International Census of Marine Microbes 1st Annual Meeting
Summary Report
June 12th-15th, 2006
NH Leeuwenhorst Noordwijkerhout, The Netherlands

The International Census of Marine Microbes (ICoMM) is an ocean realm field project of the Census of Marine Life that seeks to determine what is known, what is unknown but knowable, and what may never be known about the biodiversity of marine micro-organisms. An estimated 3.6×10^{30} microbial cells with cellular carbon of $\sim 3 \times 10^{17}$ grams may account for as much as 90 percent of the total oceanic biomass. In recent years, there have been spectacular revelations about marine microbial diversity but there has been no organized effort to identify the different kinds and relative abundance of microbial taxa throughout marine systems. ICoMM's general goal is to learn the identities and roles of key members within microbial consortia in all marine environments.

ICoMM operates as a world-wide research coordination network that seeks to organize the efforts of marine microbiologists in a survey of microbial diversity in the open oceans, coastal systems, the benthos and the sub-seafloor. Organizational meetings held during the first half of 2005 included the *Benthic Systems Working Group* chaired by Dr. Katrina J. Edwards (WHOI) and held in the Southampton Oceanography Center on January 14-15, 2005, the *Technology Working Group* chaired by Dr. Rudolf Amman (MPI) and held at the Max Planck Institute for Marine Microbiology in Bremen on January 31st-February 1st 2005, and the *Open Ocean and Coastal Systems Working Group* chaired by Dr. David Karl and held at the University of Hawaii at Manoa in May 10th -11th, 2005. The ICoMM *Scientific Advisory Council (SAC)* headed by Dr. John Baross of the University of Washington, met for the first time at The Royal Netherlands Academy of Sciences, Amsterdam, The Netherlands, on February 7 and 8th, 2005. Collectively these meetings identified the major challenges and requirements for embarking on a census of microbes in the oceans. In response to those meetings we initiated several new activities including the addition of a Working Group for Informatics and Data Management (that met in September 25-26th, 2006), submission of grant proposals to support the census, completion of a pilot project with very significant new findings that appeared in the Proceedings of the National Academy of Sciences in July 2006, and the hosting of the meeting that met in Noordwijkerhout, The Netherlands June 13-15, 2006. The intention of this meeting was to bring together some of the representatives from all of the ICoMM working groups, the Science Advisory Committee and marine microbiologists from around the world who have expressed an interest in the ICoMM initiative. We have labeled this the first annual meeting of ICoMM as we hope that we will be able to continue obtaining support for an annual meeting. We welcome ideas and suggestions on ways to continue expanding and improving our research coordination network.

The ICoMM Secretariat

Meeting Participants:
Dr. Silvia Gonzalez Acinas, NIOO-KNAW The Netherlands, Institute of Ecology Yerseke, The Netherlands
Dr. Linda Amaral Zettler, ICoMM secretariat, The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, MBL Woods Hole, MA USA
Dr. John Baross, ICoMM Scientific Advisory Council Chair, University of Washington, Seattle WA USA
Dr. Judith van Bleijswijk, The Netherlands Institute for Sea Research Texel, Den Burg, The Netherlands
Dr. Antje Boetius, ICoMM Scientific Advisory Council Member, Max Planck Institut für Marine Mikrobiologie, Bremen, Germany
Dr. Henry Boumann, The Netherlands Institute for Sea Research, Texel, The Netherlands
Dr. Peter Burkhill, ICoMM Open Ocean & Coastal Systems Working Group Member, Southampton Oceanography Centre Southampton, United Kingdom
Dr. David Caron, University of Southern California, Los Angeles, CA, USA
Dr. D. Chandramohan, ICoMM Scientific Advisory Council Member, The National Institute of Oceanography, Chennai, India
Dr. Steven D'Hondt, ICoMM Benthic Systems Working Group Member, University of Rhode Island, Narragansett RI USA
Dr. Jan de Leeuw, ICoMM Co-I, The Royal Netherlands Institute for Sea Research, Texel, The Netherlands
Dr. Virginia Edgcomb, Woods Hole Oceanographic Institution, Woods Hole, MA, USA
Dr. Katrina Edwards, ICoMM Benthic Systems Working Group Chair, Woods Hole Oceanographic Institution, Woods Hole, MA, USA
Dr. Slava Epstein, Northeastern University, Boston, MA, USA
Dr. Carola Espinoza, Center for Oceanographic Research in the Eastern South Pacific, Concepción, Chile
Dr. Isabel Ferrera, Portland State University, Portland, Oregon USA
Dr. Victor Ariel Gallardo, Center for Oceanographic Research in the Eastern South Pacific, Concepción, Chile
Dr. Steve Giovannoni, Oregon State University, Corvallis, OR, USA
Dr. Frank Oliver Gloeckner, ICoMM Technology Working Group Member; ICoMM Informatics and Data Management Working Group Member, Max Planck Institut für Marine Mikrobiologie, Bremen, Germany
Dr. Martha Liliana Gómez García, INVEMAR - Instituto de Investigaciones Marinas y Costeras José Benito Vives de Andréis, Santa Marta, Colombia
Dr. Maria-Judith B.D. Gonsalves, National Institute of Oceanography, Dona Paula, Goa, India
Dr. John Heidelberg, ICoMM Technology Working Group Member, The Institute for Genomic Research, Rockville, MD, USA
Dr. Gerhard Herndl, ICoMM Scientific Organizing Committee Member, The Netherlands Institute for Sea Research, Texel, The Netherlands
Dr. Julie Huber, The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, MBL and the NASA Astrobiology Institute, Woods Hole, MA USA

