and oligonucleotide probe design based on 23S rRNA genes of *Alphaproteobacteria* from North Sea bacterioplankton, *Syst. Appl. Microbiol.* **27** (2004), pp. 573-580.

[28] T. Pommier, B. Canbäck, L. Riemann, K. H. Boström, K. Simu, P. Lundberg, A. Tunlid, Å. Hagström, Global patterns of diversity and community structure in marine bacterioplankton, *Mol. Ecol.* **16** (2007), pp. 867-880.

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

[30] E. Pruesse, C. Quast, P. Yilmaz, W. Ludwig, J. Peplies, F. O. Glöckner. (2011) SILVA: comprehensive databases for quality checked and aligned ribosomal RNA sequence data compatible with ARB. In: de Bruijn, F. J. (Eds.) Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches. John Wiley & Sons, Incorporated, pp.

[31] P. Rijk, Y. Peer, I. Broeck, R. Wachter, Evolution according to large tribosomal subunit RNA, *J. Mol. Evol.* **41** (1995), pp. 366-375.

[32] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran,

H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, J. C. Venter, The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific, *PLoS Biol.* **5** (2007), pp. e77.

[33] P. D. Schloss, A High-Throughput DNA Sequence Aligner for Microbial Ecology Studies, *PLoS ONE* **4** (2009), pp. e8230.

[34] R. Seshadri, S. A. Kravitz, L. Smarr, P. Gilna, M. Frazier, CAMERA: a community resource for metagenomics, *PLoS Biol.* **5** (2007), pp. e75.

[35] M. L. Sogin, H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, G. J. Herndl, Microbial diversity in the deep sea and the

underexplored "rare biosphere", *Proc. Nat. Acad. Sci. USA* **103** (2006), pp. 12115-12120.

[36] S. G. Tringe, P. Hugenholtz, A renaissance for the pioneering 16S rRNA gene, *Curr. Opin. Microbiol.* (2008), pp.

[37] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, J. I. Gordon, The Human Microbiome Project, *Nature* **449** (2007), pp. 804-810.

[38] G. Van Camp, S. Chapelle, R. De Wachter, Amplification and sequencing of variable regions in bacterial 23S ribosomal RNA genes with conserved primer sequences, *Curr. Microbiol.* **27** (1993), pp. 147-151.

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

[40] T. M. Vogel, P. Simonet, J. K. Jansson, P. R. Hirsch, J. M. Tiedje, J. D. van Elsas, M. J. Bailey, R. Nalin, L. Philippot, TerraGenome: a consortium for the sequencing of a soil metagenome, *Nat. Rev. Microbiol.* **7** (2009), pp. 252-252.

[41] J. C. Wooley, A. Godzik, I. Friedberg, A primer on metagenomics, *PLoS Comput. Biol.* **6** (2010), pp. e1000667.

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

[43] C. Yeates, A. M. Saunders, G. R. Crocetti, L. L. Blackall, Limitations of the widely used GAM42a and BET42a probes targeting bacteria in the *Gammaproteobacteria* radiation, *Microbiology* **149** (2003), pp. 1239-1247.

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

VIII. Ecological structuring of bacterial and archaeal taxa in ocean surface waters

Authors: Pelin Yilmaz, Wolfgang Hankeln, Renzo Kottmann, Christian Quast and Frank Oliver Glöckner

Awaiting revised manuscript after receipt of peer reviews at FEMS Microbiology Ecology on 05-09-2011

Contribution: designed and performed the research, analyzed the data and wrote the manuscript

Ecological structuring of bacterial and archaeal taxa in ocean surface waters

Pelin Yilmaz [1,2], Wolfgang Hankeln[1,2], Renzo Kottmann[1], Christian Quast[1], Frank Oliver Glöckner [1,2*]

1 Max Planck Institute for Marine Microbiology, Celsiusstr. 1, 28359 Bremen, Germany

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

* Correspondence should be addressed to FOG (fog@mpi-bremen.de)

ABSTRACT

The Global Ocean Sampling (GOS) expedition is currently the largest and geographically most comprehensive metagenomic dataset, including samples from the Atlantic, Pacific and Indian Oceans. This study makes use of the wide range of environmental conditions and habitats encompassed within the GOS sampling sites in order to uncover the ecological structuring of bacterial and archaeal taxon ranks. Community structures, based on taxonomically classified 16S ribosomal RNA (rRNA) gene fragments at phylum, class, order, family, and genus rank levels were examined using multivariate statistical analysis and the results were inspected in the context of oceanographic environmental variables, and structured habitat classifications.

At all taxon rank levels, community structures of neritic, oceanic, estuarine biomes, as well as other exotic biomes (salt marsh, lake, mangrove) were readily distinguishable from each other. A strong structuring of the communities with chlorophyll a concentration, and a weaker yet significant structuring with temperature and salinity was observed. Furthermore, there were significant correlations between community structures and habitat classification. These results can assist further probing of one-to-one relationships between taxa and environment, and can help to shed light on ecological preferences of both cultured and uncultured bacterial and archaeal clades.

Introduction

Ecological structuring of *Bacteria* and *Archaea* from a range of habitats, at genera or even species level, is nowadays routinely investigated (Lauber, *et al.*, 2009, Andersson, *et al.*, 2010, Kirchman, *et al.*, 2010). For the human nature it is tempting to characterize and categorize “objects”, and assign them to “containers” - big or small - that reflect particular characteristics of all these objects (Philippot, *et al.*, 2010). Still, there is controversial debate about bacterial and archaeal ecologically coherent containers, mainly due to the vast genetic and physiological diversity contained at high level ranks. However there is also striking evidence that correlations between taxonomy and functions exist.

For example, several studies were able to associate phyla or classes with r- or K- type life strategies; in marine systems members of SAR11 and *Bacteroidetes* were identified as K-strategists, and in soil systems *Betaprotebacteria* were found to be r-strategists (Alonso-Sáez, *et al.*, 2006, Fierer, *et al.*, 2007). Other studies demonstrated specific taxa-habitat associations, either via cross-or within-habitat comparisons (Glöckner, *et al.*, 1999, von Mering, *et al.*, 2007, Nemergut, *et al.*, 2010). Finally, investigations of responses of specific taxa to changing environmental conditions showed supportive results (Fuhrman, *et al.*, 2006, Pommier, *et al.*, 2007, Philippot, *et al.*, 2009).

The Global Ocean Sampling (GOS) (Rusch, *et al.*, 2007, Yooseph, *et al.*, 2010) provides a range of marine and aquatic habitats, enabling both inter- and intra-habitat comparisons. The sequences are associated with a relatively rich set of associated data (contextual or metadata), such as geographical coordinates and environmental variables, thus making the GOS expedition suitable for a metaanalysis. This study explores the possible ecological cohesions in high level taxa of surface ocean *Bacteria* and *Archaea*, using taxonomically classified 16S ribosomal RNA gene fragments (rRNA) from the GOS metagenomes. We evaluated the results in the framework of comparing high level taxa ranks to low level taxa, annotating sampling sites with ontological habitat classifications (Environment Ontology, www.environmentontology.org/) at three different granularity levels (biome, feature, material), and correlating community structures to qualitative and quantitative environmental variables.

Materials and methods

Retrieval and alignment of SSU rRNA fragments

Unassembled metagenomic reads for 80 GOS sample datasets were downloaded as a FASTA file from the CAMERA website (Seshadri, *et al.*, 2007) on September, 2009. A total of 10,085,737 reads, with an average read length of 822 bp, were processed with a custom-tailored configuration of the SILVA pipeline (Pruesse, *et al.*, 2007) in order to retrieve SSU rRNA gene fragments.

Firstly, a quality inspection was conducted. Reads composed of more than 2% of ambiguous bases, or more than 2% of homopolymeric stretches longer than four bases were rejected. Additionally, reads having more than 5% identity to vector sequences

based on BLASTN hits were excluded (Altschul, *et al.*, 1990). The database used for vector contamination checking was a combined vector sequence database based on the EMVEC and UniVec databases Secondly, the SILVA INcremental Aligner (SINA) was used to spot and align actual SSU rRNA gene fragment regions (Pruesse, *et al.*, 2011). Finally, the sequences were imported into ARB (Ludwig, *et al.*, 2004) for further analysis.

Taxonomic classification of fragments

Aligned SSU rRNA gene fragments were added to the guide tree included in the SSU dataset of the SILVA Reference (Ref) release 104 using the ARB Parsimony tool. Fragments having 100500 bases within the rRNA gene boundaries were added to the guide tree using individual *Bacteria*, *Archaea* and *Eukarya* filters, excluding highly variable positions between 1 and 7 leaving 1391 out of 1444 positions. For fragments with more than 500 aligned bases, sequences were added with the same positional variability filters but excluding highly variable positions between 1 and 9 leaving 1224 valid positions. Taxonomic assignments were based on membership of the fragments to the existing clades of the SILVA taxonomy. Manual refinement of the taxonomic groups was performed after addition of all GOS rRNA gene sequences. Taxonomic path assignments were stored in the 'tax_slv' field of ARB files using the taxonomy(n) function of ARB Command Interpreter.

Environmental parameters and habitat assignments

Temperature, salinity and chlorophyll a concentration values were taken from *in situ* measurements, where available. For dissolved oxygen, nitrate, phosphate, silicate concentrations, as well as in all cases where the three former parameters were missing, World Ocean Atlas 2005, World Ocean Database 2005, and SeaWifs Chlorophyll data were used for interpolation of these parameters at the geographic locations using the GIS tools of the megx.net portal (Kottmann, *et al.*, 2010) (Table S1). Habitat assignments for GOS sites were manually curated using an edit version of Environment Ontology (EnvO) terms (http://obo.cvs.sourceforge.net/viewvc/obo/obo/ontology/environmental/envo-edit.obo), at three different levels as biome, feature and material (Table S2).

Statistical analysis

Absolute abundances of taxa were standardized using the mean abundance of a set of

'single-copy domains' (SCDs), namely: b5, ef_ts, pnpase, rbfa, rrf, ribosomal_l12, ribosomal_l15, ribosomal_l16, ribosomal_l19, ribosomal_l20, ribosomal_l21p, ribosomal_l27, ribosomal_l29, ribosomal_l9_n, ribosomal_s16, ribosomal_s20p, ribosomal_s3_c, ribosomal_s3_n, srp_spb, smpb, upf0054, trna_m1g_mt. These SCDs were found to occur only once in a set of 43 completely sequenced genomes of both marine and non-marine isolates (Table S3). GOS metagenomes were queried by hidden Markov models (HMMs) belonging to these SCDs present in the Pfam 23.0 database (Finn, *et al.*, 2008) using a single TimeLogic DeCypher card (Active Motif, Inc., Carlsbad, CA), and hits with E-values below $1*10^{-10}$ were used for the standardization. The standardized absolute abundances were then converted to relative abundances by dividing them by the total count of all 16S rRNA gene fragments with more than 100 bases at each GOS sampling site (Table S4).

Following the standardization, relative abundances were converted into a sites*species matrix, complemented by a sites*parameters matrix, and imported into the R statistical computing environment (R Development Core Team, 2010) for further statistical analysis and visualization purposes.

The R package vegan v1.17-3 (Oksanen *et al.*, 2011) was used for all the numerical ecology analyses. Bray-Curtis dissimilarities were calculated between GOS sites based on Wisconsin and square root standardized sites*species matrices of different taxonomic ranks levels (phylum, class, order, family, genus), and used in the metaMDS() non-metric dimensional scaling (NMDS) procedure. Taxa-environment relationships were studied using least squares linear vector fitting, after the variables were subjected to z-score standardization. For categorical environmental variables, or factors, (habitat assignments), centroids (average scores) with standard deviations were calculated. Significance of the fitted vectors or factors was determined by permutations (n=9999), and a Pr(>r) value less than 0.01 was judged to be significant. Additionally, a generalized additive model (GAM) surface fit was visualized as smooth, non-parametric isoclines with significance tested by permutation tests (n=9999) and ANOVA. The coefficient of determination (R2) was used as a goodness-of-fit measure for fitted vectors, factors, and non-parametric surfaces.

Results and Discussion

Overview of the taxonomic makeup

A total of 142,783 small subunit (SSU) rRNA (1.4% of total reads) gene fragments were retrieved and aligned using SINA. With the 100 aligned bases cut-off, the dataset was reduced to 12,742 SSU (9% of total SSU) rRNA gene fragments. The majority of the excluded sequences (>98%) contained less than 50 aligned bases, and since sequences this short do not carry sufficient phylogenetic information; we expect no significant loss of information. Sampling sites GS038 through GS046, and GS050 had less than five rRNA gene fragments of sufficient length and were excluded from further analysis. These sites contained, on average, only 700 total reads, explaining the low number of rRNA gene fragments. Additionally, no rRNA gene fragments were retrieved from the MOVE858 sample, which was obtained using 0.002-0.22μm filters, representing the viral metagenome fraction. Only few eukaryotic (253), mitochondrial (56), and chloroplast (112) sequences were found after taxonomic classification, which were also excluded from further analysis.

The remaining 12,313 rRNA sequences were classified into 536 distinct taxa, of which 10 were at phylum rank, 21 at class, one at subclass, 42 at order, two at suborder, 90 at family, and 370 at genus rank. It is important to note that we regarded clades consisting entirely of sequences from uncultured organisms, and Candidatus taxa as “artificial” taxa, and assigned them to taxonomic ranks based solely on their position within the taxonomic hierarchy to avoid confusion with validly described taxa. It should be also noted that rank assignments maybe depending on treeing methods. Of the 536 distinct taxa, 280 were observed only at a single sampling site, hence endemic. 149 of these endemic taxa belonged to *Proteobacteria,* 37 to *Bacteroidetes* and 18 to *Actinobacteria* (Table S5). No taxon was truly pandemic; occurring at all sampling sites. However, SAR11 Surface 1 group was detected at 59 out of 61 sites.

