

Report from the
EC-US Workshop on

**GENOMIC APPROACHES
FOR STUDYING THE
MARINE ENVIRONMENT
AND RESOURCES**

**12-13 May 2005
Bremen, Germany**

Organized by:
R. Amann · C. Boyen · M. A. Moran

Preface

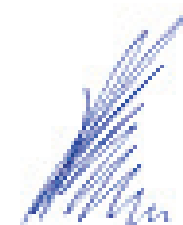
This report summarizes the presentations and discussions held at the EC-US Workshop on Genomic Approaches for Studying the Marine Environment and Resources held on 12-13 May 2005 at the Max Planck Institute for Marine Microbiology in Bremen, Germany under the auspices of the EC-US Task Force on Biotechnology Research. It brought together 21 scientists from member states of the European Union, Iceland, and the United States.

The participants at the workshop discussed the state-of-the-art research and identified knowledge gaps, bioinformatics infrastructure needs, and opportunities for collaborations to promote the future advancement of this emerging field. The workshop was organized around four themes: (1) bacteria and other unicellular microbes; (2) multicellular organisms; (3) bioinformatics; and (4) the perspectives of industry and a private foundation.

Researchers are just beginning to understand the complexity and breadth of opportunities of the marine environment. The new tools of genomics offer a means of understanding this complexity. Using these tools, we are learning how the marine environment functions as an intricate part of earth's ecosystem. We are also just beginning to appreciate the vast diversity of the oceans' genetic resources that can be of great benefit to human health and bio-based industries throughout the world. This workshop highlighted opportunities for international collaborations that will promote this exciting new field of science.

The workshop was organized by Dr. Rudolf Amann from the Max Planck Institute for Marine Microbiology in Bremen, Germany, Dr. Catherine Boyen, from the CNRS Biological Station in Roscoff, France, and Dr. Mary Ann Moran from the University of Georgia, USA. They also compiled and edited this report, which is also available on the EC-US Task Force web site, http://ec.europa.eu/research/biotechnology/ec-us/index_en.html. We would like to thank them for their outstanding efforts.

The views expressed in this document are those of the workshop participants, and do not necessarily reflect the views of the sponsors or governments.



Christian Paternmann, EC Chairperson



Kathie L. Olsen, US Chairperson

EC-US Task Force on Biotechnology Research



Contents

1	Executive Summary
2	List of Participants
4	Program
5	SUBREPORT FROM SESSION I: Bacteria and other Unicellular Microbes
6	INDIVIDUAL CONTRIBUTIONS IN SESSION I - <i>Genomics of Ecologically-Relevant Marine Bacteria</i>
7	- <i>Total Genome Sequences of Marine Cyanobacteria: Novel Genes and Ecotype-Specific Adaptation to the Environment</i>
9	- <i>Environmental Diversity and Organizing Marine Environmental Genomics</i> - <i>Environmental Genomics, Sampling, Scope and Potential Constraints</i>
11	- <i>A Metagenome Analysis of an Extreme Microbial Symbiosis</i>
12	- <i>Molecular and Genomic Tools to Study the Ecology of Marine Protists</i>
13	SUBREPORT FROM SESSION II: Multicellular Organisms
14	INDIVIDUAL CONTRIBUTIONS IN SESSION II - <i>The Ectocarpus Genome Project</i>
15	- <i>Marine Environmental Genomics: Integrating Complex Multicellular Genome Response to Changing Environment</i>
16	- <i>The Complex Response of Marine Crustaceans and Bivalves to Bacterial Pathogens</i>
17	- <i>Genomics Approaches to Fish and Shellfish</i>
18	- <i>Environmental Genomics of Polar Organisms</i>
19	- <i>Interacting Genomes in the Squid-Vibrio Symbiosis</i>
20	SUBREPORT FROM SESSION III: Bioinformatics
22	INDIVIDUAL CONTRIBUTIONS IN SESSION III - <i>High-Throughput Marine Meta-Genomics: Challenges from the Bioinformatics Perspective</i>
23	- <i>Marine Ecological Genomics Approaches Sequence Paradise</i> - <i>Bioinformatics in Marine Genomics</i>
24	- <i>The SEED and GenDB</i>
26	- <i>GenDB and The Seed</i> - <i>The UniProt Knowledgebase</i>
28	SUBREPORT FROM SESSION IV: The Industry / Foundation perspective - <i>Industry & Small/Medium Enterprises</i> - <i>Mining the Marine Diversity of Extremophile for Novel Enzymes and Products</i> - <i>Private Foundations</i>
29	- <i>Genome Resource Banks, Marine Genomic and Conservation</i>

Executive Summary

The global oceans harbour a tremendous diversity of microscopic and macroscopic life. The interaction of marine organisms with the environment profoundly influences the Earth's geo- and atmosphere. For the layman it might be surprising that, nevertheless, the vast majority of marine organisms have yet to be identified and remain unculturable outside the marine environment. Even for known organisms, there is insufficient knowledge to permit their intelligent management and application. Genomic approaches are now providing new keys for studying the marine environment and resources. They offer the ability to address environmental problems and to mine the diversity of the marine world for products and processes that will contribute to the welfare of mankind.

Thus far, microbes have been the focus of much of the research in the newly emerging field of "marine genomics". Efforts of interdisciplinary teams to use information based on genome sequencing and other high throughput techniques have already resulted in outstanding new fundamental knowledge of the microbial world. It is timely that these approaches should be applied to other studies of the marine environment and of its resources.

Consequently, interest in applying genomics to the study of multicellular organisms and their interactions is growing rapidly as the technical feasibility increases. The tools and advances making this possible include the increasing speed and declining costs of DNA sequencing, along with new applications of other high-throughput technologies (transcriptomics, proteomics, metabolomics). The new opportunities are leading scientists to apply genomics technologies to the study of large-scale environmental processes in the sea and are suggesting a wide range of applications in biotechnology. Vast quantities of data are already now generated (>100 genomes of marine bacteria and archaea; several "metagenomes", mostly of bacterioplankton). Generation/curation of publicly available databases and the continuous refinement of bioinformatics tools for analyzing the flood of data will be decisive for rapid implementation and the success of Marine Genomics.

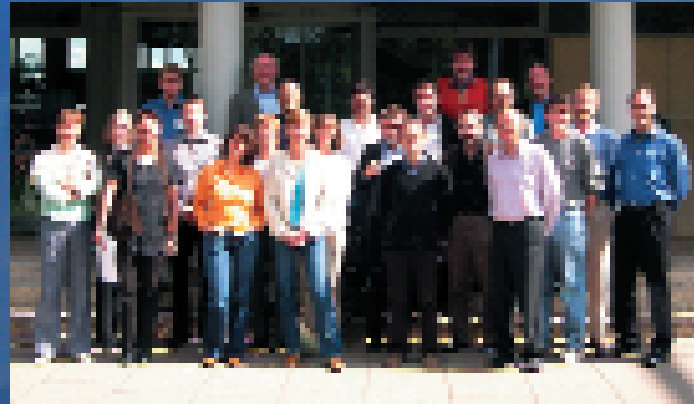
In May 2005 the EC-US Task Force on Biotechnology Research sponsored a workshop that has brought together a total of 24 scientists and science administrators at the Max Planck Institute for Marine Microbiology in Bremen, Germany. This included experts applying these technologies to the study of microorganisms and investigators studying multicellular marine life. Experts in other disciplines such as ecology and bioinformatics were included to help develop a shared vision of the future of Marine Genomics. Representative from a US foundation and a European company provided their vision on how genomic tools will facilitate the sustainable exploitation of marine resources.

Major projects have been initiated in Marine Genomics in the recent past by both EC and US funding agencies as well as private foundations. This includes, e.g., the 10 million € Network of Excellence "Marine Genomics Europe", and the Marine Microbiology Initiative of the Moore Foundation. The Bremen workshop was a forum for open discussions of first results and future plans among the scientists involved in some of these large-scale projects. It resulted in **detailed recommendations** with respect to the specific requirements of marine biology for infrastructure of genomics and bioinformatics (see pages 5, 13, 21, 28 and 29).

Consensus was reached that (1) **EC-US coordination of large-scale projects** like **genome analysis of multicellular organisms** and joint developments in **gene expression analysis** (postgenomics, i.e. transcriptomics) and **bioinformatics** is required to make full and wise use of the financial resources available for marine genomics.

(2) A **strong need** was identified for **interdisciplinary training in marine genomics** at the post-graduate level. EC and US agencies should (3) significantly expand their **support for scientific exchange across the Atlantic** and attempt to (4) **co-ordinate parallel, yet complementary research programs in marine genomics**. Implementation of these recommendations could be promoted by the foundation of an **EC-US Working Group for Marine Genomics**, under the auspices of the Task Force.

List of Participants



Microbial Expertise

Rudolf Amann

Max Planck Institute of Marine Microbiology
Celsiusstr. 1, 28359 Bremen, Germany
Phone +49 421 2028 930
Fax +49 421 2028 580
E-mail ramann@mpi-bremen.de
www.mpi-bremen.de

Stephen Craig Cary

College of Marine Studies, University of Delaware
Lewes DE 19958, USA
Phone +1 302 645-4078
Fax +1 302 645-4007
E-mail caryc@udel.edu

Arthur R. Grossman

The Carnegie Institution of Washington
Department of Biology,
Department of Plant Biology, Stanford University
260 Panama Street, Herrin Hall
Stanford CA 94305, USA
Phone +1 650 325-1521
E-mail arthurg@stanford.edu

Ian M. Head

School of Civil Engineering and Geosciences
Centre for Molecular Ecology and Institute for
Research on the Environment and Sustainability
The Drummond Building, University of Newcastle
Newcastle-upon-Tyne, NE1 7RU, UK
Phone +44 191 246-4806
Fax +44 191 222-5431
E-mail i.m.head@ncl.ac.uk

Wolfgang Hess

University of Freiburg
Inst. Biology II, Experimental Bioinformatics
and Center for Biological Systems Analysis
Schaenzlestr. 1, 79104 Freiburg, Germany
Phone +49 761 203 27 96
Fax +49 761 203 26 01
E-mail wolfgang.hess@biologie.uni-freiburg.de

Ramon Massana

Institut de Ciències del Mar (CSIC)
Passeig Marítim de la Barceloneta 37-49
08003 Barcelona, Catalonia, Spain
Phone +34 93 230 95 00
E-mail ramonm@icm.csic.es

Mary Ann Moran

Department of Marine Sciences
University of Georgia, USA
Athens GA 30602-3636
Phone +1 706 542-6481
E-mail mmoran@uga.edu

Multicellular Organisms Expertise

Catherine Boyen

UMR 7139 CNRS/UPMC
Végétaux Marins et biomolécules
Station Biologique BP 74
F29682 Roscoff Cedex, France
Phone +33 298 29 23 31
Fax +33 298 29 23 85
boyen@sb-roscoff.fr
www.marine-genomics-europe.org

Karen G. Burnett

Hollings Marine Laboratory
331 Fort Johnson Road
Charleston SC 29412, USA
Phone +1 843 762-8933
Fax +1 843 762-8737
E-mail burnettk@cofc.edu
www.cofc.edu/~burnettk

Adelino V. M. Canário

Centre of Marine Sciences
University of Algarve
Campus de Gambelas, 8005-139 Faro, Portugal
Phone +351 289 800 925
E-mail acanario@ualg.pt

Douglas L. Crawford

Rosenstiel School of Marine
and Atmospheric Science
University of Miami
4600 Rickenbacker Causeway
Miami FL 33149-1098, USA
Phone +1 305 421-4121
E-mail dcrawford@rsmas.miami.edu

Alex Rogers

British Antarctic Survey
High Cross, Madingley Road
Cambridge CB3 0ET, GB
Phone +44 1223 221 604
E-mail ADR2@bas.ac.uk

Edward Ruby

Department of Medical Microbiology
and Immunology, University of Wisconsin
1300 University Avenue
Madison WI 53706, USA
Phone +1 608 262-5911
E-mail egruby@wisc.edu

List of Participants



Bioinformatics Expertise

Ildefonso Cases

Centro Nacional de Biotecnología
Campus de Cantoblanco, E-28049 Madrid, Spain
Phone +34 915 85 54 51
Fax +34 915 85 45 06
E-mail icases@cnb.uam.es

Rob Edwards

Fellowship for Interpretation of Genomes
Chicago IL and San Diego State University
San Diego CA 92183, USA
E-mail RobE@thefig.info

Frank-Oliver Glöckner

MPI for Marine Microbiology
Celsiusstr. 1, 28359 Bremen, Germany
Phone +49 421 2028-938
Fax +49 421 2028-580
E-mail fog@mpi-bremen.de

Tania Lima, Ph.D.

Swiss Institute of Bioinformatics (ISB-SIB)
Swiss-Prot Group
1 rue Michel-Servet, 1211 Geneva 4, Switzerland
Phone +41 22 379-5853
Fax +41 22 379-5858
E-mail Tania.Lima@isb-sib.ch

Folker Meyer

CeBiTec/BRF, Universität Bielefeld
33594 Bielefeld, Germany
Phone +49 521 106 4827
Fax +49 521 106 6419
E-mail fm@cebitec.uni-bielefeld.de

Karin A. Remington, Ph.D.

