Proceedings of the International Workshop on Ribosomal RNA Technology, April 7-9, 2008, Bremen, Germany

Linda Amaral-Zettler[1], Jörg Peplies[2], Alban Ramette[3], Bernhard Fuchs[4], Wolfgang Ludwig[5] and Frank Oliver Glöckner[6,7*]

[1] The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543 (USA)

[2] Ribocon GmbH, D-28359 Bremen (Germany)

[3] Habitat Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen (Germany)

[4] Dept. of Molecular Ecology, Max Planck Institute for Marine Microbiology, D-28359 Bremen (Germany)

[5] Lehrstuhl für Mikrobiologie, Technische Universität München, D-85350 Freising (Germany)

[6] Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen (Germany)

[7] Jacobs University Bremen gGmbH, D-28759 (Germany)

* Corresponding author:

Frank Oliver Glöckner
Max Planck Institute for Marine Microbiology
Celsiusstrasse 1, D-28359 Bremen (Germany)
Tel: +49 421 2028970
Fax: +49 421 2028580
Email: fog@mpi-bremen.de

Keywords
ribosomal RNA, workshop proceedings, databases, phylogeny, biogeography, technology, diversity, ecology

Abstract

Thirty years have passed since Carl Woese proposed three primary domains of life based on the phylogenetic analysis of ribosomal RNA genes. Adopted by researchers worldwide, ribosomal RNA has become the "gold-standard" for molecular taxonomy, biodiversity analysis and the identification of microorganisms. The more than 700,000 rRNA sequences in public databases constitute an unprecedented hallmark of the richness of microbial biodiversity on earth. The International Workshop on Ribosomal RNA Technology convened on April 7-9, 2008 in Bremen, Germany (http://www.arb-silva.de/rrna-workshop) to summarize the current status of the field and strategize on the best ways of proceeding on both biological and technological fronts. In five sessions, 26 leading international speakers and ~120 participants representing diverse disciplines discussed new technological approaches to address three basic ecological questions: "Who is out there?" "How many are there?" and "What are they doing?"

Introduction

Carl Woese's discovery of the third domain of life inferred using ribosomal RNA (rRNA) molecules 30 years ago lead to the emergence of rRNA-based technologies that would prove to transform the field of Microbiology. With this discovery began the dawn of a new era in molecular taxonomy – flooding molecular databases with a deluge of rRNA gene sequences totalling over 700,000 today (Figure 1). Ribosomal RNA-based gene phylogenies have largely stood the test of time in describing the evolutionary relationships between organisms, phylogenetic probes based on fluorescently-labeled oligonucleotides complementary to rRNA inside cells have provided microbiologists with a quantitative means of assessing microbial diversity in nature, and rRNA gene-based tag pyrosequencing has enabled microbial biogeography and ecological diversity studies on a scale never before imagined.

The International Workshop on Ribosomal RNA Technology revisited the impact that rRNA has had on the fields of phylogenetics, bioinformatics, biogeography, technology, microbial diversity and ecology (for the list of speakers and affiliations, see supplementary Table 1). Among the advances discussed included updated and improved databases and software allowing for enhanced alignment of rRNA gene sequences, newer and faster algorithms that permit the construction of large scale phylogenetic trees on the order of tens of thousands of sequences at one time, and improved microscopic methods that open the window into structure-function analyses of microbial consortia. It is likely that few other molecules have been as successful in bringing together so many different kinds of scientists in addressing such a large diversity of questions. With all the recent advances in the "omics" of molecular biotechnology, one might imagine that we have exhausted the possibilities that lie ahead with respect to the development and application of new rRNA technologies of the future. However, judging from topics emerging from this workshop, that day seems to still be in the distant future.

Databases

Sequencing rRNA genes is currently the method of choice for phylogenetic reconstruction, nucleic acid-based detection and quantification of microbial diversity. The resulting exponential increase of publicly available rRNA sequences demands specialized databases and advanced data integration technologies. Three main database projects provide access to large datasets of rRNA sequences and alignments. All projects offer at least one small subunit (SSU) rRNA dataset, but vary in the volume of sequences, quality checking, alignments, and frequency of updates.

The Ribosomal Database Project II (RDP II, http://rdp.cme.msu.edu/) at Michigan State University in East Lansing, MI [5], focuses on bacterial and archaeal SSU rRNA sequences. Navigation through sequence space is supported by an advanced taxonomic browser. The RDP II maintains web-based tools such as the RDP classifier, seqmatch, probe match, library compare and tree builder to allow researchers to analyze their sequences. They have recently added an interactive heatmap tool for visualizing the relationships between thousands of sequences at one time. The myRDP space allows users to maintain and analyse their own sequences e.g. with high-throughput sequence processing pipelines for Sanger and massively parallel sequencing technologies like pyrosequencing.

Greengenes (http://greengenes.lbl.gov/), maintained by the Lawrence Berkeley National Laboratory in Berkeley, CA [8], has its roots in a combination of early RDP datasets and ARB alignments. Their intention was to build a chimera-checked and aligned database for taxonomic microarrays. The database hosts only nearly full length (>1250 bases) SSU rRNA sequences of bacterial and archaeal origin. The sequences can be accessed on the webpage by search and browse functions or downloaded as a database compatible with the ARB software suite. Multiple taxonomic classifications are available for each sequence entry. Currently, a PhyloChip with more than one million probes and over 30,000 OTUs is available using the Affymetrix technology. This, in combination with the new Phylotrac software

(www.phylotrac.org), will allow quantitative tracking of microbial communities in the environment.