Dr. Elena Ivars-Martinez, Universidad Miguel Hernandez de Elche, Elche, Spain
Dr. Alexandra Kraberg, Stiftung Alfred-Wegener der Helmholtz-Gemeinschaft Institut für Polar and Meeresforschung in Helgoland, Germany
Dr. William Li, ICoMM Open Ocean & Coastal Systems Working Group Member, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada
Dr. Debbie Lindell, Masschusetts Institute of Technology, Cambridge, MA, USA
Dr. Jose Lopez, Harbor Branch Oceanographic, Ft. Pierce, FL, USA
Dr. Connie Lovejoy, Université Laval, Québec, Canada
Dr. Alexander Loy, University of Vienna, Wien, Austria
Dr. Ana Belén Martín Cuadrado, Universidad Miguel Hernandez de Elche, Elche, Spain
Dr. Allison Murray, Desert Research Institute, Nevada, USA
Mr. Phillip Neal, ICoMM IT Specialist, The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, MBL Woods Hole, MA USA
Dr. Rodolfo Paranhos, Universidade do Brasil – UFRJ A1-071, Rio de Janeiro, Brasil
Dr. David Patterson, ICoMM Scientific Organizing Committee, The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA USA
Dr. Carles Pedrós-Alió, ICoMM Scientifc Advisory Council, Institut de Ciències del Mar, Barcelona, Spain
Mr. Wim Pool, The Royal Netherlands Institute for Sea Research, Texel, The Netherlands
Dr. Alban Ramette, Max Planck Institut für Marine Mikrobiologie, Bremen, Germany
Dr. Thomas Reinthaler, The Royal Netherlands Institute for Sea Research, Texel, The Netherlands
Dr. Anna-Louise Reysenbach, ICoMM Benthic Systems Working Group Member, Portland State University, OR, USA
Dr. Margaret Riley, University of Massachusetts, Amherst, MA, USA
Dr. Francisco Rodriguez-Valera, ICoMM Scientific Advisory Council Member; ICoMM Informatics and Data Management Working Group, Universidad Miguel Hernandez de Elche, Elche, Spain
Dr. Stefan Schouten, ICoMM Scientific Organizing Committee; Nederlands Instituut voor Onderzoek der Zee, Texel, The Netherlands
Dr. Mitchell Sogin, ICoMM Principal Investigator, The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, MBL Woods Hole, MA USA
Dr. Lucas Stal, ICoMM Scientific Organizing Committee; Netherlands Institute of Ecology, Yerseke, The Netherlands
Dr. Mike Taylor, University of Vienna, Wien, Austria
Dr. Marcel van de Meer, ICoMM IT support, The Royal Netherlands Institute for Sea Research, Texel, The Netherlands
Dr. Edward vanden Berghe, ICoMM Informatics and Data Management Working Group Member, Flanders Marine Institute, Oostende, Belgium
Dr. Bess Ward, ICoMM Open Ocean & Coastal Systems Working Group Member, Princeton University, Princeton, NJ, USA
Dr. Xiao Tian, Institute of Oceanology, Shandong, China
Dr. Shi Ning Zhou, Sun Yat-Sen (Zhongshan) University Guangzhou, Guangdong, China