Vice President, Bioinformatics Research
The Venter Institute
9704 Medical Center Drive, 4th Floor
Rockville MD 20850, USA
Phone +1 240 268-2762
Fax +1 240 268-4000
E-mail kremington@venterininstitute.org

Industry/Foundations

Viggó Thór Marteinsson

Prokaria Ltd.
Gylfaflöt 5, IS-112 Reykjavik, Iceland
Phone +354 570-7924
Fax +354 570-7901
E-mail vigg@prokaria.com

Daniel L. Distel

Ocean Genome Legacy Foundation
New England Biolabs
32 Tozer Rd., Beverly MA 01915, USA
Phone +1 978 998-7425
E-mail distel@oglf.org

Funding Agencies

Catherine Eccles

European Commission
DG Research I-3, CDMA 3/157
Rue de Loi 200, 1049 Bruxelles, Belgium
Phone +32 2 299 46 95
Fax +32 2 295 05 68
E-mail catherine.eccles@cec.eu.int

Maurice Lex

EU Brussels, EC, GD XII E-1, SDME 9/38
Rue de Loi 200, 1049 Bruxelles, Belgium
Phone +32 2 296 5619
Fax +32 2 299 1860

Maryanna Henkart

Division of Molecular and Cellular Biosciences
National Science Foundation Washington
4201 Wilson Boulevard
Arlington VA 22230, USA
Phone +1 703 292-8440
Fax +1 703 292-9061
E-mail mbenkart@nsf.gov



Program

SUBREPORT FROM SESSION I: Bacteria and other Unicellular Microbes

Day 1: Thursday, May 12

Session I: Bacteria (Unicellular Microbes) Chair Rudolf Amann

8:30–10:30 Welcome address followed by
6 talks, 15-20 min each

Mary Ann Moran (University of Georgia),
Wolfgang Hess (University of Freiburg),
Arthur Grossman (Carnegie Washington),
Ian Head (University of Newcastle),
Craig Cary (University of Delaware),
Ramon Massana (CSIC Barcelona)

10:30–11:00 Coffee break

11:00–12:00 General discussion of Session I

12:00–13:00 Lunch break

Session II: Eukaryotes (Multicellular Organisms) Chair Cathrine Boyen

13:00–15:00 Opening remarks by chairperson
followed by 6 talks,
15 to 20 min each

Douglas Crawford (Rosenstiel School),
Cathrine Boyen (CNRS Roscoff),
Karen Burnett (Hollings Marine Laboratory),
Alex Rogers (British Antarctic Survey),
Ned Ruby (University of Wisconsin),
Adelino Canario (University of Faro)

15:00–15:30 Coffee break

15:30–16:30 General discussion of Session II

17:00 Departure from Hotel Atlantic for
a tour of historic Bremen, dinner

Day 2: Friday, May 13

Session III: Bioinformatics Chair Mary Ann Moran

8:30–10:30 Welcome address followed by
6 talks, 15-20 min each

Karin Remington (Venter Institute),
Frank Oliver Glöckner (MPI),
Ildefonso Cases (CSIC Madrid),
Rob Edwards (San Diego State University),
Tania Lima (Swiss Prot, Geneva)

10:30–11:00 Coffee break

11:00–12:00 General discussion of Session I

Session IV: The Industry/Foundation Perspective Chair Arthur Grossman

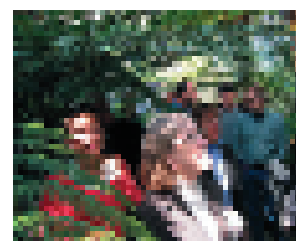
13:00–14:00 Opening remarks by Chair person
followed by 2 talks,
15 to 20 min each

Dan Distel (Ocean Legacy Foundation),
Viggo Marteinson (Prokaria Ltd.)

14:00–14:30 Coffee break

14:30–16:30 General discussion of future
perspectives in Marine Genomics

17:00 Departure for biodiversity
excursion to the "Rhododendron
Park", dinner



Rapporteurs:
Mary Ann Moran and Wolfgang Hess

SUMMARY:

Genome sequencing of unicellular microbes proceeds at a high rate – close to 400 genome sequences of unicellular organisms (largely of prokaryotes) are now publicly available and >100 more will be added in the next 12 months. Indeed, the pace of microbial genome sequence generation is so rapid as to make a tally of available sequences obsolete before it is finished. For marine microbiologists, the research opportunities afforded by genomic sciences have arrived as a revolution rather than an evolution. Virtually no sequences of environmentally relevant marine microbes existed just a few years ago. By the end of 2005, the number will pass 100.

The enormous promise of microbial genome sequences for understanding the marine environment and the wise utilization of marine resources is obvious. Yet the infrastructure, computational resources, and training opportunities needed to take advantage of this potential has lagged behind. Six scientists working in the field of marine microbial genomics took part in the Bremen Workshop. What follows is a summary of their ideas on the progress and potential of the field, followed by individual reports of how each is using genomic science to address specific questions in marine ecology and biogeochemistry.

RECOMMENDATIONS:

In order to fully exploit the upcoming wealth of sequence information from marine organisms, it is necessary to develop, coordinate, and efficiently apply new tools for the comparative and functional analysis of microbial genomes. The following four areas were identified by the subgroup for unicellular microbes as being critical for continued progress in the field:

(1) Resources

- Develop genetic systems for non-traditional model organisms and systems, including representatives from the major marine bacterioplankton taxa.
- Increase marine functional genomics efforts to the level necessary to address environmental and resource management issues. Proteomic and microarray capabilities lag far behind sequence availability.
- Support research that directly links genomics and biogeochemistry and specifically addresses the driving forces for marine diversity;
- Improve funding for environmental sequencing (including metagenomic sequencing and ecotype cluster sequencing) and for functional genomics.

(2) Bioinformatics

- Develop better informatics tools for the assembly of metagenomic data;
- Set standards for archiving environmental sequence data, including relevant environmental metadata;
- Develop advanced tools for annotating genes and regulatory elements not found by classical prediction programs (e.g., ncRNAs, transcription factor binding sites);
- Develop improved comparative genomics tools with capabilities for analysis of ecotypes and mixed microbial communities;
- Develop improved tools for the functional characterization of unknown genes;

Transatlantic training (3) and research coordination (4)!

See also executive summary!



INDIVIDUAL CONTRIBUTIONS IN SESSION I

Genomics of Ecologically-Relevant Marine Bacteria

Mary Ann Moran

University of Georgia
Department of Marine Sciences
Athens, Georgia, USA

Genomic approaches are becoming increasingly important for marine microbiologists because of their potential to provide insights into the physiological underpinnings of ocean biogeochemical cycles. 'Metagenomic' approaches retrieve a relatively unbiased sample of community DNA, and therefore provide the most ecologically relevant observations of marine microbial communities. Whole genome sequencing of cultured marine bacteria is limited to those marine microbes that can be successfully cultured from seawater. However, it has the advantage of providing greater access to physiological information compared to metagenomics, including how key microbial processes are regulated and co-regulated, and what combinations of metabolic abilities are packaged within single cells.

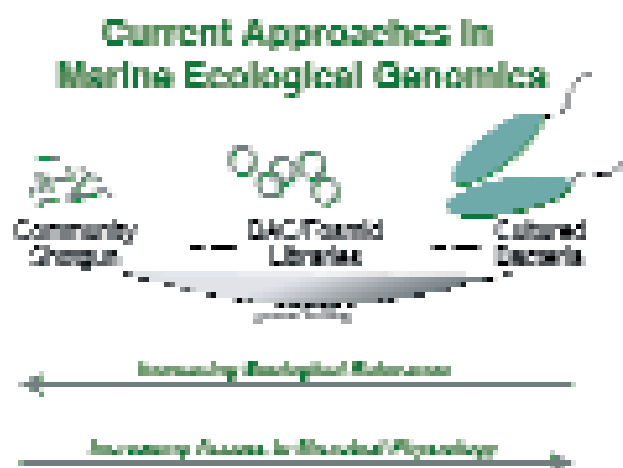


Figure 1. Genomic analyses of cultured bacteria complement culture-independent genomic methods for studying marine microbial communities. The culture independent approaches provide more ecologically relevant views of marine communities, but cultured bacteria provide greater access to microbial physiology and regulation.

The Roseobacter group is a major lineage of marine bacteria that typically accounts for 20% of coastal ocean bacterioplankton assemblages and 5-10% of open ocean assemblages. Since members of this group are amenable to culturing, Roseobacters are an excellent focus for genomic studies of marine bacterioplankton. Comparative analyses of three Roseobacter genomes (*Silicibacter pomeroyi*, *Silicibacter* sp. TM1040, and *Jannaschia* sp. CCS1) are providing new perspectives on the marine carbon, nitrogen, and sulfur cycles. All three organisms possess genes for the oxidation of carbon monoxide, an indirect greenhouse gas of considerable interest for global warming, and therefore are valuable model organisms for understanding the balance between biological consumption and ocean-atmosphere flux of carbon monoxide in marine waters. Roseobacter genomes are also providing insights into the sources of nitrogen that fuel marine bacterioplankton communities and the importance of inorganic sulfur oxidation in aerobic marine surface waters.

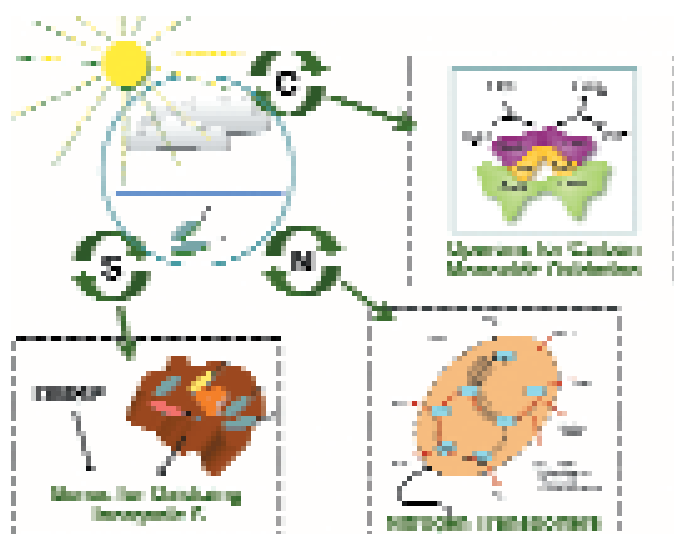


Figure 2. Roseobacter genome sequences are generating insights into marine biogeochemical cycles, including the role of heterotrophic carbon monoxide oxidation in ocean surface waters, the organic nitrogen compounds typically transported across bacterial membranes, and the presence of genes for oxidizing inorganic sulfur compounds in marine surface waters.

SESSION I

To take advantage of insights that arise from genomics, *in silico* analyses must be followed by functional genomic studies. Roseobacters and other cultured marine bacteria pose special challenges since genetic systems (for identifying and verifying function of key genes) are lacking and cell biomass (for obtaining sufficient proteins and nucleic acids for proteomics and microarray studies) can be dif-

ficult to obtain. Nonetheless, studies of these organisms have the potential to generate hypotheses about the vast assemblage of uncultured marine bacteria; to stimulate discovery of novel activities and interactions; and to provide insights into bacterial regulation and packaging of biogeochemically important activities in the ocean.

Total Genome Sequences of Marine Cyanobacteria: Novel Genes and Ecotype-specific Adaptation to the Environment

Wolfgang R. Hess

University Freiburg, Inst. Biology II
Experimental Bioinformatics, Freiburg, Germany

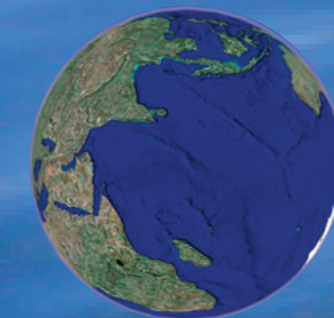
By the end of 2005 there have been more than 30 sequenced genomes from marine cyanobacteria. These analyses are the result of combined effort of groups in the U.S., mainly resulting in sequencing projects performed at the Joint Genome Institute and more recently the initiative supported by the Gordon and Betty Moore Foundation, but also projects in Europe, here mainly at the Genoscope in Paris.

Marine cyanobacteria play an important role in the marine bio-geochemical cycle because of their efficient light-harvesting strategies through which they convert solar energy into organic biomass by oxygenic photosynthesis. Their adaptation to widely differing ecological conditions and their numerical abundance makes them the most important primary producers in the world's oceans. Several oceanographic studies have shown that among the group of marine cyanobacteria, those belonging to only two genera, *Prochlorococcus* and *Synechococcus*, constitute the most important phytoplanktonic primary producers. Independently, metagenomic sequencing of the Sargasso Sea recently showed the dominance of *Prochlorococcus* over other oxyphototrophic organisms. Therefore, for marine cyanobacteria, emphasis has been on genome projects in

this group so far. These studies demonstrated several unexpected findings that have implications for current and future genome analyses not only of cyanobacteria, but more general of marine bacterioplankton species, including non-oxyphototrophic bacteria as well as of certain eukaryotic species as well.