The SILVA system (from Latin *silva*, forest, www.arb-silva.de), hosted by the Max Planck Institute for Marine Microbiology in Bremen, Germany [31], is a comprehensive web resource for up to date, quality controlled databases of aligned SSU and large subunit (LSU) rRNA sequences from *Bacteria, Archaea* and *Eukarya* that are fully compatible with the ARB software suite. All sequences are checked for anomalies and carry a rich set of metadata. An intuitive ranking system allows the user to get a rapid overview of sequence quality. SILVA integrates multiple taxonomic classifications and the latest validly described nomenclature for every entry. Sequences are flagged if they belong to a cultivated organism, a type strain, or a genome project. The online automated aligner SINA (Silva IncremeNtal Aligner) allows rapid and accurate alignment of user sequences. A taxonomic browser and advanced search functions can be used for sequence retrieval in aligned FASTA or ARB database formats. The ARB software suite (www.arb-home.de) provides extended analysis functions including phylogenetic tree reconstructions, alignments, similarity searches, probe design/probe match and improved visualisation tools.

The StrainInfo.net bioportal (www.straininfo.net) hosted by Ghent University [6] concentrates on establishing automated ways to collect and integrate all information that is available for microorganisms deposited into a global network of Biological Resource Centers (BRCs). It helps bridge the gap between the genotypic and phenotypic world and enables an integrative approach to the ecological and biogeographical distribution of species. The bioportal offers advanced search functions and data crawling of over more than 50 BRCs with extensive link-outs to EMBL, GenBank and DDBJ, as well as the SILVA rRNA databases. Features including predefined workflows for retrieving all strains subjected to genome sequencing, as well as web services are offered for integrated access to biological data.

In summary the opening session reviewed an impressive variety of high level rRNA technology-based resources. New technologies and pipelines have been designed and implemented that help us cope with the deluge of data entering public databases every day. Broad integrative approaches promise a comprehensive picture about the diversity and function of microorganisms with respect to their environmental surroundings and distributions.

Phylogeny

During the last decades, comparative rRNA sequence analysis ‘evolved’ from an expensive specialist’s technique to a cost-effective routine procedure for elucidating phylogenetic relationships that has helped to transform microbial taxonomy and identification. The second session of the workshop critically examined and evaluated the power and limitations of the rRNA approach, presented possible supplementary phylogenetic markers and tools, introduced new powerful tree-building approaches, and reviewed new resources for diagnostic rRNA targeted probes and primers.

A brief overview of the methodological history of comparative rRNA sequencing revealed the importance of appropriate data analysis software packages and pipelines, as well as the need for comprehensive, regularly-maintained integrated databases of curated and annotated sequence data. One such resource is the ARB software package that has been maintained and improved-upon for the past 15 years and comprises tools for database management, sequence analysis, phylogeny reconstruction, definition and evaluation of diagnostic sequence features [24]. The major components of the ARB software package were reviewed, recent software developments discussed and the power and limitations of the rRNA approach were summarized.

Since the introduction of comparative rRNA sequencing, there has been a continuous debate concerning the justification of a single marker molecule for inferring organismal phylogeny

and assigning taxonomy based on this underlying phylogeny [23]. The data coming out of completed and ongoing full-genome sequencing projects allow for the discovery and evaluation of alternative marker genes and molecules. There exists only a small set of genes fulfilling the requirements of universal phylogenetic markers representing the conserved core of the genome. Examples are genes that code for translation initiation, elongation and release factors, RNA polymerases, heat shock proteins, proton-translocating ATPases, recA, and few others. Despite differences in the results obtained by comparative phylogenetic analyses of such alternative markers, there is still global support of the rRNA-based picture of the major phylogenetic groups. The sparsely populated databases of alternative genes hamper analyses based on alternative markers. Consequently, rRNA-based approaches will remain the gold standard for phylogeny, taxonomy and identification into the next generation [22].

In modern metagenomics projects, rRNA genes provide valuable information for assembly of individual sequence contigs and assignment of contigs to taxonomic entities. However, the power of the rRNA approach is severely limited by the fact that only about 1% of the retrieved metagenomic sequences include rRNA genes. Thus, the number of contigs that can be assigned to a taxon is rather limited. MLTreeMap (http://mltreemap.embl.de/;[38]) uses a maximum likelihood procedure to derive the underlying taxonomic composition of the organisms represented by the sequences sampled in such experiments. Multiple protein-coding marker genes are used to map the respective fragments representing environmental organisms to a tree based on concatenated marker sequences available from fully sequenced genomes. A selection of protein-coding genes can be used to complement the rRNA markers for taxon assignment of environmental contigs, particularly when focusing on taxonomically informative, universal and highly conserved proteins. The method can handle fragmented open reading frames and limited assembly, and allows tracking of microbial lineages through various environments targeted by metagenomics studies. Combined with PCR-based studies using rRNA genes, current metagenomics data seem to indicate that microbial lineages have

pronounced and stable habitat preferences – more so than what would be expected from the study of easily cultivable microbial ‘generalists’.

The rapid increase of available sequence data in general, and rRNA primary structures in particular, requires efficient phylogenetic inference methods. Until recently, limitations in computing speed and processing power hindered the widespread application of Maximum Likelihood (ML)-based approaches to large datasets. These barriers have been recently overcome by new powerful tree-building methods such as RAxML (Randomized Axelerated Maximum Likelihood; http://icwww.epfl.ch/~stamatak/; [35]). One of the most time-consuming operations in tree reconstruction is the computation of support values for tree topologies. To overcome this limitation, novel rapid bootstrap heuristics were developed and implemented in RAxML. These heuristics provide qualitatively comparable results while accelerating the search process 15-fold. ML inference can be further accelerated by efficient parallelization on platforms such as IBM BlueGene supercomputer architecture, as well as on Linux clusters. Integration of a rapid bootstrap procedure and application of fast approximations of the phylogenetic ML function further increase speed.

The rRNA-targeted probe and PCR technology for species identification is currently routinely applied in a variety of formats. Along with the rapidly growing databases, continuous *in silico* evaluation of the specificity for already established or recently designed probes and primers is essential. The Probe Library and Evaluation System (PLEASE!, http://please.arb-home.de/webstart), a new Client-/Server System, addresses these needs. PLEASE! is a major advancement of the old ARB Probe Library that implements functions for the retrieval of all potential taxon and group specific signature sequences for all phylogenetic levels extracted from the underlying database and tree. It provides a comprehensive Graphical User Interface (GUI) driven application for the *in silico* evaluation of the probe sequence specificity against several local or remotely curated rRNA databases such as ARB/SILVA, and Greengenes. As a Java Webstart application, PLEASE! does not require time consuming proprietary hardware

or software installation. A new web application ARB - ProbeMatchOnline (PMO; http://pmo.arb-home.de) allows the fast search of probe sequences in expert maintained secondary rRNA databases.

