



EU-US Task Force on Biotechnology Research

Marine Genomics: Next Generation Sequencing

CONFERENCE PROCEEDINGS



Washington, DC, USA  
October 10-12, 2010

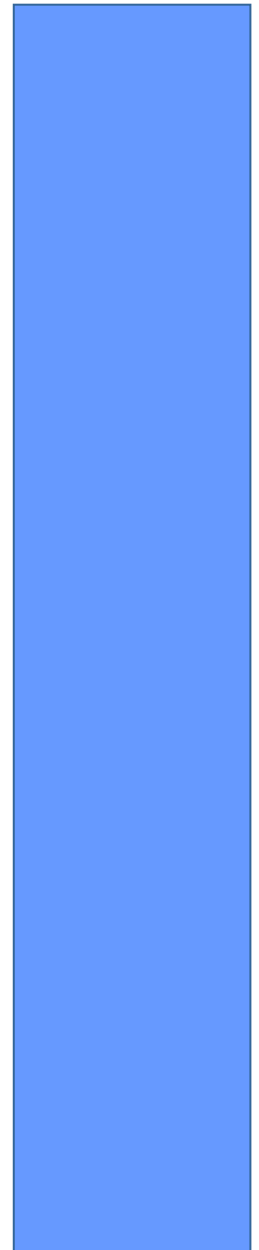




**TABLE OF CONTENTS**

<b>PREFACE</b>	<b>3</b>
<b>EXECUTIVE SUMMARY</b>	<b>7</b>
<b>AGENDA</b>	<b>9</b>
<b>SESSION-1</b> <b>Bioinformatic Tools to Process</b> <b>Large Sequence Data Sets</b>	<b>13</b>
<b>SESSION-2</b> <b>Obtaining Ever Larger Sequence Data Sets:</b> <b>Current and Future Technologies</b>	<b>29</b>
<b>SESSION-3</b> <b>Connecting Genotypes with Function</b>	<b>33</b>
<b>SESSION-4</b> <b>Training Marine Microbiologists Today:</b> <b>Culturing Versus Unix</b>	<b>51</b>
<b>List of Participants</b>	<b>57</b>

Cover page photo credits:  
Top image: Juli Trtanj  
Bottom image: John Wooley







## PREFACE

Since its creation in June 1990, the **US-EU Task Force on Biotechnology Research** has been coordinating transatlantic efforts to guide and exploit the ongoing revolution in biotechnology and the life sciences. The Task Force was established by the European Commission and the US Office of Science and Technology Policy and has since then acted as an effective forum for discussion, coordination and development of new ideas.

The Marine Genomics Working Group is one of several Working Groups under the auspices of the Task Force, and the sponsor of the *US-EU Task Force on Biotechnology Research Workshop on “Marine Genomics: Next Generation Scientist for Next Generation Sequencing”* held 10-12 October 2010 in Washington, DC, USA. Like all the activities of the Task Force, it was designed to create synergies and enhance collaboration between leading EU and US scientists in this cutting-edge field of research. The US National Oceanographic and Atmospheric Administration, the US National Science Foundation and the Directorate General for Research and Innovation of the European Commission, provided the administrative and financial support for this workshop. Doug Bartlett of Scripps Institution of Oceanography, University of California, San Diego and Frank Oliver Glöckner of the Max Planck Institute for Marine Microbiology, Bremen were the scientific conveners of this event. Twenty-two internationally renowned scientists from EU and the USA were invited to contribute to the workshop. Observers, also, related to the Task Force were present.

The workshop was organized as a high-level US-EU discussion on the applications and limitations of new sequencing technologies, new biological questions that these developments raise and could help to address, and the resulting bioinformatic bottlenecks. A round table with industry representatives allowed an open discussion about the different sequencing technologies, their strengths and their capabilities to address new marine genomic challenges. Finally, since a joint short course in marine bioinformatics was a key recommendation of the last Marine Genomics workshop, participants to this meeting were asked to contribute to a draft plan for such a course to meet the needs of both the US and EU scientific community.

The discussions during and final recommendations of the workshop demonstrate the need for the Marine Genomics working group to focus future activities in high throughput technologies and related opportunities and challenges in marine (meta) genomics. This includes optimizing existing data utilization and training the next generation of scientist to work across disciplines and have functional knowledge of basic marine bioinformatics.

It was also clear that the Marine Genomics working group efforts should expand to address the application of the marine genomics tools for societal benefits (such as health, conservation); to strengthen the link between marine biotechnology and environmental and ecological concerns, including activities such as the cross fertilization between marine biotechnology and marine biodiversity efforts which promises to result in a series of innovative and transformative technologies.

We would like to thank Drs. Doug Bartlett and Frank Oliver Glöckner, who convened the meeting and contributed to this report, for their outstanding efforts. The coordinators of the activity were Juli Trtanj (US National Oceanographic and Atmospheric Administration) and Garbiñe Guiu (European Commission).

### **US Task Force Co-Chair**

#### ***Judith St. John***

Associate Administrator  
Food Agricultural Research Services  
U.S. Department of Agriculture

### **EU Task Force Co-Chair**

#### ***Maive Rute***

Director Biotechnologies, Agriculture and Food  
DG Research and Innovation  
European Commission



Photo credit: John Wooley



## Executive Summary

This report arose from a United States – European Union workshop on marine genomics entitled “Next Generation Scientists for Next Generation Sequencing.” It was held in Washington D.C. October 10-12, 2010. A key objective was to develop a specific set of recommendations for the development of an advanced graduate level training course in marine (meta)genomics. But, before finalizing these pedagogical considerations, the participants first addressed the latest science and technology being applied in the field. This included investigations ranging from one or a few species to studies of highly complex and dynamic ecosystems. The meeting was divided into four sessions: 1) bioinformatics tools to process large data sets, 2) current and future sequencing technologies, 3) the challenges to connect the growing amount of sequences with functional properties, and finally 4) advanced training considerations that integrate physiology, biochemistry and genetics with *in silico* approaches needed to optimize and prioritize experiments.

This workshop also served to facilitate further US/EU collaborations in marine (meta)genomics and biotechnology. Some of the more applied considerations included the integration of genomics technologies into early warning systems (i.e., for the detection of toxic red tides) and the use of genome-based approaches to discover biosynthetic processes of possible biomedical significance. The participants of the workshop included leading US and EC researchers from academia, research centers and private companies. This included ten academics from both the US and the EU, two program directors, and four company representatives.

Some of the key recommendations arising from this meeting and the discussions that preceded and followed from it are as follows:

1. Continued advances are needed in the utilization of sequence data, including unknown gene characterization, metagenome binning, and the integration of environmental data and metadata.
2. A greater emphasis is needed on the application of high throughput sequencing to uncover positive and negative interactions within and across all domains of life.
3. Incorporation of (meta)genomics into studies addressing more of the biological diversity (i.e., picoeukaryotes and viruses) needs to continue to be a priority.
4. Connecting relatively unknown marine microbes with their functional attributes is a major challenge that needs ongoing support.
5. There is a great need for more reference genomes made available from creative culturing, microcosms and single-cell manipulations.

6. Computational analyses of metabolic fluxes and physiological properties should be extended to a larger collection of organisms and even to metagenomes.
7. More environmental perturbation experiments in microcosms and mesocosms are needed to facilitate the discovery of new functions.
8. It should be recognized that opportunities for major scientific breakthroughs will continue for both large-scale studies across great distances involving complex and dynamic ecosystems and for highly focused small-scale endeavors involving one or a few key species.
9. A continuing bottleneck in marine (meta)genomics is the training of biologists in bioinformatics. An intensive training course lasting at least five days is needed.\*

\*Since the end of this workshop a proposal for a marine genomics training course has been prepared and will be submitted to various US funding agencies and EC for consideration in the near future.



Photo credit: Juli Trtanj



**US-EU Workshop on Marine Genomics:  
Next Generation Scientists for Next Generation Sequencing**

**Washington D.C.  
October 10-12, 2010**

**AGENDA**

**Day 1. Sunday, October 10, 2010**

**12:00 PM Workshop lunch**

12:50 PM Welcome and charge: Juli Trtanj, Garbine Guiu, Doug Bartlett, Frank Oliver Glöckner

**Session 1:  
Bioinformatic Tools to Process Large Sequence Data Sets**

Session chair: Guy Cochrane  
Rapporteur: Lynette Hirschman /US

1:00 - 3:00 PM Participant presentations

- |   |                                   |   |
|---|-----------------------------------|---|
| 1 | Guy Cochrane (EBI, International) | Trends in data submission to EBI: How does the 'marine' perform?  |
| 2 | Lynette Hirschman, MITRE          | linking text mining tools with ontology and systems biology   |
| 3 | Christopher Quince (Glasgow, UK)  | Accurate determination of microbial diversity from 454 pyrosequencing data  |
| 4 | Folker Meyer, SoM-UMD             | Bioinformatics tools from DNA to proteins and from metabolic reconstruction to the use of clouds for metagenomics |
| 5 | Alice McHardy (MPI, Germany)      | Binning of metagenomic data   |
| 6 | Gail Rosen, Drexel Univ           | Mining metagenomes  |

3:00 PM *Coffee break*

3:30 - 5:00 PM Facilitated discussion moderated by Guy Cochrane

6:15 – 8:30 PM Opening reception with buffet (Zen Den, Topaz Bar)

**Day 2. Monday, October 11, 2010**

8:00 AM Continental breakfast

**Session 2:**

**Obtaining Ever Larger Sequence Data Sets: Current and Future Technologies**

Session chair: Frank Oliver Glöckner

Rapporteur: Folker Meyer/US

9:00 - 10:30 AM Participant presentations

- |   |                                    |   |
|---|------------------------------------|---|
| 1 | Dale Yuzuki (ABI/US)               | News about the ABI SOLID system                         |
| 2 | James Knight (Roche/US)            | News about the Roche (454) system                       |
| 3 | Gerald Nyakatura (LGC Genomics/UK) | News about LGC Genomics Next Generation Sequencing      |
| 4 | Brian Kelly (Pacific Biosciences)  | News about Single Molecule Real Time (SMRT™) sequencing |

10:30 – 11:00 AM *Coffee break*

11:00 AM - 12:30 PM Roundtable with industry representatives moderated by Frank Oliver Glöckner

12:30 PM Workshop lunch

**Session 3:**

**Connecting Genotypes with Function**

Session chair: Doug Bartlett

Rapporteur: Daniel Vaultot/EU

2:00 – 5:00 PM Participant presentations

- |   |   |  |
|---|---|--|
| 1 | <b>Josefa</b> Antón (Alicante, Spain)     | Ecology and evolution of microorganism                                 |
| 2 | <b>Victoria</b> Orphan, Cal Tech          | Connecting genomes, genes and functions in anaerobic methane oxidation |
| 3 | <b>Daniel</b> Vaultot (CNRS, France)      | Marine eukaryotic metagenomics   |
| 4 | <b>Nathan</b> Price (pending; Univ. Ill.) | Combining Computational modeling and systems biology.                  |
| 5 | <b>Jack</b> Gilbert (PML, UK)             | Marine metatranscriptomics   |
| 6 | <b>Ian</b> Hewson, Cornell Univ           | Marine viromics  |



- 4:00 PM Coffee break (20 mins)
- 7 Georgios Kotoulas (HCMR-IMBG, Greece) 'Omics' vs. Ecology?
- 8 Karen Nelson, JCVI Microbiome metagenomics
- 5:00 - 6:00 PM Facilitated discussion moderated by Doug Bartlett
- 7:00 PM Workshop dinner

**Day 3. Tuesday, October 12, 2010**

- 8:00 AM Continental breakfast

**Session 4:  
Training Marine Microbiologists Today: Culturing Versus Unix**  
Session chair: Jennifer Biddle  
Rapporteur: Roderic Guigo/EU

- 9:00 - 11:00 AM Participant presentations
- |   |   |  |
|---|---|--|
| 1 | Jörg Peplies (Ribocon GmbH)                     | Bioinformatics workshops for biologists: lessons learned |
| 2 | Jennifer Biddle (University of Delaware)        | Inspiring young science in genomics and geomicrobiology  |
| 3 | Roderic Guigo (CRG, Spain)                      | Training the next generation of biologists               |
| 4 | Frank Stewart (Georgia Institute of Technology) | Teaching microbiology from symbioses to metagenomes      |
- 11:00 AM Coffee break
- 11:30 AM Open floor discussion on structure/content of bioinformatics course, Jennifer Biddle
- 1:00 PM Actions items for workshop participants, Juli Trtanj and Garbine Guiu
- 1:10 PM Workshop lunch
- 2:00 PM End of workshop





# Report from Session 1

## **B**ioinformatic Tools to Process Large Sequence Data Sets Guy Cochrane (Session Chair)<sup>1</sup>, Lynette Hirschman (Rapporteur)<sup>2</sup>, Christopher Quince<sup>3</sup>, Alice McHardy<sup>4</sup>, Folker Meyer<sup>5</sup>, Gail Rosen<sup>6</sup>

Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom, <sup>2</sup>Information Technology Center, The MITRE Corporation, Bedford, MA, USA, <sup>3</sup>Department of Civil Engineering Glasgow University, Glasgow G128LT, United Kingdom, <sup>4</sup>Max-Planck-Institut für Informatik, Computational Genomics and Epidemiology Campus E1 4 66123 Saarbrücken, Germany, <sup>5</sup>Argonne National Laboratory, Argonne, IL and the University of Chicago, Chicago, IL, USA, <sup>6</sup>Drexel University, Philadelphia, PA, USA.

**Questions:** *What are the tools in use today for handling large sequence data sets? How accessible and understandable are they to the broader scientific community? What improvements are needed?*

### **Session 1 recommendations**

- 1. Funding agencies should provide support for a workshop that brings together experts in binning approaches for metagenomic data analysis.*
- 2. Funding agencies should provide support for a competitive challenge in which binning experts benchmark and compare their pipelines. The challenge design will be developed in the workshop proposed in recommendation 1.*
- 3. Funding agencies should require compliance with metadata standards, metadata being defined as information relating to sample context, sampling, sample processing, experimental design, library construction, machine configuration, etc. For marine genomics and metagenomics, the Genomics Standards Consortium has brought together community experts over many years and has developed a family of appropriate minimal metadata standards that are supported by tools and services to facilitate compliance.*

### Background

The first session of the workshop captured a mood of deep interest and commitment from delegates in the further development of marine genomics tools and resources and framed dynamic and fruitful discussions around how the field must proceed and where focus must be applied. The aim of the session was to address specific questions that had been proposed by the workshop organizers (see box I).

The first part of the session took the form of a series of presentations covering marine genomics resources and surrounding services, opinions on the state and the needs of the field and a selection of development efforts in data analysis tools at the leading edge of marine genomics research. Discussion of challenges and needs in marine genomics arose immediately after the presentations. In the second part of the session, a facilitated discussion focused the groups' thinking onto a range of themes that had arisen during the first part of the session and that were foreseen as being important in advance of the session. Discussion around these themes led to the development of a set of needs and a prioritization exercise, the results of which were briefly discussed later in the meeting and were presented to the group. Three specific recommendations were made during the workshop to EC and US funding agency delegates (see Box II).