There were no unclassified sequences at domain, phylum and class ranks; however there were 15 at order, 14 at family, and 37 unclassified sequences at genus rank. 67% of all rRNA fragments were classified as *Proteobacteria*, which can be further broken down into 65% *Alphaprotebacteria,* 28% *Gammaproteobacteria*, 4% *Betaproteobacteria*, 3% *Deltaprotebacteria*, and 0.2% *Epsilonproteobacteria*. Besides *Proteobacteria*, other dominant phyla were as follows: *Actinobacteria* (9%), *Bacteroidetes* (8%),

Cyanobacteria (7%), *Deferribacteres* (3%), *Euryarchaeota* (2%), *Verrucomicrobia* (0.7%), *Crenarchaeota* (0.6%), and *Chloroflexi* (0.5%). The rest of the phyla were divided into four fractions having between 0.1%-0.5%, 0.05%-0.1%, 0.01%-0.05%, and less than 0.01% relative abundance. The first fraction contains fragments classified as Candidate division OD1, *Planctomycetes, Acidobacteria* and *Firmicutes*. The second fraction consists of Candidate division OP11, *Chlamydiae*, BD1-5; a clade composed of clones isolated mainly from deep sea sediment, aquatic, or biofilm samples; *Deinococcus-Thermus*, *Gemmatimonadetes*, and *Tenericutes*. The remaining two fragments consist of TM6, *Chlorobi*, Candidate division SR1, Candidate division OP3, *Lentisphaerae, Fusobacteria, Spirochaetes*, Candidate division TM7 for the 0.01-0.05% fraction; and Candidate division OP10, *Aquificae*, *Nitrospirae*, *Thermodesulfobacteria*, Candidate division WS3, and RF3 for the last fraction. The two clades, TM6 and RF3, include clones obtained from aquatic water and sediment samples, as well as clones obtained from various gut, sludge and biofilm samples. This overall assessment of the taxonomic makeup is in agreement with previous studies on the GOS metagenome (Rusch, *et al.*, 2007, Biers, *et al.*, 2009, Yooseph, *et al.*, 2010), as well as with expectations of ocean surface microbial communities (Fuhrman & Hagström, 2008). Compared to previous assessments of the GOS metagenome taxonomic makeup, we have observed more sequences belonging to the Candidate divisions. For example, only Candidate division OD1 is acknowledged in the study by Biers and colleagues (Biers, *et al.*, 2009), whereas we report the occurrence of TM7, WS3, OP3, SR1, and OP10. Clone sequences belonging to these divisions are isolated from a wide variety of sources; including sludge, soil, human or other host tissues, deep-sea sediments, lakes, or biofilms (Hugenholtz, *et al.*, 1998, Pace, 2009). In the GOS metagenome, their distribution was limited to, except for Candidate division OD1, coastal, estuarine, brackish, hypersaline waters, as well as freshwater environments. The absence of these divisions from surface open ocean waters is congruent with previous observations, whereas the presence in estuarine and coastal waters could be indicative of anthropogenic inputs considering their prevalence in wastewater/sludge type environments, while the fresh, brackish, and hypersaline water prevalence is in line with potentially differing metabolic capabilities in comparison to surface ocean communities. Candidate division OD1 was the most widespread candidate division within the GOS metagenome, and in addition to the

previously listed locations, this taxon was also observed in ocean waters from sites GS000a, GS114, and GS117 (Fig. S1). Although the OD1 is environmentally widespread, our survey of previous isolation sources did not encounter any other surface ocean clones.

Community structures at different taxonomic rank levels

Spatial, and temporal patterns; along with ecological coherence of higher bacterial and archaeal taxonomic ranks have been discussed previously (Philippot, *et al.*, 2009, Philippot, *et al.*, 2010). Although the GOS metagenome is only composed of surface water samples, the "surfaces" sampled have an interesting variety of contrasting habitats, such as estuary vs. open ocean, or hypersaline vs. freshwater. We used this diversity of habitats in order to reveal how well these habitat differences will be reflected in community structures composed of bacterial and archaeal taxa at different rank levels.

The sites*species data for ordination analysis consisted of standardized relative abundances at five different rank levels. The phylum level consisted of 33 distinct taxa, with 12,313 sequences classified into these taxa. The class level had 55 distinct taxa, with 12,222 sequences; order level 107 distinct taxa and 12,049 sequences, family level 201 distinct taxa and 10,616 sequences, and finally genus level had 363 distinct taxa and 4270 sequences. At any taxonomic rank level, the NMDS analysis showed that certain sites have remarkably different community structures (Fig 1). To be more specific, a recurring trend was a halo of coastal (GS013), estuary (GS011-GS012), hypersaline (GS033), freshwater (GS020), mangrove (GS032), fringing reef (GS025), and some open ocean (GS00a-c) sites, surrounding a cluster of mainly open ocean sites. In addition to the aforementioned sites, another set of coastal/estuary (GS002-GS010), warm seep (GS030), coral reef (GS048), and coastal upwelling (GS031) sites were to some extent distinguishable from the open ocean cluster. The relative distances between sites at different ranks were not always the same. For example, GS011-GS012 couple was placed at varying distances from each other, but nevertheless they retained their general distinctness from the rest. Another one is GS013 and GS025 couple; clearly different from the rest of the sites, but conspicuously placed too near each other at phyla rank level. The taxa composition and relative abundances change with each rank level, therefore such conformation changes are expected.

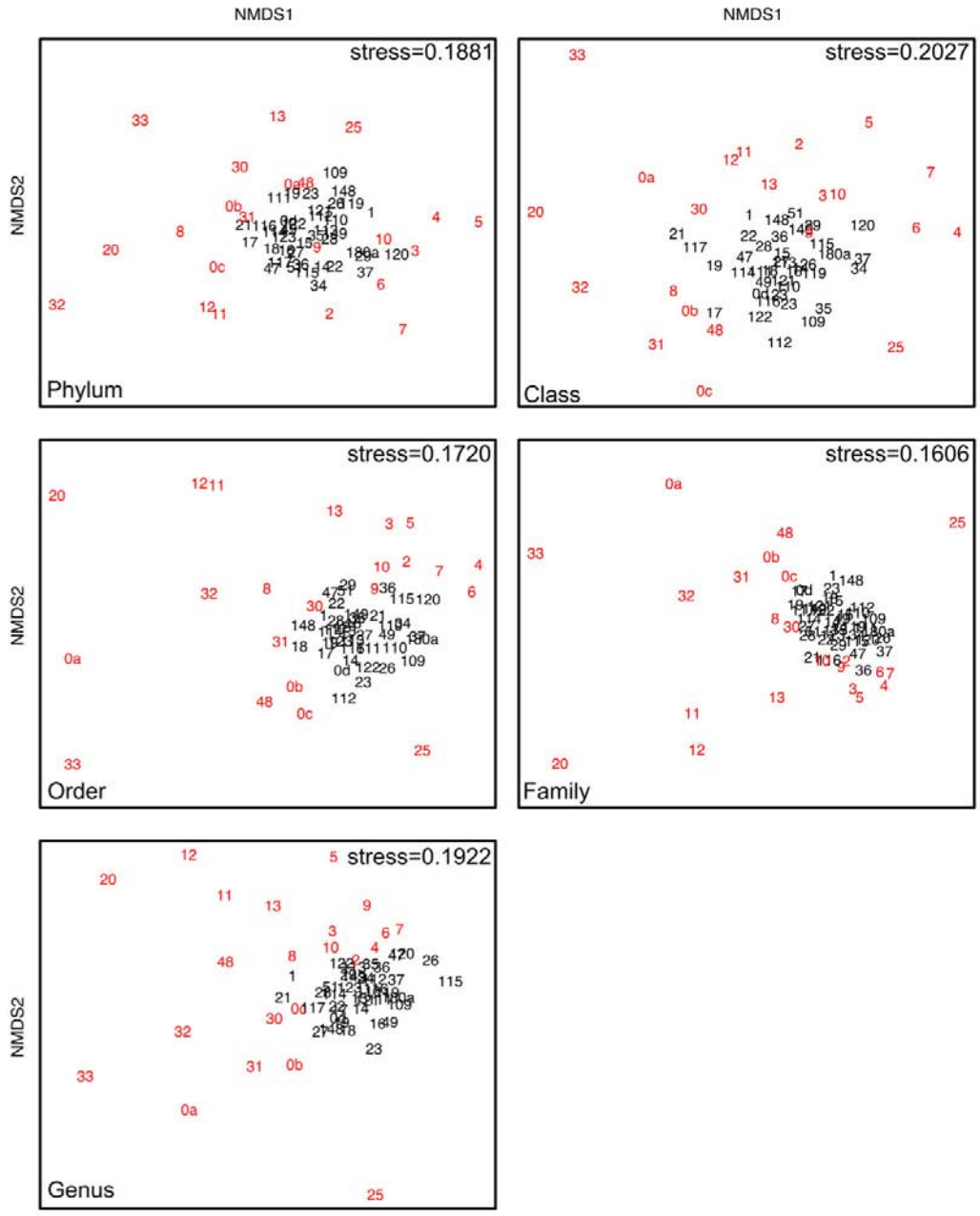


Figure 1. Panel figure showing NMDS analysis for each taxonomic rank level. The community dissimilarities were calculated based on taxa standardized relative abundances. For visibility, the "GS" prefix was omitted from sampling site names. Stress values are indicated at the top-right corner of each figure, whereas the ranks are indicated at the bottom-left corner. Sampling sites mentioned results and discussion are highlighted in red.

Since a systematic classification of the habitat types can extend these observations further, all NMDS plots were annotated with EnvO biome, feature, and material terms

(Fig 2). The three different levels of EnvO terms provide an increasing order of granularity to habitat description of the sampling sites; the first level biome broadly establishes the system that defines the scope of potential ecological inputs that a biological entity may be subjected to, whereas an environmental feature describes a range of biotic and abiotic entities and phenomena that are more local to that entity than its biome, and finally, material is understood as the substance immediately surrounding that entity and acting as the primary transmitter of ecological forces to and from it.

All EnvO term types produced significant correlations to the ordinations, however biome and material, overall, produced 1.5-2 times higher correlations, compared to feature terms. Although contrasting biome or material types, such as lake vs. oceanic, or hypersaline vs. estuarine water, were distinguishable on all rank levels, meaningful associations of identical terms started to appear at the class rank and improved at lower rank levels. At phylum and class level, for example, oceanic epipelagic zone and neritic epipelagic zone biomes, or coastal water and ocean water sites were intermixed, while from order level on these two biomes were more separated and formed their own clusters. These two clusters were by no means compact; with some neritic/coastal sites trespassing into the oceanic cluster. These sites are however, islands, and although the ontologically correct annotation would be neritic epipelagic zone biome or coastal water, biologically the open ocean effect is evidently more dominant. The two estuarine biomes were placed together on all rank level ordinations, however a third one occurred with neritic biomes, and separated from the former two. This deviation can be explained by the locality; the former two sites are located at Delaware Bay and Chesapeake Bay, respectively, whereas the latter site is listed as Bay of Fundy, which are drastically different estuaries (Lotze, *et al.*, 2006), suggesting that despite the same habitat type, higher and lower level taxon ranks reflect these differences. A number of other biomes were sampled during the GOS expedition, namely marine coral reef, marine reef, warm seep, and marginal sea. The ordinations did not reveal these biomes as being different from oceanic and neritic sites, although it is known that specific groups of bacteria are known to be associated with corals (Rohwer, *et al.*, 2002, Pantos, *et al.*, 2003, Bourne & Munn, 2005), and a low similarity between Pacific Ocean, and Caribbean Sea samples has been observed previously (Lee & Fuhrman, 1991).

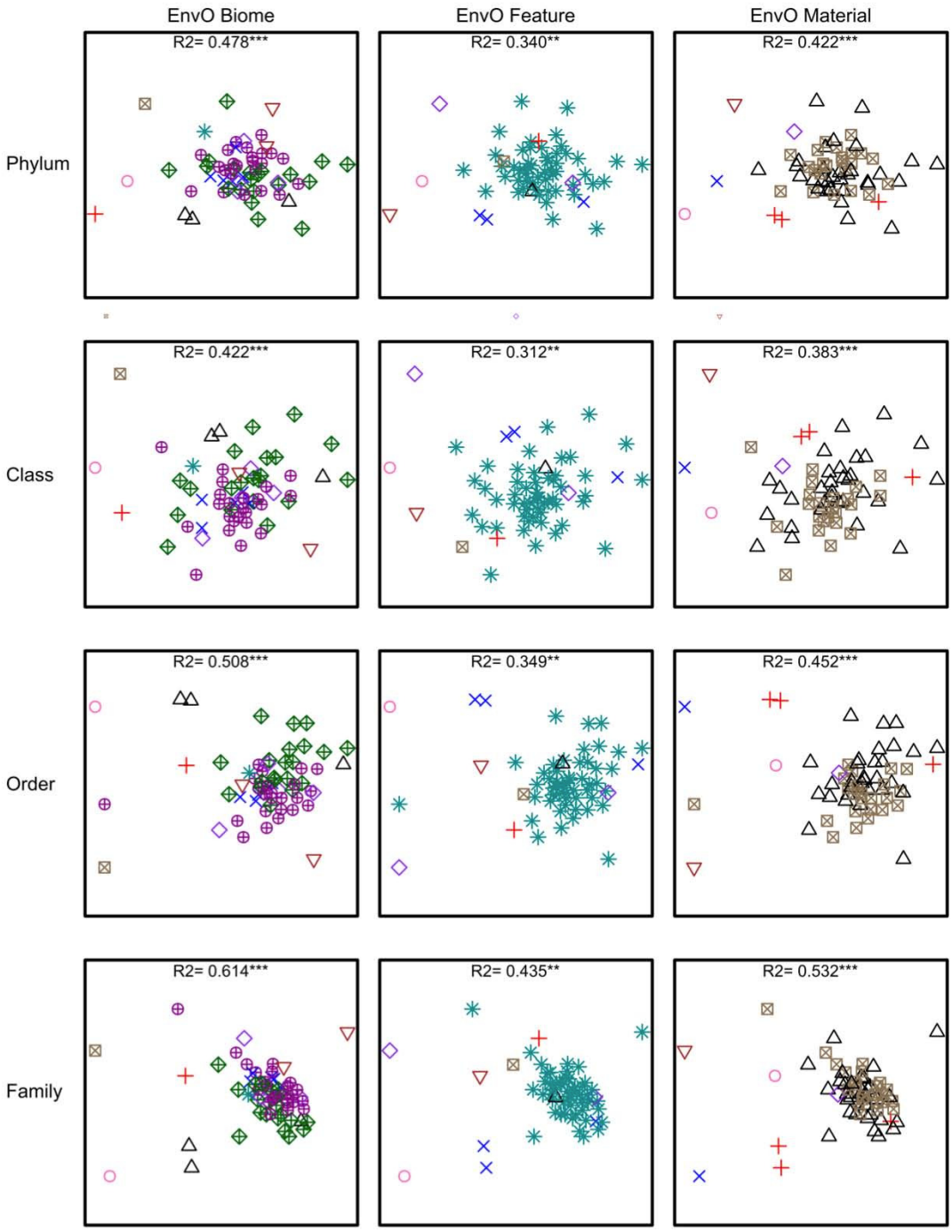


Figure 2. Panel figure showing NMDS analysis for each taxonomic rank level. The configuration of sites are the same as figure 1, however now sites are annotated with EnvO terms at three different levels; biome, feature and material. Rows indicate different rank levels, whereas columns indicate different sets of terms. Legends at the bottom of each column show the color and shape code of EnvO terms. Goodness-of-fit of term levels to the ordination are indicated by the R2 values, and the significances by asterisk symbols (0 '***' 0.001 '**' 0.01).