Agenda:

Monday, June 12th, 2006

All day Arrival at NH Leeuwenhorst, Noordwijkerhout
1700: ICoMM Scientific Organizing Committee (SOC) Meeting
1700-1900: Poster Set-up
1900-2100: Icebreaker party/reception with drinks and food.

Tuesday, June 13th, 2006

0830: Welcome, logistics, purpose of the meeting (**Baross, Sogin, de Leeuw, Amaral Zettler**)
0900: **Sogin/de Leeuw**: Meeting the ICoMM Challenge, Unfathomable microbial diversity in the deep sea: an unexplored "rare biosphere"
1000: Coffee Break
1030 **Carles Pedrós-Alió:** Marine microbial diversity: can it be determined? (30 minutes)
1100: **Slava Epstein:** Rarefaction (30 minutes)
1130: **Gerhard Herndl**: Integrating prokaryotic microdiversity into ecosystems function theory (30 minutes)
1200: Lunch Break
1300: **Steve Giovannoni:** High throughput cultivation of microbes (30 minutes)
1330: **John Heidelberg** Sorcerer II (30 minutes)
1400: Discussion of working group objectives
1430: Working Group Session I
1700-1900: Poster Session
1900: Dinner

Wednesday, **June 14th, 2006:**

0830: **Peg Riley**: Lateral gene transfer (1 hour)
0930: Plenary Session
1130: Working Group Session II
1215: Lunch Break
1330: **Katrina Edwards/Julie Huber:** Seamounts (1 hour)
1430: Working Group Session III
1600: Plenary Short Reports
1700: End of second day
1715: ICoMM Scientific Advisory Council (SAC) meeting
1900: Dinner

Thursday, **June 15th, 2006:**

0830: **Steve D'Hondt**: IODP (1 hour)
0930: **John Baross** – SAC report to meeting
1030: Working Group Session (Synthesis)
1215: Lunch Break
1330: Antje Boetius**: Microbial diversity in marine sediments – what do we know about vertical, horizontal and temporal distribution patterns? (30 minutes)**
1400: Complete writing
1500: John Baross to present SAC recommendations / **Plenary Session** to make decisions and planning the future led by SAC.
1700: End of Meeting

Workshop Summary:
The workshop consisted of a series of invited science presentations interlaced with breakout group discussions that focused on scientific investigations of marine microbial diversity, the concept of the "rare biosphere" and database management issues. The presentation of novel tag sequence information generated by massively parallel pyro-sequencing of rRNA hypervariable regions afforded by 454-Life Sciences technology was the central theme of the meeting. This technology was the basis of a proposal from ICoMM to the W. M. Keck Foundation which has now been funded. This strategy has revealed unanticipated levels of marine microbial diversity and the existence of an under-explored "rare biosphere". There was considerable discussion of the "rare biosphere" and how it might influence development of the ICoMM initiative. There were four breakout groups and they tackled a series of questions designed to develop cost effective strategies that can aggressively sample microbial communities including low-abundance populations. Each working group considered the importance of community microbial structures as it pertains to population and community evolution, ecological function, biogeochemical processes and metagenomics. Their discussion will provide the basis for developing new research proposals and structuring the ICoMM program supported by the Alfred P. Sloan Foundation.