Data Analysis and Biogeography

The session “Data Analysis and Biogeography” provided an overview of analytical strategies that focus on the contextual interpretation of SSU rRNA sequence data taking into account environmental, spatial or temporal parameters. This session presented established methods and introduced new unpublished techniques that show promising applications for ecological data analysis and biogeographic modelling using large sequence datasets.

The Generalized Regression Analysis and Spatial Prediction (GRASP) package, developed in Splus (commercial) and R (open source) statistical packages, allows for integrative analyses based on generalized regression and spatial predictions [17]. The method enables the transformation of point observations of species distribution into spatial predictions. It produces several outputs to visualize the selected models, variable contributions, and to perform cross-validation of the models. Generalized Additive Models (GAM) used in GRASP offer a good compromise between Generalized Linear Models (GLM) and Neural Networks (NN). In addition, a data-mining system that correlates genetic patterns in genomes and metagenomes with contextual environmental marine data (Megx.net and www.metafunctions.org [19]) was showcased with OceanDB, an environmental database describing global oceans with links to metagenomic data.

The contextual interpretation of a large data set based on short DNA sequence tags, e.g. obtained via 454 massive tag pyrosequencing of the V6 hypervariable region of bacterial SSU rRNA genes was illustrated on a dataset consisting of sixteen coastal sediment samples of ~ 15,000-20,000 sequence tags per sample (Gobet et al. unpublished data). Multivariate analyses including variation partitioning, partial canonical analyses, redundancy analyses, and

non-metric multidimensional scaling successfully extracted ecological patterns from this complex dataset and related them to a large number of contextual parameters. After the effects of co-varying factors were determined, the main causes of the temporal and spatial variation in the large dataset were identified. Overall, the study highlighted the usefulness of the 454 massive tag sequencing approach in generating large sequence datasets in a targeted ecosystem and of using multivariate analyses to interpret diversity patterns in their ecological context.

A new integrated tool, RAMI (Latin, *branches*), aims to help reveal the phylogenetic and spatial structure of microbial communities based on closely related sequences, i.e. microdiverse clusters (www.acgt.se/online.html, Pommier et al. unpublished data). RAMI uses phylogenetic tree-derived patristic distances (branch lengths) to identify microdiverse clusters, to characterize their structure and genetic variation, and to evaluate inter- and intra-cluster relationships. To demonstrate RAMI's ability to efficiently identify and characterize microdiverse clusters, several clone libraries based on 16S rRNA gene sequences from coastal samples distributed worldwide were analyzed to give a biogeographic perspective to the structure of marine microbial communities.

Multivariate tools are available for analyzing high-throughput sequence data in an environmental context, but better integration of existing methods in microbial ecology are required to make further advances in the field of microbial ecology [32]. To this end, a new analytical framework that quantifies and tests the significance of the structuring factors affecting community diversity at multiple taxonomic levels provides promise (Ramette, unpublished). This flexible analytical framework was illustrated on a pre-existing SSU rRNA-gene dataset obtained from microbial communities associated with obesity in mice [18]. This new approach delivers finer ecological and evolutionary insights compared with traditional statistical tools used by microbial ecologists to compare clone libraries. Future applications of this strategy are anticipated, particularly in ecology, evolution and taxonomy.

Overall, the increasing number of methods and strategies coming online to data mine large numbers of samples and DNA sequences is an indication that microbial ecologists are moving towards better exploitation of the data treasure they possess. Although the task at hand seems daunting at first given the unprecedented volume and speed with which data are accumulating in public databases, existing methodological knowledge and new analytical approaches offer a bright future to studies in microbial biogeography employing SSU rRNA sequence datasets. Emphasis on research and teaching are urgently needed that will further develop statistical tools for analyzing large sequence datasets in the context of associated contextual data.

Technology

About two decades ago, the pioneering studies of Pace, Olsen, Stahl, Giovannoni, and Ward [13, 26, 27, 34, 39] initiated the age of rRNA-based technologies in environmental microbiology. In the early years, primarily few rRNA genes from environmental samples were cloned and sequenced to access microbial diversity of uncultivated organisms, however, today we have a broad spectrum of powerful molecular tools available to not only address molecular diversity, but also functional aspects of microbial communities.

New tag sequencing technologies producing hundreds of thousands of short tag sequences of a selected marker gene in just a few hours offer a completely new scale of microbial diversity analysis and can provide much more comprehensive answers compared to former cultivation-independent approaches. Such large amounts of data demand sophisticated bioinformatics tools and appropriate computer hardware for processing. Standards for data description and storage must also be re-evaluated. The most prominent of these new techniques is pyrosequencing. An initial application of this technology based on sequencing of hypervariable regions of the bacterial V6 SSU rRNA enabled the discovery of the “rare biosphere” [33]. Now we must answer the question of what role these highly-divergent, low-

abundance organisms play in the environment. In doing so, we are knocking on the door of one of the central questions in environmental microbiology.

Sequencing of genetic markers for identification purposes (also known as “DNA barcoding”) has been widespread in the field of environmental microbiology since the late 1980’s, largely driven by limitations in the cultivation of the vast majority of the microorganisms. In contrast, sequencing of genetic markers in fields outside of microbiology has only recently become more mainstream. The Consortium for the Barcode of Life (CBOL; www.barcoding.si.edu) has been a major driver in this effort targeting primarily the mitochondrial Cytochrome Oxidase One (COI) genes for animals but also other phylum-specific markers as needed. Eukaryotic DNA barcoding campaigns, as well as the global iBOL programme, have agreed on global standards in data quality, vouchering (both DNA and specimen), taxonomy, and databasing, developed and implemented by CBOL. This resulted in INSDC’s (International Nucleotide Sequence Database Collaboration) adoption of the keyword “BARCODE”, reserved for entries in compliance with all these criteria.