The ribosomal RNA sequences of these four cyanobacteria are less than 3% divergent, yet there is an unprecedented degree of genomic variation among them, however, the set of genes common to all four cyanobacteria is only somewhere between 1,000 and 1,200, whereas the genome size and gene number per genome is quite small with 1.6 – 2.3 Mbp, and 1,716 – 2,525 genes. Thus, here it has become possible for the first time for an ecologically important prokaryote to elucidate how the specific gene content has diverged to reflect the particular ecological niche of each strain. This points to a phenomenon that might be typical for a wide variety of open ocean bacteria: the genome under study is highly streamlined and the individual number of genes in a single genome can be very small, but these bacteria act and respond to changes in the environment with their large population size. As a consequence, it is very likely that the total gene space (the "pan genome") available to these bacteria is substantially larger than just several thousands.

The goal of our analyses is to characterize the ecotype-specific genes and the regulatory apparatus controlling their expression. Indeed, the evolution and ecotype-specific distribution of genes for light-harvesting proteins and their pigments is in agreement with this ecotype concept: a spectrally tuned



SESSION I

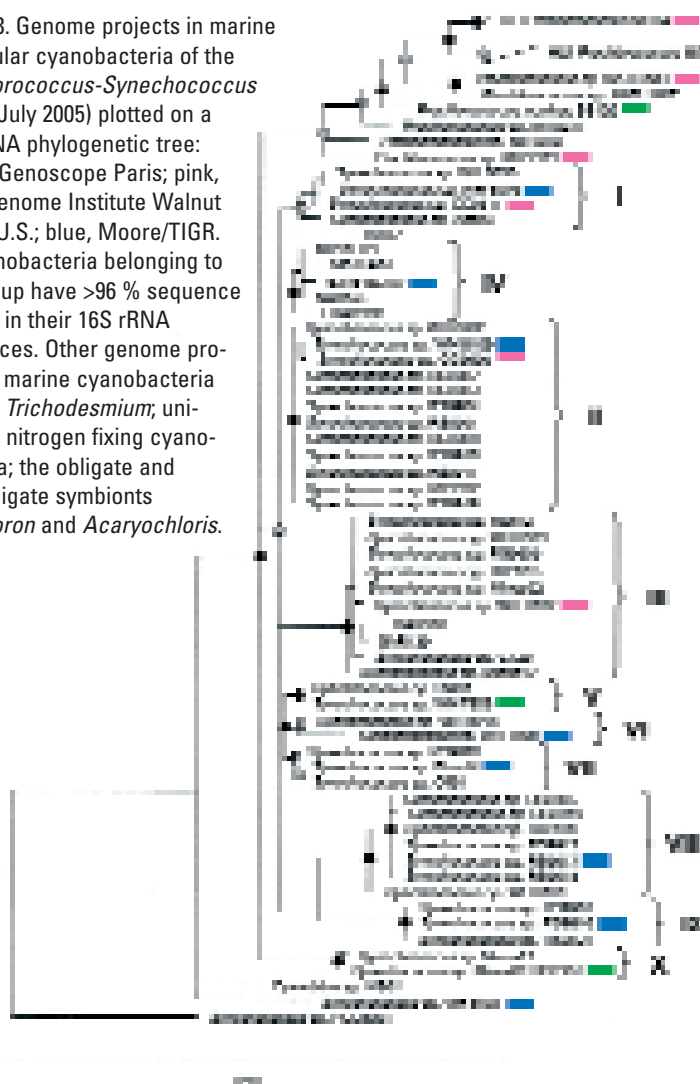
phycoerythrin is present in *Prochlorococcus*. It is a green-light absorbing pigment in isolates from close to the water surface and a mainly blue-light absorbing pigment in strains isolated from deeper water layers. The analysis of natural populations clearly showed that these facts, which were initially based on the total genome data, can be generalized to natural populations of *Prochlorococcus*. More important for photosynthesis is the evolution of a particular light-harvesting system in the *Prochlorococcus* group and in some other, not closely related cyanobacteria. Again, there is clear-cut evidence that the evolution of the genes for the

Prochlorococcus light-harvesting system, the *pcb* genes, is driven by the irradiance levels available in the parts of the euphotic zone. Intriguingly, the amplification of this gene family in low light-adapted strains has not only influenced the number of genes but also contributed to their functional diversification, i.e. specialized gene products for either photosystem I or photosystem II or expressed under certain conditions only.

Despite these exciting examples there is a whole group of genes that has remained completely overlooked during genome analysis and annotation. Recent data from the analysis of model organisms

have shown that aside from regulatory proteins, bacteria also possess a significant number of regulatory non-coding RNAs (ncRNAs). Some of these ncRNAs are key regulatory molecules acting on transcription factors, sigma factors and mediate such important regulatory processes such as iron homeostasis or redox control. Using a comparative genomics approach, we have developed algorithms to predict ncRNA genes in these organisms and subsequently tested their expression under various growth and stress conditions that are encountered in the natural environment. This has resulted in the identification of several new ncRNAs, riboswitches and possible attenuator regions in marine cyanobacteria.

Figure 3. Genome projects in marine unicellular cyanobacteria of the *Prochlorococcus-Synechococcus* group (July 2005) plotted on a 16S rRNA phylogenetic tree: Green, Genoscope Paris; pink, Joint Genome Institute Walnut Creek, U.S.; blue, Moore/TIGR. All cyanobacteria belonging to this group have >96 % sequence identity in their 16S rRNA sequences. Other genome projects in marine cyanobacteria include *Trichodesmium*; unicellular nitrogen fixing cyanobacteria; the obligate and non-obligate symbionts *Prochloron* and *Acaryochloris*.



SESSION I

Environmental Diversity and Organizing Marine Environmental Genomics

Arthur R. Grossman

Carnegie Institution of Washington
Department of Plant Biology
Stanford, California, USA

The Hot Spring Environment: There are numerous natural habitats within hot spring environments with important differences in their structures and the factors that influence growth and interactions among component organisms; these factors include light levels and quality, temperature and nutrient conditions. One habitat that has attracted our attention is that of the thermal-tolerant, microbial mats. These mats are generally highly structured with respect to microbial composition and distribution, and often dominate hot-spring environments (extending up to temperatures of 88°C). A project directed toward understanding genetic diversity within mat ecosystems and adaptation of the organisms of the mat to their environment has been funded by an NSF-FIBR grant¹.

Hot-spring microbial mats are natural biofilms in which photosynthetic and non-photosynthetic organisms are distributed along horizontal thermal and vertical light gradients. We sequenced the genomes of two unicellular thermophilic cyanobacterial isolates (*Synechococcus*) that occupy overlapping thermal niches (one from 40°C to 60°C and the other from 40°C to 65°C) within the microbial mat community of well-studied hot springs in Yellowstone National Park. While these genomes have highly similar gene content, they are remarkably different with respect to gene arrangement. There is essentially no synteny between the genomes of these organisms, in spite of the fact that there is an approximate 90% similarity with respect to putative orthologous genes between the genomes. Furthermore, both genomes harbor a full complement of genes for nitrogen-fixation, a process reported not to occur in the high temperature zones of these mats, and not yet reported for unicellular,

thermophilic fresh-water cyanobacteria. These and other findings provide insights into the biological processes that have evolved in and that distinguish these two cyanobacterial ecotypes, and that enable them to thrive the thermal and nutrient conditions of the hot-spring environment.

Marine Genomics: The marine environment can vary from rich eutrophic coastal waters, to relatively nutrient-poor waters of the open oceans, to the intertidal zone where organisms can experience desiccation and physical abuse caused by wave action. There are many free-living organisms that populate these environments, but a diversity of symbiotic associations is also apparent. We have recently initiated a project focused on defining acclimation processes in marine cyanobacteria, including strains of *Prochlorococcus* and *Synechococcus*.

¹NSF-FIBR grant awarded to Dave Ward, Devaki Bbaya, Arthur Grossman, Fred Cohan and John Heidelberg.

Environmental Genomics, Sampling, Scope and Potential Constraints

Ian M. Head

University of Newcastle
School of Civil Engineering and Geosciences,
Centre for Molecular Ecology and Institute for
Research on the Environment and Sustainability
Newcastle-upon-Tyne, UK

Marine genomics is an exciting growth area in the biological sciences and is benefiting from initiatives such as the J. Craig Venter Sorcerer cruise, NSF Microbial Observatories Program, the Moore Foundation Marine Microbes Sequencing Initiative, and national and international efforts within the EU. These include "Marine Genomics Europe", the UK Natural Environment Research Council thematic programmes on "Environmental Genomics" (Figure 4) and "Post Genomics and Proteomics".



SESSION I

In Silico analysis of published genome sequences to inform environmental genomic studies. The increasing number of completed and draft prokaryote genome which are or will be available in the near future offers the opportunity to conduct *in silico* analysis of artificially constructed microbial communities that may offer insights into the potential scope and limits of metagenomic analysis and identify the merits of particular experimental approaches. For example a preliminary analysis of published genome sequences suggests that a limited number of genes (classified on the basis of GO terms from the EBI genome digest) are found in the vicinity of ribosomal RNA genes. This is important for the strategy adopted in real metagenomic analysis and suggests that using rRNA genes as a phylogenetic anchor is likely to facilitate useful comparative analysis since the genes within 100 kb of rRNA genes do not appear to represent a random sampling of all possible genes present in completed genome sequences. In contrast, if the purpose of an experiment is to capture as broad a sample of genes from the metagenome then an approach that does not rely on co-localization with an rRNA gene may be more appropriate.

Enabling comparative metagenomic analysis by reducing sample complexity. As available metagenome data grows, many opportunities will emerge for comparative metagenome analysis. Sufficiently comprehensive sampling of a given metagenome is required for meaningful comparative analysis. There is a danger that undersampling may result in differences between samples or environments being inferred, not because genuine differences exist but because the metagenome has been poorly sampled. Recent reports suggest that even with relatively poor sampling there may be some detectable signal that can, for example, be associated with different environments. Comprehensive sampling can be obtained when the complexity of

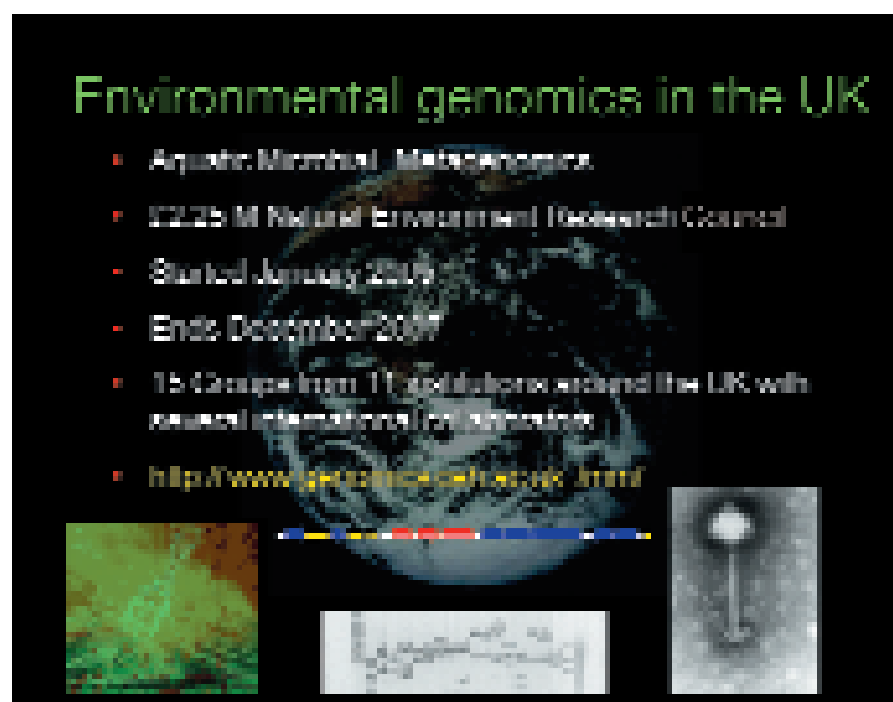


Figure 4. Environmental genomics in the UK is funded by NERC.

the microbial communities analysed is low. Not only does this permit robust comparative analysis but also facilitates the application of post-genomic approaches such as proteomics-based analysis because the sequence data will be available to reliably identify unknown proteins and examine the expression of ORFs which currently have no known function, in response to changing environmental conditions.

For more complex environments there are a number of strategies that may be suitable for targeting specific components of a complex microbial community. These include the use of stable isotope probing using ^{13}C -labelled substrates and selective purification of ^{13}C -labelled DNA from the subset of the community that has incorporated carbon from the labelled compound. In addition specific populations of cells sorted by flow cytometry using a range of different labelling approaches could be targeted for metagenomic analysis.