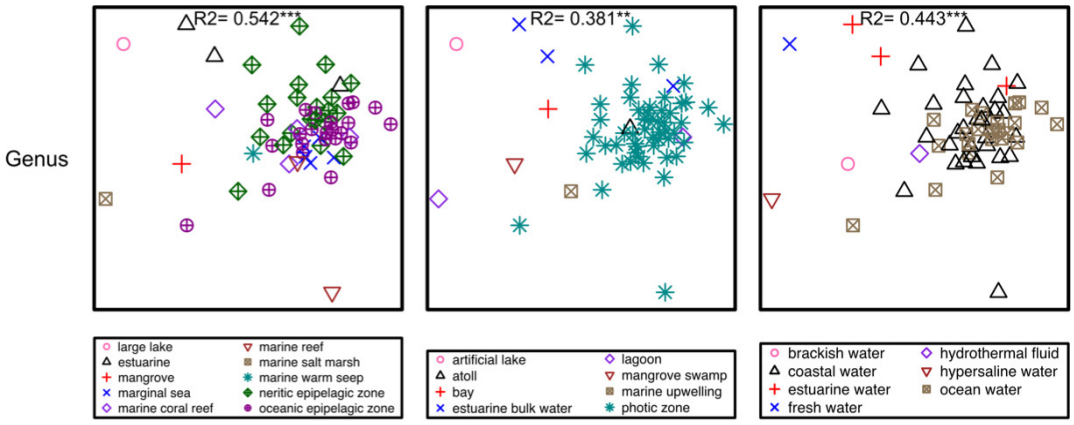


Figure 2. cont'd

The possible explanation for this observation can be that the habitat annotations are misleading, and that the prominent feature of these sites is being neritic or oceanic biomes, rather than being reef or marginal sea biomes. Another clue supporting this argument is recognized with EnvO feature annotations; the sample feature of the majority of these biomes is photic zone, and with this annotation they are clustered with sites sharing this feature.

In summary, these results support that higher taxonomic rank levels such as phylum or class do provide enough information to distinguish between highly contrasting habitat types, or in other words, it is possible to pinpoint ecological preferences to phyla or classes. However, at the interface of two habitats, like coastal/estuary and open ocean, it is necessary to have more resolution for discriminatory power. A basic example is the case of *Betaproteobacteria*; at phylum level, this freshwater-preferring class will be accounted as *Proteobacteria*, hence leading to a poor ordination. Nevertheless, these ordinations do not provide clear-cut habitat clusters, but a certain amount of fuzziness is observed even at genus level. The GOS metagenome does not have enough sequencing coverage to identify such fine scale differences, and higher coverage surveys would be more suited for that purpose. Finally, it should also be noted that the habitat bins were not equal sized, and more even sampling of different habitats could have produced different results.

To test whether bacterial and archaeal taxa distribution is related to environmental conditions, we fitted vectors and non-parametrically smoothed surfaces of seven

environmental variables (Virtanen, *et al.*, 2006), which were obtained both *in situ* and by interpolation. The combined interpretation of variable vectors and fitted surfaces is to be made as follows; the vector arrow points to the direction of most rapid change in the environmental variable, or the direction of the gradient, and the length of the arrow is proportional to the correlation between ordination and variable. A planar fitted surface indicates that the response of the community to the variable is linear, and the surface R2 will be equal to or close to the R2 of the vector; but if the response is non-linear, R2 for the surface will be higher than for the vector. Of the seven variables, only three; namely temperature, salinity, and chlorophyll a concentration, were found to be produce significant correlations (Fig 3). Temperature did not produce a significant correlation at phylum level and class level, but was significant for other rank levels. The strength of the correlation was high at order level (0.394), but a small decrease at family level was observed (0.213), which was followed by an increase at genus level (0.313). In any case, the response of the community structure to temperature was non-linear, as indicated by higher surface correlations. Salinity was a significant variable at phylum, order and family levels, and correlation was highest at order level (0.417). As with temperature, the effect of salinity was also non-linear. Chlorophyll a concentration correlations were significant at all levels, except at phylum level, and produced the strongest correlation of all the three variables. Additionally, a linear effect was observed at class level, although this effect changed to non-linear at lower rank levels. Temperature, salinity, and chlorophyll a concentrations are environmental variables with known effects on structuring the marine bacterial and archaeal communities (Nold & Zwart, 1998, Brown, *et al.*, 2005, Kirchman, *et al.*, 2005, Fuhrman, *et al.*, 2006, Pommier, *et al.*, 2007), therefore our observations are not unfamiliar, but having established these correlations on a much larger study they do second the known effects of these variables. Furthermore, they also underline the ecological structuring of high taxon levels by the environment.

These correlations can provide fresh indications about the relations of variable gradients to taxa, which can come useful especially in the case of taxa with few cultured members, or the “artificial” taxa or clades with no cultured members. For example, clades BD1-5 and TM6 (phylum level) can be ascribed as lower-salinity preferring clades, whereas RF8 appears at extreme salinity levels.

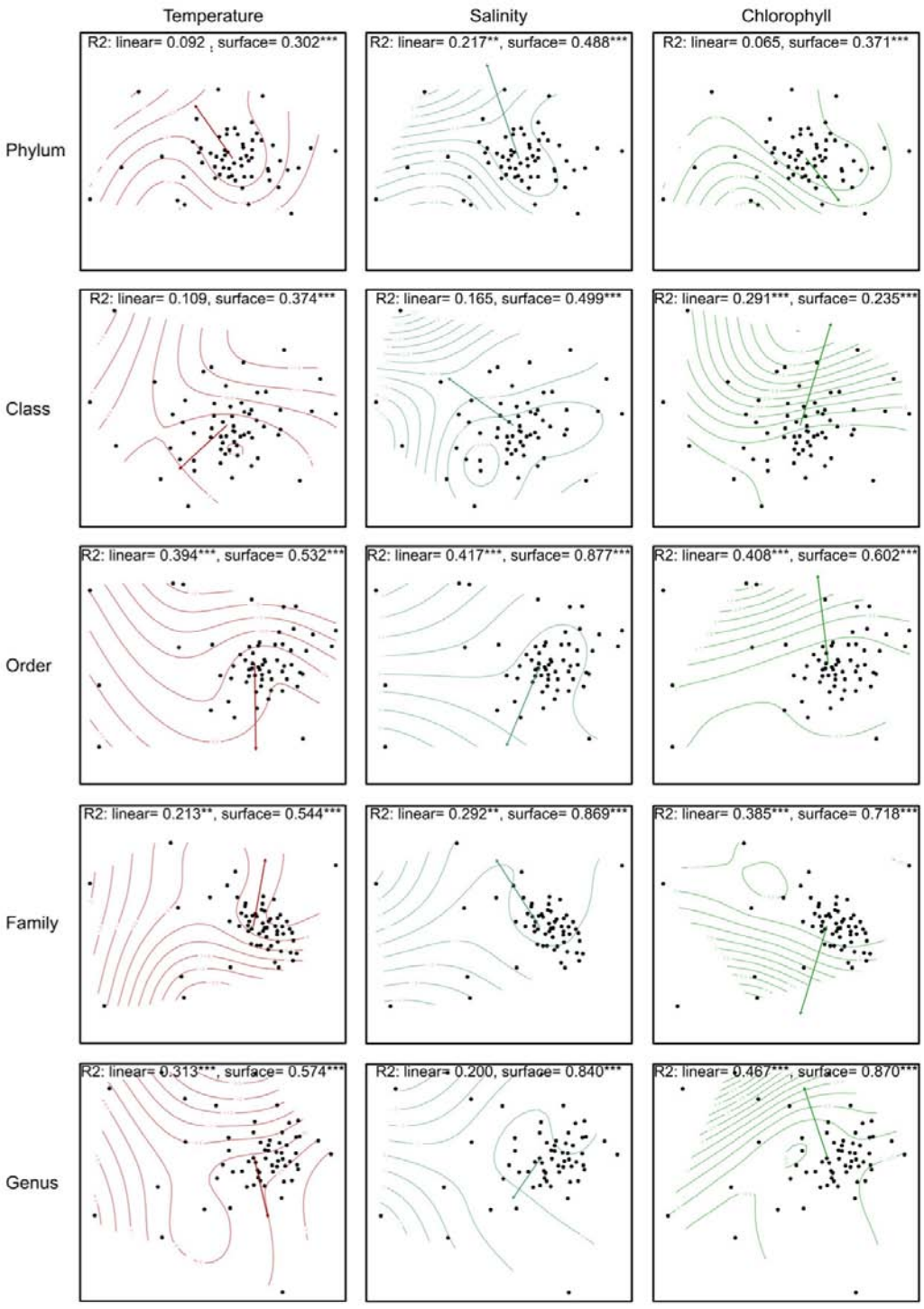


Figure 3. Panel figure showing NMDS analysis for each taxonomic rank level with fitted environmental variable vectors and non-parametric surfaces. Rows indicate different rank levels, whereas columns indicate different variables. All variable values were z-score standardized prior to vector and surface fitting, hence the isocline values reflect the z-scores. Goodness-of-fit of vectors and surfaces are again indicated by the R2 values, R2 linear, and R2 surface, respectively, while significances by asterisks (0 '***' 0.001 '**' 0.01).

At class level, the clade JTB23, belonging to *Proteobacteria*, appears to prefer higher extremes of the temperature gradient, while being on the lower extreme of the chlorophyll a gradient. On the contrary, *Acidobacteria* appears at lower extremes temperatures, but at higher extremes of chlorophyll a gradients (Fig S2).

Despite encouraging results, caution in interpretation is necessary. This study is a meta-analysis; with habitat annotations based on our conclusions from geographic maps or locale descriptions, and with environmental conditions deduced from a limited set of i*n situ* data, which were complemented by interpolation. Although the GOS metagenome is geographically vast, and environmental gradients exist, they are not continuous.

Conclusions

This study shows that inspecting bacterial and archaeal communities at high level taxonomic ranks provides enough information to distinguish between surface ocean sampling sites of the GOS metagenome. Furthermore, application of ontological annotations to these sampling sites provided a better overview of the ecological structuring of the high level taxa communities by providing a context to the community structure differences. Finally, it was possible to delineate the distribution of the taxa to environmental conditions. Our observations, along with previous evidence (Fierer, *et al.*, 2007, Philippot, *et al.*, 2009), indicate that there can be an ecological cohesion at high taxonomic levels.

Acknowledgements

We would like to thank Mar Fernández Méndez, Petra Pjevac and Elmar Prüsse for their assistance in phylogenetic and taxonomic analysis of the dataset. We would also like to thank Pier Luigi Buttigieg for assistance in ontological annotation of the dataset and for helpful suggestions. This study was supported by the Max Planck Society.

References

Alonso-Sáez L, Gasol JM, Lefort T, Hofer J & Sommaruga R (2006) Effect of natural sunlight on bacterial activity and differential sensitivity of natural bacterioplankton groups in northwestern Mediterranean coastal waters. *Appl Environ Microbiol* **72**: 5806-5813.

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

Andersson AF, Riemann L & Bertilsson S (2010) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISME J* **4**: 171-181.

Biers EJ, Sun SL & Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the Global Ocean Sampling metagenome. *Appl Environ Microbiol* **75**: 2221-2229.

Bourne DG & Munn CB (2005) Diversity of bacteria associated with the coral Pocillopora damicornis from the Great Barrier Reef. *Environ Microbiol* **7**: 1162-1174.

Brown MV, Schwalbach MS, Hewson I & Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ Microbiol* **7**: 1466-1479.

Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354-1364.

Finn RD, Tate J, Mistry J *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res* **36**: D281-288.

Fuhrman J & Hagström Å (2008) Bacterial and archaeal community structure and its patterns. *Microbial ecology of the oceans*, (Kirchman DL, ed.), pp 45-90. Wiley-Blackwell, New York.

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV & Naeem S (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc*

Nat Acad Sci USA **103**: 13104-13109.

Glöckner FO, Fuchs BM & Amann R (1999) Bacterioplankton compositions in lakes and oceans: a first comparison based on fluorescence *in situ* hybridization. *Appl Environ Microbiol* **65**: 3721-3726.

Hugenholtz P, Goebel B & Pace N (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**: 4765-4774.

Kirchman DL, Cottrell MT & Lovejoy C (2010) The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol* **12**: 1132-1143.

Kirchman DL, Dittel AI, Malmstrom RR & Cottrell MT (2005) Biogeography of major bacterial groups in the Delaware Estuary. *Limnol Oceanogr* **50**: 16971706.

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, Waldmann J & Glöckner FO (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38**: D391-D395.

Lauber CL, Hamady M, Knight R & Fierer N (2009) Soil pH as a predictor of soil bacterial community structure at the continental scale: A pyrosequencingbased assessment. *Appl Environ Microbiol* **75**: 5111-5120.

Lee SH & Fuhrman JA (1991) Spatial and temporal variation of natural bacterioplankton assemblages studied by total genomic DNA cross-hybridization. *Limnol Oceanogr* **36**: 1277-1287.

Lotze HK, Lenihan HS, Bourque BJ, Bradbury RH, Cooke RG, Kay MC, Kidwell SM, Kirby MX, Peterson CH & Jackson JBC (2006) Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science* **312**: 1806-1809.

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

Nemergut DR, Costello EK, Hamady M *et al.* (2010) Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13**: 135-144.

Nold SC & Zwart G (1998) Patterns and governing forces in aquatic microbial

communities. *Aquat Ecol* **32**: 17-35.

Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH and Wagner H (2011) vegan: Community Ecology Package. R package version 1.17-6. http://CRAN.Rproject.org/package=vegan

Pace NR (2009) Mapping the Tree of Life: Progress and Prospects. *Microbiol Mol Biol Rev* **73**: 565-576.

Pantos O, Cooney RP, Le Tissier MDA, Barer MR, O'Donnell AG & Bythell JC (2003) The bacterial ecology of a plague-like disease affecting the Caribbean coral Montastrea annularis. *Environ Microbiol* **5**: 370-382.

Philippot L, Bru D, Saby NPA, Čuhel J, Arrouays D, Šimek M & Hallin S (2009) Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial tree. *Environ Microbiol* **11**: 1462-2920.

Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB & Hallin S (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Micro* **8**: 523-529.

Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, Tunlid A & Hagström Å (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**: 867-880.

Pruesse E, Quast C, Yilmaz P, Ludwig W, Peplies J & Glöckner FO (2011) SILVA: comprehensive databases for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, (de Bruijn FJ, ed.), pp 393-398. John Wiley & Sons, Incorporated.