One of the most powerful tools among the molecular techniques for the investigation of microbial communities is represented by rRNA-targeted Fluorescence *in situ* Hybridization (FISH). It combines molecular identification with microscopic visualization of selected populations and even enables a quantification of these populations on the cellular level. While in the early years, sensitivity was a major constraint, recent improvements achieved by enzyme-mediated signal amplification now allow for high detection efficiency of microbial cells in oligotrophic habitats or in cases of high background signals [30]. Also, detailed systematic evaluations of selected parameters influencing a hybridization reaction such as target accessibility are available [4].

Another central aspect of rRNA-based *in situ* hybridization is probe quality in terms of specificity. The rRNA databases are dramatically growing over time and with this comes the question “Are commonly used probes - often designed nearly a decade ago - still valid?” The

answer is “more than could be expected”, but nevertheless, we should periodically evaluate probe efficacy including the use of 23S rRNA as a target [2, 20, 21]. High-quality, curated databases required for this task are now available [31].

Besides the questions “Who is out there?” and “How many are there?”, the questions “What are they doing?” and “How are they interacting?” are of major interest, since only knowledge of the function of the organisms will ultimately allow us to understand an ecosystem. In recent years, various techniques have been introduced for the use of combined FISH identification and functional analysis of so far uncultivated microorganisms using substrate-mediated labeling techniques. Raman microspectroscopy based on stable-isotope-labeling of cells combined with FISH [14] is among the most recent developments and offers some clear advantages such as quantitative detection of incorporation rates and information on label incorporation into certain compound classes. A potential further development is represented by the combination of DNA microarray technology (PhyloChips) and Raman microspectroscopy for high-throughput analysis.

The session closed with contributions from two company representatives who act as technology-providers for academia and industry: biomers.net and Zeiss MicroImaging. The company biomers.net (Ulm, Germany; www.biomers.net) specializes in custom-made synthesis of modified and unmodified biopolymers. This basic biochemistry is an integral aspect of modern molecular microbiology since specific PCR or *in situ* hybridization experiments rely on successful synthesis of defined oligonucleotides. However, users are often not aware of the complexity of the synthesis process.

Microscopy is another essential technology in microbiology which also represents a primary cornerstone of microbiology. Zeiss MicroImaging (Munich, Germany; www.zeiss.de/mikro) is offering hardware and software for microscopic analysis and is constantly working on new developments. Future technologies such as REversible Saturable OpticaL Fluorescence Transitions (RESOLFT), Photo-Activated Localization Microscopy (PALM), and Array

Tomography will provide resolution beyond the borders defined by Abbe's equation, achieved by modern computer-based image analysis. Bringing together service providers and scientists to exchange perspectives and simply restore knowledge on "basic" tools, is an important yet often neglected activity in the daily routine of cutting-edge scientific research.

The pioneering work of Carl Woese on microbial evolution [41] has enabled sequence-based "molecular environmental microbiology". Many new rRNA-based technologies have been introduced since then, and the field has advanced significantly. However, we have only scratched the surface of microbial diversity, and our view of the true composition of microbial communities is utterly incomplete. What steps are required next? Of course, new techniques will evolve but we should also pay attention to standardization, data integration, and combinations of techniques, including the "traditional" methods such as cultivation. Every single technique has its limitations, but taken together cultivation-based and cultivation-independent methods do now allow us to proceed to new levels of knowledge! Appropriate applications of our technological resources will yield optimal progress in the understanding of microbial populations and ecosystems.

Diversity and Ecology (I)

A central focus of the "Diversity and Ecology" session was to discuss how the sheer amount of data accumulated in different databases is transformed into knowledge about microbial life. In 2000 the Sloan Foundation established a decadal program called the "Census of Marine Life"-network (CoML) to catalogue the diversity, distribution and abundance of marine life. Since then, this scientific initiative has grown to a global network of researchers engaging more than 80 nations. The European Census of Marine Life (EuroCoML) is one of twelve national and regional committees formed within the network and has been operational for almost three years now. EuroCoML promotes public awareness of marine biodiversity by establishing partnerships, coordinating with relevant European programmes and organisations

and engaging in education and outreach activities. The workshop on Ribosomal RNA technology in Bremen is just one of several workshops that EuroCoML has funded in recent years.

With decreasing DNA sequencing costs, SSU rRNA gene sequencing is becoming increasingly cost-effective. Partial sequencing of rRNA genes and comparing the resulting fragments with publicly available databases is rapidly displacing fingerprinting techniques like T-RFLP or DGGE commonly used in microbial diversity research. A seminal tool for the first-pass identification of rRNA gene sequences is BLAST [1]. However, phylogenetic analyses with online tools like those provided via the RDP II [5], Greengenes [8] or the ARB database [31] provide more detailed analyses and minimize false classifications sometimes encountered with BLAST hits.

An environmental study comprising more than 5,000 partial 16S rRNA gene sequences of cultured strains and clones presented some of the pitfalls with currently available tools. Initial BLAST analysis of about 1,800 cultures indicated moderate to high similarities with described species. The results of the phylogenetic positioning of sequences ~ 450 bp in length using automated Greengenes/SINA/ARB systems allowed the comparison of the phylogenetic position with BLAST similarities. In most of the cases a high degree of correlation could be found, but in some cases expert knowledge was needed to correct for false phylogenetic placement. This example suggests that new online tools with better alignment capabilities should be developed to confidently align and ultimately identify those problematic sequences. The SINA webaligner provided through the SILVA webpage is a major step forward in this direction.