SESSION I

A Metagenome Analysis of an Extreme Microbial Symbiosis¹

S. Craig Cary

University of Delaware, College of Marine Studies
Lewes, Delaware, USA

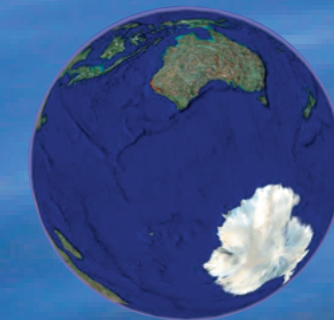
Early efforts in community level environmental genomics require tractable biological systems that can be surveyed in an environmental context. The hydrothermal vent polychaete *Alvinella pompejana* is characterized by a dense, specific epibiotic microflora associated with the worms' dorsal integument. Now considered the most thermotolerant and eurythermal metazoan, *A. pompejana* has been found on the walls of high temperature vent chimneys from 13°N to 32°S along the Eastern Pacific Rise (EPR). *A. pompejana* adult worms are found in the hottest sections of black smoker chimneys, where they construct tubes directly on the chimney wall. The environment of diffuse flow habitats surrounding the worms' tubes has been characterized not only by its temperature (ranges from 2°C to in excess of 100°C), but also by high levels of total (free + complexed) hydrogen sulfide (>1 mM) and high concentrations of heavy metals (0.3-200 μM). These diffuse flow sites, where *A. pompejana* flourishes, consist of areas of intense mixing of end-member type fluids with ambient seawater (2°C) creating thermo-chemical gradients that are unmatched anywhere else on the planet. Past studies suggest that the epibionts of *A. pompejana* inhabit an environment where an unparalleled temperature gradient exists within the space of only a few centimeters. The posterior symbionts may experience environmental temperatures over 80°C while the anterior symbionts encounter 20°C or lower - an apparent environmental gradient of almost 60°C. Genetic analysis showed that the symbiotic community of *A. pompejana* is composed of closely related (<98% on 16S rRNA), though diverse assemblage of 20-40 phylotypes constrained within a single bacterial subdivision, the epsilonproteobacteria.

We undertook an metagenomic analysis of the *Alvinella* epibiont community to address the following hypotheses. 1) Under the constraints imposed

by the geochemical environment the epsilon epibiotic consortia employ a core basic metabolic strategy, 2) The epibiotic association can be defined as a mutualistic partnership enabling survival in this hostile environment, 3) The epibiont community employs an extensive range of sensing capabilities to change and respond to their dynamic environment, 4) The epibiont community detoxifies the harsh chemical environment experienced by the worm, 5) Protein eurythermalism is a common adaptation of the *Alvinella* symbiont proteome. We constructed a small insert (1500-3000 bp) shotgun library from a single worm's epibionts and sequenced over 250,000 clones in the forward and reverse direction. Over 377,000 sequencing reads were performed resulting in over 134 Mb retained after quality trimming. This amounts to 54 epsilonproteobacterial genome equivalents (2.5 Mb). Our gene finder has identified over 210,000 ORFs for further annotation. Analysis of the 16S rRNA and rpoB genes in the metagenome library has confirmed the diversity of the community to be solely epsilonproteobacteria dominated by two phylotypes. The analysis has revealed the diversity to be considerably greater (over 110 genomes) than estimated from previous studies resulting in severe problems in the assembly of large contigs.

Preliminary sequence and gene frequency analysis has immediately identified the prevalence well represented complete pathways in the metagenome. Genes diagnostic for certain carbohydrate metabolisms (glycolysis, gluconeogenesis, rTGA, mannose uptake) are represented in high frequency in the library. For nitrogen nitrate, nitrite, nitric and nitrous oxide reductases are all probable pathways for nitrogen metabolism. An abundance of metal detoxification genes and pathways (Ag efflux pump, telluride resistance gene, Co/Zn/Cd cation transporters, Co/Zn/Cd efflux pump) have been discovered consistent with the environment in which this community thrives. These data have allowed us to begin to form a preliminary 'core' metabolic model.

¹With co-authors B. J. Campbell, R. Feldman, G. Gao, J. Grzymalski, M. Kaplarevic, and A. E. Murray.



SESSION I

Molecular and Genomic Tools to Study the Ecology of Marine Protists

Ramon Massana

Institut de Ciències del Mar (CSIC)
Barcelona, Spain

Microorganisms play central roles in marine ecosystem functioning due to their huge abundance and high metabolic rates. Molecular studies have revealed the identity of the main groups within bacteria and archaea and their putative function in the sea. Marine protists also play significant roles in global carbon and mineral cycles, and their species identification often require the use of molecular tools too. Compared with prokaryotes, smaller metabolic versatility is expected for marine protists and therefore genomic analysis could seem less exciting: some are phototrophs (primary producers) and others are heterotrophs (grazers or parasites). Nevertheless, each species has a particular suit of adaptive traits that modulates its life strategy (generalist versus specialist), trophic mode (pigment composition, prey spectrum, host range) and response to biotic and abiotic factors. All these traits, that finally explain the success of each species in the sea and therefore its distributional range, are obviously codified by genetic information. In addition, some species could contain interesting metabolites for biotechnological applications. Therefore, marine protists represent a significant reservoir of marine diversity that awaits to be explored.

Molecular surveys for protist identification have unveiled a huge diversity and the presence of novel phylogenetic lineages. Some of these novel groups seem to represent a robust component of marine assemblages that have escaped scrutiny by science so far. Such is the case of MAST-4 (marine stramenopiles group 4), whose 18S rDNA sequences have been retrieved from almost all marine systems investigated (Fig. 5A). Using molecular probes targeting MAST-4 cells, it has been shown that these uncultured protists are globally distributed bacterivorous flagellates of very small size (Fig. 5B). The large temperature range of this group could be explained

by phenotypic plasticity of the same genotype or the existence of genetically determined ecotypes adapted to different temperatures. Indeed, the genetic variability within MAST-4 sequences would favor the second explanation. Parallel to these environmental studies, it is of high priority to isolate a MAST-4 representative. The genome size in these cells is predicted to be small and therefore a genomic project would be feasible and very interesting both for evolutionary (gene organization on very small eukaryotes) and ecological (success in marine systems) aspects. Finally, since the genetic variability in natural communities is considerable, it is not realistic to envision a genomic project of each variant, so metagenomic analysis remains as the only possibility to access the variability of different genes in the environment.

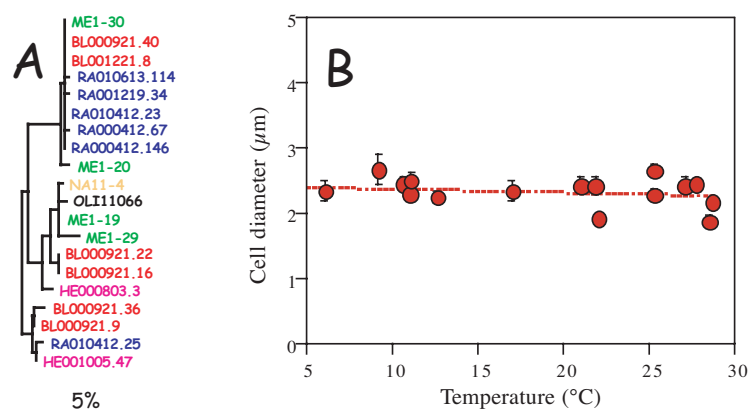


Figure 5.
(A) Phylogenetic tree of 18S rDNA sequences of MAST-4 organisms. Clones from different codes and colors come from different genetic libraries.
(B) Cell diameter of MAST-4 cells observed by FISH in a variety of marine environments comprising a large temperature range.

SUBREPORT FROM SESSION II: Multicellular Organisms

Rapporteurs: Catherine Boyen
and Douglas Crawford

SUMMARY:

Although eukaryotes comprise a large number of independent lineages, only five have evolved complex multicellularity in parallel: animals, green plants/algae, fungi, brown algae and red algae. Marine organisms are present in all of these lineages with the exception of fungi. Eukaryotic genomes are often orders of magnitude larger than those of bacteria. To date, few genomic studies in marine eukaryotes have been initiated. They are often directed by small, rather national groups of investigators. One notable exception (and success) has been the establishment of international consortia conducting the production of expressed sequence tag (EST) libraries for aqua-cultured species of oyster and catfish. Unfortunately aqua-cultured species represent a limited resource for addressing questions of environmental or evolutionary importance.

As a consequence, genomic research involving marine multicellular eukaryotes has fallen far behind its microbial counterpart in terms of communication among research groups, technological capability, and last, but perhaps most critically, in terms of accomplishments.

It is only recently that access to large sequencing facilities was opened for research groups working on organisms not related to health, biotechnology or agriculture. It is now timely to initiate a better coordination between the US and Europe in order to implement genomic high through-put approaches that will allow us to increase our knowledge of the biology of key organisms fundamental to the functioning of marine ecosystems.

RECOMMENDATIONS:

Large scale projects such as the sequencing of complete multi-cellular eukaryotes and/or the generation of massive number of EST are costly and require the establishment of robust international consortia ensuring that a large community of scientists

will be involved in expert annotation and exploitation of data. A prerequisite to the implementation of such large sequencing projects is agreement on the choice of organisms. Several criteria can be considered such as phylogenetic relevance, economic interest, etc. Workshop participants also felt that, despite its high profile in research with marine organisms, the possible role that meta-genomics studies might play in future research involving eukaryotes is difficult to predict and/or prioritize at the present time.

(1) Coordination/Harmonization: Our first recommendation is to **set up an EU-US steering committee** with the tasks of (i) guiding identification of target organisms for sequencing, (ii) promoting synergism among individual researchers using genomic approaches to study marine eukaryotes and (iii) avoiding redundancy among research projects.

(2) Education: More cross-training is needed to promote the use of genomic/bioinformatics approaches in research at the undergraduate, graduate and postdoctoral levels. However, such emphasis on genomic approaches should always be balanced by an appropriate appreciation for interdisciplinary investigation.

(3) Technology: Technologies are only beginning to come on-line suggesting that whole genomic sequencing of large numbers of eukaryotic organisms may become feasible and affordable. More resources are required to develop such rapid, cost-effective sequencing technologies, as well as appropriate tools for representing and analyzing the resulting data. At the same time, there is also a great need to support the development of sequencing, microarray analysis and bioinformatic tools for expressed sequence libraries from a broader range of eukaryotic marine organisms. Research involving microarray techniques is progressing rapidly, with application to the study of organismal biology, population biology, evolutionary biology and even to environmental/ecosystem health.



INDIVIDUAL CONTRIBUTIONS IN SESSION II

The Ectocarpus Genome Project

Mark Cock and Catherine Boyen

CNRS & UPMC, Station Biologique de Roscoff
France

As a result of impressive developments in the fields of genetics and molecular biology over the past decades, we now have a very detailed understanding of the biology of a few selected model organisms. However, although the choice to concentrate on a limited number of model systems was essential to the success of this approach, it has not allowed full profit to be drawn from the large amount of informative diversity inherent the marine system. More recent developments in genomics and high-throughput methods are changing this situation, opening up the possibility of developing additional models at strategic positions in the phylogenetic tree of life as a means to gain access to this rich source of biological information.

One phylogenetic lineage that deserves greater attention is the brown algae as this group represent one of only five eukaryotic lineages that have evolved complex multicellularity (the four others being animals, green plants/algae, fungi and red algae; Baldauf et al., 2000; Figure 6).

Surprisingly this feature of the brown algae remains almost totally unexplored at the genomic and genetic levels. Brown algae are also of interest for a number of other reasons ranging from their relevance to fundamental evolutionary questions to their enormous potential as sources of bioactive molecules.

Based on a recent survey of a large range of brown algal species, the Algal Genetics Group in Roscoff has decided to develop the filamentous alga *Ectocarpus siliculosus* as a model species for this group (Figure 7). The choice of *Ectocarpus* was based on several characteristics including its small size, the fact that the entire life cycle can be completed in Petri dishes in the laboratory, its high fertility and rapid growth (the life cycle can be completed

in 2 months), the ease with which genetic crosses can be carried out and the relatively small size of the genome (214 Mbp compared with 1095 and 640 Mbp for *Fucus serratus* and *Laminaria digitata* respectively). Moreover, the Ectocarpales are closely related to the most economically important brown algal group, the Laminariales.