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

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3900051-07-0, URL http://www.Rproject.org.

Rohwer F, Seguritan V, Azam F & Knowlton N (2002) Diversity and distribution of coral-associated bacteria. *Mar Ecol Prog Ser* **243**.

Rusch DB, Halpern AL, Sutton G *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

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

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

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N & Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126-1130.

Yooseph S, Nealson KH, Rusch DB *et al.* (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**: 60-66.

Supplementary Information

Table S1. Environmental variables (non-standardized) used in vector and surface fitting to NMDS ordinations. Values are *in situ*, or where this was not possible, interpolated with megx.net GIS tools. NA: not available.

Site	Temperature (^{o}C)	Salinity (PSU)	Dissolved oxygen (mL L^{-1})	Nitrate (µmol L^{-1})	Phosphate (µmol L^{-1})	Silicate (µmol L^{-1})	Chlorophyll (µg kg^{-1})
GS000a	20.5	36.7	5.12	0.11	0.06	0.79	0.17
GS000b	20.5	36.7	5.14	0.24	0.06	0.81	0.17
GS000c	19.8	36.7	5.18	0.38	0.06	0.96	0.17
GS000d	20	36.6	5.12	0.11	0.06	0.79	0.17
GS001	22.9	36.7	5.14	0.1	0.05	0.91	0.1
GS002	18.2	29.2	5.88	0.29	0.21	2.18	1.4
GS003	11.7	29.9	5.84	0.33	0.21	2.21	1.4
GS004	13.86	28.3	6.58	0.05	0.09	0.55	0.4
GS005	15	30.2	6.58	0.07	0.12	0.71	6
GS006	11.2	28.9	0.15	0.07	0.13	0.81	2.8
GS007	17.9	31.7	6.21	0.35	0.28	2.86	1.4
GS008	9.4	26.5	2.94	0.34	0.6	0.76	2.2
GS009	11	31	3.91	1.39	0.63	0.83	4
GS010	12	31	5.86	1.95	0.48	1.02	2
GS011	11	19.94	4.33	3.33	0.53	0.67	4.8
GS012	1	3.5	NA	NA	NA	NA	21
GS013	9.3	33.95	5.37	1.06	0.25	1.27	3
GS014	18.6	36.04	5.08	0.15	0.2	1.14	1.7
GS015	25	36	4.84	0.95	0.04	1.22	0.2
GS016	26.4	35.8	4.86	0.59	0.04	1.32	0.16
GS017	27	35.8	4.59	0.31	0.14	1.82	0.13

GS018	27.4	35.4	4.63	0.45	0.1	2.21	0.14
GS019	27.7	35.4	4.59	0	0.05	2.25	0.23
GS020	28.6	0.1	4.85	0.02	0.52	3.55	NA
GS021	27.6	30.7	4.64	0.01	0.21	2.71	0.5
GS022	29.3	32.3	4.65	1.86	0.33	2.17	0.33
GS023	28.7	32.6	4.71	6.5	0.69	4.5	0.07
GS025	28.3	31.4	4.71	6.63	0.71	4.42	0.11
GS026	27.8	32.6	4.72	0.69	0.28	3.02	0.22
GS027	25.5	34.9	4.47	3.2	0.52	6.42	0.4
GS028	25.22	34.39	4.48	3.17	0.52	6.37	0.35
GS029	26.2	34.5	4.4	1.95	0.48	6.45	0.4
GS030	26.9	34.4	4.45	4.95	0.48	6.91	NA
GS031	18.6	29.07	3.48	0.87	0.12	0.5	0.35
GS032	25.4	29.47	3.85	0.56	0.09	0.36	NA
GS033	37.6	63.4	4.5	2.89	0.49	5.97	NA
GS034	27.5	34.23	4.4	1.95	0.48	6.45	0.36
GS035	21.8	34.5	4.61	2.55	0.5	3.22	0.28
GS036	25.8	34.6	4.26	2.69	0.64	1.1	0.65
GS037	28	34.38	4.75	5.61	0.56	4.83	0.21
GS047	28.6	37.3	4.7	2.14	0.4	1.35	NA
GS048	28.9	35.1	4.69	0.01	0.19	0.8	0.1
GS049	28.8	32.6	4.69	0.01	0.19	0.8	0.1
GS051	27.3	34.2	4.51	0.08	0.24	0.73	NA
GS108a	25.8	32.4	4.52	0.02	0.21	1.43	0.11
GS109	27.2	32.6	4.43	0.03	0.13	2.15	0.14
GS110	27	32.7	4.61	0.12	0.11	3.44	0.13
GS111	26.4	32.3	4.69	0.15	0.08	2.6	0.2

GS112	26.6	32.5	4.62	0.2	0.05	2.97	0.13
GS113	27.5	33.3	4.5	0.3	0.16	4.37	0.24
GS114	28.2	33.1	4.55	0.11	0.23	2.75	0.14
GS115	27.9	33.2	4.46	0.13	0.27	4.15	0.14
GS116	26.2	33.1	4.76	0.18	0.13	2.93	0.29
GS117	26.4	35.5	4.75	0.25	0.19	2.29	0.21
GS119	23.8	35.4	5.37	0.17	0.17	2.93	0.08
GS120	22.5	35.6	5.32	0.12	0.2	3.18	0.12
GS121	23.1	35.4	5.29	0.72	0.17	3.52	0.14
GS122	20.2	35.8	5.4	1.01	0.15	2.71	0.15
GS123	20.4	35.8	5.13	0.15	0.22	3.55	0.23
GS148	21.27	29.28	4.19	NA	0.16	3	NA
GS149	21.27	29.28	4.19	NA	0.16	3	NA

Table S2. EnvO term annotations of GOS sampling sites at three different levels, along with tags. NA: not available.

Site	Biome	Feature	Material	Tags
GS000a	oceanic epipelagic zone biome	photic zone	ocean water	marine wind mixed layer
GS000b	oceanic epipelagic zone biome	photic zone	ocean water	marine wind mixed layer
GS000c	oceanic epipelagic zone biome	photic zone	ocean water	marine wind mixed layer
GS000d	oceanic epipelagic zone biome	photic zone	ocean water	marine wind mixed layer
GS001	oceanic epipelagic zone biome	photic zone	ocean water	marine wind mixed layer
GS002	neritic epipelagic zone biome	photic zone	coastal water	eastern boundary current, marine wind mixed layer
GS003	neritic epipelagic zone biome	photic zone	coastal water	marine reef, eastern boundary current, marine wind mixed layer
GS004	neritic epipelagic zone biome	photic zone	coastal water	eastern boundary current, marine wind mixed layer
GS005	neritic epipelagic zone biome	photic zone	coastal water	bay, marine wind mixed layer
GS006	estuarine biome	estuarine bulk water	estuarine water	bay
GS007	neritic epipelagic zone biome	photic zone	coastal water	eastern boundary current, marine wind mixed layer
GS008	neritic epipelagic zone biome	photic zone	coastal water	naturalharbour, eastern boundary current, marine wind mixed layer
GS009	neritic epipelagic zone biome	photic zone	coastal water	eastern boundary current, island, marine wind mixed layer

GS010	neritic epipelagic zone biome	photic zone	coastal water	eastern boundary current, western boundary current, peninsula, marine wind mixed layer
GS011	estuarine biome	estuarine bulk water	estuarine water	bay, natural harbour
GS012	estuarine biome	estuarine bulk water	estuarine water	NA
GS013	neritic epipelagic zone biome	photic zone	coastal water	western boundary current, marine wind mixed layer
GS014	neritic epipelagic zone biome	photic zone	coastal water	western boundary current, marine wind mixed layer
GS015	marginal sea biome	photic zone	coastal water	neritic epipelagic zone biome, marine wind mixed layer
GS016	marginal sea biome	photic zone	coastal water	oceanic epipelagic zone biome, marine wind mixed layer
GS017	marginal sea biome	photic zone	coastal water	oceanic epipelagic zone biome, marine wind mixed layer
GS018	marginal sea biome	photic zone	coastal water	oceanic epipelagic zone biome, marine wind mixed layer
GS019	marginal sea biome	photic zone	coastal water	oceanic epipelagic zone biome, marine wind mixed layer
GS020	Large lake biome	artificial lake	fresh water	NA
GS021	neritic epipelagic zone biome	photic zone	coastal water	NA
GS022	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS023	oceanic epipelagic zone biome	photic zone	ocean water	island
GS025	marine reef biome	photic zone	coastal water	NA
GS026	oceanic epipelagic	photic zone	ocean water	NA

	zone biome			
GS027	marine coral reef biome	photic zone	coastal water	island
GS028	neritic epipelagic zone biome	photic zone	coastal water	island
GS029	neritic epipelagic zone biome	photic zone	coastal water	island, bay
GS030	marine warm seep biome	photic zone	hydrothermal fluid	NA
GS031	neritic epipelagic zone biome	marine upwelling	coastal water	island
GS032	mangrove biome	mangrove swamp	brackish water	coastal water
GS033	marine salt marsh biome	lagoon	hypersaline water	island
GS034	neritic epipelagic zone biome	photic zone	coastal water	island
GS035	neritic epipelagic zone biome	photic zone	coastal water	island
GS036	neritic epipelagic zone biome	photic zone	coastal water	island
GS037	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS047	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS048	marine coral reef biome	bay	coastal water	island
GS049	neritic epipelagic zone biome	photic zone	coastal water	island
GS051	marine coral reef biome	atoll	coastal water	NA

GS108a	marine coral reef biome	lagoon	coastal water	NA
GS109	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS110	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS111	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS112	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS113	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS114	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS115	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS116	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS117	neritic epipelagic zone biome	photic zone	coastal water	island
GS119	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS120	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS121	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS122	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS123	oceanic epipelagic zone biome	photic zone	ocean water	NA
GS148	marine reef biome	photic zone	coastal water	NA

GS149	neritic epipelagic zone biome	photic zone	coastal water	NA

Table S3. Whole organismal genomes of both marine and non-marine isolates from which the single copy domains (SCDs) were selected.

Alteromonas macleodii 'Deep ecotype'	*Psychromonas ingrahamii* 37
"*Candidatus* Pelagibacter ubique" HTCC1062	*Rhodobacter sphaeroides* ATCC 17025
"*Candidatus* Protochlamydia amoebophila" UWE25	*Rhodopirellula baltica* SH 1
Chlorobium phaeobacteroides DSM 266	*Rickettsia felis* URRWXCal2
Coxiella burnetii RSA 493	*Shewanella amazonensis* SB2B
Erythrobacter litoralis HTCC2594	*Shewanella baltica* OS185
Flavobacterium johnsoniae UW101	*Shewanella baltica* OS195
Flavobacterium psychrophilum JIP02/86	*Shewanella oneidensis* MR-1
Geobacillus kaustophilus HTA426	*Shewanella sediminis* HAW-EB3
Geobacillus thermodenitrificans NG80-2	*Shewanella* sp. ANA-3
Gramella forsetii KT0803	*Shewanella* sp. MR-7
Idiomarina loihiensis L2TR	*Shewanella* sp. W3-18-1
Magnetospirillum magneticum AMB-1	*Shewanella woodyi* ATCC 51908
Marinobacter aquaeolei VT8	*Silicibacter pomeroyi* DSS-3
Nitrosococcus oceani ATCC 19707	*Sphingomonas wittichii* RW1
Nostoc punctiforme PCC 73102	*Sphingopyxis alaskensis* RB2256
Novosphingobium aromaticivorans DSM 12444	*Synechocystis* sp. PCC 6803
Pelobacter carbinolicus DSM 2380	*Thiomicrospira crunogena* XCL-2
Photobacterium profundum SS9	*Trichodesmium erythraeum* IMS101
Prosthecochloris vibrioformis DSM 265	*Vibrio fischeri* ES114
Pseudoalteromonas atlantica T6c	*Vibrio parahaemolyticus* RIMD 2210633
Pseudoalteromonas haloplanktis TAC125	--

Table S4. Counts of 16S rRNA gene fragments longer than 100 bases retrieved from each sampling site.

Site	Count	Site	Count
GS000a	767	**GS031**	567
GS000b	454	**GS032**	202
GS000c	455	**GS033**	853
GS000d	438	**GS034**	212
GS001	266	**GS035**	190
GS002	112	**GS036**	129
GS003	80	**GS037**	85
GS004	85	**GS039**	2
GS005	46	**GS040**	1
GS006	59	**GS045**	3
GS007	73	**GS047**	109
GS008	296	**GS048**	187
GS009	100	**GS049**	123
GS010	119	**GS050**	2
GS011	177	**GS051**	162
GS012	146	**GS108a**	56
GS013	149	**GS109**	79
GS014	155	**GS110**	151
GS015	166	**GS111**	91
GS016	158	**GS112**	152
GS017	337	**GS113**	151
GS018	197	**GS114**	402
GS019	196	**GS115**	68
GS020	313	**GS116**	102
GS021	152	**GS117**	490

GS022	164	**GS119**	106
GS023	188	**GS120**	52
GS025	77	**GS121**	156
GS026	103	**GS122**	178
GS027	322	**GS123**	145
GS028	251	**GS148**	148
GS029	162	**GS149**	179
GS030	446	--	--

Table S5. Table showing number of endemic taxa at each phylum.

Taxa	Count
TM6	1
Candidate division WS3	1
Deinococcus-Thermus	1
RF3	1
Chlorobi	1
Thermodesulfobacteria	1
Aquificae	1
Nitrospirae	1
Lentisphaerae	1
Crenarchaeota	1
Candidate division OP10	1
Fusobacteria	2
Spirochaetes	2
Tenericutes	2
Deferribacteres	2
Gemmatimonadetes	4
Planctomycetes	5
Cyanobacteria	5
Chloroflexi	5
Chlamydiae	5
Acidobacteria	7
Euryarchaeota	9
Verrucomicrobia	9
Firmicutes	18
Actinobacteria	18

Bacteroidetes	37
Proteobacteria	149

Figure S1. The relative abundances of different phyla (rows) at each GOS sampling site (columns). Presence of a spot indicates the presence of a phylum, and the area of a spot represents the relative abundance.