Many early molecular studies of microbial distribution were based on the use of DNA probes targeting specific phylogenetic groups. The rational behind this approach was justified due to expense – methods were not available to resolve individual species - and the assumption that a natural phylogeny of the organism reflects physiology and ecology. Early studies also

focused on functional assemblages showing reasonable association between phylotype and function (ecotype). This was demonstrated with the description and quantification of the cellulolytic genus *Fibrobacter* in the bovine rumen in the early 1990's [3], related to a phylogenetic assemblage more recently found in the hindgut of a wood eating termite [40]. *Nitrosopumilus maritimus* was recently isolated from a marine aquarium as the first representative of a ubiquitous marine clade of Crenarchaeota [16]. In contrast to what is known so far from cultured Crenarchaeota, mostly sulphur-metabolizing thermophiles, *Nitrosopumilus* oxidizes ammonium and fixes carbon dioxide. It is tempting to believe that a major part of the Crenarchaeota living in the deep ocean is relying primarily on ammonium oxidation [15]. However this needs to be proven by future studies. Culture independent approaches will continue to play a major role in studies of the biogeochemical significance of this discovery.

We are currently facing an enormous amount of new bacterial and archaeal species descriptions per year. In 2007, 614 species were newly classified representing about 8.16% of a total of 7,521 validly published names at that time. In parallel, the information content of gene sequence databases is exponentially growing, with a current doubling rate of ~18 months. In January 2008 out of the 109,626,755 gene entries in EMBL release 93, 1,200,423 were attributed to ribosomal RNAs gene sequences. However, just 20,754 were obtained from pure cultures grown in the laboratory, and 9,889 of them corresponded to sequences assigned to type strains (according to SILVA release 93). Publicly available entries frequently contain errors in strain assignment and nomenclature, as well as low sequence quality. In addition for many species, redundant 16S rRNA gene sequence information often of different lengths and quality exist. Supported by the journal "Systematic and Applied Microbiology", "The Living Tree" project [42]) will provide a reliable phylogenetic 16S rRNA tree comprising all classified type strains listed in the List of Prokaryotic names with Standing in Nomenclature (www.bacterio.cict.fr). So far 6,800 type strain sequences have been selected from the SILVA

SSURef database. The final tree will serve as a guide tree for confident classification for newly retrieved strains and will be updated twice a year together with the complete 16S rRNA alignment (www.arb-silva.de/living-tree).

Diversity and Ecology (II):

The final session of the workshop and the second half of the Diversity and Ecology session highlighted discoveries in microbial ecology enabled by rRNA technologies.

Members of the heterotrophic alphaproteobacterial lineage SAR11 were first reported from the Sargasso Sea in 1990 [12]. In addition to being among the most abundant bacteria in the sea accounting for approximately 25% of the biomass and 50% of the cell abundance, they are noteworthy due to the size of their genome – a mere 1,308,759 base pairs. The ubiquity of *Candidatus* Pelagibacter ubique, yet to be formally described, was first quantified using combined FISH and rRNA-gene based phylogenetic methods. Six years and three genomes later, *Candidatus* Pelagibacter ubique continues to provide a wealth of information about microbial ecological processes and genomics in the ocean. What can we learn about studying SAR11? Giovannoni and colleagues have recently added to our understanding about the contribution of this abundant microbe by documenting SAR11's requirement for reduced sulfur [37]. With a highly reduced set of genes, SAR11 also lacks the ability to reduce sulfur independently – instead taking advantage of reduced sulfur excreted by other cells. More surprises lie in SAR11's extraordinarily high allelic variation and genome rearrangement. Evidence for SAR11 ecotypes from the Bermuda Atlantic Time Series (BATS) study explain the repeatable patterns in microbial distributions. At the center of these patterns lies genomic recombination which likely allows the different ecotypes of SAR11 to adapt to a given environment.

Discovery in the open ocean is not limited to the pelagic zone. Since the detection of the first hydrothermal vents of the Rose Gardens off of the Galapagos in 1977 – exploration of the

deep sea has forever been transformed. A major question that still confronts us in these systems is the temporal and spatial patterns of microbial diversity at deep-sea vents. Reysenbach and colleagues have been exploring these patterns in their research [28]. They find that development from immature to mature chimneys is accompanied by chemical shifts over time, as well as differences in the fragility of the chimney structures. Mature chimneys tend to harbour a higher microbial diversity. These are often environments dominated by epsilon proteobacteria, but different archaeal chemolithotrophs often prevail as well. Research in this area has seen advances thanks to the ease of SSU rRNA gene sequencing and high throughput methods such as DGGE that allow for statistically relevant sample sizes to be collected and community structures compared. Future challenges include untangling the separate and combined effects of mineralogical changes, geochemical differences and microbial activity on microbial diversity in these geochemically diverse environments.

A combined approach of rRNA-based and genomic technologies was used to elucidate the microbial community structure and activity in the meso- and bathypelagic zones of the ocean where microbes drive ocean biogeochemistry [36]. In the North Atlantic Deep Water mass, bacterial and archaeal populations are highly stratified both vertically and latitudinally. Marine Group I Crenarchaeota decrease in relative abundance closer to the equator to be outcompeted by members of the SAR202 bacterial clade. Likewise, Marine Group I crenarchaeotal concentrations are higher in mesoplagic than bathypelagic zones. Alternative substrate utilization in the form of D-amino acids is responsible for these differences in the bathypelagic waters as revealed by single-cell analysis of microautoradiography combined with CARD-FISH. A combination of microbial ecology and genomics approaches will likely continue to shed light on the “dark ocean’s” microbial consortia and determine the role particulate matter might be playing in the microbial food loop of the deep.

“Everything is everywhere, but not equally happy” summarizes Jakob Pernthaler’s take home message of the Diversity and Ecology session at the rRNA technology workshop. rRNA-

based microscopic methods coupled with assessment of cell growth or substrate uptake via pulse-labeling experiments enable us to go beyond questions of “Who’s there?” to “What are they doing?”. Microautoradiography combined with FISH can be automated to overcome the lab-intensive task of obtaining cell concentrations while simultaneously measuring substrate incorporation. In many lacustrine settings, horizontal and vertical gradients are generated by varying oxygen concentrations with depth and concentrations of humic substances along the surface. Differential utilization of glucose and acetate for biomass production can explain differences in species compositions at oxic and anoxic zones in humic lakes. The physiological performance of bacteria in different habitats likely helps shape their genomic constitution and their biogeographic distributions in nature.