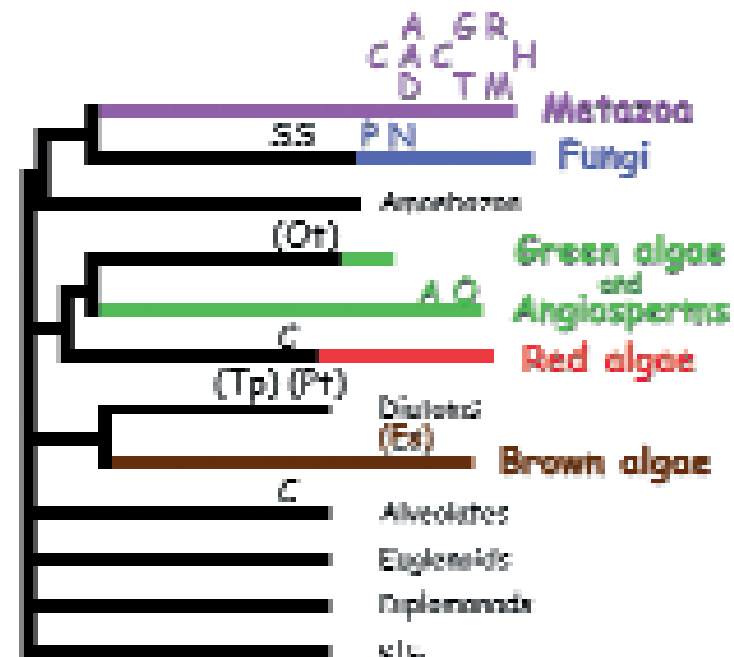


Figure 6. Phylogenetic tree of the eukaryotes showing the five major lineages that have evolved complex multicellularity (in colour). Species for which whole genome sequences are available or in progress (the latter in brackets) are indicated on the tree (multicellular organisms are indicated in colour). Es, *Ectocarpus siliculosus* would be the first Brown algae.

The Roscoff laboratory is currently involved in a collaborative project with groups in Germany (N. Delaroque, MPICE, Jena) and England (C. Brownlee, MBA, Plymouth) that aims to develop a transformation protocol and RNAi technology for *Ectocarpus*. A large number of transgene constructs have been made and several approaches are being tested for gene delivery including biolistics and microinjection.

SESSION II

A proposal requesting 10x coverage of the *Ectocarpus* genome (4,280,000 sequencing reads) has been accepted and is now in progress at Genoscope. The project also includes large-scale sequencing (100,000 reads) of clones from *Ectocarpus* cDNA libraries corresponding to different developmental and physiological conditions.

Marine Environmental Genomics: Integrating Complex Multicellular Genome Response to Changing Environment

Douglas L. Crawford

Rosenstiel School of Marine and Atmospheric
Science, University of Miami, USA

Marine multicellular organisms represent a vast diversity of macrofauna. These include organisms with little cellular-tissue organization, to organisms that are diploblastic or triploblastic (three primary cellular layers: endo-, meso- and ectoderm). Developmentally, marine multicellular organisms include the full diversity of anterior-posterior or dorsal-ventral patterns of development. They encompass organism that resemble humans in having a complex brain, hormonal, cardio-vascular, and immune systems and thus are important for medical research. For example, the integration of different organ systems to produce an adaptive response to hypoxia or any other environmental insult can only be studied in multi-cellular organisms. These variations provide both an important evolutionary and physiological perspective not available from terrestrial macro-fauna nor from unicellular life.

Marine Environmental Genomics explores how the genome and proteome interacts with and integrates cues from the environment in producing appropriate and adaptive responses. This has been achieved in some so-called 'model' species in laboratory situations, notably yeast, but a far richer understanding comes from characterizing responses in species exposed routinely to environmental stress in natural situations and within an appropriate ecological context. Such organisms frequently possess more powerful responses that are thus clearly displayed and analyzed. Characterizing natural responses to environments requires organisms that are 'wild' and that have 'felt the filter of natural selection' and achieved success. It also benefits from comparisons between alternative evolutionary solutions. For example, fishes represent an important group in the evolution of vertebrates. They are the most speciose of vertebrates (28,600 extant species,

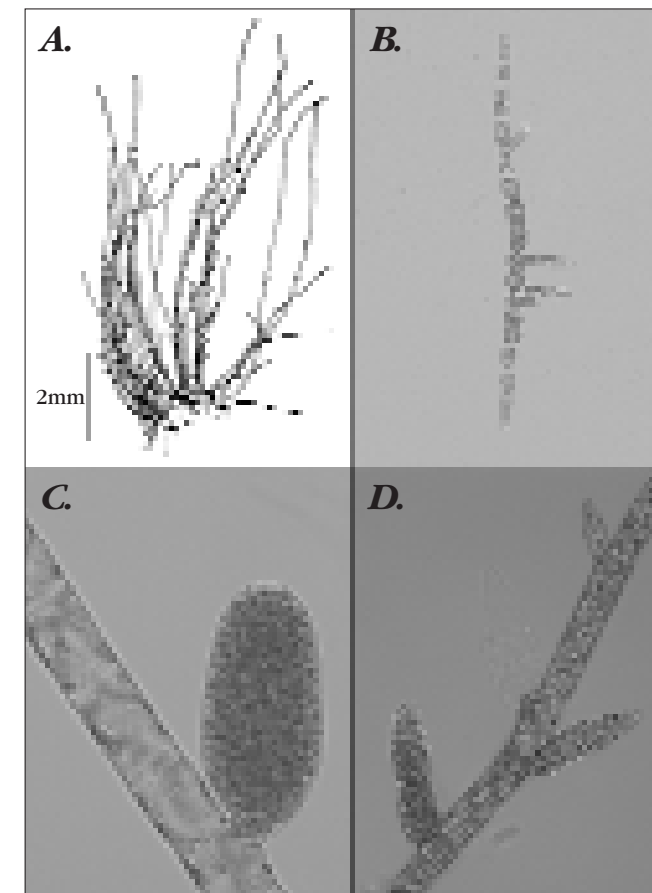
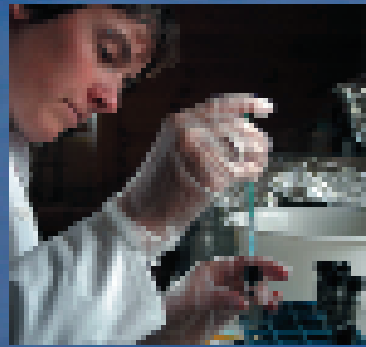


Figure 7. The filamentous brown alga *Ectocarpus siliculosus*.

- A. Mature sporophyte (cartoon)
- B. Sporophyte germling
- C. Unilocular sporangium on a sporophyte
- D. Plurilocular gametangia on a gametophyte

SESSION II



versus 4,629 in mammals and 9,946 in birds). This diversity of species, many of them closely related, provides for powerful phylogenetic comparisons. Currently, among ray-finned fishes there are approximately 1.8 million nucleotides that code for approximately 80,000 proteins, from 8,700 taxonomic groups. There are 174 fish genomes in the NCBI database. Most of these are mitochondrial genomes rather than nuclear ones. 44,538 of fish gene sequences are UniGenes. Unigenes attempt to distill the large number of ESTs encoding the same protein to a single entry. Unigenes do reduce the number of "genes" but fail to put all appropriate sequences together. The number of Unigenes is an interesting statistic because NCBI has to make the decision to analyze the collection of ESTs from a species. The relative lack of fish Unigenes reflects NCBI priorities. These data suggest that there is considerable DNA sequence information for fishes, sequence information that could be used by fish physiologists, toxicologists and biochemists and thus provide functional information to genomes.

The Complex Response of Marine Crustaceans and Bivalves to Bacterial Pathogens

Karen Burnett

Hollings Marine Laboratory
Charleston, South Carolina, USA

The relationship of marine multicellular organisms with their microbial community is maintained by interdependent physiology, immunological and molecular mechanisms. For example, the immune response of the penaeid shrimp *Litopenaeus vannamei* and the blue crab, *Callinectes sapidus* (Fig. 8), to the bacterial pathogen *Vibrio campbellii* has dramatic physiological consequences. Hemocytes aggregate in the presence of bacteria and become trapped in the gill, leading to depression of oxygen uptake and oxidative metabolism. Field studies confirm that blue crabs harboring low level bacteri-

THE FUTURE:

What needs to be established is the utility of these genomic and EST sequence to study the environmental impact on marine organisms. Currently, there are only a very few microarrays or well annotated ESTs to provide the bioinformatics or the physical tools for genome analyses. Additionally, there is a significant complexity among marine macrofauna: many homologous genes (genes evolved from common ancestor) have many different names in many different taxa, and many genes have recent duplication and tissue specializations. This duplicity in names, and duplication of similar genes for different function, provide a challenge not seen among unicellular organisms. For marine genomics of multicellular organisms to achieve success in defining how organisms adapt and how changes in the environment impact our oceans much more research needs to be done to: (1) to provide a common annotation of homologous genes across organism, (2) EST and genomic sequences of both vertebrate and non-vertebrate organisms and (3) experimental studies of changes in gene expression relative to changes in the environment.

al infections are functionally hypoxic and may suffer diminished ability to forage and avoid predators. In another approach to understanding the host: pathogen relationship in crustaceans and oysters, we are using quantitative real time PCR to track tissue distribution and fate of *V. campbellii* as well as monitoring hemocyte trafficking and the induction of hemocyte-specific genes in response to the pathogen.

Microarrays built from oyster and shrimp EST libraries may provide a more suitable platform for studying the complex host: pathogen relationship and its response to environmental change. In collaboration with the Marine Genomics Core at the Hollings Marine Laboratory in Charleston, SC, we will be generating transcriptional profiles of oyster and shrimp tissues over the course of infection under a range of environmental conditions. Research involving microarray techniques is progressing rapidly,

SESSION II

with application to the study of organismal biology, population biology, evolutionary biology and even to environmental/ecosystem health.



Communication with this project is available through the www.marinegenomics.org, a web-based interface bioinformatics data management and data analysis pipeline developed by informatician at the Medical University of South Carolina for the Marine Genomics Consortium at the Hollings Marine Laboratory in Charleston, SC. The pipeline automates processing, maintenance, storage and analysis for 16S RNA and microarray experiments

Genomics Approaches to Fish and Shellfish

Adelino V. M. Canário

Centre of Marine Sciences, University of Algarve
Faro, Portugal

A renewed interest in the Oceans, sparked by promises of "hidden" economic treasures, rapid changes in technologies and lower costs are facilitating the spreading of Genomics approaches to Marine Sciences. And although only a handful of marine metazoans have been until now targets for development of genomics tools, this number is expected to increase in the near future with important immediate contributions to aquaculture, fisheries, population biology, ecotoxicology and biotechnology.

Already, several aquaculture projects take advantage of the wealth of information and potential offered by the growing number of genomics resources. These projects are especially in the fields of animal health and ultimately aim at selective breeding of

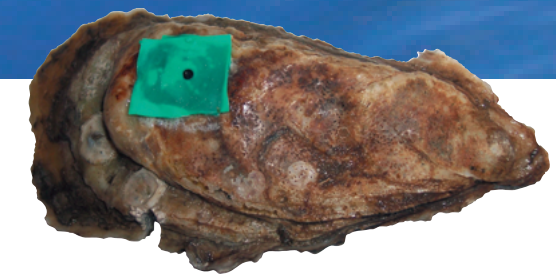


Figure 8. Shrimp, blue crab and oyster are studied by microarrays.

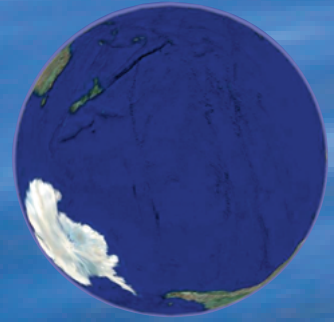


for more than 26 marine species (eukaryotic and prokaryotic). Species data are maintained by registered users from local and remote locations in Europe and South American, as well as the United States.

stress or disease resistant strains using genetic markers to speed up the selection process (e.g. marker assisted selection). Summarized examples from fish and oysters illustrate these efforts.

Oysters (*Crassostrea gigas*, *Ostrea edulis*) suffer heavy losses from "summer mortality" the mechanism of which still has to be elucidated. However, there is differential sensitivity and resistance to this phenomenon which is being investigated by using multiple approaches to gene expression in different tissues and environmental conditions, including suppressive subtractive hybridization and microarrays. Polymorphic markers, such as SNPs from regulatory and coding regions of candidate genes, are being developed and used to map the Quantitative Trait Loci (QTL). On a more global scale a series of genomics tools, including linkage maps, BAC libraries and a growing number of EST are now available. An Oyster Genome Consortium promotes resource development and aims at sequencing the Pacific oyster (*C. gigas*) genome.

The main cultivated marine fish species in Europe



SESSION II

are the sea bass (*Dicentrarchus labrax*) and sea bream (*Sparus auratus*) with a production of over 12,000 tons / year. Genomics projects have been directed largely at stress tolerance and disease resistance. Medium density genetic maps, radiation hybrid maps, BAC libraries, microarrays and a range of cDNA libraries are available. Marine Genomics Europe (MGE) has just sequenced 60,000 EST from the two species and is developing large microarrays. As an objective MGE aims to develop a range of genomics tools that can be applied to aquaculture, fisheries and environmental management.