Working Group A. Field investigations: Exploratory sampling and low abundance taxa.
Participants:

Katrina Edwards (Group leader)
John Baross
Henry Boumann
David Caron
Isabel Ferrera
Victor Ariel Gallardo
Bill Li
Connie Lovejoy
Allison Murray
Rodolfo Paranhos
Stefan Schouten
Lucas Stal
Mike Taylor (rapporteur)
Bess Ward

The use of 454 technology to generate sequence tags from rRNA hypervariable regions has confirmed the existence of a rare biosphere in which low-abundance organisms represent most of the genetic diversity in microbial communities. At the same time, massively parallel high throughput DNA sequencing provides tools to explore population genetics and evolution of microbial communities in natural environments. We are optimistic about securing funding for a series of ICoMM-related high throughput DNA sequencing studies including the meso- and bathypelagic waters of the North Atlantic, several seamounts, sub-seafloor cores, and samples from the HOT station. This

breakout group considered how an additional 800 samples might be allocated to other ICoMM related projects. Key questions that they addressed include:

A-1. What are the most compelling ecological justifications for selecting study sites?

A-2. What is the required level of resolution – i.e., spatial, temporal and amount of sampling required to describe the rare biosphere at each site?

A-3. Which sites should receive the highest priority for investigation?

A-4. Are there really fundamental differences in the patterns of microbial distribution for the water column versus the sediments? For example, a relatively small number of taxa dominate the water column and diffuse flow communities whereas thousands of phylotypes make up the rare biosphere. Will sediments show more even distributions of diverse microbial taxa comparable to soil communities? Or will a small number of phylotypes dominate these ecosystems?

Selection and Prioritization of Study Sites: Working Group A recommended that existing or planned capillary 16S surveys and compelling science should drive the selection of study sites. Areas of particular interest included existing long term study sites and locations where diversity gradients are likely to be found. The rationale for choosing long-term study sites was the advantage of preexisting infrastructure at such locations, as well as the potential for a diversity of contextual data that the 454 data could relate to. The Hawaii Ocean Time-Series (HOT) and the Bermuda Atlantic Time-Series Study (BATS) are of particular interest because of the large amount of available chemical/biological information at appropriate spatial/temporal scales. The required level of resolution will be habitat-dependent and relevant experts should be consulted in designing specific sampling strategies for each given habitat. It was noted that rare microbes could easily be missed due to inappropriate sampling (e.g. by taking bulk water and missing particle-associated microorganisms).

The group also provided a detailed set of recommendations about study sites that should be prioritized for inclusion in the large scale 454 proposal pending at the W.M. Keck Foundation. Those priorities will be detailed in the ICoMM 2006 submission to the Sloan Foundation for continued support of ICoMM. 'Missing' habitats that should be included in 454 surveys: coastal sediments and estuaries, host-associated microbes, igneous basement sub seafloor, Polar regions – Antarctica and Arctic, Oxygen Minimum Zones and Upwelling areas, suboxic basins, hypersaline / extreme environments, open ocean, high diversity "hot spots", and cold seeps. As a guiding principal for the census, the group recommendation was to try to cover as many sites as possible, and to scale the number of samples devoted to 454 analyses to the complexity and depth of current or intended samplings with respect to other molecular, chemical, and physical parameters. Additionally ICoMM should be as "inclusive" as possible in an international sense with respect to contributing investigators and country affiliation. For example, by inviting sample submission from scientists in countries not represented in the sites already identified in the Keck proposal, it will be possible to sample more widely in the various oceanic provinces.

Rare microbe issue: There was a lot of discussion about the concept of the rare biosphere including challenges about its importance. If, at some time and place, these rare microorganisms become abundant (due to changing environmental conditions, removal of competitors, etc.) they are clearly of great importance. They might be keystone organisms and have a disproportionate, community-changing effect. It would be interesting to know if keystone taxa are well-represented among the rare taxa. It is also possible that members of the rare biosphere represent low abundance taxa in a sampled habitat (e.g. seawater) but highly abundant in adjacent microhabitats (e.g. particles)? Or, they are rare at our sample sites but very abundant in another (unsampled) geographic area? These questions reiterate the importance of developing a proper sampling strategy. If a member of the rare biosphere becomes abundant, does its ecological role change? i.e. does it suddenly have an enormous effect on the community / ecosystem processes? If so, why don't we see this more often in nature? (probably depends on what parameter you are looking at - maybe it is happening but we miss it).