A list of tools and databases that have been presented on the workshop is available at: http://www.arb-silva.de/rrna-workshop/tools/.

Conclusions:

The workshop emphasized the rapid progress that has been gained in rRNA technology over several decades. Twenty years ago, the sequencing of 150 bases was a challenge that kept researchers busy for weeks, but now several thousands of full-length 16S rRNA sequences or greater than 400,000 pyrosequencing tags can be readily generated within a week (Figure 2). Data production has become a routine procedure and powerful tools and software packages are available to process, store and visualize data and interpret them for transfer into biological knowledge. A remaining critical step in achieving a holistic picture of microbial diversity and function is to analyze genes and genomes in the context of their surrounding environment [7, 19, 25]. To reach this goal it is now necessary to emend our sequence collection with more contextual (meta)data. The Genomic Standards Consortium (GSC) has recently been established to promote the inclusion of metadata alongside submission of genomes and metagenomes [10, 11]. The Minimum Information about a Metagenome Sequence (MIMS)

intends to standardize contextual data acquisition by requiring at least GPS coordinates plus depth/altitude and sampling time (see http://gensc.org) [9]. The Minimum Information about an ENvironmental Sequence (MIENS) has been proposed as a natural extension to MIGS and MIMS (http://gensc.org/gc_wiki/index.php/MIGS/MIMS_for_16S) targeted at rRNA gene sequences [29].

Ribosomal RNA technology has become a mature science, but is still far from being exhausted of innovative applications. Over the years we have witnessed important steps in delineating the diversity on our planet using molecular markers with rRNA in the forefront. What can we expect from rRNA technologies thirty years from now? Hand-held devices that allow anyone to investigate and report biodiversity in real-time from the environment? Geotagging and geoblogging (http://en.wikipedia.org/wiki/Geotagging) of information is currently emerging and cell phones with GPS devices are already available. Therefore, it seems just a matter of time before pocket sequencing machines leave the realm of fiction and join our future collection of portable communication equipment.

Acknowledgement

We thank all speakers for proofreading and helpful comments. A special thanks to David Todd for all the efforts with the workshop logo and the timeline figure. The workshop was a joint collaborative effort of the Max Planck Institute in Bremen and the Ribocon GmbH Bremen, the Technical University Munich, the International Census of Marine Microbes (ICoMM) and the European Census of Marine Life (EuroCoML). The workshop was further sponsored by the Operon and BioCat biotechnology companies.

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) 403-410.

[2] R. Amann, B.M. Fuchs, Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques, Nature Rev. Microbiol. 6 (2008) 339-348.

[3] R.I. Amann, L. Krumholz, D.A. Stahl, Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology, J. Bacteriol. 172 (1990) 762-770.

[4] S. Behrens, C. Ruhland, J. Inacio, H. Huber, A. Fonseca, I. Spencer-Martins, B.M. Fuchs, R. Amann, In situ accessibility of small-subunit rRNA of members of the domains Bacteria, Archaea, and Eucarya to Cy3-labeled oligonucleotide probes, Appl. Environ. Microbiol. 69 (2003) 1748-1758.

[5] J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, A.M. Bandela, E. Cardenas, G.M. Garrity, J.M. Tiedje, The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data., Nucleic Acid Res. 35 (2007) D169-172.

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

[7] E.F. DeLong, D.M. Karl, Genomic perspectives in microbial oceanography, Nature. 437 (2005) 336-342.

[8] T.Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G.L. Andersen, Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, Appl. Environ. Microbiol. 72 (2006) 5069-5072.

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

[10] D. Field, G. Garrity, T. Gray, J. Selengut, P. Sterk, N. Thomson, T. Tatusova, G. Cochrane, F.O. Glöckner, R. Kottmann, A.L. Lister, Y. Tateno, R. Vaughan, eGenomics: Cataloguing our complete genome collection III, Comp. and Funct. Genomics. 2007 (2007) 1-7.

[11] D. Field, G. Garrity, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, T. Gray, M. Ashburner, S. Baldauf, J. Boore, G. Cochrane, J. Cole, C. dePamphilis, R. Edwards, N. Faruque, R. Feldmann, F.O. Glöckner, e. al., Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification, Nat. Biotechnol. 26 (2008) 541-547.

[12] S.J. Giovannoni, T.B. Britschgi, C.L. Moyer, K.G. Field, Genetic diversity in Sargasso Sea bacterioplankton, Nature. 345 (1990) 60-63.

[13] S.J. Giovannoni, E.F. DeLong, G.J. Olsen, N.R. Pace, Phylogenetic groupspecific oligodeoxynucleotide probes for identification of single microbial cells, J. Bacteriol. 170 (1988) 720-726.

[14] W.E. Huang, K. Stoecker, R. Griffiths, L. Newbold, H. Daims, A.S. Whiteley, M. Wagner, Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function, Environ. Microbiol. 9 (2007) 1878-1889.

[15] A.E. Ingalls, S.R. Shah, R.L. Hansman, L.I. Aluwihare, G.M. Santos, E.R.M. Druffel, A. Pearson, Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon, Proc. Natl. Acad. Sci. USA. 103 (2006) 6442-6447.

[16] M. Konneke, A.E. Bernhard, J.R. de la Torre, C.B. Walker, J.B. Waterbury, D.A. Stahl, Isolation of an autotrophic ammonia-oxidizing marine archaeon, Nature. 437 (2005) 543-546.

[17] A. Lehmann, J.M. Overton, J.R. Leathwick, GRASP: generalized regression analysis and spatial prediction, Ecological Modelling. 157 (2002) 189-207.

[18] R.E. Ley, F. Backhed, P. Turnbaugh, C.A. Lozupone, R.D. Knight, J.I. Gordon, Obesity alters gut microbial ecology, Proc. Natl. Acad. Sci. USA. 102 (2005) 11070-11075.

[19] T. Lombardot, R. Kottmann, H. Pfeffer, M. Richter, H. Teeling, C. Quast, F.O. Glöckner, Megx.net - database resource for marine ecological genomics, Nucleic Acid Res. 34 (2006) D390-D393.