In this report, summaries of the presentations are given and a full list of needs that were identified during discussion is provided.



Photo credit: John Wooley



## European Nucleotide Archive

The first presentation, “Big data services – how’s the ‘marine’ coping?” given by **Dr. Guy Cochrane** from the European Bioinformatics Institute in Cambridge, United Kingdom, began with an outline of the speaker’s institution, EMBL-EBI (<http://www.ebi.ac.uk/>), which provides broad biomolecular and related data resources, tools, services and research. A European context was laid out in which the ELIXIR initiative (<http://www.elixir-europe.org/page.php>), a funded initiative within the EC ESFRI infrastructure development program coordinated at EMBL-EBI, sets out to construct a plan for the operation of a sustainable infrastructure for biological information in Europe. The 32-member ELIXIR consortium engages many of Europe’s main bioinformatics funding agencies and research institutes.

Dr. Cochrane proceeded with an introduction to the project for which he is responsible, the European Nucleotide Archive (ENA; see figure 1, Leinonen et al 2011.), a repository that provides open access to comprehensive global raw data, assembly information and submitted functional annotation. Along with its close partners at the US NIH National Center for Biotechnology Information and the DNA Data Bank of Japan, the ENA acts under the auspices of the International Nucleotide Sequence Collaboration (INSDC). The ENA provides a range of services for a diversity of users, including submissions interfaces, text and sequence similarity search, browsers and programmatic data retrieval (<http://www.ebi.ac.uk/ena/>).

A current development focus of the ENA lies in the Sequence Read Archive, a repository for raw next generation sequence data, which is entering a second phase of its development that involves developing useful data presentations to support robust usage, including sequence similarity search for unassembled short read data sets and alignment servers to support genomic reference coordinate look-ups. Dr. Cochrane illustrated the scale of the challenge in this task, highlighting explosive rates of growth of public domain data with respect to technological growth in network, compute and storage. The ENA strategy for approaching this challenge, that calls on community-informed data reduction in combination with sophisticated reference-based compression, gives the potential to contain storage costs for next generation data within acceptable and bounded annual budgets.

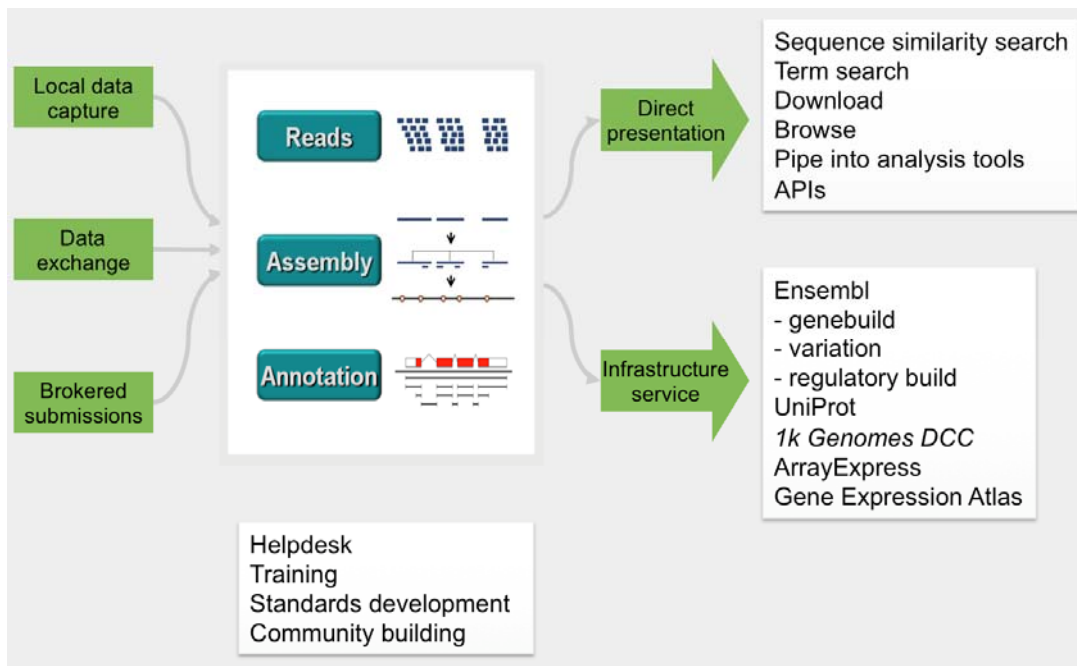


Figure I: The European Nucleotide Archive (from Cochrane)

Dr. Cochrane then moved on to cover the penetrance of marine data into ENA content. Raw metagenomic data are estimated to make up only a small portion of total public domain next generation sequence data, the major component being human resequencing data. However, it is surprising, perhaps, given the intense activity in marine metagenomics, that of total public domain metagenomic raw next generation data, from ENA records, we estimate that only 10% of the data is derived from marine samples. One key reason for this, proposed Dr. Cochrane, is that while there exists a very small number of well annotated records, a lack of systematic capture and presentation of sample contextual and methodological information reduces utility and perceived need for deposition of raw data. Contextual data do not flow easily between the many partners that need to be involved, leading to a lack of flow of interpretations (gene calls, functional annotation, etc.) into the public domain. A low cost point per sequenced base has broadened the number of applications for sequencing and this new complexity needs input from domain experts. On the positive side, Dr. Cochrane noted the significance of framework technologies, such as GCDML and the Investigation-Study-Assay (ISA) Infrastructure, developed under such organizations as the Genomic Standards Consortium (GSC). A model for ENA was presented in which active collaboration with expert marine science communities would allow for the development of the necessary tools and pipelines to bring tools such as GCDML and ISA into full usefulness.

In concluding, Dr. Cochrane raised the importance of data integration, citing ENA's involvement in taxonomic mapping work (with Species2000/Catalogue of Life), and the need for georeference-aware data presentation services.



## Metagenomics Analysis Server MG-RAST

**Folker Meyer** (Argonne National Laboratory and University of Chicago), continued the theme of assessment of the *status quo* and needs for marine science, in a presentation entitled 'Bioinformatics tools from DNA to proteins and from metabolic reconstruction to the use of clouds for metagenomics - from the perspective of MG-RAST'. Despite the fact that metagenomics is a young domain, there exist already 8,492 metagenome studies in the MG-RAST service from more than 500 data generating groups, with 20GB of data flowing through the submission pipeline each week. Dr. Meyer asserted that the community is not ready for this volume and that tool providers and funding agencies must move rapidly to offer greater readiness. One cause for optimism is that there are specialist metagenomics resources in the domain, including IMG/M, CAMERA, MEGAN, Galaxy and Metarep that will play into the new ecosystem. Key capabilities needed by the community include support for large-scale comparative metagenomics and the necessary integration of many data sets, standardized community sets for benchmarking and better 'slicing and dicing' of the corpus of data to support specific query approaches (such as requests for 'all proteobacterial reads underlying purine metabolism pathways' in the union of five metagenomics studies).

Following a presentation of the classes of metagenomics activity, from broad shotgun metagenomics, through functional metagenomics (sequencing of clones within environmental nucleic acid) to gene surveys (such as 16 ribosomal marker gene studies), Dr. Meyer touched on the key questions asked in metagenomics: which organisms are in the sample and what functions are present?

Table I: Current large metagenomics datasets (from Meyer)

Project	Data volume	Environment	Provider	Number of samples
Cow rumen	250Gbp	Cow rumen	Joint Genomes Institute	1
DeepSoil	100GBp	Soil	DeepSoil Consortium	2
MetaHit	0.5TBp	Gut microbiome	MetaHit Consortium	>100
HMP	5.7TBp	Human microbiomes	Human Microbiome Project	~700

Dr. Meyer then moved on to express the major concern in metagenomics that a meeting of the balance between data production and data analysis capacity will require new approaches. Evidence for democratization of sequence data generation includes the fact that more than 70% of Illumina sequencing machines are now purchased by 'small', non-sequencing centre users as bench-top machines. Data are growing quickly; in

## EU-US TASK FORCE ON BIOTECHNOLOGY RESEARCH

2004, the Global Ocean Survey was 600MBp and contributed half of the non-redundant protein sequences to the NCBI non-redundant database, while 'large' datasets today reach the Tbp scale (see table I). At the same time, sequencing costs are dropping rapidly. Dangerously, compute costs increase rapidly with growing data (see figure II).

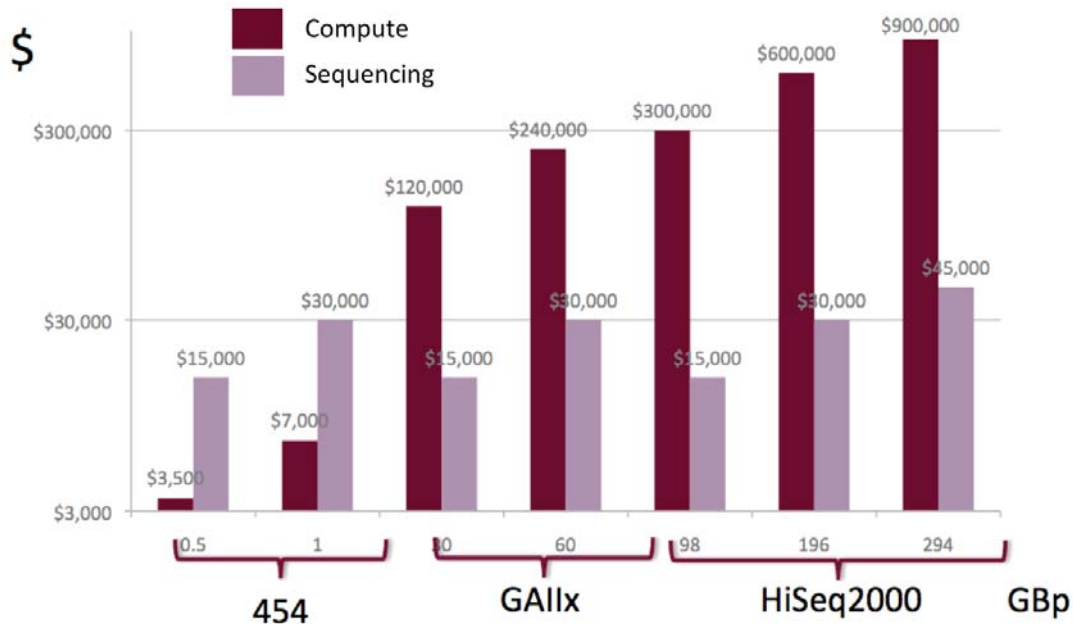


Figure 2: Sequencing and compute costs (from Meyer)

In this scenario where computing becomes the dominant cost for metagenomics, Dr. Meyer introduced the area of cloud computing. While clouds provide an alternative means to organize computation in metagenomics, it was noted that cloud computing is expensive; indeed at current costs on the Amazon EC2 public instance, analysis of an Illumina HiSeq sequencing run could cost \$900,000. Private clouds are unlikely to be a complete solution as they lack sufficient software support. Importantly, although clouds do not offer fundamentally different ways of provisioning compute resource, they bring novel advantages. First, they expose real and clear costs to their users and therefore bring to bear a very direct impact upon those who run analysis. Second, they expose design flaws and opportunities for performance tuning in analysis pipelines, pressuring the community to improve analysis methods and outputs. One impact of this would certainly be a clearer understanding of algorithmic bottlenecks and a strong motivation for computer scientists to focus attention on metagenomics' palette of analytical methods. Finally, clouds expose redundant computation across previously dispersed analysis groups and serve to improve overall integrated use of limited global resources.

Dr. Meyer finished his presentation with a discussion of the centrality of rich metadata, quoting Dr. Dawn Field of the Genomics Standards Consortium that 'metadata unlocks data analysis' (Field et al. 2008). Indeed, there are examples of work in this area; GSC, for example, has environment packages that enable better metadata reporting and use in analysis. Finally, Dr. Meyer called for the development of Standard Operating



Procedures (SOPs) to allow facilities to interoperate and a greater public availability of code to allow greater understanding and fluidity of analysis.

## Text mining

Unexpectedly and unfortunately, **Dr. Lynette Hirschman** (Mitre Corporation) was unable to attend the workshop in person, but was able to contribute opinions to the discussion and provided this report in place of her presentation.

Metagenomics relies heavily on many levels of metadata to integrate and interpret the rich data sets now being produced in the marine domain. It is critical that these layers of metadata be captured consistently, in a computable format, as completely as possible. Particularly for qualitative information (e.g., description of sample source, or biological evidence for gene/protein function), this will depend on the development and use of shared, structured vocabularies or ontologies. The establishment of minimal information standards such as MIMARKS (Minimal Information about a Marker Gene Sequence) is an important first step, as are tools that facilitate deposition of the minimal information into various repositories (e.g., ISATab). However, where vocabularies are lacking (or not widely adopted or too difficult to use), information will either be lost or will be recorded as free text – making it much less computable.

This problem can be encountered (and addressed) at several different points in the data collection/analysis pipeline. At data deposit time, the researcher (generator of the data/metadata) is available to resolve ambiguities, fill in missing data, and validate any encodings or linkage of the data to ontologies or terminologies – provided that there are interfaces/tools to support this interaction. Text mining can assist by helping to navigate the ontology and/or suggesting mappings into the ontology based on free text entries. At article submission time, the researcher could be required to deposit minimal data/metadata as part of the submission process, but this requires extensive cooperation with publishers. As the FEBS Letters experiment has shown, many authors/researchers are willing to provide metadata, but some aspects of biological curation are difficult or onerous for them, because they lack the necessary training to navigate the bioinformatics resources (e.g., finding the correct EntrezGene or UniProt identifier for the genes/proteins under investigation). Again, text mining tools can provide assistance by nominating candidate annotations that would allow the researcher/author to select the correct annotation based on associated biological information. Finally, there is the “post-publication” curation of metadata by expert curators from the published literature. While this provides high quality annotation, it is ultimately unsustainable in terms of cost and unscalable in terms of throughput. Interactive text mining tools can aid curators in prioritizing articles for curation, extracting biological entities and linking them to the appropriate resources, and extracting key biological facts represented in the article.

The capture of metadata for metagenomics can be significantly improved by:

- Understanding the whole metadata acquisition process: who collects metadata, what they collect, when they collect it, how they record it, and where/when they deposit it;
- Identifying the bottlenecks in this process;
- Identifying the tools and infrastructure necessary to relieve those bottlenecks, such as usable ontologies or structured vocabularies; text mining tools to assist in mapping free text into appropriate standardized formats; text mining tools to facilitate construction of usable ontologies or structured vocabularies;
- Training the next generation of researchers in the importance of metadata capture, the use of the standard bioinformatics resources and biological databases, and the proper encoding of metadata.

### Noise in pyrosequencing data

**Dr. Christopher Quince** (School of Engineering, University of Glasgow) presented his work on the 'Accurate determination of microbial diversity from 454 pyrosequencing data'.