Despite important advances, few marine metazoan (e.g. sea squirt, puffer fish) have been or are being sequenced. Moreover, the planned genome sequencing still has important gaps in the animal evolutionary scale. For example, the crustaceans are not contemplated. Although there are crustaceans with smaller genomes, we suggest that the green crab, *Carcinus maenas*, a decapod crustacean, should be sequenced. It has a genome slightly larger than that of the puffer fish. In addition to its evolutionary interest, there are biological, economic and ecological reasons to target it for development of genomic resources. It is a species native of European waters which occupies a wide range ecological niches, it is highly adaptable, it is invasive in North America and other regions, and it has a long history as an object of scientific study by different disciplines.

Some of the European Commission funded projects, ongoing or under completion, focusing on genomics approaches to aquaculture have the acronyms: Bassmap (www.bassmap.org)
Bridge-map (www.bridgemap.tuc.gr)
Aquafirst (aquafirst.vitamib.com)

Environmental Genomics of Polar Organisms

Alex Rogers

British Antarctic Survey, Cambridge, UK

Environmental genomics is the study of genomic responses to environmental variation over a range of timescales. Molecular studies to date have demonstrated that single point mutations leading to a single amino-acid change can have profound consequences in how organisms interact with the environment. At high latitudes, physical conditions are extreme and evolutionary adaptation has occurred conferring greater flexibility in a range of proteins. Similar adaptations have been detected in temperate organisms, however, permanent changes in gene expression have also been identified as a mechanism of thermal adaptation. Recent work in the Antarctic suggests that permanent up-regulation of stress response proteins may have also occurred in polar organisms. Work within the British Antarctic Survey BIOFLAME (Biodiversity: Functions, limits and adaptation from molecules to ecosystems) programme is trying to integrate research on marine animal on the scales of genome/proteome, individuals/populations/species and habitats/ecosystems and the system Earth.

Diversity: Gene to ecosystems



Figure 9. Research in the BIOFLAME project involving genomics, taxonomy and ecology.

SESSION II

Interacting Genomes in the Squid-Vibrio Symbiosis

Edward Ruby and Margaret McFall-Ngai

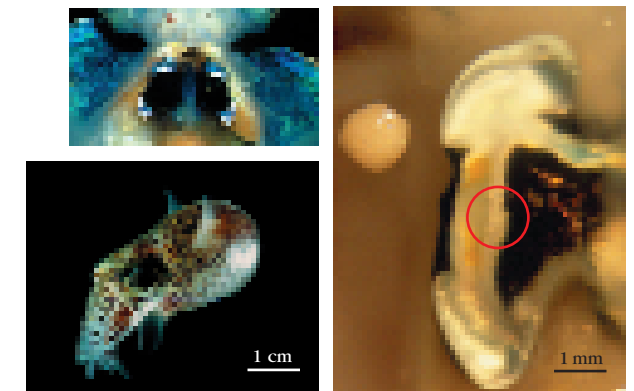
Department of Medical Microbiology and Immunology, University of Wisconsin Madison, Wisconsin, USA

Many microbial species, and essential all multicellular animals and plants, exist in symbiotic associations with each other. Marine animals and plants live their whole lives bathed by seawater, which contains a vast diversity of microbial life, yet we are just beginning to understand the ways in which these associations initiate and persist.

A number of marine animals, such as the bioluminescent bobtail squids, have developed monospecific associations with particular bacteria that provide an easily recognized function to the host. Some of these symbioses have provided model systems for discovering the mechanisms by which a eukaryotic genome and a prokaryotic genome learn to communicate. I will discuss recent work examining the global transcriptional responses of both *Euprymna scolopes* and *Vibrio fischeri* as they initiate the development of a specific, stable bioluminescent organ symbiosis (Fig. 10). The results indicate the role of host nutrients and bacterial light emission in controlling gene expression in these partners, and point out the value of genomic approaches to discovering the role of preadaptation in the development of symbioses.

The squid light organ and its bacterial symbionts

Euprymna scolopes



Vibrio fischeri

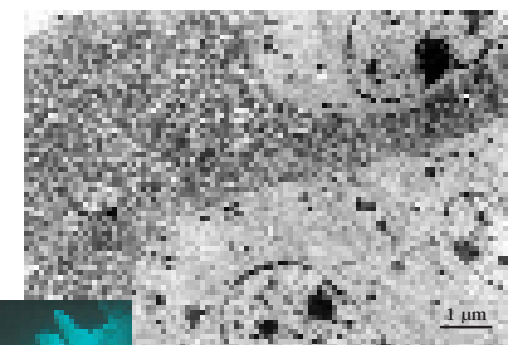
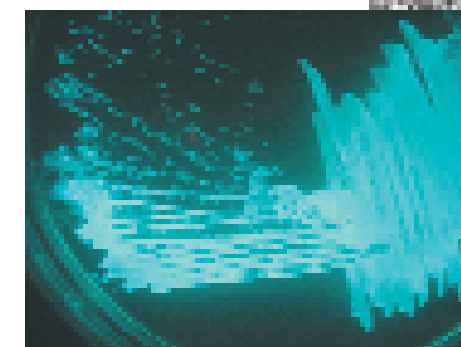
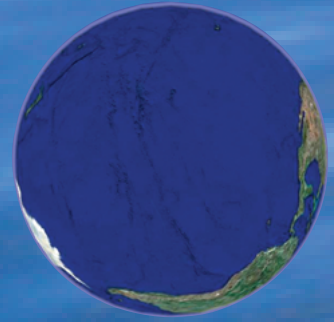


Figure 10. The light organ of the squid *E. scolopes* harbours chemiluminescent bacteria.





SUBREPORT FROM SESSION III: Bioinformatics

Rapporteurs:

**Karin Remington and
Frank Oliver Glöckner**

SUMMARY:

Molecular biology and bioinformatics have undergone a co-evolution. Triggered by large sequencing projects such as the human genome project, as well as the availability of hundreds of microbial genomes, the pressure to apply and develop bioinformatics for data storage and processing has rapidly increased. Often seen in the beginning as a necessary but annoying addition to the daily work of lab scientists, bioinformatics is now a well established and maturing field of research. Today, bioinformatics is used to generate hypotheses or to find new specific targets for further wet-lab evaluation. In the field of Marine Genomics sequence data are currently the most promising key to access to the wealth of genes and functions, thereby addressing questions of niche adaptations and resource utilization. Coupled with post-genomic techniques this should lead in the future to a more complete understanding of marine ecosystems functioning.

Assembly and ORF prediction

The Venter Sargasso Sea metagenomic approach has initially provided 1.3 Mio genes and data from the Sorcerer II expedition will give access to more than 13 Mio additional genes. Since a big part of the data in marine genomics is currently provided by metagenome projects, an important question arising from discussions was how to adapt traditional genome assembly processes to this type of data. Current assembly tools were designed with the goal of assembling whole genomes. Assembly parameters vary from group to group, complicating efforts to compare results. For this reason, there was broad agreement that, at least, parameters such as thresholds for overlaps and the quality values used have to be made more transparent. It is also important to make raw (unassembled) data more readily available, since an optimum set of assembly parameters cannot be defined, and varying these parameters may provide valuable insight that may not be revealed with any static published assembly.

Conditions for assembly would likely vary with the question being asked. There was a clear agreement that benchmarking is needed (e.g. artificial fragmentation of genomes from environmental organisms and subsequent reassembly, comparing the results of various assemblers and parameter selections). Furthermore, new techniques, such as intrinsic genomic signatures, should be integrated to aid in the assignment of fragments from community sequencing approaches (see report of Dr. Glöckner). A major concern of the discussion participants was that development of new assemblers specific for metagenomic studies would not be supported at sufficient funding levels because it is considered a "solved problem". Similarly, development of open reading frame (ORF) calling programs is not supported at the moment, although it was clearly illustrated that for environmental organisms further developments in this area are necessary.

Annotation

The presentation from Dr. Edwards highlighted the difference between traditional and subsystems-based annotations. In the traditional annotation model each protein is considered individually with little if any reference to its location on the chromosome. In this approach, a single annotator works methodically through each genome, and is consequently unlikely to be an expert in many of the genes that they are annotating. This approach is currently implemented in the software system GenDB (see report of Dr. Meyer). In contrast, the subsystems approach, first proposed by the Fellowship for the Interpretation of Genomes (FIG), uses experts in each area to annotate a subsystem, or group of genes, across all available genomes – and, if necessary, metagenomes. The SEED database and tools developed by FIG are explicitly designed to support these annotations. Both strategies have their pros and cons and therefore a combination of both systems is currently preferred. Plans are underway to merge GenDB and The Seed to provide such an integrated framework. Furthermore, the quality of the annotation in the public databases was identified as a crucial issue. Most databases act just as repositories of sequence data, and erroneous

Subreport from Session III: Bioinformatics

data and incorrect annotation are widely propagated by any tools reliant on these repositories. Dr. Lima pointed out that Swiss-Prot can fill in this gap by providing a well curated and reliable database.

The **release of data** was briefly raised by the participants. There are different policies, often depending on the funding agencies. Policies need to be clearly communicated. An important point in this context is how genomic data are exchanged. Currently, this is mainly done via **public databases** like Genbank or EMBL by providing flat-file formats. This mode leads to a significant loss of primary data and the conversion of data from different systems is problematic in terms of data consistency. Dr. Edwards advised the group of the Interoperability Working Group of the Bioinformatics Resource Centers. This group has been charged with exchanging information between different annotation centres associated with the NIH funded program for annotating pathogens. It was generally agreed that a common exchange platform would be preferred and standards have to be defined. Additionally, the proposal to establish a clearinghouse for the central exchange of data was discussed. Under this model researchers would be free to publish their data to a central clearinghouse in a unified format, and the data could then be downloaded and incorporated into any clearinghouse-enabled annotation or database system. A first approach in this direction has already been established within the EU-funded Network of Excellence "Marine Genomics Europe" where the GenDB system acts as a central data repository.

To fully exploit the information provided by the millions of sequence data in the context of the contribution of the organisms to the environment and their specific adaptations, **contextual data/metadata (e.g. location, biological, chemical and physical measurements)** are urgently needed (see report of Dr. Cases). Researchers submitting their sequence data (genomic-, metagenomic- or single gene-data e.g. rRNA) to public databases have to be more strongly encouraged to provide at least a minimum set of contextual information.

Further advancement in the field of marine genomics is hampered by the lack of training programs to help bridge the gap between biologists, bioinformaticians, computer scientists and statisticians. The group proposes that the EU and NSF consider a program to encourage molecular biologists to enter marine biology, and vice-versa. It was noted that there are several ad hoc programs for this, such as those being run by Dr. Cary in Maryland, Dr. Robwer in San Diego and the summer schools and workshops organised by the NoE "Marine Genomics Europe".

Recommendations:

- (1) Funding agencies need to be aware of the central role of assembly and ORF prediction which especially for metagenomics require additional research.
- (2) Long term financial support is needed for the reannotation of genomes, data cleansing and well curated databases such as Swiss-Prot.
- (3) Controlled vocabularies (e.g. those of FIG and the Gene Ontology consortium) with clear entity-relationship models should be further developed and used in annotation.
- (4) Funding for the sustainable development of a unified storage and exchange system for genomics must have priority.
- (5) Standardization efforts by JGI (www.jgi.doe.gov/16s/) and the Int. Census of Marine Microbes (<http://icomm.mbl.edu/>) should be promoted.
- (6) Support for crossdisciplinary training in marine biology/bioinformatics; mobility funds for junior scientists.



INDIVIDUAL CONTRIBUTIONS IN SESSION III

High-Throughput Marine Meta-Genomics: Challenges from the Bioinformatics Perspective

Karin A. Remington

The Venter Institute, Rockville, Maryland, USA

Toward our goal of engineering microbiological systems for carbon neutral energy production, the research team at the Venter Institute is looking to the environment to identify key genes and pathways to provide the basic machinery for these systems. Marine microorganisms, including eukaryotes, prokaryotes and viruses, influence the fate and transport of carbon in the world's oceans by means of the biological carbon pump. This pump is responsible for transporting carbon dioxide from the surface to the deep ocean, essentially sequestering carbon away from the atmosphere, thereby impacting global climate. However, our inability to culture the vast majority of these organisms has hindered our ability to study their activity in any detail. Beyond identifying what species are present in various ecosystems, of even greater interest are the roles they play and functions they provide in the complex web of life.

The strategy of our environmental program is therefore to broaden the view of the scientific community from a few model organisms and systems to the entire gene complement of various communities, using comparative analyses to gain a better understanding of the biological processes at play. With a broad survey of metagenomic data, complemented by a growing body of sequenced reference genomes, we will be able to investigate variation and conservation within a single gene or gene family, as well as the distribution of organisms and metagenomic variation across both small- and large-scale ecosystems. Understanding the functional space at play in nature will help inform the selection of pathways to harness in our energy production system.