[20] A. Loy, R. Arnold, P. Tischler, T. Rattei, M. Wagner, M. Horn, probeCheck - a central resource for evaluating oligonucleotide probe coverage and specificity, Environ. Microbiol. (2008) in press.

[21] A. Loy, F. Maixner, M. Wagner, M. Horn, probeBase - an online resource for rRNA-targeted oligonucleotide probes: new features 2007, Nucleic Acid Res. 35 (2007) D800-D804.

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

[23] W. Ludwig, K.H. Schleifer, Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes in: J. Sapp (Ed.), Microbial phylogeny and evolution, concepts and controversies, Oxford university press, New York, 2005, pp. 70-98.

[24] 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 Acid Res. 32 (2004) 1363-1371.

[25] V.M. Markowitz, Microbial genome data resources, Curr. Opin. Biotechnol. 18 (2007) 267-272.

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

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

[28] A. Page, M.K. Tivey, D.S. Stakes, A.L. Reysenbach, Temporal and spatial archaeal colonization of hydrothermal vent deposits, Environ. Microbiol. 10 (2008) 874-884.

[29] J. Peplies, R. Kottmann, W. Ludwig, F.O. Glöckner, A Standard Operating Procedure for Phylogenetic Inference (SOPPI) using (rRNA) marker genes Syst. Appl. Microbiol. XX (2008) XX-XX.

[30] A. Pernthaler, J. Pernthaler, R. Amann, Fluorescence In Situ Hybridization and Catalyzed Reporter Deposition for the Identification of Marine Bacteria, Appl. Environ. Microbiol. 68 (2002) 3094-3101.

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

[32] A. Ramette, Multivariate analyses in microbial ecology, FEMS Microbiol. Ecol. 62 (2007) 142-160.

[33] M.L. Sogin, H.G. Morrison, J.A. Huber, D.M. 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. Natl. Acad. Sci. USA. 103 (2006) 12115-12120.

[34] D.A. Stahl, Analysis of hydrothermal vent associated symbionts by ribosomal RNA sequences, Science. 224 (1984) 409-411.

[35] A. Stamatakis, RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics. 22 (2006) 2688-2690.

[36] E. Teira, P. Lebaron, H. van Aken, G.J. Herndl, Distribution and activity of Bacteria and Archaea in the deep water masses of the North Atlantic, Limnol. Oceanogr. 51 (2006) 2131-2144.

[37] H.J. Tripp, J.B. Kitner, M.S. Schwalbach, J.W.H. Dacey, L.J. Wilhelm, S.J. Giovannoni, SAR11 marine bacteria require exogenous reduced sulphur for growth, Nature. 452 (2008) 741-744.

[38] C. von Mering, P. Hugenholtz, J. Raes, S.G. Tringe, T. Doerks, L.J. Jensen, N. Ward, P. Bork, Quantitative phylogenetic assessment of microbial communities in diverse environments, Science. 315 (2007) 1126-1130.

[39] D.M. Ward, R. Weller, M.M. Bateson, 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community, Nature. 345 (1990) 63-65.

[40] F. Warnecke, P. Luginbuhl, N. Ivanova, M. Ghassemian, T.H. Richardson, J.T. Stege, M. Cayouette, A.C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S.G. Tringe, M. Podar, H.G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N.C. Kyrpides, E.G. Matson, E.A. Ottesen, X.N. Zhang, M. Hernandez, C. Murillo, L.G. Acosta, I. Rigoutsos, G. Tamayo, B.D. Green, C. Chang, E.M. Rubin, E.J. Mathur, D.E. Robertson, P. Hugenholtz, J.R. Leadbetter, Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite, Nature. 450 (2007) 560-565.

[41] C.R. Woese, E. Fox, Phylogenetic structure of the procaryotic domain: The primary kingdoms, Proc. of the Natl. Acad. Sci. USA. 74 (1977) 5088-5090.

[42] P. Yarza, M. Richter, J. Peplies, J. Euzeby, R. Amann, K.H. Schleifer, W. Ludwig, F.O. Glöckner, R. Rossello-Mora, The All-Species Living Tree Project: a 16S rRNA-based phylogenetic tree of all sequenced type strains, Syst. Appl. Microbiol. XX (2008) XX-XX.

Figure Legends

Figure 1:

Growth of the ribosomal RNA databases since 1992 measured by RDP II and SILVA. The databases show an exponential grows phase with a doubling time of around 15 to 18 months. Light grey: statistics by RDP, dark grey: statistics taken from the latest SILVA releases.