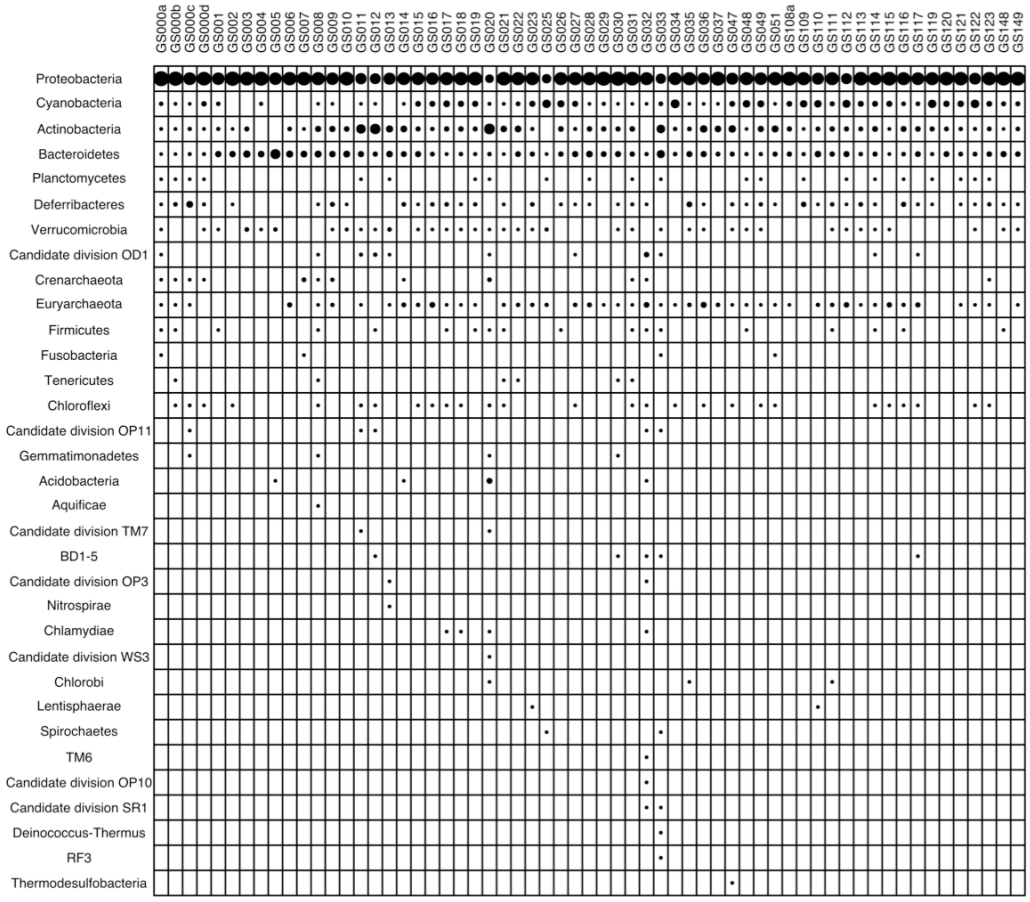
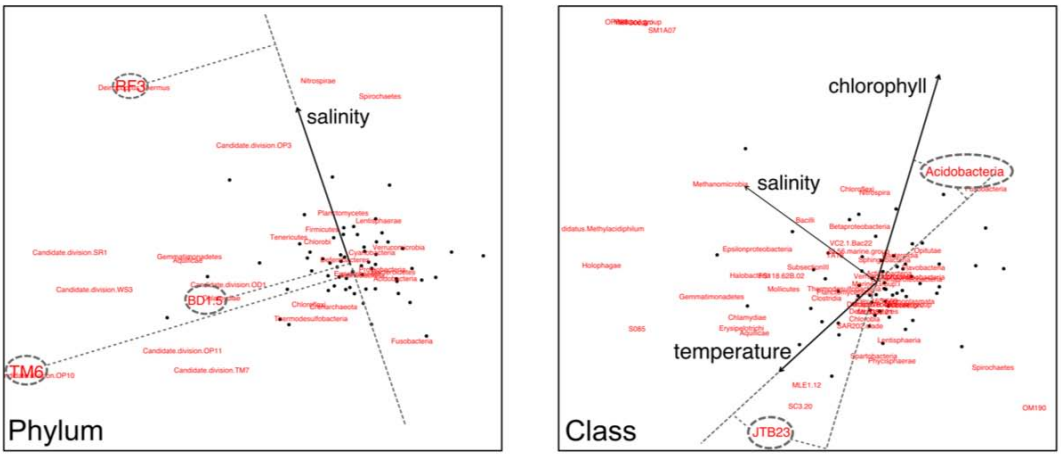


Figure S2. NMDS ordinations at phylum and class levels, with "species" weighted average scores plotted, alongside sampling sites (black dots). Perpendicular projections of some taxa on to environmental variable vectors are shown by grey-dotted lines in order to indicate their placement on the variable gradient.



SUMMARY

The work accomplished during this Ph.D. thesis produced several scientific research articles, which effectively tackled the challenges of biocuration in improving the usage of rRNA gene in microbiology and microbial ecology studies presented in the research aims and motivation section. Specifically, papers I through IV addressed the problem of biocuration due to lacking standards, and their subsequent implementation for rRNA contextual data, while papers V and VI were focused on improving the quality and quantity of the sequence data provided by already established bioinformatics resources via curation and integration efforts. The final papers, VII and VIII, presented research use cases for these value-added datasets.

Contextual Data Standards for Biocuration

Minimum information about any (x) sequence (MIxS)

Developments and understanding in microbiology have been tightly connected to molecular biology methods since the early days of bacterial genetics. Episodes of seminal discoveries in microbiology were preceded by development of methods like cloning, PCR, or sequencing. Once again, microbiology, and especially microbial ecology is going through a transformation with the rapidly improving sequencing technologies. Researchers have rapid access to rRNA, functional gene, genome, or even collections of hundreds of genomes in the form of metagenome sequences. Hopes are high; an integrated understanding of microbial diversity, community structure, functions, as well as the relationship of these parameters to the workings of our environment is promised. In papers I and II, we tried to demonstrate to the community that the right way to achieve this goal is not only through the sheer quantity of sequence data, but additionally through the quality of sequence data in terms of additional context attached to them. Our main argument was that the contextual data would enable different microbial studies to be

integrated easily, over locations, time, habitats and over different types of sequence data, and will incidentally facilitate cross-comparison and meta-analysis of global microbial communities. In paper I, we took the first step towards this contextual-data centric view of nucleic acid sequences by developing a minimum information checklist for marker gene sequences (Minimum Information abour a MARKer gene Sequence-MIMARKS), and harmonizing this newly developed standard with the previously developed standards for genomes (Minimum Information about a Genome Sequence-MIGS) and metagenomes (Minimum Information about a Metagenome Sequence-MIMS) as MIxS. The development premises for a standard that is to effectively serve the diverse research interests in microbiology were complex, but not unachievable. The preliminary design considerations were:

- Contextual data fields should be selected based on community agreement
- Contextual data fields should be comprehensive but not overwhelming
- Fields should be well-defined to minimize the risk of differing value formats
- Contextual data fields should be extended to cover more environmental parameters
- The standard should focus on, but not be limited to rRNA sequences
- The standard should focus on, but not be limited to microbial organisms
- The standard should not be limited by amplification or sequencing methods
- The standard should have a flexible framework to facilitate potential future changes

The MIMARKS standard is designed to be universal, therefore there are no restrictions on studies targeted towards a specific organism, gene or gene locus, or technology. All these variables are recognized as a contextual data field (target gene or locus, sequencing method) or the type of MIMARKS checklist (survey or specimen).

During the initial design phase of MIMARKS, contextual data fields to be used in extending the MIGS/MIMS standards to marker genes were carefully selected by conducting user surveys, and surveying most reported fields from both publications and

INSDC databases. This approach ensured that the fields included in the MIMARKS standard were in fact what the microbiology community deemed necessary and sufficient.

The extended MIMARKS checklist contains 42 contextual data fields, which may seem plenty at first glance, however by using a mixture of requirement levels (mandatory, conditional mandatory, optional), the very core of MIMARKS is reduced to just 13 fields, effectively meeting the requirements of being both comprehensive but not overwhelming.

In the original publication, the MIMARKS fields were associated with a clear description of the field. Additionally, in the spreadsheet format of MIMARKS fields, the fields were also associated with an “expected value” and a “syntax” definition, the former being the pseudo-syntax for value format, and the latter the real computational syntax for the value of a field. Furthermore, several examples of compliant datasets provided with the publication also ensured a better understanding of the correct value formats for contextual data fields.

The most important addition to existing MIGS/MIMS standards, which came with the development of MIMARKS, is the environmental packages. These include an extensive list of additional fields/parameters, specific to studies performed in a particular environment. The packages not only extend the basic MIMARKS fields to a richer set of environmental contextual data fields (Figure 4), but also provide a simple way to introduce new sets of fields in the future, in the form of new environmental packages.

The flexible framework is accomplished by collecting all GSC standards under a single umbrella term; MIxS, and sharing all the main contextual data fields, as well as environmental packages (Figure 4). Any future developments can be integrated into this framework, and will automatically inherit already developed fields and environmental packages. To enable such extension requests from within or outside the GSC community, a trac-based ticketing system was implemented, facilitating a transparent request system that is fully open to public.

To conclude, the new features implemented, as well as the changes made to the old standards during the development of the MIMARKS standards fully met the design considerations stated, and these were reflected in paper I.

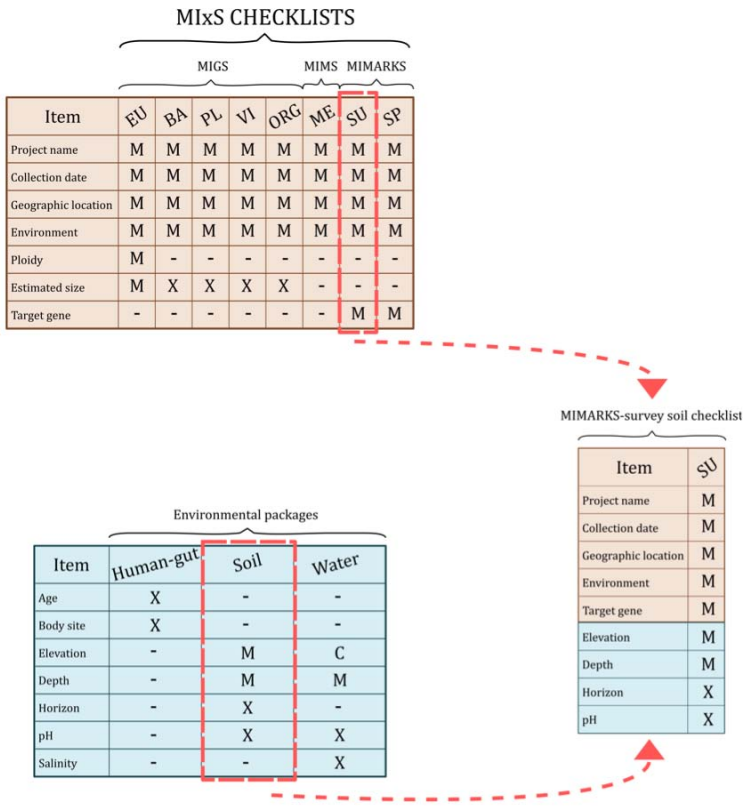


Figure 2. GSC checklists working together; overview of the MIxS (MIGS, MIMS and MIMARKS) checklists (brown) and their combination with specific environmental packages (blue).

Paper II, unlike paper I, was not a research and development paper, but a compact recapitulation of the results and suggestions of paper I. It is however, an important publication and helped to reach the research aims, since it was published in a major microbial ecology journal, directly aiming at the largest research community that will make use of the MIxS standards.

Implementation of MIxS

A simple way for a study to be MIxS-compliant would be including a contextual data as a table, or supplementary material. However, this does not ensure full machine-readability or enriching of sequence databases with contextual data. Ideally, the sequences submitted by researchers should include, at the time of submission, MIxS contextual data fields. Therefore, the development of MIxS was tightly coupled to the development of software to aid submission of the contextual data to sequence databases. Papers III and IV

illustrate this effort, which has resulted in the tools MetaBar and CDinFusion. Both tools implement essentially the same final output; creating a contextual data table ready for submission to GenBank (Figure 5), which would be placed as a structured comment within the sequence entry.

MetaBar

```
COMMENT     ##MIENS-Data-START##
            collection_date     :: 2009-10-14
            collection_time     :: 09:38:00
            lat_lon             :: 55.02814 8.45248
            geodetic_datum      :: WGS84
            lat_long_details    :: 11 m recorded accuracy
            site                :: German Wadden Sea, Sylt
            depth               :: -1.0 m
            samp_size           :: 1000.0 ml
            temperature         :: 10.155 degrees Celsius
            container           :: 10 ml glass exetainers
            environment         :: Temperate shelf and sea biome
                                   [ENVO:00000895], coastal water body
                                   [ENVO:02000049], coastal water
                                   [ENVO:00002150]
            alt_elev            :: 0 m
            country             :: Germany
            investigation_type  :: miens-survey
            project_name        :: Marine Microbiology (MarMic) class 2013 field
                                   excursion to Sylt, 2009
            sequencing_meth     :: Sanger
            target_gene         :: 16S rRNA
            MetaBar_barcode     :: 1000009000039
            ##MIENS-Data-END##
```

CDinFusion

```
COMMENT     ##MIMARKS:3.0-Data-START##
            investigation_type  :: mimarks-survey
            project_name        :: Water sample from Germany
            samp_size           :: 25 ml
            collection_date     :: 2009-09-14T09:25
            depth               :: 2 m
            biome               :: [ENVO:00000895] temperate shelf and sea
                                   biome
            feature             :: [ENVO:02000049] coastal water body
            material            :: [ENVO:00002150] coastal water
            alt_elev            :: 0 m
            geo_loc_name        :: Germany, Wadden Sea
            lat_lon             :: 55.02461 N 8.45042 E
            lib_reads_seqd      :: 0
            lib_size            :: 0
            nucl_acid_amp       :: PCR, Taq Polymerase
            pcr_cond            :: 94 degC denat., 1 min; 55 degC annealing, 1
                                   min; 72 degC elong., 3 min
            rel_to_oxygen       :: aerobe
            samp_collect_device :: 50 ml Falcon tube
            samp_mat_process    :: Plankton sample collected with 80 micrometre
                                   mesh
            seq_quality_check   :: manually edited
            seq_meth            :: Sanger
            target_gene         :: 16S rRNA
            env_package         :: water
            temp                :: 10 degC
            created with        :: CDinFusion
            ##MIMARKS:3.0-Data-END##
```

Figure 5. Structured comments of two different sequence entries, created by MetaBar (left), and CDinFusion (right). The main differences shown in red boxes are the “metabar_barcode” and “created_with” stamps, and the additional fields introduced by Metabar.

The overall idea behind each tool is unique; MetaBar is designed to cover the whole data collection process from sample to sequence submission with its barcode concept as a sample management system. The creation of GenBank submission files is a side product of MetaBar, the main product is the consistent acquisition and storage of all contextual data, either MIxS compliant or not, associated with a sample. On the other hand, CdinFusion is the last stop before submission of sequences. Here, the main focus is solely the merging of contextual data with existing sequences, and creation of submission files. Having two software products with different foci serves different use case scenarios. MetaBar supports a researcher through the whole workflow of sample acquisition, storage, processing and sequencing, with the contextual data being attached to sample and incidentally the sequence at all times. CdinFusion covers the cases where the sequences and contextual data are decoupled, by providing means to merging them.