By way of introduction, he noted that pyrosequencing of 16S ribosomal RNA genes has revolutionized microbial ecology, bringing an unprecedented depth to our understanding of species richness. Modern techniques permit the study of many tens of thousands of species. Interestingly, the observed abundance distribution appears skewed and opportunities for re-sequencing, which would otherwise help to untangle true diversity from noise, are limited. Numerous studies have addressed noise in 454 data, particularly in the context of species richness estimation and algorithms have been developed to reduce noise. The interest of Dr. Quince is to assess whether or not these algorithms can remove noise and determine accurately total microbial diversity from pyrosequencing data.

In test data sets of known diversity, noise leads to dramatic inflation of operational taxonomic units (OTUs). In pyrosequencing approaches, noise arises from sequencing and from PCR amplification. Current noise removal pipelines comprises three steps: Filtering of reads with features associated with noise (e.g. length, noisy base calls), clustering of noisy reads onto the sequence from which they were generated and the removal of chimaeras. Several tools have appeared in 2009 and 2010, that Dr. Quince discussed (see table II).

In concluding, Dr. Quince commented on a number of future challenges in this area. The development of appropriate metrics for newly emerging sequencing platforms in a computationally tractable way will provide the most important component in effective noise removal. Provisioning computational resources to run iterative clustering methods will become an issue. Biases in PCR amplification need attention and an understanding of why chimera frequencies vary so greatly across samples must be developed.



A detailed presentation of Dr. Quince's work in this area has been published since the workshop (Quince et al. 2001).

Table 2: Published clustering strategies (from Quince)

Algorithm	Distance metric	Method
PyroNoise Quince <i>et al.</i> (2009) <i>Nat Methods</i> 6:639-41	Flowgram	Probabilistic - iterative
Pyrotagger Kunin <i>et al.</i> (2010) <i>The Open Journal</i> 1:1	Sequence	Minimum distance threshold – one pass
Single-linkage preclustering (SLP) Huse <i>et al.</i> (2010) <i>Environ Microbiol</i> 12(7):1889-98	Sequence	Minimum distance threshold – one pass
DeNoiser Reeder and Knight (2010) <i>Nat Methods</i> 2010, 7(9):668-69	Flowgram	Minimum distance threshold
AmpliconNoise Quince <i>et al.</i> (2011) <i>BMC Bioinformatics</i> 2011, 12:38	PyroNoise – flowgram SeqNoise - sequence	Probabilistic - iterative

## Rapid analyses of large sequence data sets

**Dr. Alice McHardy** (Max-Planck Institute for Informatics, Saarbruecken) presented her work on the 'Taxonomic characterization of metagenome samples (and other methods)'. Advances in sequencing technologies allow scientists to survey extensively genome-wide genetic diversity of microbial communities, as well as of individual populations from all domains of life. As a consequence, data are accumulating at an ever-increasing pace, and the capability to analyse rapidly thousands of large-scale data sets will become a critical requirement. However, as such technologies are currently not available; it is foreseeable that this will become a major bottleneck in sample analysis in the near future. One of the biggest challenges in the field is thus the development of computational methods for the rapid analysis of sequence data sets of gigabase or even terabase scales. The corresponding tools need to be developed and updated from what is currently available for every step in the standard metagenome processing pipeline; namely assembly, taxonomic characterization of sequences (also known as binning), gene prediction, sequence similarity searches, functional annotation (requiring computationally expensive sequence similarity searches for protein domains and homologous genes), as well as process-level annotation.

In metagenomics analysis, the taxonomic origin of sequences is an *a priori* unknown and fragments may stem from any of the organisms of the sampled community. Inferring fragment origins is non-trivial, as due to our inability to culture most microbial species, complete reference genomes of closely related organisms can only rarely be obtained and used for fragment assignment based on sequence homology. Computational methods that utilise composition of sequence fragments in terms of short oligomers or more remote homology information can be used for taxonomic characterization, as Dr. McHardy had previously shown (McHardy and Rigoutsos 2007). Composition-based assignment is an attractive option for the analysis of large-scale data sets, as it does not rely on computationally expensive sequence similarity

searches, and furthermore allows the user to characterize taxa from deep-branching phyla, where reference information is sparse, as model construction requires comparatively little reference data. Dr. McHardy and her team have developed a novel fast binning method which shows on real-life datasets performance competitive with existing techniques. The method allows the accurate taxonomic assignment of metagenome fragments and rapid processing of data sets of several gigabases in size. However, model construction requires expert input for the best results and is thus currently not easy to automate. In future, methods which enable automated and accurate high-throughput taxonomic characterization will be needed.

Another major challenge continues to be the development of computational methods for the inference of function and functional relationships of protein families. Improved methods for functional inference will be the key to further leveraging the value of the data. However, currently, the functions for the majority of protein families discovered in metagenomics are unknown. As this is not a novel problem, it is likely in particular that methods will be successful which utilize innovative concepts and multiple sources of information. For instance, methods should use other information besides sequence homology or structural similarity to proteins of known function. Dr. McHardy's group is working on new methods for inference of functional relationships for protein families. For instance, they have developed a method for direct inference of functionally coupled groups of protein families based on their co-occurrence profiles, which identifies with comparable accuracy sets of interactions that are largely distinct from the those derived using standard procedures.

### Additional taxonomy analysis pipelines

**Dr. Gail Rosen** (Drexel University) presented her work on 'Metagenomic classification pipelines'. Starting with a comparison of existing tools, Dr. Rosen reviewed some key user requirements for classification tools, that include classification of known species, novel taxa, up-to-date reference databases and sufficiently comprehensive database coverage. Useful for these tools would be recommended parameterization for specific ecological contexts, such as marine, soil or human body compartment examples, although *a priori* assumptions about what is expected in a sample may lead to biases in findings. Finally, the output from these tools should provide details of organisms found and genes identified.

Dr. Rosen discussed the nature of novelty in metagenomics findings. At least 90% of species cannot be cultured and an environmental study typically reveals new strains for known species, unknown species belonging to a known genus and new phyla of unknown domain. The concept of 'known', then, must be considered in light of a given taxonomic rank. A caveat for all users must be that any classification depends on correct use of taxonomy in reference data sets used in the analysis, which cannot always be guaranteed.

The taxonomic analysis pipeline under development as an online community analysis tool at Drexel University centers upon the Naïve Bayes Classifier tool (NBC) developed



by Dr. Rosen's group, drawing on machine learning approaches. BLAST-based approaches, which are based on a dynamic programming algorithm tuned for sequence similarity search and not specifically optimized for taxonomy, are used in several tools, including MEGAN, CARMA and SORT-ITEMS. Composition-based approaches include NBC, which performs well at Genus and species level, Phylopythia, a Support Vector Machine that works well at class and phylum level and Phymm, a tool based on Markov Models. Hybrid approaches, such as PhymmBL, combine BLAST and composition-based methods for optimal performance.

Using a previously benchmarked Biogas reactor dataset of some several hundred thousand reads of length ~230 base pairs, MG-RAST, Galaxy, CAMERA, WebCARMA and NBC were compared (see figure III). For *Bacillus* and *Clostridium*, all methods were in agreement that there is high representation in the sample studied. For all tools but Galaxy, the Genus *Methanoculleus* also featured as prominent. Importantly, several Genera consistent with the sample under scrutiny were found to be prominent in NBC analysis alone, suggesting that the tool may be able to identify Genera that would otherwise be missed. Finally, it was noted that where the training set lacks representation of a particular group, learning-based tools are disadvantaged.

A detailed presentation of Dr. Rosen's work in this area has been published since the workshop (Rosen et al. 2011).

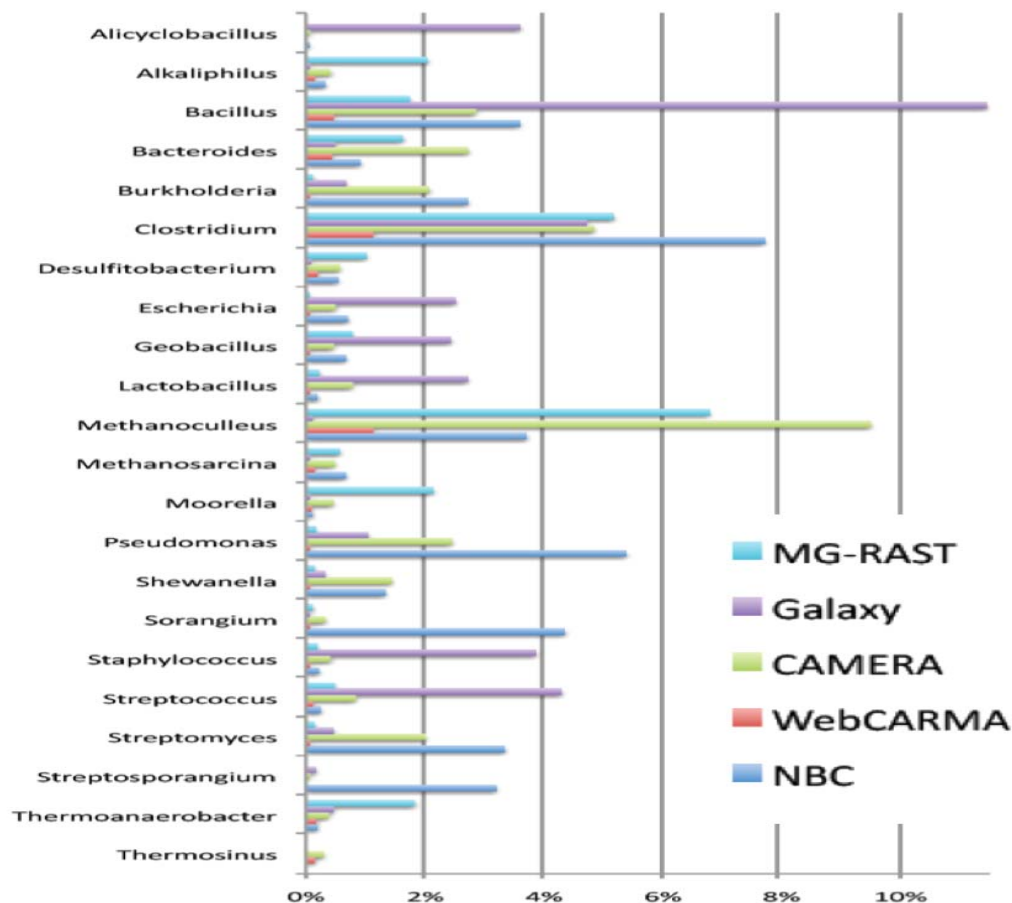


Figure 3. Percentage of reads ranking in top Genera for each tested analysis tool (from Rosen)

### Facilitated Discussion

The facilitated discussion took the structure of a number of pre-prepared propositions and questions that had been designed to provoke discussion.

*Proposition: 'Many of the components required for metagenomics analysis exist but interoperability and connectivity are lacking.'*

This proposition met with the greatest discussion and was, for the most part, accepted as true. The discussion began with comments on difficulties in attracting funding for infrastructure that brings about tool generation for research purposes, but fails to address the robust engineering requirements to present these tools as broad infrastructure services. It was recognized that larger groups have been able to operate portals and that these are helping to build community services in the area, but that more support for this was required. An advantage of more holistic infrastructure funding would be strategic thinking and consistent rather than *ad hoc* comings together of like-minded people; such an approach would clearly promote avoidance of repetitive redundant analyses, which is in the current state of affairs common for such processes as search and alignment of raw reads to reference sequences.

In an attempt to develop a common set of criteria that would define a useful infrastructure, the interrelated issues of standardization of methods and interoperability were considered separately. The first, standardization of methods, must be treated cautiously; a mature field of science may converge on a number of optimal methods, but this should not be constrained artificially. The second, interoperability (and standardization of interfaces), is absolutely required to allow tools and users to interoperate optimally and involves formats, well described standard operating procedures, tool specifications, minimal reporting standards, etc. To this end, the need to be aggressive in driving forward interoperability is paramount for the success of marine science. It was noted that in contrast to data, metadata around which much of the standardization work must be carried out are lightweight and do not bring the enormous challenges in portability and storage that are seen in data handling.

It was felt that there have been positive developments: There already exist well-developed standards and strong initiatives to extend and develop standardization concepts. Furthermore, there is a growing body of tools to support these standards, including MIMARKS, MIGS/MIMS and the Genomic Contextual Data Markup Language (GCDML) of the Genomics Standards Consortium. The issue of addressing legacy data is a challenge; clearly not only will effort be required to return to existing data to improve metadata, but in many cases, no improvement will be possible as metadata simply were not collected with sufficient granularity or resolution.

Finally, the discussion turned to the fact that while funding agencies have developed policy to require reporting of raw data and interpretations of those data, the requirements are far less developed for metadata. The group felt strongly that funding agencies might address this issue.



The discussion then turned to an analysis of the gaps in tools required for metagenomics. While there does exist a large number of tools, many of these are specific for limited groups. Typically, eukaryotic and viral studies are less supported with rich tool environments.

*Proposition: 'Taxa that we identify and relate to the environments that we study are sufficiently well recorded that we know when we keep seeing the same taxon in different settings.'*

This proposition was largely rejected as it is clear that work must be carried out in this area. Clearly the complexity of definitions of taxonomic concepts across the domains of life and the foundations of taxonomy in the study of cultured organisms or isolated specimens create many challenges in developing new ways of describing and recording observations of the vast biodiversity that is not represented in cultured and preserved specimen collections. How, for example, should we refer and record novel species, such as the SAR11 group? Here, the need for an authoritative reference akin to Bergey's manual was indicated. Taxonomic classification systems perform differently to each other, as we have seen in the presentations, and it is not yet clear that the outputs from the different tools available can be used for comparative studies across metagenomic analysis outputs.

*Proposition: 'When it becomes possible to run programmatic comparative queries that span molecular and non-molecular domains, existing comparable services outside the molecular domain will have been developed.'*

In this proposition, the discussion turned to readiness in marine science domains beyond those of genomics and metagenomics to contribute data and services to cross-domain analyses, such as bringing oceanographic and remote sensing data to molecular findings to relate physical phenomena to community composition and function. While there are some areas that are well treated, such as in data from some of the time series observatories, it was felt by the group that there are barriers to success here. Clearly there is a high level of digitization in marine science beyond the molecular domain and services exist, but as yet unmet needs in mapping between data models necessitate heavy manual work. This is an area that is felt to need more attention and communication across all of marine science is an appropriate approach to developing capability.

*Proposition: 'Molecular techniques are under-applied in marine science.'*

This proposition met with a mixed response, perhaps because the extent to which other domains have adopted molecular tools was unclear and no baseline is apparent. However, it is clear that projects such as Tara Oceans will vastly enhance uptake of these techniques.

## Identification of needs and priorities

Based on themes that had emerged during presentations and discussion within the session, a number of needs were collated and defined by the group (see Table 3). An online Survey Monkey survey was set up to allow the group to prioritize needs. Seventeen members of the group completed the survey by identifying four priority needs from the nine that were presented. The results of this survey are presented in Figure 4. The two needs felt to be of highest priority were promoted to specific and immediate recommendations to funding agencies (see session 1 recommendations).