454 sequencing to document diversity and its challenges: This breakout group also considered initial experiments that would better inform us about the effectiveness of 454 approaches. There was concern that the proper controls be used for future experiments applying this technology. Controls that might be explored included applying the 454 approach to a "constructed microbial community" and a well-characterized low-diversity environment where we already have a lot of data on diversity such as Iron Mountain, Lost City, or salterns, with replication experiments to assess reproducibility (within and among samples). There was also concern about the PCR-basis for 454 creating a bias in its effectiveness as a quantitative technique and recommended that this might be validated with FISH and/or qPCR approaches.

The group agreed that this fantastic new technology will generate enormous amounts of sequence data but it is not clear how to interpret all of this new information. It will be imperative to develop ways to relate this data back to the existing information we have about our study organisms. It will be critical to try and fit these data into the context of known (and yet unknown) ecological/morphological/physiological characters. It is important to continue Sanger sequencing efforts in order to provide a better context for the 454 data. This also highlights a more general issue about the importance of considering 16S data in the context of ecological/environmental parameters.

Working Group B. Experimental Paradigms: Linking technological advances to ecological theory.

Participants:
Steve Giovannoni (Chair)
Linda Amaral-Zettler
Isabel Ferrera
Martha Liliana Gómez García
Gerhard Herndl
Debbie Lindell (rapporteur)
Jose Lopez

Carlos Pedros-Alio
Alban Ramette
Thomas Reinthaler
Anna-Louise Reysenbach
Peg Riley
Mitch Sogin
Xiao Tian
Maria-Judith B.D. Gonsalves

Massively parallel, high-throughput DNA sequencing of hypervariable regions in rRNA genes, DNA micro-arrays or other high throughput-technologies are only the first steps for defining and exploring the rare biosphere. This working group discussed experimental strategies to test the ecological and evolutionary consequences of diversity and population shifts within the rare biosphere. Priorities still need to be established for these next steps. Key questions addressed by this group include:

B-1.What are the most efficient strategies for linking highly divergent, low-abundant taxa with our current understanding of microbial evolution and physiology?

B-2. How can we culture members of the rare biosphere?

B-3. How does the rare biosphere concept influence theory and investigations of: 1) evolution of new phylotypes via mutation, lateral gene transfer, etc 2) microdiversity, 3) connections between community composition and function, 4) small scale local community dynamics, 5) biogeography of microbes (does the rare biosphere exhibit a small geographic range? If so, ecological theory based on investigations on macroorganisms would predict a high extinction rate), 6) large scale global change

B-4.What are the most powerful systems for exploring the ecological and functional roles of highly diverse and low-abundant taxa? Are these same systems suitable for studying the evolution and changes in the population numbers of low-abundance taxa? (both laboratory and field based systems)

B-5. How can we leverage high-resolution community profiles to interpret the distribution of lipid and pigment profiles in extant environments and corresponding biomarkers in the geological record?

Defining and Explaining the Rare Biosphere: Working Group B posed hypotheses that related to questions B-1 thru B-5: They described the kinds of organisms that might represent the rare biosphere. These included:

Opportunists – characterized as slow growers under ambient conditions, but are capable of rapid growth under certain specialized conditions.

Specialists – found to be abundant under very specialized conditions. e.g. in association with animal hosts – symbionts, commensals; extremophiles.

Persistent rare types – slow growers in all environments; specialized to utilize rare resources or sources that yield very little energy.

Allochthonous – imported from adjacent systems such as via atmospheric input, river run-off, , soil, etc

The working group discussed possible experimental approaches to distinguish between the hypothesized rare biosphere members including sampling different environments or altering microbial environments in artificial systems. To explore where **opportunists** are members of the rare biosphere and ask questions such as, "What changes enable opportunists to grow rapidly in "perturbed" seawater; i.e., what causes "bottle effects? and "What are the opportunities that opportunists exploit in nature?", good experimental choices are available to address these questions. For example, *Alteromonas* is a common marine opportunist. It is known to become a dominant bacterium when natural communities are confined in bottles or mesocosms. 454 sequencing from multiple samples will reveal natural abundance of *Alteromonas*. Experiments could be designed to test specific hypotheses: eg., "Are they surface associated?" or "Do they respond to turbulence?"