It is also worthwhile to mention that several other tools supporting MIxS compliancy and submissions exist. For example, the MG-RAST platform [86] provides webforms for input of MIMS data[5], the RDP-II MIMARKS GoogleSheet[6] supports compliance with MIMARKS data, and QIIME web platform[7] supports all MIxS standards.

[5] metagenomics.anl.gov
[6] tinyurl.com/RDPSheet
[7] www.microbio.me/qiime

Curation of rRNA Datasets

SILVA taxonomy

Taxonomic classification of rRNA sequences is one of the integral components of an rRNA database, alongside alignments. Whether a user is performing probe design for an environmental sample, or classifying thousands of unknown variable region tag sequences, taxonomy is under the spotlight. After all, the purpose of rRNA sequences, especially regarding *Bacteria* and *Archaea*, is the identification of organisms, and taxonomic classification identifies an unknown organism and attaches the previously described properties of a taxon. Despite the importance of taxonomic classification, the SILVA rRNA database project has been lacking a standardized classification up until release 98 in 2009. An extensive manual effort has been invested in fixing this downside. Paper V illustrates this effort in a compact form, which will be elaborated in this section.

The first line of action in curation of SILVA taxonomy was the bacterial and archaeal sequences in the SSU Ref dataset. Later on, the classifications were extended to cover bacterial and archaeal LSU Ref sequences, and finally eukaryotic SSU Ref sequences. The bacterial and archaeal classification are based on Bergey's Taxonomic Outlines. Specifically, *Archaea*, *Cyanobacteria*, *Chloroflexi* and *Chlorobi* are based on Volume 1, *Proteobacteria* on Volume 2, *Firmicutes* on Volume 3, *Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes* on Volume 4, and finally *Actinobacteria* on Volume 5. Since taxonomy and species are dynamic entities, news and changes are rapid and resources other than Bergey's Outlines are required. In these cases, name changes and taxonomic outlines are adapted from "Classification of domains and phyla - Hierarchical classification of prokaryotes (bacteria)" at the web resource List of Prokaryotic Names with Standing in Nomenclature (LPSN) [87]. Although the classification is mainly based on these authoritative resources, deviations from their recommendations do exist, since the classification is a phylogenetic tree-based process and differences from the original description and classifications are to be expected. For example, the genus *Ahrensia* (type species accession: D88524) is classified under family *Rhodobacteraceae* of *Alphaproteobacteria*, however in the SSU

Ref guide tree, this genus is grouped together with members of family *Phyllobacteriaceae*. Normally, such discrepancies are accommodated by introducing non-monophyletic groups, however in this case genus *Ahrensia* is kept under *Phyllobacteriaceae* due to high sequence identities (>94%) observed with other members of this family.

The LPSN resource is also used to track down names without standing in nomenclature (not-validly published taxa) and *Candidatus* taxa. The inclusion of the two latter categories is a specialty of the SILVA classification; the RDP-II project only includes validly published taxa, and there is no documentation from the Greengenes resource on whether such taxa are included. Due to this extensive resource curation, SILVA taxonomy contains the highest amount of taxa among all three rRNA databases (Figure 6).

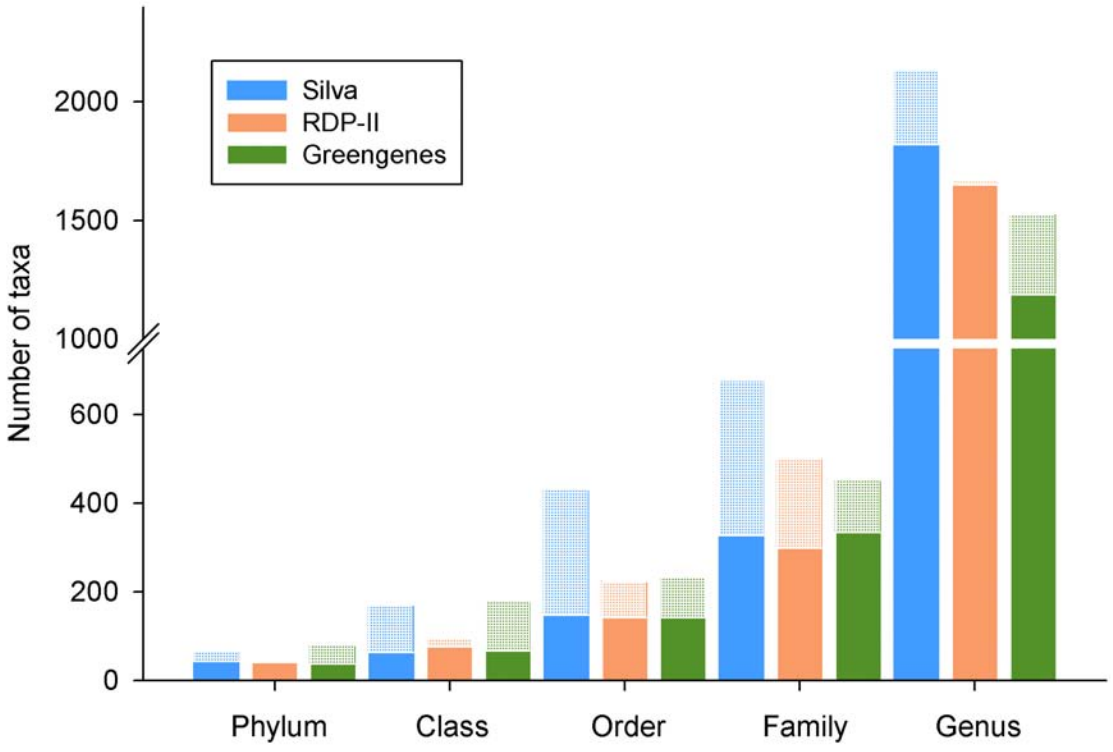


Figure 6. Comparison of number of taxa at different rank levels from three rRNA databases. The solid colored portion of bars indicate amount of cultured taxa (can include valid or non-valid taxa), while the shaded portion indicates the amount of environmental clades (uncultured groups and *Candidatus* taxa).

Figure 6 also indicates that in most cases SILVA has the highest amount of environmental clades. This is in part due to inclusion of *Candidatus* taxa in this category, but is also due to the extensive effort taken in representing uncultured clades from published resources by establishing successful collaborations with experts in relevant fields. A number of examples are; OCS116 clade [88], SAGMC and SAGME groups [89], and termite clusters [90].

Despite the differences between curation methods used, the three databases still share a sizeable number of taxa with each other (Figure 7).

Figure 7. Venn diagram showing the number of shared taxa at genus level between SILVA, RDP-II, and Greengenes. Only cultured taxa are included in this comparison. The middle overlapping part shows the number of all taxa jointly shared by all three databases, the other overlaps show taxa shared between two databases, but not the third. In this case, SILVA and Greengenes share no other taxa in addition to the 1046 shared jointly by three databases.

The RDP-II project and SILVA share the highest amount of taxa at genus level, while SILVA and Greengenes share the least. Furthermore, SILVA has the highest number of unique taxa included in the classification, followed by Greengenes and RDP-II. This comparison illustrates the differences in the curation procedure; SILVA and RDP-II use the same resources, and follow the same guidelines, whereas Greengenes has a less formalized approach. Additionally, SILVA taxonomy includes non-validly published taxa, as well as being updated more frequently with each release of new datasets.

The newest addition to the SILVA taxonomy is the *Eukaryota*. Like uncultured clades of *Bacteria* and *Archaea*, this new classification is also accomplished by external collaboration. The taxonomic hierarchy provided follows the requirements of "The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists" [91]. The taxonomic assignments from reference sequences to query sequences were made by KeyDNATools[8]. The taxonomic paths are standardized to seven ranks; domain, supergroup, division, class, order, family, and genus.

Georeferenced rRNA diversity in megx.net

The basic idea that gave life to the MIxS standards was the inclination towards having spatially and temporally referenced (x-latitude, y-longitude, z-depth/altitude/elevation, t-time) genome sequence datasets [84]. Incidentally, the most important component of the MIMARKS standard is the x, y, z, t fields as well. In order to showcase the value of these data, SSU and LSU rRNA gene sequence data from SILVA were integrated into megx.net, and georeferenced sequences are displayed alongside genomes and metagenomes on the mapserver tool, which is described in paper VI. The georeferenced sequences constitute 10% of 1,157,764 SSU and LSU rRNA gene sequences (of SILVA Parc release 102), which is low, but does not hinder the usefulness of the service. In a theoretical use case scenario, a researcher is able to address the geographic distribution of an organism, and investigate this distribution with respect to oceanographic parameters, or with respect to genomes, metagenomes, and genes.

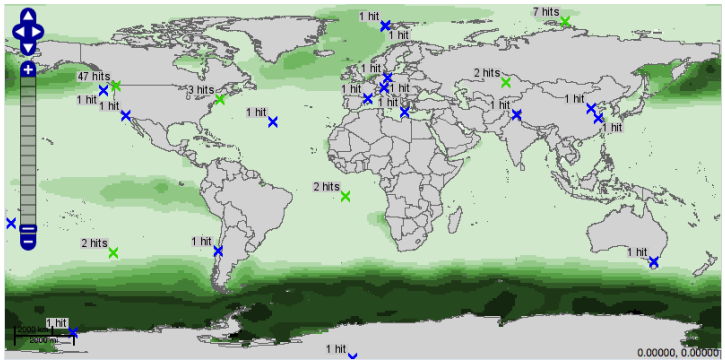


Figure 8. Geographic-BLAST results displayed on the world map with annual average nitrate concentration layer. The crosses (blue, 1 hit; green more than 1 hit) represent BLASTN target hit sequences

[8] keydnatools.com

To conclude, the integration of SSU and LSU rRNA sequence data into the unique services provided by the megx.net platform enables researchers to address questions like, “Have my organisms been seen before, where, with whom, and under which environmental conditions?”

Usage of Curated Datasets in Microbial Ecology

Investigating the potential of 23S rRNA genes in metagenomes

Paper VII is a study of the 23S rRNA gene sequences in the Global Ocean Sampling (GOS) metagenome [92, 93], with the aim of demonstrating that 23S rRNA genes from metagenomes and studies alike are equally useful phylogenetic markers as 16S rRNA gene sequences. The knowledge that the 23S rRNA gene is a useful marker, or even better than 16S rRNA is a well known [94-96], but not a widely acknowledged fact in microbial ecology. With the comparative analysis of amount and length of 23S rRNA and 16S rRNA fragments in the GOS samples, it was shown that more abundant and longer 23S rRNA fragments are obtained from metagenomic reads, meaning that there will be more phylogenetic signal per sample when 23S rRNA is used as a marker, compared to the use of 16S rRNA. Furthermore, since 16S rRNA based taxonomic classification is the gold-standard for assessment of diversity and community structure [97], it was necessary to establish that the two markers result in the same classification. Overall taxonomic composition of the whole GOS dataset, as well as sample-by-sample comparisons showed good agreement with each other at high taxonomic rank levels, but below order level, the 23S rRNA based classifications resulted in more unclassified fragment numbers than 16S rRNA. This discrepancy was explained by the lower amount of full-length reference sequences in the 23S rRNA database, and especially the lack of reference sequences guiding environmental clade annotations. In near future though, this disadvantage of the 23S rRNA gene is going to be remedied with the help of mega-genome sequencing and single-cell genomics projects [98, 99]. In addition to comparisons of 16S and 23S rRNA gene based approaches, 23S rRNA primer and probe evaluation was also carried out to stimulate discussion in the design of novel primers and probes. This survey revealed that most primers for 23S rRNA gene had low coverage, and were in need of redesign. Of the probes tested, BET42A (*Betaproteobacteria*) yielded comparable results to previous evaluations, but this was not the case with probe GAM42A (*Gammaproteobacteria*), where clade-specific mismatches distributed over different groups of *Gammaproteobacteria* were observed.

In the grand scheme of research aims, this paper has made use of the taxonomy curation effort spent on the LSU Ref dataset of SILVA by verifying that 23S and 16S rRNA based diversity studies in metagenomes provide comparable results, and by primer/probe evaluations. Additionally, evaluation of primers and probes on non-PCR amplified sequence environmental datasets was shown to be a valuable approach, since it provides a much more novel and diverse test set.

Ecological structuring of bacterial and archaeal taxa in the marine environment

Paper VIII attempts to uncover distribution patterns in bacterial and archaeal high taxon ranks (phylum, class, order, and family), and to explain these patterns with respect to habitat and environmental condition differences. As with paper VII, the dataset of choice was GOS metagenome; since GOS is still the most geographically comprehensive marine dataset, and has the added value of each sampling site being associated with geographic coordinates and a minimal but consistent set of environmental contextual data.

The study makes use of the rich environmental clade annotations from SILVA SSU Ref, which strengthens the analysis, since the paper is dealing with the distribution of high taxon rank levels, where most marine microbial diversity resides in the form of such clades. By selecting a dataset that is minimally rich in contextual data for a meta-analysis, enrichment of the contextual data from other resources like megx.net and Environment Ontology (EnvO) was possible. The reported environmental parameters were complemented with an additional four; dissolved oxygen, nitrate, phosphate, and silicate concentrations, from interpolated World Ocean Atlas 05 values. This was done to study the potential effects of these parameters on the distribution of taxa. Although these parameters did not correlate significantly with community structures, it still serves as a proof-of-principle that if certain contextual data is present (x, y, z, t); integration of more data to the original dataset from diverse resources is possible. Moreover, using the originally provided sampling site descriptions, ontological habitat terms of EnvO at three different granularity levels were added to the available data. Visualization of these terms on non-metric multidimensional analysis plots revealed that even at phylum or class level community structures from contrasting habitats (estuarine water vs. ocean water) are very

different. On the contrary, sampling sites annotated with different habitat terms (marine coral reef vs. oceanic epipelagic zone) did not always reveal differing community structures, suggesting that the annotation of the sampling site was inaccurate. Although some satisfactory conclusions were reached at this study, regarding the ecological significance of high taxon rank levels, some questions remained unanswered. For example, it was not possible to distinguish between oceanic epipelagic zone habitats of geographically vast sampling sites neither at higher or lower ranks. Likewise, the correlations of environmental parameters to community structures were significant, but not very strong (maximum 0.467). These observations suggest either that these habitats are similar in their taxa composition and other environmental conditions are responsible for shaping the taxa composition, or that this dataset is not perfectly suitable for answering these questions. High-throughput tag sequencing may be more suited for capturing differences arising from geographical effects for similar habitats, and datasets with more continuous environmental gradients sampled may change the strength of the correlations observed.