Table 3. Needs identified from facilitated discussion

Need	Description
Metadata	Improving capture and availability of metadata (information relating to sampling procedure, sample processing and treatment, experimental design, library creation, sequencing machine configuration, etc.)
Interoperability	Interoperability (formats, schemas, services) between data analysis services, intermediates and outputs.
Robustness	Robustness and compute-scalability of component analysis tools that exist.
New tools	Development of new tools for sequence assembly.
Non-linear reference	Develop data models, tools and data that build on the notion that biological communities will not be understood by trying to re-create quantised individual 'clonal' genomes that co-exist in the community, but rather by considering that genomes in the community are unique at the level of individuals and that gradients of continuous sequence variation exist across individuals.
Viruses and eukaryotes	Specialist methods for eukaryotic and viral analysis - including gene prediction, annotation, binning, etc.
Classification	Sequence classification systems - binning by taxonomic lineage and by function - are a problem particular to metagenomics and must be addressed.
Communication	Communication with marine scientists outside the molecular domain to assure interoperability and to provide genuine cross-domain access to marine science.
References	Increasing the breadth and availability of reference genomes

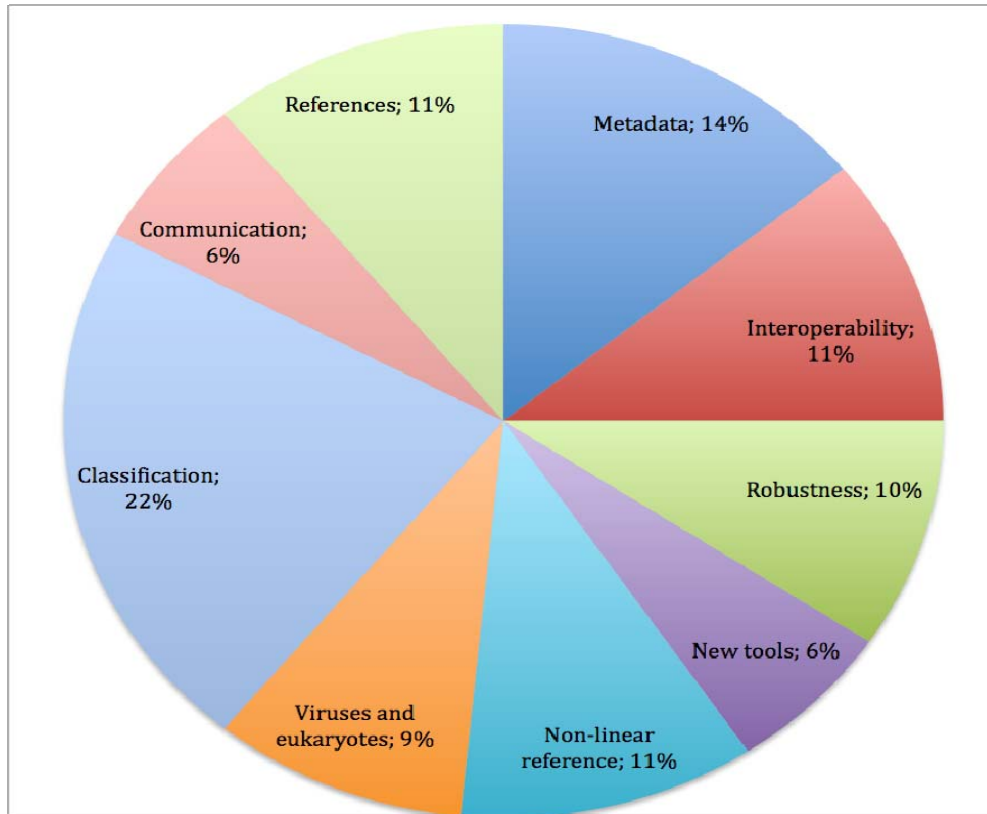


Figure 4. Summary of group opinion on prioritization of needs

## Literature Cited

Field, D, Garrity, G, Gray, T et al. (70 additional authors). 2008. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol* 26: 541-547.

Krallinger M, Valencia A and Hirschman, L. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 9: S8.

Quince C, Lanzen A, Davenport RJ and Turnbaugh PJ. 2011. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics* 12:38.

Leinonen R, Akhtar R, Birney E et al. (18 additional authors). 2011. The European Nucleotide Archive. *Nucleic Acids Res.* 39:D28-D31.

McHardy AC and Rigoutsos I. 2007. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol.* 10:499-503.

Rosen, GL, Reichenberger ER and Rosenfeld AM 2011. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27:127-129.



Photo credit: John Wooley



## Report from Session 2

### **O**btaining Ever Larger Sequence Data Sets: Current and Future Technologies (Roundtable discussion with industry representatives and workshop participants) Frank Oliver Glöckner (Session Chair)<sup>1</sup> and Folker Meyer (Rapporteur)<sup>2</sup>

<sup>1</sup>Max Planck Institute for Marine Microbiology, Bremen, Germany, <sup>2</sup>Argonne National Laboratory, Argonne, IL and the University of Chicago, Chicago, IL, USA.

Report based on presentations by Dale Yuzuki, Applied Biosystems, Life Sciences Corporation. Carlsbad, CA, USA, James Knight, 454 Life Sciences, a Roche Company, Branford, CT, USA, Gerald Nyakatura, LGC Genomics, UK and Brian Kelly, Pacific Biosciences, Inc. Menlo Park, CA, USA.

**Questions:** *What are the new next generation sequencing technologies and what are their strengths and weaknesses?  
What capabilities are needed for progress in marine genomics?*

#### **Session 2 recommendations**

- 1. Appropriate experimental design is crucial to be able to deal with the deluge of sequencing data. Quality management procedures need to be implemented for sampling, data acquisition and structured electronic storage of contextual (meta)data as well as sequencing.*
- 2. Appropriate experimental design and access to metadata must be enforced by funding agencies, reviewers and journals.*
- 3. New kinds of bioinformatic tools need to be developed that enable researchers to connect themselves to the ever growing data steam and allow investigating and integrating data on a global scale.*
- 4. Researchers need to be educated to be able to find the appropriate technology and tools to solve their biological questions.*

With the introduction of the Next Generation Sequencing (NGS) technologies biology faces a quantum leap in sequence data production. The presentations by representatives of the companies AB life technologies, Roche 454 sequencing, LGC and Pacific Biosciences showed impressively that the new technologies are still in the start up phase and a rapid increase in throughput can be expected within the next year(s) (Stratton et al. 2009). With the so called 3<sup>rd</sup> generation sequencing technologies life sciences will have the capacity to not only “shotgun the whole ocean” as J Craig Venter stated some years ago, but to sequence “everything” as long as it contains DNA (Kahvejian et al. 2008).

To cope with market demands the companies follow two general strategies in developing new sequencing machines. The first one is to keep the read length short (at around 100 – 150 bases) but to increase the amount of bases per run significantly by more parallel reactions and the reduction of the time needed per sequencing run. As an example the new SOLID 4hq system is expected to produce up to 200 gigabases per run in only half of the time (6 days) compared to the former SOLID 4 system. Additionally, the accuracy of each base will grow to 99.99%. Illumina, although not present at the workshop, follows a complementary development in terms of throughput and accuracy with the new HiSeq generation. The fields of application for these ultra high throughput 3<sup>rd</sup> generation machines are mainly re-sequencing of genomes (from small bacterial to the human genome), detection of single nucleotide polymorphisms and transcriptomics.

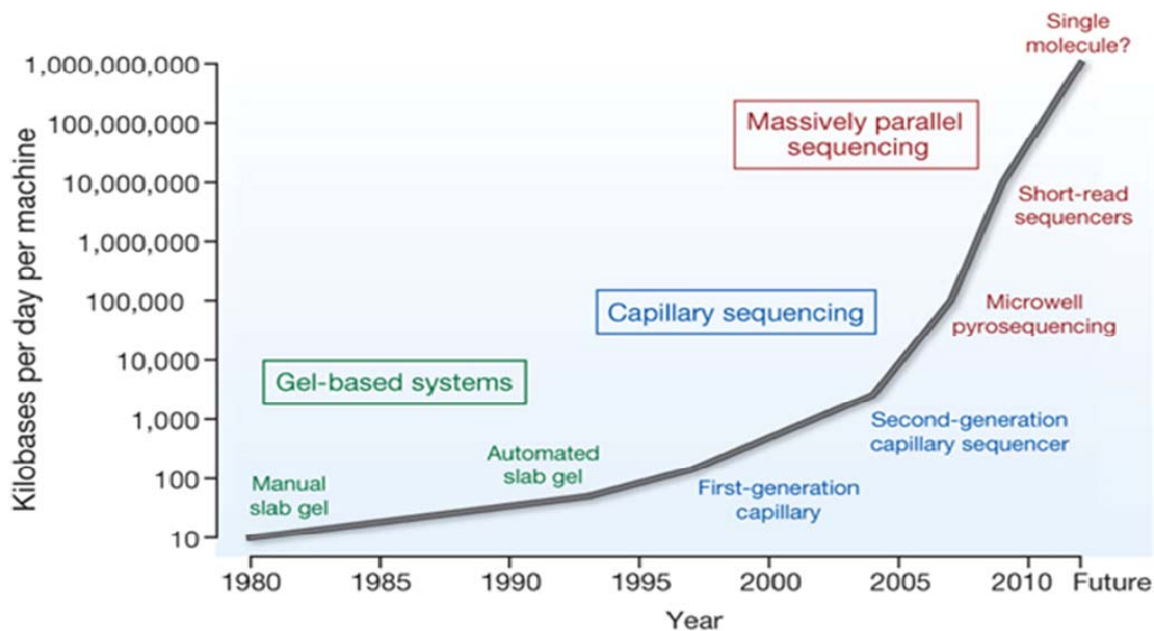
The second strategy is to develop new technologies that improve throughput by longer read length. The pioneer in this respect is Roche with the 454 sequencing technology. With the introduction of the 454 FLX Ti(tanium) an average of 400 bases per run could be gained routinely. With the improved version 454 FLX Ti 1K read length of 700 bases and more will be made available to customers in 2011. According to Pacific Biosciences the next generation of long read length sequencers is based on single molecular real time sequencing. In theory this method is able to produce reads with several kilobases per “reaction cell” while hundreds of such cells can run in parallel. The first commercial system, expected for mid 2011, will produce a majority of reads with a length of 1000 bases and up to 5% of reads will be longer than 3000 bases. The application of long read sequencing technologies are de novo sequencing even of large and complex genomes, sequencing of complex environmental samples and metatranscriptomics. Long reads will facilitate assembly and therefore promise to turn down the bioinformatic burden to deal with billions of short fragments.

An interesting development currently followed by AB life technologies and Roche 454 is the deployment of NGS bench top sequencers. By throughput they position themselves between classical capillary sequencers and the large scale 2<sup>nd</sup> or 3<sup>rd</sup> generation of sequencing machines. They promise a democratization of sequencing by offering an unprecedented flexibility in sequencing even for small labs. Applications range from diversity analysis to genomics and transcriptomics as well as small scale metagenomics and metatranscriptomics. They promise to be perfect machines for method development and evaluation in biotechnology, medicine as well as environmental surveys.



Although this brave new world of nearly unlimited sequencing capacities offers unprecedented possibilities for new discoveries and research, all manufacturers agreed that the bioinformatic processing and analysis of the sequence data is a major burden for the users. To avoid unnecessary workload when dealing with “big data” the experimental design is crucial. This includes not only the selection of the appropriate sequencing technology (long reads vs. short reads), but also the availability of biological replicates for sound statistical analysis. To be able to interpret and compare billions of AGCT’s, it is crucial to follow standardized sampling and sample processing protocols and electronically record a decent set of contextual or metadata that describe the sampling site and sample. The Genomics Standards Consortium has recently published a set of minimal checklists for genomes, metagenomes and marker genes (Field et al. 2008; Yilmaz et al. 2011) that provide a framework of the minimal set of contextual data that should be assigned to each sequencing event. The development of long read sequencing technologies will help in this respect by rendering data processing and analysis much easier. Nevertheless, appropriate education of researchers with respect to the fast evolving technologies and bioinformatics will be a cornerstone to be able to handle the data deluge.

The capacities of the new sequencing technologies have the potential to push biology into a data intensive science with a dense network of biological data from all around the globe. By enabling researchers to easily connect themselves to these global data streams new kinds of statistically sound, large scale comparative and integrative analysis will be possible. This will paint a new picture of the diversity and function of biological life on our planet.



Improvements in the rate of DNA sequencing over the past 30 years and into the future.  
Taken from MR Stratton et al., 2009, Nature 458, 719-724.

### Literature Cited

Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* 458:719-724.

Kahvejian A, Quackenbush J, Thompson JF (2008) What would you do if you could sequence everything? *Nat. Biotechnol.* 26:1125-1133.

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

Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* (in press).

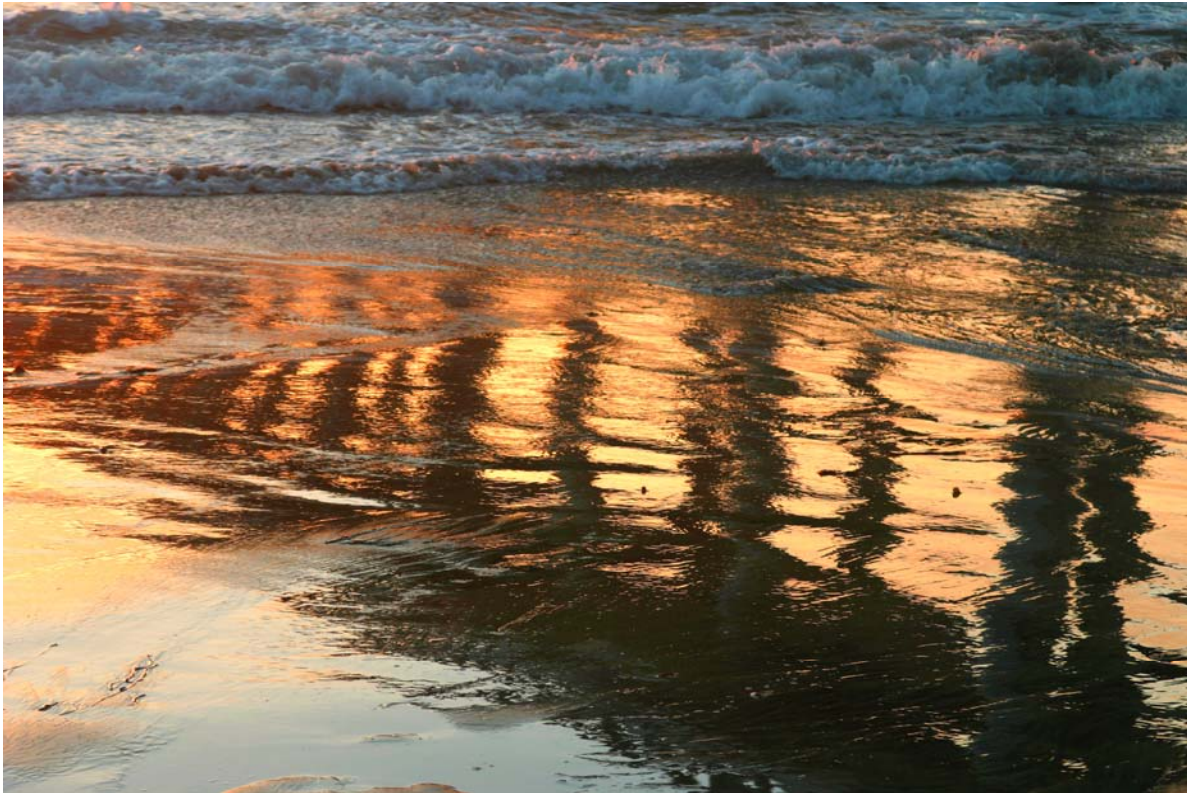


Photo credit: John Wooley

## Report from Session 3

### **C**onnecting Genotypes with Function

Douglas H. Bartlett (Session Chair)<sup>1</sup>, Daniel Vaultot (Session Rapporteur)<sup>2</sup>, Josefa Ant3n<sup>3</sup>, Jack Gilbert<sup>4</sup>, Ian Hewson<sup>5</sup>, Georgios Kotoulas<sup>6</sup>, Karen Nelson<sup>7</sup>, Victoria Orphan<sup>8</sup>, Nathan Price<sup>9</sup>

<sup>1</sup>Scripps Institution of Oceanography, La Jolla, CA, USA, <sup>2</sup>CNRS, Station Biologique, Roscoff Cx France, <sup>3</sup>University of Alicante, Alicante, Spain, <sup>4</sup>Plymouth Marine Laboratory, Plymouth, United Kingdom, <sup>5</sup>Cornell University, Ithaca NY, USA, <sup>6</sup>Hellenic Center For Marine Research, Greece, <sup>7</sup>J. Craig Venter Research Institute, Rockville, MD, USA, <sup>8</sup>California Institute of Technology, Pasadena, CA, USA, <sup>9</sup>University of Illinois, Urbana, Ill., USA.