With inspiration from the voyages of H.M.S. Beagle in the early nineteenth century and the Challenger Expedition in the 1870's, the Venter Institute's Sorcerer II Expedition embarked in 2003 on a global sailing voyage focused around "whole environment" marine microbial genomics. The goals of this Global Ocean Survey (GOS) are to discover new spe-

cies; to better understand microbial biodiversity; to discover new genes of ecological importance; and to establish a freely shared, global environmental genomics database that can be used by scientists around the world. Already our pilot study (Venter et al. (2004) *Science* 304, 66-74) revealed new insights into what had been considered a well-studied marine environment, the oligotrophic ocean gyre of the Sargasso Sea. We have now sequenced and begun to analyze data from an additional 34 sites along the initial leg of the Sorcerer Expedition. This broader study has allowed us to evaluate comparative genomic techniques and the dispersal and variation of genes and organisms of critical importance to biogeochemical cycling and local and global scales. Rather than finding "everything everywhere", we have found distinct environmental signatures, comparative correlations, and substantially new genes and gene variants with each sequencing increment. These findings suggest discovery has just begun, and build a compelling case for the extension of sequencing and analysis to sites extending beyond the Galapagos, across the Pacific and Indian Oceans to coastal Africa.

The benefit of scale and efficiency owed to recent technology improvements has led to an explosion in the genomic and meta-genomics data available to marine scientists. Even the savvy are struggling to come to grips with this data, while efficient, comprehensive and usable databases and analysis tools are still on the drawing board. The bioinformatics challenges are apparent. There is the long-standing need to address traditional informatics requirements such as standardized data formats and interfaces. Progress in this area is vital to facilitate "plug and play" tools that can be customized as scientific insight progress. There are also more complicated issues which may not appear on the surface to be "bioinformatics" challenges "per se", but which have serious ramifications for bioinformatics. Standardization (and rigorous use) of nomenclature is a challenge across the biological sciences, and sample collection protocols and metrics are particularly challenging in environmental studies. Even the evolving definition of "species" necessitates a new approach to sequence assembly and annotation, one that must be flexible enough to evolve to accommodate new developments in the science.

SESSION III

Marine Ecological Genomics Approaches Sequence Paradise

Frank Oliver Glöckner

Microbial Genomics Group

*Max Planck Institute for Marine Microbiology
and International University Bremen, Germany*

The quantum leap in sequence production has surpassed our current ability to store, analyse and classify these abundant data. Although gene prediction in prokaryotes seems to be solved by well known tools like Glimmer, Critica or Zcurve, our experience with environmental organisms has shown a lack in sensitivity and specificity of all these gene finders. To meet this challenge, a meta-ORF-finder named MORFind was developed to integrate the different tools in an intelligent way. First benchmarks with the marine bacterium *Gramella forsetii* indicate a nearly quantitative yield of ORFs without a loss of specificity. To support and speed up the manual annotation process, automatic annotation systems are urgently needed. The tool MicHanThi is able to automatically generate annotations for all predicted genes based on the observations returned by a set of the individual search tools. A comparison between the manually curated genome annotation of *G. forsetii* and MicHanThi showed perfect matches for nearly 60% of the genes.

In the field of Metagenomics it is an urgent problem to classify the sequenced fragments. Often, BAC or Fosmid clones that carry metabolic genes of interest lack suitable markers for their phylogenetic classification. Measures such as the average G+C content of the fragments, the best BLAST hits, and the codon usage of the corresponding coding regions are commonly used to provide further hints for their affiliation. These measures, however, can produce ambiguous or even misleading results, and have to be supplemented by tools taking intrinsic genomic signatures into account. Our tool TETRA (www.megx.net/tetra) allows the assignment of metagenomic fragments based on oligonucleotide frequencies and Chaos Game representations in a highly reliable way.

To best proceed in the field of marine ecological

genomics, we must find a way to ensure that the ability to store, analyse, and integrate all the data coming from genomics, metagenomics and post-genomics, is available not only to computer scientists but to biologists as well. This means creating interfaces to enable researchers in the lab to access and work with the data on a routine basis and to train both biologists and computer scientist to share their questions and solutions. Bioinformatics is perfectly suited to close the gap between the two disciplines but this can only be achieved by a long term support of interdisciplinary projects bringing together biologists, bioinformaticians and computer scientists.

Further information can be found at www.microbial-genomics.de and www.megx.net

Bioinformatics in Marine Genomics

Ildefonso Cases

CNB, CSIC, Madrid, Spain

Bioinformatics has traditionally been seen by experimentalists as just an additional tool among many. However, it is important to note that bioinformatics, by means of comparative genomics, can also be used to deduce new scientific questions and to generate additional knowledge, per se.

While the search for specific genes or set of genes related to some biological features have rendered very few results, more general but still relevant properties of genomes like genome composition or life style descriptors (Figure 11) have been obtained by comparative genomics. In the field of marine microbiology, the ability to correlate genomic and metagenomic sequences with other types of data (physiological, geographical, physico-chemical) could open the possibility to formulate completely new questions or get more detailed answer to old ones. Just as an example, one can envision things like metagenomic libraries obtained following a depth profile. In silico metabolic reconstruction per-



SESSION III

formed at different depths could provide relevant information about the distribution of metabolic activities, and thus ecological relevance of each stratus. This genomic data could then be combined with other physico-chemical data, providing relevant clues about the interaction between the environment and the biota.

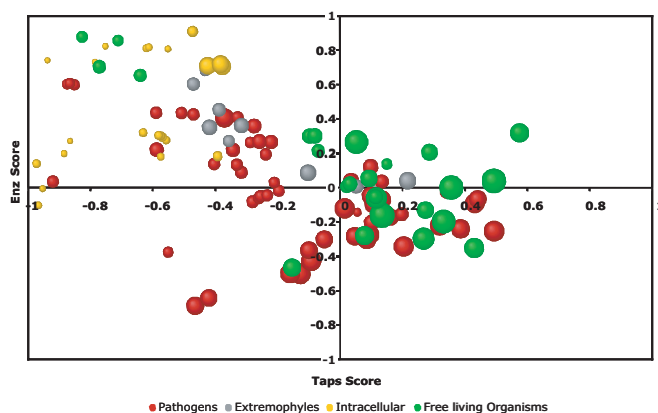


Figure 11. Bioinformatic life style descriptors extracted from 60 genomes.

However, the widespread use of this kind of approaches will require changes in several aspects of ecological research:

- **Training:** There is a serious lack of well-trained personnel in both microbial ecology and bioinformatics. A strong effort has to be made to approach these two disciplines, and overcome the "language" problem. Facilitating the exchange of personnel between labs and career redirection for bioinformaticians could be a good start.

- **Multidisciplinary Teams:** The current model of collaboration between different research groups has to be replaced by the integration of bioinformaticians in Marine Ecology Groups. On the one hand, this will promote the development of better, more specific and productive tools. But, more importantly, once bioinformatics workflows are integrated in scientific groups, it will help to formulate new hypotheses, which could in turn drive experimental research. This trend can already be observed in other fields of biological research which address complex systems, as immunology and cancer research.

- **Data Accessibility and Integration:** Physiological and physico-chemical data need to be stored in well-designed databases. Proper ontologies and controlled vocabularies have to be developed to encapsulate ecologically relevant information, so it can be automatically retrieved and mined. This information has to be connected to genomic data, and in the case of metagenomics libraries, further connected to sampling information such as dates, locations, etc. Again, a proper, standardized information structures have to be developed to facilitate data mining. These approaches are working very successfully in other biological fields. The mouse community, for example, now has detailed ontologies and databases for anatomy, diseases and phenotypes. The computational mining of this data in relation with gene expression and genetic variability data is providing extremely useful information for researchers.

The SEED and GenDB

Rob Edwards

Fellowship for Interpretation of Genomes, Chicago and San Diego State University, USA

Two presentations at the workshop highlighted how international cooperation can benefit teams on both sides of the Atlantic and support the efforts to unravel the biological complexity of the Oceans.

GenDB is a project developed by Dr. Folker Meyer's team at Bielefeld University, Germany. In his presentation, Frank Oliver Glöckner demonstrated the strength of this software in analyzing genome sequences. An early step in understanding how organisms work is deciphering which proteins are encoded in their DNA. GenDB provides an integrated suite of software for understanding which proteins are encoded in pieces of DNA, and for accurately determining the locations of those genes in the genome.

The SEED (Fig. 12) is a complementary technology that is designed for identifying and denoting the

SESSION III

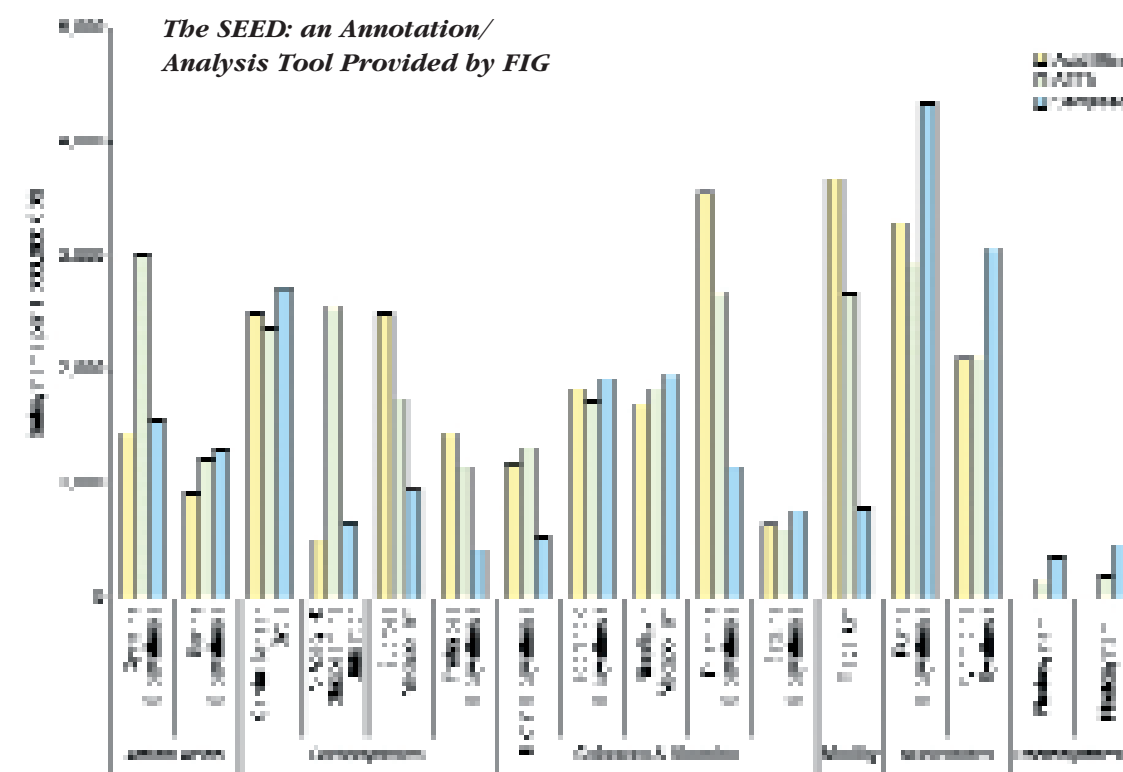


Figure 12. The Subsystem Forum of "The Seed" facilitates (re-)annotation of marine genomes.

functions of the proteins encoded by the DNA. The SEED is developed in the US by researchers at the Fellowship for Interpretation of Genomes (FIG), a non-profit organization dedicated to advancing our understanding of genomes through comparative analysis of sequences.

Researchers from FIG and the Meyer's team quickly realized the synergy between their respective projects, and although they were developed independently and maintain their identity, the teams have been working towards a joint integration of their software packages. The interactions have been bidirectional, with researchers from both teams visiting the sister sites. In the spring of 2004 there was a joint meeting between developers in Bielefeld. Later that year, several students and researchers attended a three-day joint workshop at Argonne National Laboratory in Chicago.

The latest technological advances, including the

Access Grid Internet video conferencing technology, facilitate interactions between the two teams. Weekly videoconferences are used to discuss the current status and future directions of the systems. These videoconferences are occasionally extended to other interested researchers in locations outside of Chicago or Bielefeld. Other technologies are leveraged to ensure individual researchers remain in contact almost daily.

For marine genomics, the joint GenDB and SEED platforms can handle the most pressing needs immediately, and yet provide the scalability and to extend and adapt to releases of more data. For example, the databases currently integrate all freely available genome sequence and metagenome sequence data, and serve the relationships between datasets. Truly understanding the marine genomics is going to be a massive undertaking, but by using the latest technology, and promoting successful transatlantic collaborations, teams like those at FIG and Dr. Meyer's team at University of Bielefeld are poised to fully exploit the genomic revolution.



SESSION III

GenDB and The Seed

Folker Meyer

University of Bielefeld, Germany

The amount of sequence data available from marine systems is growing steadily, yet to a large extent the marine community still has to learn to optimize the data mining process. Some developments in bioinformatics will help foster a greater understanding of the opportunities and the limitations of learning new biology from sequence data: purpose driven portals (like the Marine Genomics Europe portal for all High-throughput data) for large user communities, platforms that enable web-driven data mining in large scale data sources (like the SEED system for mining genomes as well as environmental samples). These software and data platforms, together with specialized training courses, could quickly enable a growing number of marine scientists to grasp the full potential of high throughput data and the corresponding analysis techniques required.