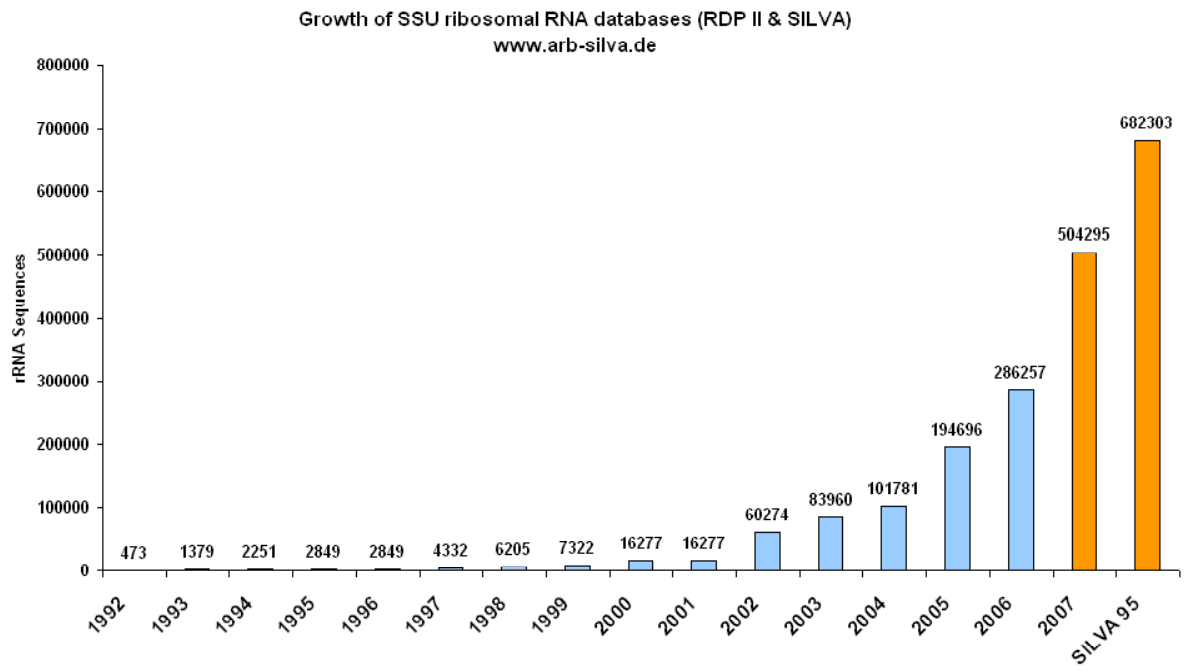
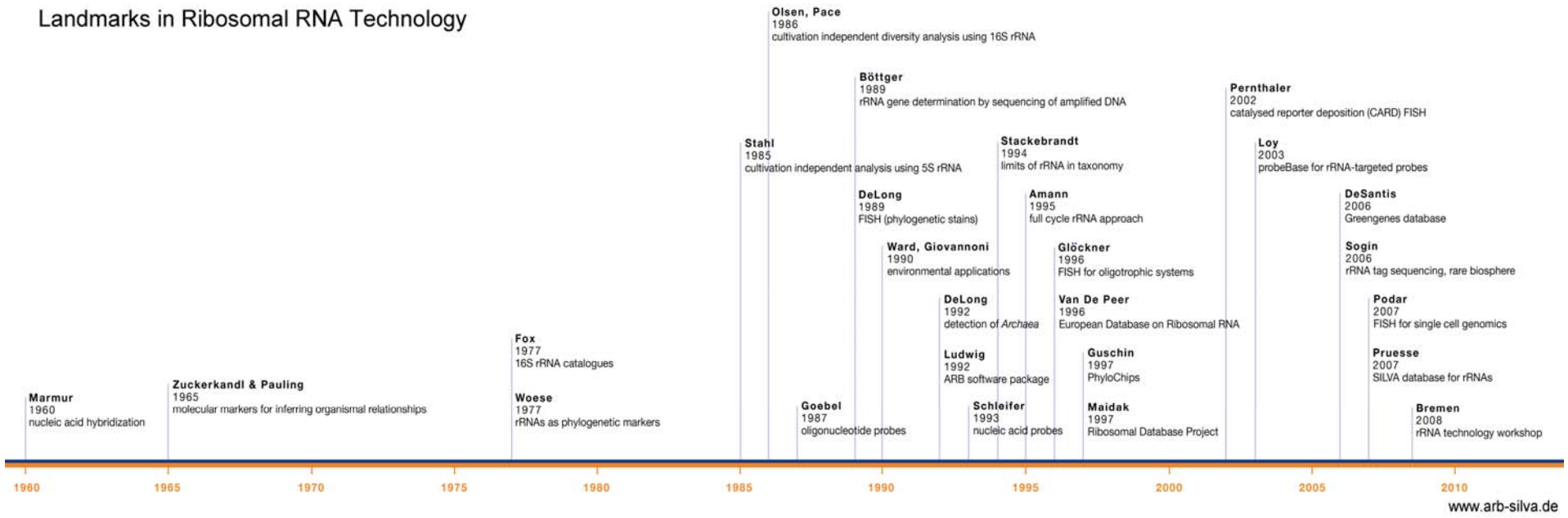


Figure 2:

Timeline of the landmarks in ribosomal RNA technology over the last decade



Supplementary Table 1.

List of speakers at the International Workshop on ribosomal RNA Technology, April 7-9, 2008, in Bremen, Germany

Speaker	Affiliation
1. Databases	
James Cole	Michigan State University, USA
Todd DeSantis	Lawrence Berkeley National Laboratory, USA
Peter Dawyndt	University of Ghent, Belgium
Frank Oliver Glöckner	Max Planck Institute for Marine Microbiology, Bremen, Germany
2. Phylogeny	
Christian von Mering	University of Zurich, Switzerland
Wolfgang Ludwig	Technical University Munich, Germany
Alexandros Stamatakis	Ludwig Maximilians University, Munich, Germany
Harald Meier	Technical University Munich, Germany
3. Data Analysis & Biogeography	
Anthony Lehmann	University of Geneva, Switzerland
Angelique Gobet	Max Planck Institute for Marine Microbiology, Bremen, Germany
Thomas Pommier	Université Montpellier, France
Alban Ramette	Max Planck Institute for Marine Microbiology, Bremen, Germany
4. Technology	
Mitchell Sogin	Marine Biological Laboratory, Woods Hole, USA
Freek Bakker	National Herbarium and University of Wageningen, The Netherlands
Rudolf Amann	Max Planck Institute for Marine Microbiology, Bremen, Germany
Alexander Loy	University of Vienna, Austria
Barbara Pohl	biomers.net GmbH, Ulm, Germany
Wolf Malkusch	Carl Zeiss Imaging Solutions GmbH, Munich, Germany
5. Diversity & Ecology	
Pedro Martinez-Arbizu	DZMB Forschungsinstitut Senckenberg, Wilhelmshaven, Germany
Erko Stackebrandt	German Collection of Microorganisms and Cell Cultures (DSMZ), Braunschweig, Germany
Dave Stahl	University of Washington, USA
Ramon Rossello-Mora	IMEDEA (CSIC-UIB), Esporles, Illes Balears, Spain
Stephen Giovannoni	Oregon State University, USA
Anna-Louise Reysenbach	Portland State University, USA
Jakob Pernthaler	University of Zurich, Switzerland
Gerhard Herndl	Royal Netherlands Institute for Sea Research (NIOZ), Texel, The Netherlands