In the impending revision of paper VIII, it is intended to perform an in-depth analysis of the distribution patterns of specific taxa, and interpret these patterns with respect to the environmental conditions. Selection of taxa of interest is ongoing, but the emerging candidates are subgroups of the SAR11 clade (Surface1, Deep1 etc…), marine *Actinobacteria* groups, *Gammaproteobacteria* OMG clades, and *Bacteroidetes* groups. These taxa are recurring in the marine realm, but despite years of study, the driving forces behind their distribution and functions is not clear [100]. With this extended analysis, we hope to supplement the generalized conclusions of this study with more specific knowledge generation.

OUTLOOK

Past challenges in microbiology included lack of powerful microscopes, or difficulties in cultivation. As microbiology is being transformed into a data-intensive science, the field is facing new challenges, like computer power and informatics. Massive sequencing projects such as the Earth Microbiome Project[9]or the Human Microbiome Project [101] promise up to 200,000 samples sequenced and petabytes of data production. It is not hard to see that, with the current deluge of data, microbiology will immensely benefit from standard practices to data management and integration. Researchers should understand that the promised deciphering of microbial life can only be realized if every study, from any environment, is described by rich and structured contextual data, to facilitate cross-comparison and meta-analysis of global microbial communities. The ability to obtain a comprehensive answer to the question "Have my microbes been seen before, and, if so, where, with whom, and what were they doing?" will provide great assistance to microbial research.

The steps taken by the GSC in the form of MIxS family of standards is a very important contribution to sequence data integration in microbiology. With MIxS, the three largest sequence realms (genomes, metagenomes and marker genes) are covered, with the possibility of easy extensions to new sequence data types, and new environments i.e. pangenomes, or indoor environments. While such extensions will constitute an important part of the future of MIxS standards; activities directed towards measuring and increasing the adoption of the standards will constitute a bigger part of the future workload. Currently, there is no shortage in software tools helping the community to be compliant with standards, but most certainly alternative implementations for different communities will improve adoption. The biggest challenge ahead of standards adoption is still improving the understanding of contextual data fields, and reporting of values in the

[9] www.earthmicrobiome.org

correct format. The "expected value" and "syntax" descriptions serve this purpose, but including a new "example" column in MIxS spreadsheets will be useful. More importantly, sequence submission systems should include a validation step. Currently, the INSDC database submission systems only check for the presence of mandatory MIxS fields, but do not check the content. The only validation mechanism available is external compliance and submission tools, which are only effective if used by submitters. Attending to communities' requests regarding improvements and questions to the standards should also improve the adoption of the standards. This is currently managed via the MIxS ticketing system[10], which gives options to report enhancements, defects, term requests, tasks, and questions. In future, a simple help-desk style service can be set up specifically to deal with questions regarding compliance. Compliance requirement by journals, as demonstrated by the Standards in Genomic Sciences journal[11] is also a successful, albeit more enforcing means of improving adoption. If reviewers and funding agencies were to look for and require standards compliance, this will be a more effective means of improving adoption, rather than journal requirements.

In order to track the adoption of the standards, it will be desirable to implement some measures. For example, tracking the number of citations on the MIGS/MIMS and MIxS publications, or surveying the number of compliant records from GenBank or ENA. The former measure may be problematic in future, if the standards become "generic", and users simply fail to cite the original publications. The latter measure would be more time-resistant, provided that surveying is extended to other specialized sequence database resources such as MG-RAST, megx.net or CAMERA [102]. In fact, with only three years since the initial publication of MIGS/MIMS, the number of compliant records seems promising. As of 07-September-2011, there are 16,262 MIMARKS, 3 MIMS, and 616 MIGS-compliant records in GenBank. Finally, the best success measure would be seeing global meta-analyses being performed with the help of standardized contextual data provided by MIxS.

It will also be desirable to see interplay of contextual data with data curation in microbial phylogeny and taxonomy. While the benefits of contextual data may seem ecology-

[10] http://mixs.gensc.org
[11] http://standardsingenomics.org/index.php/sigen

centric, the same guidelines can be extended to systematics. Identification of new species can benefit greatly from standardized metadata available to the researcher along with the rRNA sequence data. Phylogenetic trees can be visualized based on this phenotypic, physiological or biochemical contextual data, providing a new perspective on bacterial and archaeal taxa. Furthermore, having standardized habitat descriptions attached to each environmental sequence can help in the classification and taxonomy of yet uncultured clades of *Bacteria* and *Archaea*. A habitat-based classification can provide a natural and ecological organization to these taxa, as well as reducing the sightings of the uninformative uncultured clades. Finally, specific rRNA gene datasets that address the needs of specific communities, such as the marine microbiology, or soil microbiology, can be prepared reliably. Such reductions will become significant in near future, as the amount of sequence data that individual researchers can handle is limited. With its experience in habitat-specific subsets of genome sequences, the megx.net platform can be used as a first example of such datasets by curating habitat annotations for georeferenced rRNA sequences.

Another future topic, which is further away from the contextual data realm, but within the data management and integration practices for microbiology, would be the reconciliation of classifications between three rRNA databases. The aim of such a project would be, at the very least, to provide users with the “same” taxa names, regardless of the dataset and classification method used. As illustrated in the summary section, there is good agreement on cultured taxa; therefore such reconciliation should be a trivial task mainly involving better sharing of data. Reconciliation of environmental clades (clades without described cultivated representatives), however, will require more effort. A possible roadmap may involve examination of the conventions used by SILVA, RDP-II, and Greengenes for environmental clades, and determination of stable phylogenetic clusters that all parties agree on. For these stable clusters, either taxon names that are already ascribed (for example, originating from literature like Sargasso Sea clades) can be used, or new names based on an agreed convention can be devised. In order for these environmental taxon names to be persistent, it will also be desirable to build a registry, which can include data such as the original publication naming the clade, reference sequence for the clade, habitat patterns, or name changes to clusters.

There will be ample opportunities for meta-analysis projects, similar to Papers VII-VIII, in near future. New mega sequencing projects and sequencing centers, such as the Earth Microbiome Project, TARA Oceans[12], or Beijing Genomics Institute[13] , either require rich contextual data before any sample is accepted for sequencing, or collect extensive contextual data. Therefore, both compositional and functional studies can be performed on new and interesting datasets other than GOS, which perhaps will be more suited to testing of ecological theories for microorganisms.

To recapitulate the results and conclusions, the standards developed, and the curation work undertaken in this thesis work has provided new approaches and resources to deal with the sequence data deluge, and has shown the importance of standards, biocuration, and their cooperation. Continuation of these activities both in the same and new directions will benefit microbial research.

[12] http://oceans.taraexpeditions.org
[13] http://en.genomics.cn

APPENDIX

A list of scientific publications that are relevant to this thesis work, but not discussed in detail.

Meeting report: Metagenomics, Metadata and Meta-analysis" (M3) Workshop at the Pacific Symposium on Biocomputing 2010

Authors: Lynette Hirschman, Peter Sterk, Dawn Field, John Wooley, Guy Cochrane, Jack Gilbert A., Eugene Kolker, Nikos Kyrpides, Folker Meyer, Ilene Mizrachi, Yasukazu Nakamura, Susanna-Assunta Sansone, Lynn Schriml, Tatiana Tatusova, Owen White, Pelin Yilmaz

Published in: Standards in Genomic Sciences. 2010; 2(3): 357-360

Meeting report: GSC M5 roundtable at the 13th International Society for Microbial Ecology meeting in Seattle, WA, USA August 22-27, 2010.

Authors: Jack A Gilbert, Folker Meyer, Rob Knight, Dawn Field, Nikos Kyrpides, Pelin Yilmaz, John Wooley

Published in: Standards in Genomic Sciences. 2010; 3 (3): 235-239

Data shopping in an open marketplace: Introducing the Ontograter web application for marking up data using ontologies and browsing using facets

Authors: Norman Morrison, David Hancock, Lynette Hirschman, Peter Dawyndt, Bert Verslyppe, Nikos Kyrpides, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner, Jeff Grethe, Tim Booth, Peter Sterk, Goran Nenadic, Dawn Field

Published in: Standards in Genomic Sciences. 2011; 4 (2): 286-292

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to Frank Oliver Glöckner, for conducting a very interesting interview, which was my first step to MPI-Bremen, for accepting a biologist to the Bioinformatics group first as a M.Sc. student, and then as a Ph.D student, for introducing me to the wonderful world of taxonomy, for giving me a prominent role within the Genomic Standards Consortium, and finally being an excellent supervisor in general.

Next, I would like to acknowledge my thesis committee members, Prof. Dr. Matthias Ullrich and Dr. Wolfgang Ludwig for being present at each committee meeting, supporting my work and providing their valuable insights. Special thanks goes to Wolfgang Ludwig, for being an influential figure for a young researcher, and for taking the time to listen to my questions on taxonomy and rRNA.

I have worked closely both with the SILVA and megx.net teams, therefore I would like to thank everyone for their support and great working environment. Special thanks goes to Wolfgang Hankeln; for being a great supervisor and showing me around during my first months at the Microbial Genomics and Bioinformatics Group. Renzo Kottmann has a special place as a great colleague, supervisor, travel companion, and upon all a great friend. I am also highly indebted to colleagues from the Microbial Genomics Group and the Genomic Standards Consortium.

I have not forgotten friends and family, and they should receive special acknowledgements for contributing to my well-being. Hannah Marchant, Petra Pop Ristova, Jessika Füssel and Tim Kalvelage have provided their great friendship, my mother and father have provided endless support for 25 years, and finally Morten Iversen has shared life (the good, the bad and the ugly) with me, and proved to be a great chef while I skipped my duties during writing.

BIBLIOGRAPHY

[1] Oren, A. (2010) Concepts about Phylogeny of Microorganisms-an Historical Overview. In: Oren, A. and Papke, R. T. (Eds.) Molecular Phylogeny of Microorganisms. Caister Academic Press, Norfolk, pp. 1-21.

[2] Woese, C. R., Goldenfeld, N., How the Microbial World Saved Evolution from the Scylla of Molecular Biology and the Charybdis of the Modern Synthesis, *Microbiol Mol Biol Rev* **73** (2009), pp. 14-21.

[3] Breed, R. S., The Present Status of Systematic Bacteriology, *J Bacteriol* **15** (1928), pp. 143-163.

[4] Rossello-Mora, R., Amann, R., The species concept for prokaryotes, *FEMS Microbiol Rev* **25** (2001), pp. 39-67.

[5] Sapp, J. (2005) The Bacterium's Place in Nature. In: Sapp, J. (Eds.) Microbial phylogeny and evolution : concepts and controversies. Oxford, pp. 3-52.

[6] Avery, O. T., MacLeod, C. M., McCarty, M., Studies on the chemical nature of the substance inducing transformation of pneumococcal types, *The Journal of experimental medicine* **79** (1944), pp. 137.

[7] Lederberg, J., Lederberg, E. M., Zinder, N. D., Lively, E. R. *Recombination analysis of bacterial heredity*. in *Cold Spring Harbor Symposia on Quantitative Biology*. 1951.

[8] Crick, F. H. C., The Biological Replication of Macromolecules, *Symposia of the Society for Experimental Biology* **12** (1958), pp. 138-163.

[9] Zuckerkandl, E., Pauling, L., Molecules as documents of evolutionary history, *Journal of Theoretical Biology* **8** (1965), pp. 357-366.

[10] Woese, C. R., Fox, G. E., Phylogenetic structure of the prokaryotic domain: the primary kingdoms, *Proc Nat Acad Sci USA* **74** (1977), pp. 5088-5090.

[11] Fox, G. E., Stackebrandt, E., Hespell, R. B., Gipson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., Woese, C. R., Phylogeny of Prokaryotes, *Science* **209** (1980), pp. 457-463.

[12] Woese, C. R., Kandler, O., Wheelis, M. L., Towards a natural system of organisms: proposal for the domains *Archaea, Bacteria,* and *Eucarya, Proc Nat Acad Sci USA* **87** (1990), pp. 4576-4579.

[13] Mayr, E., Two empires or three?, *Proc Natl Acad Sci USA* **95** (1998), pp. 9720-9723.

[14] Woese, C. R., Default taxonomy - Ernst Mayr's view of the microbial world, *Proc Nat Acad Sci USA* **95** (1998), pp. 11043-11046.

[15] Boone, D. R., Castenholz, R. W., Garrity, G. M., Stanley, J. T. (2001) The Archaea and the Deeply Branching and Phototrophic Bacteria. In: Bergey's Manual of Systematic Bacteriology, Springer.

[16] Brenner, D. J., Krieg, N. R., Garrity, G. M., Staley, J. T. (2005) The Proteobacteria. In: Bergey's Manual of Systematic Bacteriology, Springer.

[17] Krieg, N. R., Staley, J. T., Brown, D. R., Hedlund, B. P., Paster, B. J., Ward, N. L., Ludwig, W., Whitma, W. B. (2010) The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes. In: Bergey's Manual of Systematic Bacteriology, Springer.

[18] Vos, P. D., Garrity, G. M., Jones, D., Krieg, N. R., Ludwig, W., Rainey, F. A., Schleifer, K.-H., Whitman, W. B. (2009) The Firmicutes. In: Bergey's Manual of Systematic Bacteriology, Springer.

[19] Jones, M. L., On the Vestimentifera, new phylum: six new species, and other taxa, from hydrothermal vents and elsewhere, *Bulletin of the Biological Society of Washington* **6** (1985), pp. 117-158.

[20] Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., Stahl, D. A., Microbial ecology and evolution: a ribosomal RNA approach, *Annu Rev Microbiol* **40** (1986), pp. 337-365.

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

[22] Pace, N. R., Stahl, D. A., Olsen, G. J., Lane, D. J., Analyzing natural microbial populations by rRNA sequences, *ASM News* **51** (1985), pp. 4-12.

[23] Curtis, T. P., Sloan, W. T., Scannell, J. W., Estimating prokaryotic diversity and its limits, *Proc Nat Acad Sci USA* **99** (2002), pp. 10494-10499.

[24] Rappe, M. S., Giovannoni, S. J., The uncultured microbial majority, *Annu Rev Microbiol* **57** (2003), pp. 369-394.

[25] Amann, R. I., Ludwig, W., Schleifer, K. H., Phylogenetic identification and *in-situ* detection of individual microbial cells without cultivation, *Microbiol Rev* **59** (1995), pp. 143-169.

[26] Orcutt, B., George, D., Dayhoff, M., Protein and nucleic acid sequence database systems, *Annual Review of Biophysics and Bioengineering* **12** (1983), pp. 419-441.

[27] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., GenBank, *Nucleic Acids Res* **39** (2011), pp. D32-D37.