**Question:** *How are large sequence data sets and next generations sequencing being coupled with functional studies that assess biological and environmental significance?*

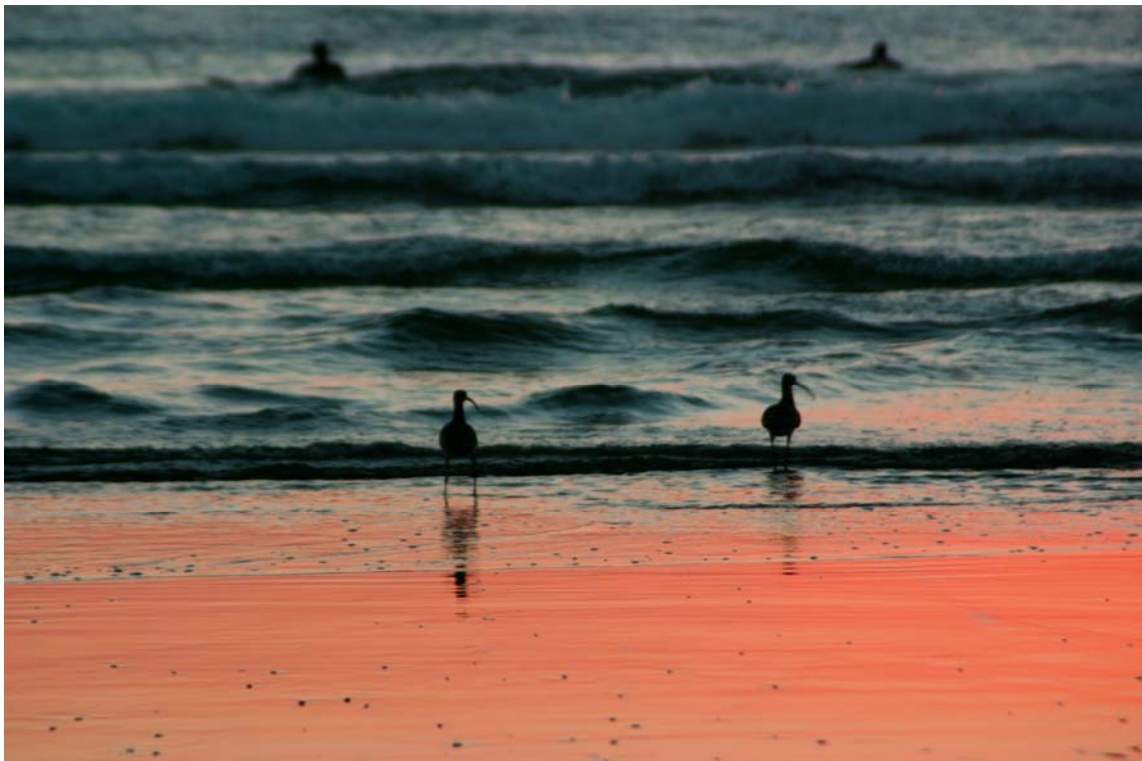


Photo credit: John Wooley

**Session 3 recommendations**

- 1. Innovative approaches are needed to assign roles to the increasing numbers of sequences with no known function.*
- 2. More reference genomes are needed from enrichment cultures, pure cultures and single-cell manipulations.*
- 3. There is a need for culture collections of phylogenetically, physiologically and environmentally diverse marine organisms.*
- 4. Environmental perturbation experiments (i.e. in microcosms and mesocosm) are needed to facilitate the discovery of new functions*
- 5. Phylogenetic groups of interests should be collected (i.e., using magnetoFISH or flow cytometry) and their functional properties assessed (i.e., using SIP, nano-SIMS, STARFISH, etc.).*
- 6. More emphasis should be placed characterizing the positive and negative interactions within and across all domains of life. This includes analyses at the levels of (meta)genomics and functional (meta)genomics.*
- 7. Advances are needed in analyses across large data sets of (meta)genomic, (meta)transcriptomic, (meta)proteomic, and (meta)metabolomic information with metadata .*
- 8. Ongoing progress in computational analyses of metabolic fluxes and physiological properties should be extended to a larger collection of organisms and even to metagenomes*
- 9. More detailed information is needed on a wide range of virus-host interactions.*
- 10. It should be recognized that opportunities for major scientific breakthroughs will continue for both large-scale studies across great distances and for highly focused small-scale endeavors.*

A growing arsenal of tools exist to bridge the gap between inferences derived from genome science and functional studies of gene products, physiology and ecology. In this session the presentations covered much of the biological diversity present in the oceans. Discussions ranged from concern about the gaps in current marine metagenomics to opportunities posed by innovative new functional approaches that will better capitalize on the growing sequence space.



## General Issues

**Professor Georgios Kotoulas** from the Hellenic Center For Marine Research in Heraklion, Greece provided some general definitions and described many of the broad-scale opportunities presented by marine (meta)genomics. These help to set the foundation for the more focused science presentation descriptions that follow.

(a) *According to Article 2 of the 1992 Convention on Biological Diversity, UN Conference on Environment and Development, held in Rio de Janeiro, "Biological diversity" means the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems. "Biotechnology" means any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for specific use.*

(b) Biodiversity structure results from the interplay between both: historical and physical, chemical and biological factors present at the time of assessment. We cannot always infer biodiversity patterns from oceanographic data. The physical environment is patchy and intertwined with the structure and dynamics of macro-biodiversity.

(c) Microbes depend on and affect the underlying macro-biodiversity. They are associated with it.

(d) The genomics of model species need to be developed across the phylogenetic tree of life.

(e) Sorting single organisms can be used to improve both genomics and phylogenetics. This approach is especially powerful if combined with morphological data such as is possible with micro CT – X-ray captured 3D images.

(f) Biogeography can be used to mutually inform ecology and genome annotation.

(g) To move from metagenome assembly to phenotypes and from the biogeography of genes to the biogeography of pathways and functions, it is necessary to know which sets of genes are found in the same cell type. To determine this advances are needed in single-cell genomics, development of culturing pipelines based in part on analyses of genome sequences, co-cultures, mesocosms and *in situ* enrichments, and populating the tree of life with reference genomes.

(h) Advances in population genomics will shed light on the demographic history of species (bottlenecks, epidemics: shape of coalescence trees, diversity levels and distributions across genomes), assessing functionally important genes by assessing their selection

- (i) (Meta)genomics can be used to help address how species' distributions are controlled by "extreme" physical parameters, competitors, presence-absence of co-evolved species and specific prey-predator relationships.
- (j) Genomic regions can be indicative of selective sweeps or more variable environmental parameters (seascape genomics).
- (k) Ecology and genome annotation mutually inform one another. Genomics offers species identification or through the application of metatranscriptomics "molecular phenotypes".
- (l) Homologous recombination and transposases are indicators of co-habitation with other species (Hooper et al. 2009). The complex interplay of microorganisms may be recorded in their genomes, even for distantly related species which cohabitate.
- (m) It may also be possible to explore the dynamics of bacteria and virus coevolution.

### The *Salinibacter ruber* Model System

In contrast to the more general issues in marine metagenomics posed by Professor Kotoulas, **Professor Josefa Antón** from the University of Alicante in Spain focused on one single model bacterial species in order to address many questions enabled by possessing multiple closely-related genome sequences. The microbe in question is the extremely halophilic bacterium *Salinibacter ruber*. In some environments *S. ruber* is very abundant, above 30% of the total counts, whereas in other settings it is a member of the "rare biosphere" under the detection limit of molecular techniques such as DGGE or FISH, but retrievable by cultivation. It is possible to find fifty *S. ruber* strains from one mil of saltern water, all of which possess different genomic patterns. Indeed, even *S. ruber* strains with identical ribosomal operons and internal transcribed spacer regions have 10% strain-specific genes, a situation observed within other bacterial species as well. It is results like these that cause microbiologists to wonder whether there is a pangenome specific to every type of environment? Clearly richness estimates don't just apply to communities of organisms, but also to the richness that exists within a single "species".

Another poorly appreciated fact is that even (or especially) very closely related bacterial strains, like *S. ruber* strains compete with one another. Where one might expect mutually beneficial "altruistic" interactions the opposite is generally true. The basis for these negative interactions are rarely known, but can include competition for resources and the production of bacteriocins and other toxins. Another possibility is differences between strains in phage production and sensitivity. In the case of *S. ruber* the viral metagenome and viral transcriptome of its hypersaline environment has been investigated and some day more detailed views of virus-*S. ruber* interaction dynamics may be available. Improved understanding of strain competition, toxin warfare and phage attack is needed better unravel the units of selection in the evolution of microbial communities.



The *S. ruber* model system also makes it possible to correlate its genes and pangenome with functional properties such as metabolites. When the metabolomes of different *S. ruber* strains were compared using high-field ion cyclotron resonance Fourier transform mass spectrometry, consistent quantitative differences were noted, mainly but not only in the extracellular fraction (Rossello-Mora et al. 2008). Until the basis of these metabolite differences are known, (meta)genomics will be missing important aspects of biology. Therefore, detailed studies are needed that extend all the way from genes to transcripts to proteins to metabolites, ultimately including biochemical studies aimed at testing hypotheses based on sequence inferences.

## Modeling the *Escherichia coli* Model System

Another model system that was discussed is the ultimate model system, *Escherichia coli*. In some respects the fine-scale resolution available for the systems biology of *E. coli* is the antithesis of the more global perspectives raised in most marine (meta)genomics projects today. However, the great accomplishments made with this and other intensively studied model systems provide valuable lessons and a useful platform to better connect genomics with function. **Professor Nathan Price** from the University of Illinois made this point while presenting information on automated methods for building integrated metabolic and regulatory networks (Feist et al. 2009). Developing these methods is essential to be able to harness the exponentially increasing amount of genome-enabled high throughput data, including for marine microbes. *In silico* models that can link genotype and phenotype to predict the effects of metabolic changes that result from genetic or environmental perturbations can guide rational design of genetically modified organisms for synthetic biology and deepen our understanding of the operating biology. A cardinal challenge in obtaining accurate predictions on transcriptional perturbations is the integration of the gene regulatory network with the corresponding metabolic network. A seamlessly integrated metabolic-regulatory network would enable us to better predict how genetic mutations and transcriptional perturbations are translated into flux responses at the metabolic level – including predictions for how genetic modification would affect marine microbes.

The Price group has developed a new method called Probabilistic Regulation of Metabolism (PROM) that integrates the transcriptome and metabolome for use in constraint based modeling. Using PROM, they constructed an integrated regulatory-metabolic network for the model organism and biotechnology workhorse, *Escherichia coli*, and demonstrated that the method accurately predicts growth phenotypes under various environmental conditions. This study incorporated data from over 900 microarrays, 1700 transcription factor–target interactions regulating over 2300 metabolic reactions in the genome-scale metabolic network of *E. coli*.

PROM requires the following: 1) reconstructed genome scale metabolic network; 2) regulatory network, consisting of transcription factors (TF) and their targets; 3) abundant gene expression data, where the transcriptome has been measured under various environmental and genetic perturbations; and 4) additional interactions involving

enzyme regulation by metabolites and proteins. It's novelty lies in the introduction of probabilities to represent gene states and gene–transcription factor interactions. These interaction probabilities are then used to constrain the fluxes through the reactions controlled by the target genes. Once the constraints are set, the optimal growth of the regulated network is determined by Flux Balance Analysis (FBA).

PROM's ability to predict the growth phenotypes of fifteen transcription factor knockouts under 125 different growth conditions was 84.3% (Table 1). PROM's accuracy is highly significant, given that it computationally quantified the interactions using high throughput data. Thus, the critical advance of PROM is that, not only does it show high accuracy, but it utilizes high throughput data and thus can be used to construct comprehensive models. The probabilistic framework used in PROM has many other advantages to model regulation, apart from the fact that it can be readily learned from high throughput data. The model framework is designed to circumvent the need for kinetic parameters for metabolic modeling, and most importantly does not assume direct correlation between enzyme activity and mRNA expression. PROM is robust to noise in high throughput data and can be easily integrated with automated algorithms for network inference. More broadly, PROM represents an important step that unifies two key biological networks - biochemical reaction networks and transcriptional regulatory networks.

Table 1: Model Features and accuracy in predicting knockout phenotypes in each organism. L – Lethal, NL – non-lethal phenotypes.

	<i>E. coli</i>
<b>Metabolic Reactions</b>	2382
<b>Regulatory Interactions</b>	1773
<b>Microarrays</b>	907
<b>Total Genes in the model</b>	1400
<b>Validation Data set</b>	1875 growth phenotypes
<b>Accuracy %</b>	84.3
<b>Sensitivity %</b>	71/89.5 (L/NL)
<b>Specificity %</b>	73.3/88.5 (L/NL)



## Modeling Complex Marine Ecosystems

The transition for modeling the behavior of pure cultures of well-studied model systems to complex marine ecosystems has already been made. **Dr. Jack Gilbert**, currently at the Argonne National Laboratory, presented just such a case based on marine metatranscriptomics data.