With the combined GenDB and EMMA systems for collaborative, web-based genome and microarray analysis, Bielefeld University has provided a set of useful tools to several national German and international networks and projects. Utilizing these tools and integrating them in the context of a portal, as in Marine Genomics, while at the same time integrating other tools such as the SEED system, is a very labor intensive process, but provides very powerful tools to the user of the combined web-platform.

Only by providing adequate infrastructure and platforms will the majority of marine biologists be able to utilize the current flow of sequence data and the flood of expression data that is clearly visible in the near future. To establish these platforms, serious effort needs to be undertaken on both sides of the Atlantic; current centers like NCBI or EBI are not focusing on providing these kinds of platforms for a niche group such as the marine biologist community.

The UniProt Knowledgebase

Tania Lima

Swiss-Prot, Swiss Institute of Bioinformatics
Geneva, Switzerland

The increasing number of completely sequenced genomes represents an unparalleled opportunity to explore the resources of the marine environment. However, the sequences themselves are not enough. It is of fundamental importance that these genomes are annotated with high quality and that the nomenclature be standardized.

We can find out a lot more about the function of proteins by studying their properties, such as the biological processes they are involved in, their post-translational modifications, their interaction with other molecules, and their location in cells and organisms, than we could ever learn by studying the DNA that encodes them. As the number of complete genomes increases, the research community is refocusing on collecting information about all the proteins encoded in those genomes. UniProt allows biologists to access and rationalize this wealth of data.

UniProtKB/Swiss-Prot contains high-quality manually annotated and non-redundant protein sequence records (Fig. 13). Manual annotation consists of analysis, comparison and merge of all available sequences for a given protein, as well as a critical review of associated data, either experimentally proven or predicted. UniProt curators extract biological information from the literature and perform numerous computational analyses.

UniProtKB/Swiss-Prot aims to provide all the known relevant information about a particular protein.

It describes, in a single record, the different protein products derived from a certain gene (or genes if the translation from different genes in a genome leads to indistinguishable proteins) from a given species, including each protein form derived by alternative splicing and/or post-translational modifications. Protein families and groups of proteins are regularly reviewed to keep up with current scientific findings.

SESSION III

It also standardizes the nomenclature of protein and gene names. Consistent nomenclature is indispensable for communication, literature searching, and entry retrieval. Life scientists are very reluctant to follow any type of nomenclature and are generally not fully aware of the consequences of this "laissez-faire" approach. Many of the difficulties in term of interconnecting life science information resources are due to the lack of standards for naming and referencing biological objects such as genes and proteins. Many species-specific commu-

te standards where none exist. If no accepted unification exists, and several alternatives are of equal frequency in the literature, we will use the one with the easiest extensibility or standardization. In addition, preference will be given to names that best reflect the common acronym or gene symbol.

The protein names provided in the description lines of the entries in the Swiss-Prot component of UniProtKB are widely used by a very wide palette of life scientists and are "mined" by genome annotation groups to propagate names in the process of

annotating gene products of new genomic sequences. Therefore any effort to provide structured description lines, clean and well organized sets of "recommended names" as well as clear naming rules and guidelines is of paramount importance. Swiss-Prot is also moving toward an increasing use of controlled vocabularies in all the sections of an entry. This will facilitate parsing and retrieval of information. We also should try to improve the use of GO terms in the data-

bases but for this we need more terms catered to prokaryotes in the GO ontology.

But, to carry out all these tasks, funding is **essential**. Right now, we feel that the funding agencies have the tendency to fund genome sequencing projects but not the infrastructure that is essential for the understanding of the data generated by these efforts. Infrastructure funding for databases in general is still very minimal and efforts should be made to increase long-term, substantial funding to relevant databases.

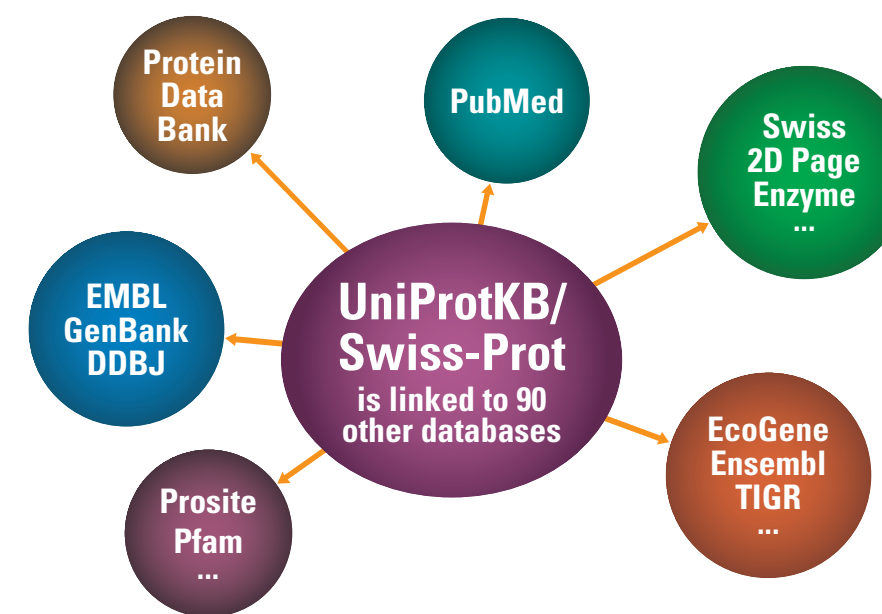


Figure 13: UniProtKB/Swiss Prot is a highly cross-linked, curated protein database.

nities have established gene nomenclature committees that try to assign consistent and, if possible, meaningful gene symbols. But there is no established organization involved in the standardization of protein names, nor are there any efforts to establish naming rules that are valid across the largest spectrum of species possible.

We strongly believe that the UniProt consortium is in a privileged position to help the user community to gradually emerge from this "terminology muddle". Whenever available, we make use of the official nomenclature defined by international committees and we will make every attempt to crea-



SUBREPORT FROM SESSION IV: The Industry / Foundation Perspective

Rapporteur: Rudolf Amann

From the very beginning it was planned to enrich this Marine Genomics workshop with expertise from the non-academic, private field. There are two very different types of stakeholders that immediately come to one's mind when discussing the exploration of biodiversity, companies and foundations.

Industry & Small/Medium Enterprises

Industry have for along time screened terrestrial and aquatic habitats for new organisms producing, e.g., antibiotics or enzymes. These on-going efforts are now complemented by a significant number of often newly founded small/medium-sized enterprises (SMEs) which are using molecular approaches, including genomics. During the Bremen workshop Dr. Viggo Marteinson from the Iceland-based start-up Prokaria Ltd. shared his visions with the participants.

Mining the Marine Diversity of Extremophiles for Novel Enzymes and Products

Viggó Thór Marteinson

Prokaria Ltd., Reykjavik, Iceland

Prokaria Ltd. is convinced that metagenomics is currently one of the most likely technologies to provide new molecules for problems where classical chemistry has little or no solutions. Patents on aspects of metagenomics have been claimed, such as expression cloning and sequence base screening. Furthermore, Prokaria is involved in the development of new genetic markers for fish and other marine life. This portfolio reflects the abundant demand for novel enzymes and biocatalysts and for a reliable traceability of marine species.

Prokaryotes and their genomes are the most easily targeted by the functional screening tools available in metagenomics. Access to novel enzymes and biocatalysts has been severely limited by the relatively small number of cultivated bacteria. Environmental diversity is high but many bacteria are recalcitrant to cultivation. In contrast, the last decade has

proven that direct cloning of microbial DNA sequences from the environment has been successful in many ways. Screening based on metagenomics is including the genomes of not yet cultivated microorganisms. Metagenomics, therefore, has the potential to provide new molecules with diverse function based on the enormous resource of uncultivated microbial diversity in the sea and elsewhere. It will promote the major roots for the "white biotechnology" in Europe and across the Atlantic.

Prokaria is exploiting some of the unique geological ecosystems of Iceland and has mainly been focusing on extremophiles living in terrestrial and marine hydrothermal vents at various depths as source of new enzymes and biocatalyst for the industry. Thus, by combine unique natural access with high-throughput DNA sequencing and bioinformatics to mine Nature's metagenomics for novel genes, Prokaria has constructed a proprietary DNA database (Fig. 13) of metagenomes that allows them to discover, develop and commercialise genes, enzymes and small molecules from natural organisms for the biotech industry and chemical, agricultural and pharmaceutical sectors.

Recommendation:

The industrial exploitation of marine diversity cannot be realized without coordinated and well-funded research and development efforts in basic and applied sciences, including genomics. A supportive and guiding political framework is needed for the economic success of the industry.

Private Foundations

In recent years the old notion that one can only protect what is known has resulted in several large projects to explore and characterize marine diversity. This included e.g. the "bar coding of life"-initiative. Especially in the US these often large endeavours are more and more funded by private foundation such as the Betty & Gordon Moore Foundation, the Alfred B. Sloan Foundation or the Ocean Genome Legacy Foundation. The latter is focusing on genome repositories and was represented in Bremen by Dr. Dan Distel, its executive director.

SESSION IV

Genome Resource Banks, Marine Genomics and Conservation

Dan Distel

Ocean Genome Legacy Foundation (OGLF)
Ipswich, Massachusetts, USA

Historically, the collection, study and preservation of physical specimens have played a central role in biology and biological repositories have served as critical resources and catalysts for discovery. Currently, with so much emphasis and excitement surrounding genomic information, little attention is being paid to the traditional activities of describing, verifying and archiving the biological source materials from which genomic data is derived. Traditional archival practices developed by biological repositories over the centuries have proven their value time and again; making it possible to return to source materials to uncover errors in identification or handling, allowing follow-up inquiries with new technologies, and providing the information that connects specimens to the natural environment and new information to the traditional literature. The omission of this important component of traditional biological practice may have grave consequences for the advancement of marine genomics beyond the enthusiasm of its youth.

To remedy this important oversight, a major effort must be expended, and significant financial support must be provided, to develop an international system of genome resource repositories. Genome resources are defined as any materials that may serve as sources of genomic information, including preserved organisms, tissues, genomic DNA, RNA, whole genome amplification products and DNA or cDNA libraries. Researchers may deposit such specimens to genome resource repositories along with detailed collection information, voucher specimens and descriptions of sequences, publications and other publicly available research products derived from them. Deposition in public repositories provides the researcher with safe centralized storage of their materials as well as increased transparency, credibility and accountability for their research. Voucher specimens provide a means to retroactively confirm or reassess the species identity of the source organisms. Detailed collection and site information provide data that place the archived materials in their proper environmental context and link them to the existing literature of tradition-

nal descriptive and experimental science. Finally, the archived tissues, amplification products, genomic DNAs and DNA libraries contain the raw genetic materials that can be isolated, sequenced, expressed and manipulated to test inferences drawn from sequence data. In these ways, public genome resource collections provide the physical materials and source information that add value to published sequence data. Similarly, published sequences and experimental data add value to the archived genome resources. Thus, public genome resource repositories leverage considerably greater value from research funds already being spent for existing collection and sequencing efforts.

An additional benefit of public genome resource repositories is their capacity to spread the benefits of research funding. Increasingly, large-scale genomic investigations require large-scale investments and expensive equipment accessible to relatively few investigators. These considerations are rapidly driving the centralization of biological research, increasingly concentrating the control of genomic research into fewer and fewer hands. While gene sequence information often becomes widely available, the source organisms, genomic DNAs and libraries used to generate these data usually do not. Public genome conservation archives can serve genomic research by placing publicly funded resources within reach of a greater number of researchers and fostering cooperation among smaller groups to utilize products created by centralized research facilities. OGLF is devoted to establishing a marine genome resource repository that not only stands alone as an important research collection, but that also serves as a model for repositories to come.

Finally, genome resource collections have considerable value for environmental and species conservation. It can assist preservation and management of endangered species.

Recommendation:

Archiving the genetic diversity of the oceans is a goal that not only requires, but also clearly deserves, substantial public support and the integrated efforts of the international scientific community. OGLF strongly recommends that EU and US funding agencies form a joint task force to explore the best means by which an **international system of genome resource repositories** might be created to archive the vast diversity of the sea.



Published by:

**Max Planck Institute
for Marine Microbiology**

Celsiusstr. 1 · 28359 Bremen · Germany

Concept and text editing:

Prof. Dr. Rudolf Amann,

Dr. Manfred Schlösser

Phone: +49 421 2028-704

Fax: +49 421 2028-790

E-Mail: mschloes@mpi-bremen.de

Internet: www.mpi-bremen.de

Photos associated with the articles were supplied by the authors, Photos showing the Earth were from Google.com. Further photos were from the MPI coworkers Arne Dübecke, Manfred Schlösser, Jens Harder and Bo Barker Jørgensen.