[28] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., Cochrane, G., The European Nucleotide Archive, *Nucleic Acids Res* **39** (2011), pp. D28-D31.

[29] Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K., Nakamura, Y., DDBJ progress report, *Nucleic Acids Res* **39** (2011), pp. D22-D27.

[30] Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y., The International Nucleotide Sequence Database Collaboration, *Nucleic Acids Res* **39** (2011), pp. D15-D18.

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

[32] Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., Tiedje, J. M., The Ribosomal Database Project: improved alignments and new tools for rRNA analysis, *Nucleic Acids Res* **37** (2009), pp. D141-145.

[33] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G. L., Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, *Appl Environ Microbiol* **72** (2006), pp. 5069-5072.

[34] Mardis, E. R., Next-Generation DNA Sequencing Methods, *Annual Review of Genomics and Human Genetics* **9** (2008), pp. 387-402.

[35] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. T., Dewell, S. B., de Winter, A., Drake, J., Du, L., Fierro, J. M., Forte, R., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Hutchison, S. K., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lee, W. L., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Reifler, M., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Willoughby, D. A., Yu, P. G., Begley, R. F., Rothberg, J. M., Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **441** (2006), pp. 120-120.

[36] Rusk, N., Torrents of sequence, *Nat Meth* **8** (2011), pp. 44-44.

[37] Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., Knight, R., Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex, *Nat Meth* **5** (2008), pp. 235-237.

[38] von Wintzingerode, F., Gobel, U. B., Stackebrandt, E., Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis, *FEMS Microbiol Rev* **21** (1997), pp. 213-229.

[39] Rochelle, P. A., Cragg, B. A., Fry, J. C., Parkes, R. J., Weightman, A. J., Effect of sample handling on estimation of bacterial diversity in marine sediments by 16S rRNA gene sequence analysis, *FEMS Microbiol Ecol* **15** (1994), pp. 215-225.

[40] Liesack, W., Weyland, H., Stackebrandt, E., Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria, *Microb Ecol* **21** (1991), pp. 191-198.

[41] Tebbe, C. C., Vahjen, W., Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast, *Appl Environ Microbiol* **59** (1993), pp. 2657-2665.

[42] Farrelly, V., Rainey, F. A., Stackebrandt, E., Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species, *Appl Environ Microbiol* **61** (1995), pp. 2798-2801.

[43] Brunk, C. F., Avaniss-Aghajani, E., Brunk, C. A., A computer analysis of primer and probe hybridization potential with bacterial small-subunit rRNA sequences, *Appl Environ Microbiol* **62** (1996), pp. 872-879.

[44] Reysenbach, A. L., Giver, L. J., Wickham, G. S., Pace, N. R., Differential amplification of rRNA genes by polymerase chain reaction, *Appl Environ Microbiol* **58** (1992), pp. 3417-3418.

[45] Chandler, D. P., Fredrickson, J. K., Brockman, F. J., Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries, *Mol Ecol* **6** (1997), pp. 475-482.

[46] Shuldiner, A. R., Nirula, A., Roth, J., Hybrid DNA artifact from PCR of closely related target sequences, *Nucleic Acids Res* **17** (1989), pp. 4409.

[47] Cariello, N. F., Thilly, W. G., Swenberg, J. A., Skopek, T. R., Deletion mutagenesis during polymerase chain reaction: dependence on DNA polymerase, *Gene* **99** (1991), pp. 105-108.

[48] Gelfand, D. H. (1992) Taq DNA polymerase, PCR technology. In: Erlich, H. A. (Eds.) Principles and Application for DNA Amplification. Freeman and Company, New York, pp. 17-22.

[49] Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., Frazer, K. A., Evaluation of next generation sequencing platforms for population targeted sequencing studies, *Genome Biol* **10** (2009), pp. R32.

[50] Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., Knight, R., UCHIME improves sensitivity and speed of chimera detection, *Bioinformatics* **27** (2011), pp. 2194-2200.

[51] Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F., Sloan, W. T., Accurate determination of microbial diversity from 454 pyrosequencing data, *Nat Meth* **6** (2009), pp. 639-641.

[52] Walters, W. A., Caporaso, J. G., Lauber, C. L., Berg-Lyons, D., Fierer, N., Knight, R., PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers, *Bioinformatics* **27** (2011), pp. 1159-1161.

[53] Lee, Z. M.-P., Bussema, C., Schmidt, T. M., rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea, *Nucleic Acids Res* **37** (2009), pp. D489-D493.

[54] Wang, Y., Zhang, Z., Ramanan, N., The actinomycete Thermobispora bispora contains two distinct types of transcriptionally active 16S rRNA genes, *J Bacteriol* **179** (1997), pp. 3270-3276.

[55] Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., Jin, Z., Lee, P., Yang, L., Poles, M., Brown, S. M., Sotero, S., DeSantis, T.,

Brodie, E., Nelson, K., Pei, Z., Diversity of 16S rRNA genes within individual prokaryotic genomes, *Appl Environ Microbiol* (2010), pp. AEM.02953-02909.

[56] Stackebrandt, E., Ebers, J., Taxonomic parameters revisited: tarnished gold standards, *Microbiology Today* **33** (2006), pp. 152-155.

[57] Ludwig, W., Schleifer, K., Phylogeny of *Bacteria* beyond the 16S rRNA standard, *ASM News* **65** (1999), pp. 752-757.

[58] Doolittle, W. F., The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them, *Philos Trans R Soc Lond B Biol Sci* **364** (2009), pp. 2221-2228.

[59] Gogarten, J. P., Doolittle, W. F., Lawrence, J. G., Prokaryotic Evolution in Light of Gene Transfer, *Mol Biol Evol* **19** (2002), pp. 2226-2238.

[60] Gest, H., Bacterial classification and taxonomy: a'primer'for the new millennium, *Microbiology Today* **26** (1999), pp. 70-72.

[61] Ludwig, W., Schleifer, K. H. (2005) Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes. In: Sapp, J. (Eds.) Microbial phylogeny and evolution, concepts and controversies. Oxford university press, New York, USA, pp. 70-98.

[62] Konstantinidis, K. T., Tiedje, J. M., Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead, *Curr Opin Microbiol* **10** (2007), pp. 504-509.

[63] Kottmann, R., Kostadinov, I., Duhaime, M. B., Buttigieg, P. L., Yilmaz, P., Hankeln, W., Waldmann, J., Glöckner, F. O., Megx.net: integrated database resource for marine ecological genomics, *Nucleic Acids Res* **38** (2010), pp. D391-D395.

[64] Rossello-Mora, R., Updating Prokaryotic Taxonomy, *J Bacteriol* **187** (2005), pp. 6255-6257.

[65] Jeraldo, P., Chia, N., Goldenfeld, N., On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys, *Environ Microbiol* (2011), pp. no-no.

[66] Tringe, S. G., Hugenholtz, P., A renaissance for the pioneering 16S rRNA gene, *Curr Opin Microbiol* (2008), pp.

[67] Brock, T. D., Brock, M. L., Relationship between environmental temperature and optimum temperature of bacteria along a hot spring thermal gradient, *J Appl Microbiol* **31** (1968), pp. 54-58.

[68] Cho, J.-C., Tiedje, J. M., Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil, *Appl Environ Microbiol* **66** (2000), pp. 5448-5456.

[69] ZoBell, C. E., Johnson, F. H., The influence of hydrostatic pressure on the growth and viability of terrestrial and marine bacteria, *J Bacteriol* **57** (1949), pp. 179.

[70] von Mering, C., Hugenholtz, P., Raes, J., Tringe, S. G., Doerks, T., Jensen, L. J., Ward, N., Bork, P., Quantitative phylogenetic assessment of microbial communities in diverse environments, *Science* **315** (2007), pp. 1126-1130.

[71] Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., Brown, J. H., A latitudinal diversity gradient in planktonic marine bacteria, *Proc Nat Acad Sci USA* **105** (2008), pp. 7774-7778.

[72] Pommier, T., Canbäck, B., Riemann, L., Boström, K. H., Simu, K., Lundberg, P., Tunlid, A., Hagström, Å., Global patterns of diversity and community structure in marine bacterioplankton, *Mol Ecol* **16** (2007), pp. 867-880.

[73] Auguet, J.-C., Barberan, A., Casamayor, E. O., Global ecological patterns in uncultured Archaea, *ISME J* **4** (2010), pp. 182-190.

[74] Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., Knight, R., Bacterial community variation in human body habitats across space and time, *Science* **326** (2009), pp. 1694-1697.

[75] Lozupone, C. A., Knight, R., Global patterns in bacterial diversity, *Proc Nat Acad Sci USA* **104** (2007), pp. 11436-11440.

[76] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., Frame, M., Data sharing by scientists: practices and perceptions, *PLoS ONE* **6** (2011), pp. e21101.

[77] Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., Brazma, A., Brinkman, R. R., Michael Clark, A., Deutsch, E. W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J. M., Hardy, N. W., Hermjakob, H., Julian, R. K., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Novere, N. L., Leebens-Mack, J., Lewis, S. E., Lord, P., Mallon, A.-M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J. M., Robertson, D. G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R. H., Schober, D., Smith, B., Snape, J., Stoeckert, C. J., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., Wiemann, S., Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project, *Nat Biotechnol* **26** (2008), pp. 889-896.

[78] Janetzki, S., Britten, C. M., Kalos, M., Levitsky, H. I., Maecker, H. T., Melief, C. J. M., Old, L. J., Romero, P., Hoos, A., Davis, M. M., MIATA Minimal Information about T Cell Assays, *Immunity* **31** (2009), pp. 527-528.

[79] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M., Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, *Nat Genet* **29** (2001), pp. 365-371.

[80] Field, D., Hughes, J., Cataloguing our current genome collection, *Microbiology* **151** (2005), pp. 1016-1019.

[81] Field, D., Morrison, N., Glöckner, F. O., Kottmann, R., Cochrane, G., Vaughan, R., Garrity, G., Cole, J., Hirschman, L., Schriml, L., Mizrachi, I., Federhen, S., Schindel, D., Miller, S., Hebert, P., Ratnasingham, S., Hanner, R., Amaral-Zettler,

L., Sogin, M., Ashburner, M., Lewis, S., Smith, B., Working together to put molecules on the map, *Nature* **453** (2008), pp. 978-978.

[82] Field, D., Sansone, S. A., A special issue on data standards, *OMICS* **10** (2006), pp. 84-93.

[83] Field, D., Wilson, G., van der Gast, C., How do we compare hundreds of bacterial genomes?, *Curr Opin Microbiol* **9** (2006), pp. 499-504.

[84] Martiny, J. B. H., Field, D., Ecological perspectives on the sequenced genome collection, *Ecology Letters* **8** (2005), pp. 1334-1345.

[85] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glockner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S. A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., Wipat, A., The minimum information about a genome sequence (MIGS) specification, *Nat Biotechnol* **26** (2008), pp. 541-547.

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

[87] Euzéby, J. P., List of Bacterial Names with Standing in Nomenclature: a Folder Available on the Internet, *Int J Syst Bacteriol* **47** (1997), pp. 590-592.

[88] Morris, R. M., Vergin, K. L., Cho, J. C., Rappe, M. S., Carlson, C. A., Giovannoni, S. J., Temporal and Spatial Response of Bacterioplankton Lineages to Annual Convective Overturn at the Bermuda Atlantic Time-Series Study Site, *Limnol Oceanogr* **50** (2005), pp. 1687-1696.

[89] Takai, K., Moser, D. P., DeFlaun, M., Onstott, T. C., Fredrickson, J. K., Archaeal diversity in waters from deep South African gold mines, *Appl Environ Microbiol* **67** (2001), pp. 5750-5760.

[90] Köhler, T., Stingl, U., Meuser, K., Brune, A., Novel lineages of Planctomycetes densely colonize the alkaline gut of soil-feeding termites (Cubitermes spp.), *Environ Microbiol* **10** (2008), pp. 1260-1270.

[91] Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G. U. Y., Fensome, R. A., Fredericq, S., James, T. Y., Karpov, S., Kugrens, P., Krug, J., Lane, C. E., Lewis, L. A., Lodge, J., Lynn, D. H., Mann, D. G., McCourt, R. M., Mendoza, L., Moestrup, Ø., Mozley-Standridge, S. E., Nerad, T. A., Shearer, C. A., Smirnov, A. V., Spiegel, F. W., Taylor, M. F. J. R., The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists, *Journal of Eukaryotic Microbiology* **52** (2005), pp. 399-451.

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

[93] Yooseph, S., Nealson, K. H., Rusch, D. B., McCrow, J. P., Dupont, C. L., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., Beeson, K., Williamson, S. J., Tovchigrechko, A., Allen, A. E., Zeigler, L. A., Sutton, G., Eisenstadt, E., Rogers,

Y.-H., Friedman, R., Frazier, M., Venter, J. C., Genomic and functional adaptation in surface ocean planktonic prokaryotes, *Nature* **468** (2010), pp. 60-66.

[94] Konstantinidis, K. T., Tiedje, J. M., Towards a genome-based taxonomy for prokaryotes, *J Bacteriol* **187** (2005), pp. 6258-6264.

[95] Ludwig, W., Klenk, H. P. (2001) A phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone, D. R. and Castenholz, R. W. (Eds.) The *Archaea* and the deeply branching and phototrophic *Bacteria*. Springer-Verlag, New York, pp. 49-65.

[96] Ludwig, W., Rossello-Mora, R., Aznar, R., Klugbauer, S., Spring, S., Reetz, K., Beimfohr, C., Brockmann, E., Kirchhof, G., Dorn, S., Bachleitner, M., Klugbauer, N., Springer, N., Lane, D., Nietupsky, R., Weizenegger, M., Schleifer, K.-H., Comparative sequence analysis of 23S rRNA from *Proteobacteria*, *Syst Appl Microbiol* **18** (1995), pp. 164-188.

[97] Wooley, J. C., Godzik, A., Friedberg, I., A primer on metagenomics, *PLoS Comput Biol* **6** (2010), pp. e1000667.

[98] Walker, A., Genome watch: Singled out, *Nat Rev Micro* **9** (2011), pp. 485-485.

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

[100] Fuhrman, J., Hagström, Å. (2008) Bacterial and archaeal community structure and its patterns. In: Kirchman, D. L. (Eds.) Microbial ecology of the oceans. 2, Wiley-Blackwell, New York, pp. 45-90.

[101] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., Gordon, J. I., The Human Microbiome Project, *Nature* **449** (2007), pp. 804-810.

[102] Sun, S., Chen, J., Li, W., Altinatas, I., Lin, A., Peltier, S., Stocks, K., Allen, E. E., Ellisman, M., Grethe, J., Wooley, J., Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource, *Nucleic Acids Res* (2010), pp.