Thanks to the development of a number of techniques for the removal of rRNA for better observation of mRNA, metatranscriptomics applications to explore microbial functional dynamics in marine ecosystems have been growing (Gilbert et al. 2008). Dr. Gilbert described metatranscriptomic analyses using the 454 pyrosequencing platform against a well-characterized sampling site in the Western English Channel. Fundamental to the ability to make sense of the metatranscriptomic data was access to corresponding and exceptionally detailed phylogenetic data and metadata.

Phylogenetic community dynamics context resulted from one million reads of 16S rRNA amplicon pyrosequencing covering 72 monthly time points between January 2003 and December 2008. The result was a description of the microbial community changes over seasons and the demonstration of an extremely robust cyclicity with richness peaking in the winter and being lowest during the summer. This dataset made it possible to describe how the communities change and respond to environmental variables. Dr. Gilbert and colleagues were able to show that variables such as temperature, nutrient availability and day-length were extremely important in structuring the richness of the microbial system. This dataset also enabled the detection of blooms of extremely rare bacteria, which may affect metatranscriptomic profiles. The absence of this dataset would of made interpretation of the metatranscriptomic data extremely difficult.

The metagenomics/metatranscriptomics study was implemented during 2008. Samples were collected during the day and night in January, April and August, representing Winter, Spring and Summer. Eight million putative proteins were found. One of the most astounding findings was that the differences in the functional potential (metagenomic) and functional response (metatranscriptomic) of these different communities was mostly due to proteins with no known functional homolog. This highlights one of the major needs in genome science today. There is a pressing need for more emphasis to be placed on protein characterization studies. The ability to annotate sequence information is the limiting factor in metagenomic and metatranscriptomic studies.

Having the aforementioned metagenomic and metatranscriptomic data available enabled the investigators to ask questions about the structure of function in the English Channel ecosystem. Strikingly, the metagenomic functional potential of each time point was nearly identical except for photosynthetic activity, that had greater gene abundance during the winter then the summer and more at night than during the day. These differences correlated with cyanobacteria, whose relative abundance changes during

the winter and at night, due to resource availability and cellular division respectively. Mapping the metatranscriptomic data onto the metagenomic data confirmed that the cyanobacterial photosynthetic genes were only transcribed during the day, and that the cellular division genes were transcribed at night. Photosynthetic transcriptional activity was only found in winter and spring, which may be a function of coverage.

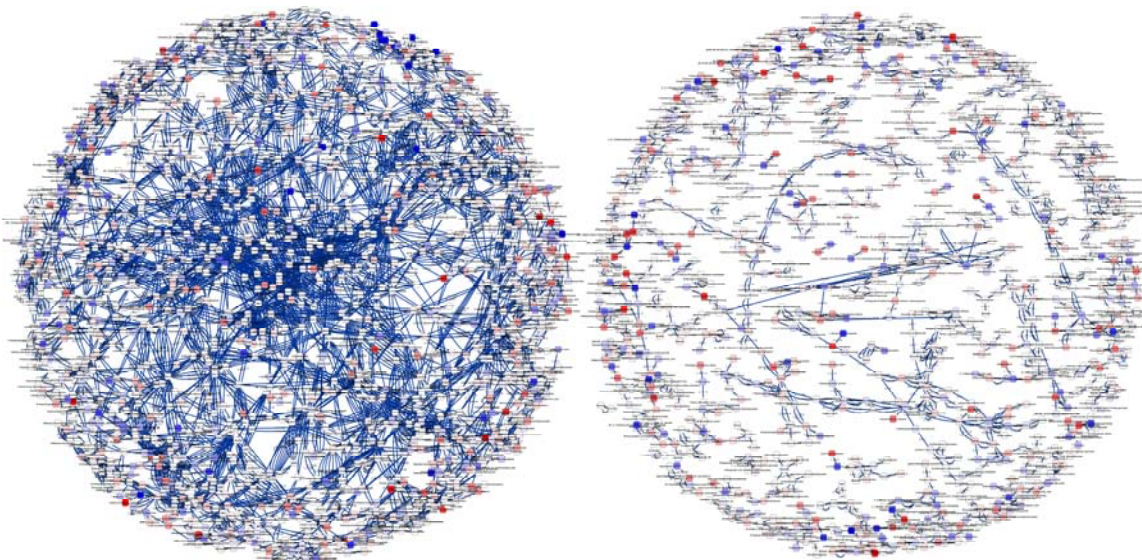


Figure 1. The modeled metabolome from the Western English Channel data. This interactive figure represents metabolites (dots) and the enzymes which connect them (lines) and can be used to predict changes in the production or consumption of specific metabolites over space and time from metatranscriptomic data.

Going still one step further the Gilbert team used the metatranscriptome data to model the metabolic potential and explore the dynamics of the ecosystem. To do this they used Predicted Metabolic Turnover (PMT, Figure 1). PMT was used to predict changes in metabolite potential between different time points that could be validated against observed chemical characteristics. For example, the predicted change in the consumption of ammonia between January and August significantly correlated with the recorded variation in actual concentration of ammonia ( $R = 0.96$ ,  $p < 0.001$ ). These modeling efforts made it possible to maximize the utility of the metatranscriptomic data and predict the metabolic outcome resulting from the exposure of the characterized communities to specific environmental changes. Without this modeling component the data would have been far less informative and useful for exploring ecological dynamics.



## Targeting and Characterizing Cells of Interest

In addition to the better known “omics” of (meta)transcriptomics, (meta)proteomics and (meta)metabolomics, other tools exist to better characterize the functional attributes of microbes. **Professor Victoria Orphan** discussed two additional methodologies, immunocapture as a means to determine what cell types develop physical connections with cells of interest, and nanometer secondary ion mass spectrometry as a means to determine cell metabolic properties. The development of methodologies and new approaches that extend our ability to link uncultured microorganisms to their specific exophysiological role(s) in nature is a top priority for the marine microbiology field in the coming decade. While metagenomic data from complex environments has expanded our view of the metabolic potential existing within a complex microbial assemblage, gene inventories are frequently incomplete and are hampered by uncertainty in assigning sequencing reads to a specific phylogenetically identified source organism. Furthermore, this mix of gene fragments, now separated from the context of the source microorganism and its spatial relationship within the community, provide little insight into potential interactions between microorganisms *in situ*.

In effort to tackle this latter issue, Orphan presented an alternative metagenomics-compatible technique called Magneto-FISH for assessing the metabolic potential of physically associated microorganisms in natural samples (Orphan 2009). Magneto-FISH is a culture-independent immuno-magnetic cell capture method which uses catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) to identify microbial cells of interest in an environmental sample, followed by the attachment of paramagnetic beads to the hybridized microorganisms using an antibody targeting the fluorochrome used in the CARD-FISH reaction. Bead-associated target cells and other physically attached microorganisms are then selectively recovered with a magnet from the environmental matrix. Genomic DNA extracted from this enriched, low complexity microbial assemblage can then be used for downstream metagenomics or targeted PCR-based gene surveys, providing unique insight into the identity of potential partner microorganisms and their metabolic potential.

Using the Magneto-FISH technique, methane-oxidizing ANME-2c Archaea and their associated syntrophic bacterial partners were selectively captured from anoxic methane seep sediment (Fig. 2). 16S rRNA gene surveys of the enriched assemblage followed by FISH experiments with the original seep sediment revealed the existence of at least two distinct sulfate-reducing bacterial partners that form physical associations with members of the ANME-2c clade. To gain further insight into the metabolic potential of this uncultured syntrophic association, genomic DNA from the captured assemblage was also used for metagenomic analysis. Predicted gene fragments from the pyrosequenced consortia confirmed the presence of metabolic pathways for both bacterial sulfate-reduction and archaeal methanogenesis/anaerobic methanotrophy. Additionally, this metagenomic survey also recovered genes linked to nitrogen fixation and nitrate utilization, prompting the development of new hypotheses regarding the role of methane-oxidizing consortia in nitrogen cycling within the methane seep ecosystem.

This type of directed environmental metagenomic analysis of a targeted low complexity assemblage of microorganisms provides a powerful mechanism for hypothesis generation regarding the functional capabilities of uncultured microorganisms/ microbial consortia. However, this gene-based information on its own is insufficient for

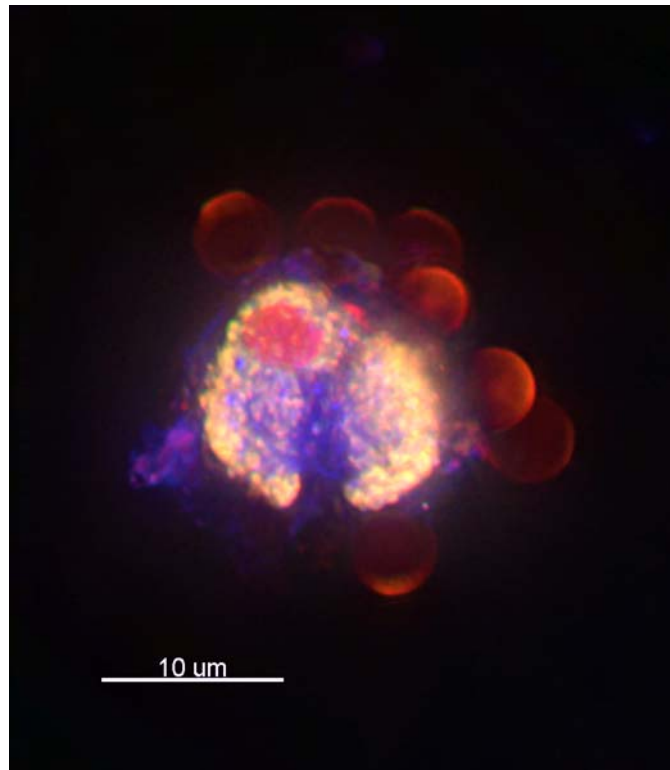


Figure 2. Magneto-FISH image of methane-oxidizing Archaea. The image is a combination of DNA-staining (blue), fluorescence in situ hybridization of the archaeal cells (yellow) and paramagnetic bead fluorescence (red). Courtesy of Victoria Orphan.

demonstrating the true metabolic capacity of these organisms *in situ* and requires the use of complimentary methodologies that enable the direct assessment of microbial activity by specific microorganisms in natural samples (see Orphan 2009).

To test the hypothesis of nitrogen-fixation by methane-oxidizing consortia of ANME-2 archaea and sulfate-reducing bacteria for example, methane seep sediment was incubated with isotopically labeled nitrogen-15 dinitrogen gas ( $^{15}\text{N}_2$ ). Cell-specific fixation of nitrogen was examined in phylogenetically identified microbial cells from the sediment incubation using a combination of fluorescence hybridization (FISH; to identify the target cells of interest) and nanometer secondary ion mass spectrometry (nanoSIMS; to enable direct imaging the incorporation of  $^{15}\text{N}$  label into FISH hybridized



cells and microbial consortia). Through the use of this type of multidisciplinary strategy, researchers are now equipped with the appropriate tools to fully maximize the utility of metagenomic data for assessing metabolism and functional relationships between microorganisms in nature.

A few categories of environmental microorganisms deserve special attention owing to the many tantalizing questions that remain unresolved, including those related to their functional genomic properties. These are eukaryotic microbes, environmental viruses and microbiome communities.

## Capturing Eukaryotic Microbes

**Dr. Daniel Vaultot** from the University of Pierre and Marie Curie and CNRS, Biological Station in Roscoff, France, described the diversity and metagenomics of small eukaryotic phytoplankton. In many oceanic waters, an important fraction of photosynthetic primary production can be assigned to very small cells, less than 3  $\mu\text{m}$  in size, called picoplankton. While prokaryotic primary producers are mostly limited to two genera of cyanobacteria, *Prochlorococcus* and *Synechococcus*, small photosynthetic eukaryotes (picoeukaryotes) are far more diverse. Photosynthetic picoeukaryotes have been obtained in pure culture for many groups including prasinophytes (e.g. *Micromonas*, *Ostreococcus* and *Pycnococcus*, see Fig. 3). These culturing efforts have resulted in the description of many novel picoeukaryotes, including *Bolidomonas*, a close relative of diatoms, and *Partenskyella*, the first picoeukaryote chlorarachniophyte. However, many picoeukaryotes resist cultivation, a very small number of species have been thoroughly described, and fewer still have had their genomes sequenced. In the last decade, the application of molecular techniques has enabled the direct characterization of some important groups of picoeukaryotes directly from natural samples, via the sequencing of the eukaryotic 18S rRNA gene ( Moon-van der Staay et al. 2001). Two important advancements in the characterization of these picoeukaryotes have been the optimization of primers for their detection (based on the 18S rRNA gene and conserved photosynthesis-associated genes) and cell targeting and physically sorting by flow cytometry. These approaches have led to the conclusion that numerous groups of photosynthetic picoeukaryotes exist, especially in the open ocean, including uncharacterized and uncultivated novel clades within the Prasinophyceae, Chrysophyceae and Haptophyta.

Metagenomic studies of picoeukaryotes pose special challenges. The “standard” metagenomic strategies successfully developed for prokaryotes are difficult to apply to eukaryotic microbes. The DNA obtained from filtered samples is dominated by the prokaryotic sequences: for example, in the Sargasso samples from the Venter study, less than 20% of the scaffolds correspond to eukaryotes (many of which are not primary producers but heterotrophs). So, innovative approaches are required. One very promising strategy is the use of flow cytometry to sort cells based on size and pigment fluorescence. DNA from the cell type of interest is then amplified by multiple displacement amplification (MDA) and sequenced. This approach has been successful

to reconstruct the genome of very small uncultured nitrogen fixing cyanobacteria lacking photosystem II, and in another study to obtain genomic information about uncultured haptophytes (Cuvelier et al. 2010).

Scientists at Roscoff have begun to apply whole genome amplification by MDA to sorted populations of coastal picoeukaryotes from the Chile upwelling. 454-generated Fig. 1.

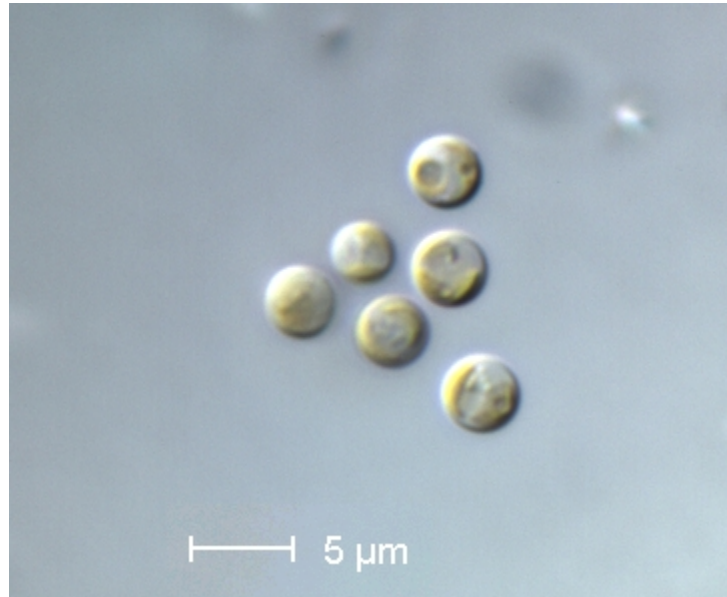


Figure 3. *Pycnococcus provasolii* (RCC 444), a prasinophyte (green alga) typical of small photosynthetic eukaryotes. Courtesy of Daniel Vaultot.

sequence data revealed that 60% of the reads matched the genome of a small prasinophyte species, *Bathycoccus*, which was recently sequenced from a Mediterranean strain. Comparison between the Pacific and Mediterranean sequences revealed amazing conservation at the DNA level. They are now sequencing other metagenomic samples from the South East Pacific in order to gain insight into the physiology and ecology of groups of haptophytes and chrysophytes without marine cultured representatives.