For understanding **specialists** with niches separated in space, dispersal theory will be important. Experimental systems are available to address these questions. In particular, symbionts provide an important example that is tractable for study. **Persistent rare species** are another class of specialists. They may be metabolically specialized to utilize substrates that are available at a low flux, for example, C1 compounds. Some geochemically significant species may fall into this class. Finally, for questions related to **allochthonous** species such as, "How long do they survive in a particular environment?" , "Do they grow?" "What is their impact on the community already at the deposition site?", one experimental strategy would be to follow deposition events (eg. Saharan dust). The Aeros tower on Bermuda, or Solas projects taking place around the world could be used to address the microbial aspects of aeaolian transport and deposition.

454 Technology. The group also discussed the experimental proofs needed to firmly establish 454 technology as a means to study marine microbial diversity. Some of the questions that were raised included: Does 454 methodology introduce rare sequences artificially? How well does it detect rare types? What are the biases? These questions can be answered with controlled experiments that begin with known mixtures of cells and carried through the entire procedure as if it were a field sample (similar to suggestions made in Group A). Use mixtures of 10-20 organisms with varying GC contents as input (i) all at equal concentrations (ii) at varying concentration over orders of magnitude difference. Do we get the same number of ribotypes out as we put in or do we get the generation of artificial ribotypes? How reproducible is the method? The introduction of artificially created sequences could be used to overcome the possibility of unknown levels of contamination causing confusion. Suggestions for further evaluation of this technology will be included in our 2006 submission for continued funding of ICoMM.

Powerful systems for exploring the rare biosphere. Finally, the group discussed the importance of determining the ecological role of rare species, the use of 454 technology for studies of population genetics, the role of rare taxa in colonization of more abundant species after environmental perturbations, and their functional stability and redundancy with other taxa. The group suggested the following follow-up technologies: (i) Develop microarray of 454 sequence tags to do more in depth census/fingerprinting of communities (drawback – you only get what you already have), (ii) Quantitative real-

time PCR. Design specific primers from within the variable region sequenced by 454 to follow specific populations, (iii) Use 454 sequence tags in conjunction with metagenomics to provide data about the functionality of uncultured microbial groups. For example, BAC or fosmid libraries can be screened for V6 tags, but are there sufficient high-throughput methods? (iv) Sort using V6 as a tag and use genome amplification to obtain genome sequences. (v) Use alternative sequencing technologies to obtain more genetic information of rare biosphere members. New Licor technology produces 20,000 nt reads, with a throughput of 14 billion nt sequenced per day. This technology will be more useful for metagenomics, but is not especially appropriate for surveys of microbial diversity.

Working Group C. Analysis: What are the most important analytical questions to address with high-throughput molecular data?

Participants:
John Heidelberg (Chair)
Judith van Bleijswijk
Antje Boetius
Ana Belén Martín Cuadrado
Steve D'Hondt
Virginia Edgcomb
Carola Espinoza
Silvia G. Acinas
Julie Huber (rapporteur)
Elena Ivars-Martinez
Renzo Kottmann
Alexander Loy
Peg Riley
Shi Ning Zhou

The detailed analysis of microbial communities using massively high throughput DNA sequencing (454 technology) or metagenomic profiling requires development of new analytical tools in the areas of bioinformatics, population genetics, graphical representations and visualization. For example, a typical experiment generates information about the relative frequency of thousands of distinct OTUs in a single sample. This presents enormous challenges when attempting to display this data (simple pie charts are woefully insufficient) yet fully resolved graphical displays are much too complex for normal publication media. Similarly, the highly resolved information about community structure presents challenges to *in situ* analysis of low abundance populations. This working group explored how to extract information from community profiles of the rare biosphere and apply that information to new *in situ* analytical procedures:

C-1. What analytical tools will be required to compare two or more complex community profiles?

C-2. What advances will be required to visualize highly-resolved molecular descriptions of microbial communities?

C-3. What is the optimal way to analyze the biogeography of microbial communities including the low-abundance taxa on a global scale?
C-4. Given advances in imaging, what levels of discrimination will be required to visualize low-abundance taxa *in situ* and what are the most important technologies to be developed?
C-5. How can we leverage the rare biosphere and 454 sequencing for metagenomic investigations?
C-6. How can we leverage the use of 454 data, MLST (Multi Locus Sequence Typing), DNA microarrays and metagenomics to carry out population genetic studies of microbes in natural environments?