The approach coupling flow cytometry sorting of selected populations (or even single cells), whole genome amplification and next generation sequencing appears much more promising than bulk sample metagenomic approaches to obtain genetic information on small photosynthetic eukaryotes that constitute one of the key component of the pelagic ocean ecosystems.

## Marine Virus puzzles

Some of the exciting new developments in the exploration of viruses in the environment was presented by **Professor Ian Hewson** from Cornell University. Viruses play crucial roles in the marine ecosystems as agents of disease (pathogens) and in terms of their biogeochemical and ecological roles in microbial mortality. They are also the most abundant biological entities on Earth (Fig. 4). Since they contain no universally conserved housekeeping genes, and they cannot be easily distinguished by morphology

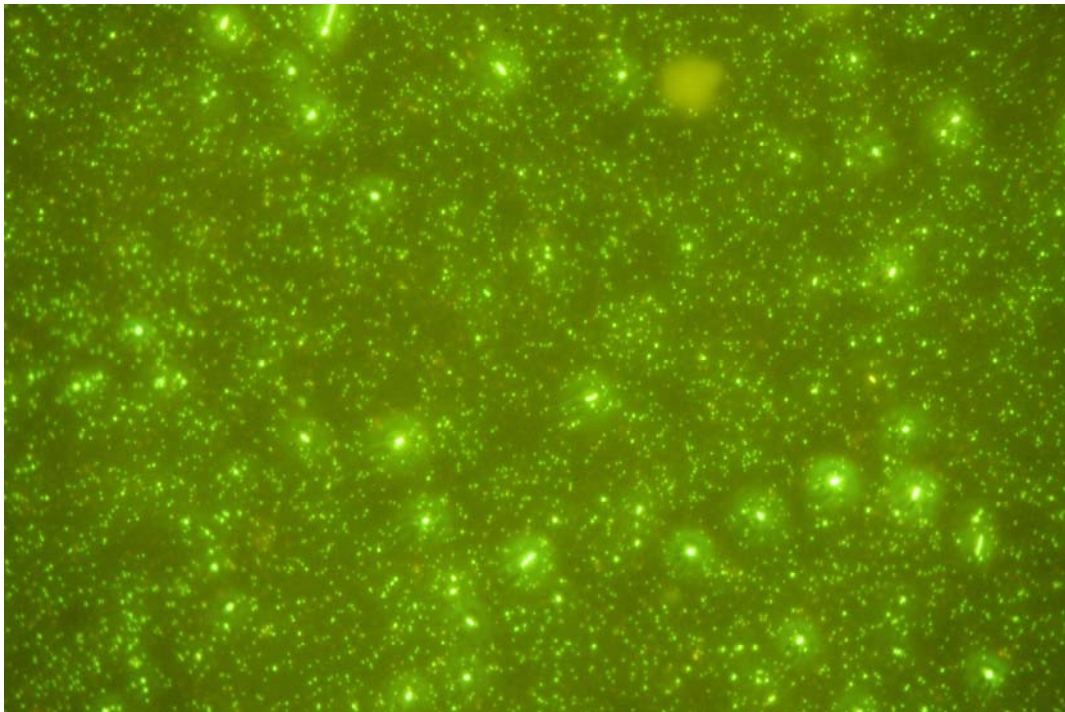


Figure 4. Fluorescence image of viruses (small particles) and prokaryotes (larger particles) in seawater. Courtesy of Ian Hewson.

or genome length, study of their diversity and ecology is particularly well suited to metaviromics (shotgun sequencing of viral communities). Early estimates based upon shotgun sequencing of planktonic and benthic viruses indicated an enormous diversity of viruses, on the order of  $7 \times 10^3$  viral genotypes per liter seawater and  $>10^4$  viral genotypes per kg of sediment. Viral genomics has also provided insight into unexpected, yet globally significant roles of viruses in ocean biogeochemistry, including their conversion of hosts with phosphorus uptake genes (*pho*) and genes involved in photosynthesis (*psa*, *psb*) (Sharon et al. 2007). Despite increasing application of viromics to various marine habitats, there remain several key questions in marine viral ecology which are only now possible due to accessibility to high throughput sequencing technologies: 1) Do [potentially latent] viruses infect and kill off bloom forming autotrophs?; 2) What types of viruses infect and kill metazoan zooplankton, especially those which play significant roles in mediating the biological carbon pump?; and 3) what is the viral role in threatened benthic invertebrate disease?

Metaviromics provides a powerful tool for elucidating potential pathogens of numerous organisms, through extraction of viral genomes directly from host tissues or lysates, then shotgun sequencing of viral genomes, re-assembly and annotation. This approach has been used in investigations of several diseases for which other pathogens have not been found. For example, metaviromic preparations from bee colonies helped elucidate potential pathogenic viruses implicated in colony collapse disorder (although in this case viruses were just one of several interacting pathogens that resulted in complete die-offs of bee colonies). More recently, metaviromics has elucidated potential viral pathogens responsible for the formation of growth anomalies on scleractinian corals of Hawaii and the Caribbean (Thurber et al. 2008). A concern of using metaviromics as a sole means by which to investigate disease-associated viral pathogens has been whether a virus associated with a diseased condition is necessarily involved in that disease. Koch's postulates are not satisfied simply by monitoring co-occurrence. Metaviromics provides a wide suite of targets for downstream studies that are unavailable in the absence of genomic information. Two examples of these approaches are quantitative PCR-based studies of metaviromically-elucidated viral genomes comparing diseased and healthy states of organisms, or detection of expressed viral genes in the host, which may indicate active infection. Viromics also provides targets for studies of epidemiology, including prevalence when combined with random sampling techniques, and examination of viral load within individual animals.

There are major challenges to the study of metaviromics which are greater than those in other groups of microorganisms. Most notably, there is very poor coverage of representative viral groups in nature, which leads to bias in comparative metaviromics, and poor annotation of viruses recovered from specific habitats or host organisms. In publically-available metatranscriptomes, for example, almost all expressed viral genes that can be annotated are cyanophage photosystem genes. However, these likely make up only a fraction of total expressed viral genes in seawater - but make up a huge proportion of those that are recognizable. Recent sequencing of the 1.2 Mbp protozoan (*Acanthamoeba polyphaga*) mimivirus and its pathogen (virophage) demonstrate that much of the viral diversity is unknown, and that many sequences retrieved in prokaryotic metagenomes may represent large viruses.

The utility of metaviromics depends on the scientific question that is raised. While metaviromics as a single tool provides useful information on the richness of viruses within an assemblage, genes carried on genomes, and some information on putative host, it is incapable in isolation of being used to address questions related to the magnitude of viral roles in aquatic ecosystems, nor viral disease epidemiology. However, metaviromic information can be used as a key component of the data needed to answer these latter questions, since it permits distinguishing features to be used in downstream analyses. As a technical aside it is also worth noting that most metaviromics is not quantitative. The majority of metaviromic studies use rolling circle amplification using the phage  $\Phi 29$  polymerase, which does not have consistent product to template ratios, and hence does not provide relative or quantitative information on constituent viral genotype abundances.



There is much work to be done in the application of (meta)viromics with viral ecology. This includes sequencing additional (meta)viromes directly from tissues or cell lines of isolated hosts, and sequencing across habitats, which may permit identification of viral ecotypes. The integration of viral diversity studies with those examining viral ecology is necessary to link the identity of viruses with their importance in ocean function. By combining viral diversity studies with host gene expression response, it should also be possible to provide indicators of viral activities within the host population.

## Superorganisms

**Dr. Karen Nelson** from the J. Craig Venter Research Institute provided the final presentation of session III. She discussed microbial communities (microbiomes) associated with primates, humans in particular. Given that the microbial world does not evenly distribute itself in aqueous environments but rather displays a strong preference for associations with other life forms in myriad symbiotic relationships, human microbiome studies have many lessons for microbial ecology. The human “superorganism” contains at least an order of magnitude more microbial cells than human cells (somatic and germ cells), and is an amalgamation of human and microbial attributes. The microbes are located on the skin, in the oral and nasal cavities, additional airways, stomach, colon, and vagina. Many relationships between the human host and microbiome remain to be determined. But, it is clear that these microbial interactions endow or enhance human processes related to development, nutrition, immunity and resistance to pathogens. Changes in the human microbiome have been linked to a variety of diseases including esophageal cancer, bacterial vaginosis and pre-term babies, neonatal microbiome/necrotizing enterocolitis, nasopharynx microbiome and vaccination in children, skin microbiome, acne and psoriasis, oral diseases including periodontitis, obesity, Crohn’s and inflammatory bowel disease, colon cancer and diabetes.

Human gastrointestinal microbiome studies have progressed at many levels including phylogenetics, metagenomics and culturing. Comprehensive 16S ribosomal DNA sequence-based characterization of the distal gut and fecal microbiota indicates a highly selective environment, dominated by just two bacterial divisions, the Bacteroidetes and the Firmicutes, and one methanogenic archaeon, *Methanobrevibacter smithii*. Gene (Clusters of Orthologous Groups, COG) diversity based on metagenomic analyses indicate that the COG richness of the human gut is greater than that of some highly restrictive environments such as acid mine drainage, but less most such as whale fall, soil, and the Sargasso Sea (Gill et al. 2006).

The human gastrointestinal microbiome provides us with the ability to utilize many plant polysaccharides that are a common part of our diet, including those rich in xylan, pectin, and arabinose carbohydrates. It also includes genes for the synthesis of essential amino acids and vitamins and the detoxification of xenobiotics. The methanogenic species present in our gut is also a valuable microbial community members as it helps

reduce the accumulation of hydrogen and thereby prevents feedback inhibition of bacterial fermentation.

Despite the obvious connection between the  $10^{13}$  to  $10^{14}$  microbes per person inside us and human health, they remain our intimate strangers. Remarkably, most members of the human microbiome have never been cultured. This may change as a result of growing interest in the human microbiome. In 2007, the National Institutes of Health (NIH) initiated the Human Microbiome Project (HMP), one component of which is the production of reference genome sequences. As a part of the HMP Dr. Nelson and colleagues recently obtained genome sequences for 178 bacterial strains cultured from the human gut (Nelson et al. 2010). These reference bacterial genomes made it possible to explore the core and strain-specific genes (pangenome properties) shared among sequenced members of the same species. In addition, these genomes provided microbial species contextual information to the available human gut microbiome metagenomic data. Forty percent of the metagenome sequence fragments could be associated with one of these cultured isolates, a major advance in linking sequence to microbial source. Despite the many hundreds of completed bacterial genome sequences, five percent of the genes annotated from these reference genomes were identified as novel. This underscores the remarkable diversity of bacterial proteins. We are still far from saturating microbial species genetic data sets.”

In this as in other areas of (meta)genomics there is a need for greater metadata, it is plagued by the issue of so many unknown gene functions, and the role of microbes present as minor fractions of the overall population remain unclear. Correlating/coordinating data cross different groups is still problematic.

### Literature Cited

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messie M, Tringe SG, Woyke T, Welsh RM, Isohey T, Lee JH, Binder BJ, DuPont CL, Latasa M, Guigand C, Buck KR, Hilton J, Thiagarajan M, Caler E, Read B, Lasken RS, Chavez FP and Worden AZ. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci. USA* 107:14679-14684.

Feist AM, Herrgard MJ, Thiele I, Reed JL, and Palsson BO. 2009. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7:129-143.

Gilbert JA, Field D, Huang Y, Edwards R, Li WZ, Gilna P, Joint I. 2008. Detection of large Numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* 3: e3042.

Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM and Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. *Science*: 1355-1359.



Hooper SD, Mavromatis K, Kyrpides NC. 2009. Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol.* 10:R45.

Moon-van der Staay SY, De Wachter R and Vaulot D. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409:607-610.

Nelson KE, Weinstock GM, Highlander SK et al. (plus additional members of the Human Microbiome Jumpstart Reference Strains Consortium). 2010. A catalog of reference genomes from the human microbiome. *Science* 328: 994-999.

Orphan VJ. 2009. Methods for unveiling cryptic microbial partnerships in nature. *Curr Opin Microbiol* 12: 231-237.

Rossello-Mora R, Lucio M, Pena A, Brito-Echeverria J, Lopez-Lopez A, Valens-Vadell M, Frommberger M, Anton J, and Schmitt-Kopplin P. 2008. Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*. *ISME J.* 2:242-253.

Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB, Yooseph S, Zeidner G, Golden SS, Mackey SR, Adir N, Weingart U, Horn D, Venter JC, Mandel-Gutfreund Y and Beja O. 2007. Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J.* 1:492-501.

Thurber RLV, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C, Edwards RA, Haynes M, Angly FE, Wegley L and Rohwer FL. 2008. Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc. Natl. Acad. Sci. USA* 105:18413-18418.



Photo credit: John Wooley



## Report from Session 4

### **T** raining Marine Microbiologists Today: Culturing Versus Unix

Jennifer Biddle (Session Chair)<sup>1</sup>, Roderic Guigo (Rapporteur)<sup>2</sup>, Jörg Peplies<sup>3</sup>, Frank Stewart<sup>4</sup>

<sup>1</sup> School of Marine Science and Policy, University of Delaware, Lewes DE USA; <sup>2</sup> Centre de Regulacio Genomica, Barcelona, Spain; <sup>3</sup> Ribocon GmbH, Bremen, Germany; <sup>4</sup> Massachusetts Institute of Technology, Cambridge, MA USA.