454 Sequencing Technology. Working group C expressed the view that 454 tag sequencing is an excellent tool for getting a snapshot (panoramic snapshot) of microbial diversity in the oceans, and it should be complemented with other molecular analyses and geochemical, expression, and activity measurements appropriate for the environment of interest. The 454 tag sequencing approach can and should be used for diversity surveys, but the experimental method must be verified with a number of control experiments that demonstrate whether or not the technique is quantitatively reliable. It will also be very important to determine the accuracy of single base insertions and deletions that account for much of the microdiversity*.

Assessing the utility of 454 versus more conventional tools. The Global Ocean Survey (Venter Institute) has collected many surface water samples that they are using to construct shotgun libraries and 16S clone libraries. The quantitative nature of the 454 method could be established using FISH or qPCR, especially on the dominant common sequences. ICoMM could use a couple of these samples to perform the V6 tag analysis, thus allowing comparison of these three methods and their utility in determining microbial diversity. Group C emphasized the importance of working towards the development of a tag approach for both archaea and eukaryotes. The working group also discussed other methodologies available for gathering highly resolved molecular descriptions of microbial communities, including the promising new technology of 454 sequencing, as well as well-established methods such as 16S/18S cloning and sequencing, microscopy, qPCR, DGGE, metagenomics, and FISH studies. Regardless of the methodology utilized, a database is needed that is flexible enough to link existing data and a wide range of future data, including 454 sequence tag information. For a meaningful interpretation of the enormous load of incoming sequence data, it is essential that a sequence database be linked to contextual information on chemistry of the environment, sampling coordinates, depth, etc. Visualizing 454 data in a manner that makes site to site and sample to sample comparisons not only statistically, but visually interpretable will require development efforts. Querying the database as to the similarity between two different sites or samples is a statistical question, not just a bioinformatics question, and as the development of this database must incorporate the input of statistics AND bioinformatics groups qualified to handle these kind of data.

---

* Single base indels do not account for much of the microdiversity in the discussed data sets – M. Sogin, unpublished data

Leveraging the rare biosphere and 454 sequencing for metagenomic studies. Because of the high-throughput and relatively low costs, the 454 V6 tag-sequencing technology will likely enable us to dramatically expand our knowledge about microbial diversity in marine environments. It will not be useful in a metagenomics context that seeks to assemble portions of microbial genomes. However, the 454 method may be very useful for complete sequencing of large insert libraries (fosmids, BACs, etc.) used to survey metagenomes of microbial communities. Not much is known on the population structure within microbial taxa, because a proper method is lacking to know populations from species and higher taxa. Evolution acts on populations. Hence, understanding microbial population structure and changes thereof are of high importance in microbial ecology. It is possible that 454 data could help discern microbial populations, but this would need thorough testing, possibly by comparative analysis of multiple genetic markers. A possible pilot experiment would be to use a combination of FISH counts, qPCR and 454 diversity, to assess fluctuation of a distinct population in an experimental system (e.g. by adding different amounts of a population to replicate water samples). If the V6 region turns out to be a good marker for the microbial population level, 454 diversity could be used to design probes followed by cell-sorting, and efforts to culture the cells, or single cell amplification. 454 diversity information could then be used to design probes for microarrays, to more economically monitor microbial populations. Finally, such primers could also be combined with mRNA analysis, to determine activity, function and metabolic capacity of populations. Such methods have the exciting potential to resolve temporal redistribution patterns of microbial populations. There are manifold field-based applications, such as following an eddy over time and test stability/succession of populations; following phytoplankton blooms in variable water depths, analyzing population patterns induced by tidal, diurnal, seasonal variations, or disturbances like storms, food falls etc.

Working Group D. Database issues: How to represent microbial diversity and distribution on a global scale?