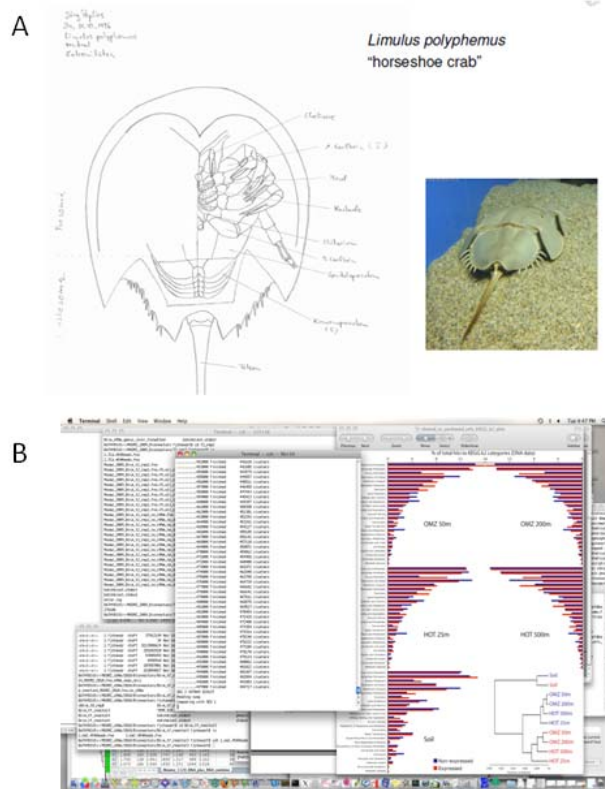
**Question:** *How do we better train next generation marine scientists in the integration of next generation sequencing with microbiology and oceanography?*

#### **Session 4 recommendations**

- 1. Students should be trained to be well rounded scientists, grounded with solid training. New technologies will be best adapted by students with cross-disciplinary skills.*
- 2. Bioinformatics training is needed across all experience levels, ranging from beginning students to established PIs.*
- 3. Outreach is needed to the computational community to intrigue them with the major questions in marine biology. Outreach is needed to the marine biological community to facilitate their acquisition of computational skills.*
- 4. A training course is needed of at least five days that addresses the analysis of real datasets from users. The output of the class should be condensed into a online, portable format.*
- 5. Funding is needed for tool development. Making the computational part of the science “easier” to use will decrease the need to teach intricate details. A revolution in biological computing infrastructure is needed.*

## Background

Marine microbiology has often been seen as a bit of a “soft science” and small datasets and limited access to resources, be it ships or cultured isolates, has allowed this label to perpetuate. However, marine microbiologists today find themselves among large datasets that span the globe and they have a large desire to process and analyze these datasets. As such, we are in a renaissance period where talented students in marine science, computational science, mathematics, geography, biology and chemistry are taking interest in the marine metagenomic datasets. To do this, they must challenge themselves to explore the world beyond the Niskin bottle and grapple with the comprehension of tools that previously were unneeded or underutilized in their study. As such, the instruction of such students in marine microbiology is poised for reorganization to allow the instruction of both classic techniques of microbiology with next-generation skills for data processing. We utilized this session to discuss the challenges facing the instruction of future and current marine metagenomicists.



Desktops of biologists in the A) past and B) present

Figure 1.

## Training marine microbiologists today

### Work from the bottom up

The brave new world of marine metagenomics and the advent of next-generation sequencing technology do not remove the requirement that students in science need basic scientific training. The next generation of scientists retains the need to be grounded in basic techniques of critical and quantitative thinking, hypothesis generation, communication and collaboration. While tools will change, the need for future biologists to understand experimental design will remain and be key to their success in science (Fig.1). New technology, such as next generation sequencing (NGS) will not solve all of our problems. The lure of new technology must be avoided in the infancy of scientific study, so students may develop a solid foundation of basic knowledge and scientific thinking needed to tackle any technological developments they may see in their careers.

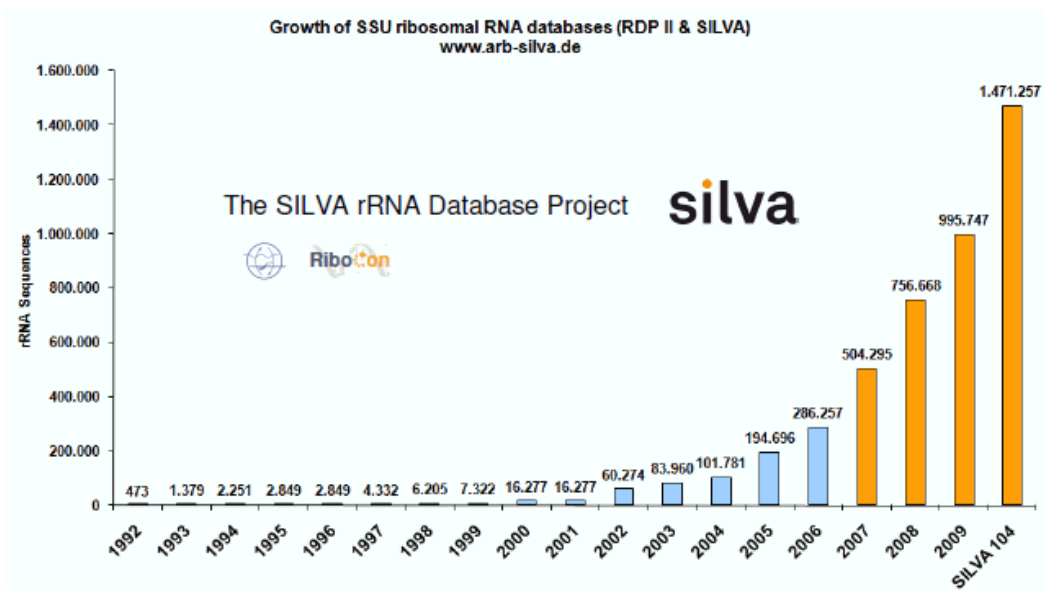


Fig. 2: Growth of ribosomal databases over time (Courtesy of J. Peplies).

### Work from the top down

An interesting challenge is faced by marine microbiologists today, in that technological developments, particularly in NGS, outpace their ability to learn and adopt these technologies (Fig. 2). As such, education also needs to be provided to established PIs in the field. Many established researchers, when faced with the different approach needed to tackle NGS data, face a “barrier of fear”. In order to move scientists past this barrier, a hypothesis-driven approach appears to be effective. Students of any age, when given a dataset they have inherent interest in, often are more willing and capable of tackling large datasets. Experience from teaching biologists command-line programs, or Unix-based programs, shows that biologists are typically not as computer savvy as they should be. Within the last five years, the service company Ribocon has trained more than 400 users from all over the world in five days workshops focusing on these

aspects. These workshops have shown the strong interest of the community in NGS and the corresponding “modern” ways of DNA sequence analysis but they also showed that often the users are quickly overstrained by missing basic computational (Unix-related) skills plus a limited overview in the field of bioinformatics. However, these basics are required for efficient data handling and to create flexibility, two important aspects in current and future DNA sequence analysis.

The need to tackle both instruction in computational techniques and biological inquiry of NGS datasets, suggests that these skills are best learned in a training course that is specific to the field. From the experience of the Ribocon workshops it is clear that users tend to avoid workshops solely focused on the basics such as the Unix command line, because it is too removed from their daily work. These aspects of training must be integrated as small modules of workshops that primarily focus on the analysis of the user-relevant data. With this approach, the users will realize the power of the skills they are developing and will be motivated to continue to advance.

### **The problem with interdisciplinarity**

A difficulty faced in the instruction of both established researchers and young students in topics such as marine metagenomics is the breadth of knowledge needed for accurate analysis of the data. Whether training from the bottom up, or the top down, students will need skills in many disciplines in order to comprehend and process large amounts of sequencing data. While students could potentially remain in school much longer to do this, a more sensible option would be to recruit students from needed course disciplines, such as microbiology, computer science and ecology, to attack the datasets and problems facing marine microbiology, utilizing their unique skills. Additionally, students, perhaps of a training course, need to be taught at the level of “contractor”. In home building, a contractor oversees the work of electricians, plumbers and carpenters. The contractor should understand the use of other’s tools and appreciate the time and effort they put into their job. As such, a marine microbiologist should not be expected to develop the next web-based metagenomics tool. Instead, they should appreciate the development needed for this tool, understand its use and interpret its output to others. As the “contractor”, students will be able to gain knowledge in many areas quickly, allowing for greater understanding of the entire framework of how these datasets are analyzed.

Scientists should also be trained in communication skills to improve the networking between disciplines and cultures. Collaborative work is the future and the field of marine metagenomics is a testament to this new reality.



## The structuring of a training course

As metagenomic analyses become more pervasive, teachers of microbiology are increasingly challenged to prepare students to handle large, multi-sequence datasets. Analysis of these datasets can be made using web-based platforms and also through command-line programs. Students in a training course should learn basic skills of utilizing Unix-based platforms. While web-based platforms for automated metagenomic analysis (e.g., MG-RAST) are rapidly improving, the implementation of project-specific analyses will likely still require that students have a working knowledge of command-line methods for manipulating text files. Students may also benefit directly from skills in modifying or writing text-management scripts (e.g., in Perl, Python). Consequently, students with a formal background in bioinformatics and coding may be at an advantage initially. However, a surprisingly large number of analyses can be implemented with a fairly modest bioinformatics skill set. The latter can be obtained through concerted teaching of command-line operations and scripting, perhaps in lab exercises as part of a traditional microbiology course, or through a concentrated workshop format.

Knowledge of metagenomic techniques is best obtained through targeted exercises that immerse a student in a real dataset (preferably their own). Fortunately, metagenomic datasets lend themselves to a seemingly limitless series of questions and analyses, many of which could constitute ideal learning exercises (e.g., parsing of specific taxonomic groups or functional gene sets, statistical analysis of hit count variance, mapping reads onto genomes). Such projects would not only serve to familiarize students with techniques but also provide an inlet to more exhaustive, follow-up explorations of specific physiological processes or taxonomic groups.

Students of metagenomics should be aware of potential biases related to database structure and sequence annotation. As an example, a highly expressed functional gene in a bacterioplankton community was overlooked by automatic parsing methods due to an oversight in how this gene is annotated in a public database (F. Stewart, unpublished results). Such results underscore a need to thoroughly understand potential limitations of current analytical methods and resources and highlight a potential role for manual verification of metagenomic trends. This also serves as a reminder that there is still a place for the study of pure culture microbes and improvements need to be made in the field of gene annotation.

Finally, students should be equipped to use metagenomics/metatranscriptomics (and microbiology in general) as an experimental tool. To date, metagenomics has been used primarily in an observational mode to better characterize the staggering amount of genetic and metabolic diversity in natural microbial communities. However, as analytical platforms improve, high throughput sequencing of community DNA and RNA is increasingly applicable to studies examining community functional responses to experimental manipulation (e.g., substrate additions, succession experiments). Applying these techniques experimentally requires that students carefully consider and (ideally) be formally trained in issues of experimental design and analysis. Notably,

biostatistical methods will be increasingly necessary for teasing apart signal from noise and for parsing out the effects of community composition shifts versus actual changes in gene expression/abundance within an organism. The expansion of metagenomics into other areas of microbiology presents a tremendous opportunity for creativity in both how we mine and analyze existing datasets and in how we apply high throughput techniques to new systems and experimental questions. Fostering this creativity in students is an ongoing challenge, but is probably best done collaboratively and through hands-on learning exercises.

### Becoming more efficient, better teachers

While strides can be made to educate marine microbiologists thoroughly, efforts should also be made to bring this education to the “masses” in the form of more usable web-based tools and online teaching modules. Overall, the scientific infrastructure for bioinformatics needs to be improved on a global scale, requiring rethinking even the most basic processes and algorithms used to be sure that the most appropriate means are taken to process NGS data.



Adelie penguins on floe, R/V LMG in background (Photo credit: F. Stewart)



## LIST OF PARTICIPANTS

Josefa Antón  
University of Alicante  
anton@ua.es  
SPAIN

Douglas H. Bartlett  
Scripps Institution of Oceanography  
dbartlett@ucsd.edu  
USA

Jen Biddle  
University of Delaware  
jfbiddle@udel.edu  
USA

Guy Cochrane  
European Bioinformatics Institute  
cochrane@ebi.ac.uk  
UK

Garbine Guiu Etxeberria  
European Commission Biotechnologies Unit  
Garbine.Guiu@ec.europa.eu  
BELGIUM

Jack Gilbert  
Plymouth Marine Laboratory  
gilbertjack@anl.gov  
UK

Frank Oliver Glöckner  
Max Planck Institute for Marine Microbiology  
fog@mpi-bremen.de  
GERMANY

Roderic Guigo  
Centre for Genomic Regulation  
rguigo@imim.es  
SPAIN

Ian Hewson  
Cornell University  
ih88@cornell.edu  
USA

Lynette Hirschman  
The MITRE Corporation  
lynette@mitre.org  
USA

Brian Kelly  
Pacific Biosciences, Inc.  
bkelly@pacificbiosciences.com  
USA

James Knight  
454 Life Sciences  
james.knight@roche.com  
USA

Georgios Kotoulas  
Hellenic Center For Marine Research  
kotoulas@her.hcmr.gr  
GREECE

Alice McHardy  
Max Planck Institute for Informatics  
mchardy@mpi-inf.mpg.de  
GERMANY

Folker Meyer  
Argonne National Laboratory  
folker@anl.gov  
USA

Karen Nelson  
J. Craig Venter Research Institute  
kenelson@jcvr.org  
USA

## EC-US TASK FORCE ON BIOTECHNOLOGY RESEARCH

---

Gerald Nyakatura  
LGC Genomics  
gerald.nyakatura@lgcgenomics.com  
UK

Victoria Orphan  
California Institute of Technology  
vorphan@gps.caltech.edu  
USA

Jörg Peplies  
Ribocon GmbH  
jpeplies@ribocon.com  
GERMANY

Nathan Price  
University of Illinois  
ndprice@illinois.edu  
Illinois  
USA

Christopher Quince  
Glasgow University  
quince@civil.gla.ac.uk  
UK

Gail Rosen  
Drexel University  
gailr@ece.drexel.edu  
USA

Michael E. Sieracki  
National Science Foundation  
msierack@nsf.gov  
USA

Frank Stewart  
Massachusetts Institute of Technology  
fstewart@MIT.EDU  
USA

Mark S. Strom  
Northwest Fisheries Science Center  
Mark.Strom@noaa.gov  
USA

Juli M. Trtanj  
NOAA Oceans and Human Health Initiative  
Juli.Trtanj@noaa.gov  
USA

Frances M. Van Dolah  
Center for Coastal Environmental Health and  
Biomolecular Research  
Fran.Vandolah@noaa.gov  
USA

Daniel Vaultot  
CNRS, Station Biologique  
vaultot@gmail.com  
FRANCE

Dale Yuzuki  
Applied Biosystems  
Dale.Yuzuki@lifetech.com  
USA



Photo credit: John Wooley



## **US-EU Task Force on Biotechnology Research The Marine Genomics Working Group**

The new age of metagenomics enables the study of the vast majority of marine microbial species which we are as yet unable to culture in the laboratory. These technologies and the analyses they enable have ushered a new era on marine genomics. The October 2010 workshop of the Marine Genomics Working Group was organized as a high-level US and EU discussion on the applications and limitations of new sequencing technologies, new biological questions that these developments raise and could help to address and the resulting bioinformatic bottlenecks. The need for interdisciplinary training of the next generation marine genomic scientists was also an important part of the discussion.

[http://ec.europa.eu/research/biotechnology/eu-us-task-force/index\\_en.cfm?pg=links](http://ec.europa.eu/research/biotechnology/eu-us-task-force/index_en.cfm?pg=links)