Participants:
Edward Vanden Berghe (Chair)
D. Chandramohan
Frank Oliver Glöckner
Alexandra Kraberg (rapporteur)
Renzo Kottmann
Phillip Neal
David Patterson
Wim Pool
Francisco Rodriguez Valera
Marcel van der Meer

The massive volumes of data from metagenomic data and new information from 454 studies will require development of databases that connect this information with legacy information and data from culture collections. Several database efforts are underway but

they are not yet linked nor do they necessarily provide answers for interpreting and sharing information about the rare biosphere. This group tackled this very complex problem and identified links to ongoing and anticipated investigations of the rare biosphere and information from cultivars. They addressed the following questions:

D1. What steps are required to integrate 454 data with existing rRNA databases representing cultivars and environmental isolates?
D2. What should be contained within ICoMM's database MICROBIS?
D3. How should MICROBIS serve molecular data for the rare biosphere?
D4. What links need to be established between MICROBIS, OBIS, MGE, MICRO MAR and the CAMERA initiative?
D5. Which areas of legacy science (Database issues, technical issues, sequence data, descriptive data, lipid data, other data) must be tackled by MICROBIS and how can we secure funding for those efforts?
D6. How should ICoMM and its data link into the Ocean Observatory programs?

The database working group held extensive discussions about the data management environment for ICoMM. It was informed by the report of the ICoMM Informatics and Data Management working group that met in September of 2005. The group discussed mechanisms for linking a diverse array of data sets using specific databases such as MEGX and EurOBIS as example for the types of resources to be networked in the context of the existing limited functionality of MICROBIS and with specific reference to new challenges arising from the emergent high-throughput molecular technologies. The group identified the need for intensive consultation with existing database initiatives and the creation of communication channels with emerging initiatives early on. The development of multiple overlapping initiatives requires innovation in services that will integrate data so that data from multiple sources can be delivered through a single portal. The group also addressed the needs to engage the community of marine microbiologists especially in the context of gathering what is 'known' about marine microbes (the legacy data).

Developing a framework for interlinking a large amount of data from multiple databases with different data models and different user needs.

Discussions covered a broad range of topics that addressed "operational issues" including resource requirements for continued development of MICROBIS. The Moore / DOE supported CAMERA initiative has only recently contacted ICoMM and there are overlapping objectives that need to be clearly defined, particularly in the aggregation of certain forms of molecular data and gathering of contextual data. However, based upon limited information, the CAMERA initiative does not appear to be concerned with integration of legacy and biogeographical data. These issues are high priorities for managers of the MICROBIS database. MICROBIS seeks to collect and organize different types of sequence data (pooled 454, 16s etc.), name based data, information about lipids, proteins and flow cytometry data. Rather than a single data base, a distributed/federated system with a common repository to avoid losses should any individual system go offline will better meet the needs of ICoMM investigators. This strategy will require decisions about protocols for exchanging information, the requirement for common identifiers for

linking disparate data sets e.g. global identifiers such as longitude/ latitude, and design of a common portal for retrieving information. The working group envisaged a pilot project that integrates several different types of information: ribosomal/ genome/ geographical and image data to demonstrate our ability to facilitate links between diverse databases with (at first glance incompatible data). Cooperating databases for this project would include resources at the MPI, MICROMAR and possibly PLANKTON*NET (to cover image data). CAMERA should also be engaged in this federation of databases. This will require the secretariat to establish a formal relationship with the organizers of CAMERA. MICROBIS should contribute data streams to OBIS and address the bioinformatics needs imposed by the emergence of high throughput technology e.g. 454 tag sequence data. MICROBIS should store minimum information for indexing purposes and link out to other existing online resources. The proposed pilot project should integrate data from a set of databases that are name, geography or sequence centric. It should create a portal via which information can be accessed. Features of the portal should include graphical (histograms etc) and mapping tools, analysis tools (e.g. calculation of diversity indices, correlation with environmental variables), and the ability to integrate data either within a single database or across multiple databases, RSS feeds, and links to metainformation. This will require a data manager to form liaisons with project partners, external collaborators and contributors. The manager must also coordinate data acquisition, entry and reconciliation tasks, including integration with existing funded databases (e.g. Ocean biodiversity initiatives), identification of existing databases in need of long term preservation strategies, e.g. collections of data for which financial support has run out and training of relevant collaborators to enable them to maintain their databases even without the need for IT training.