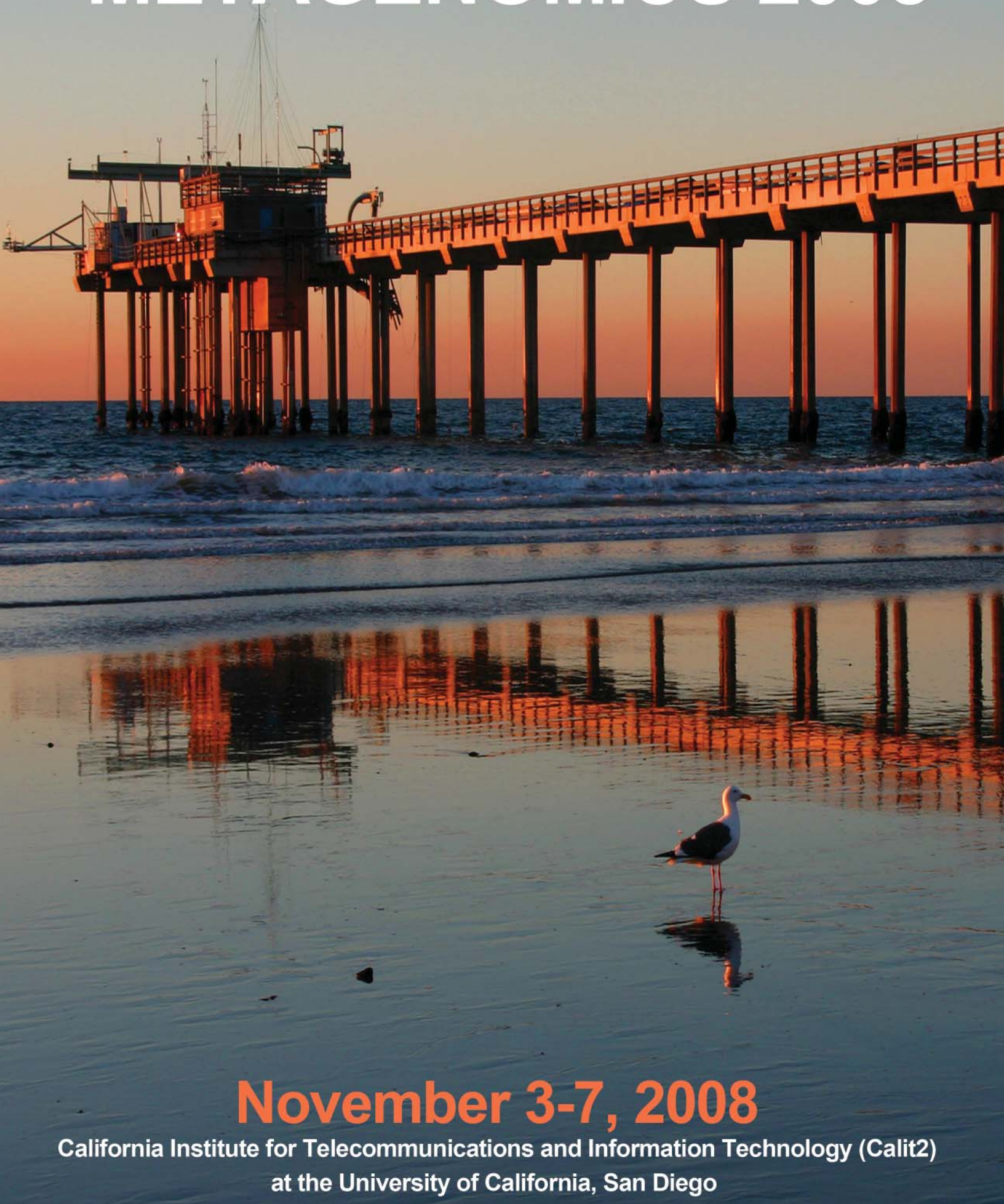


METAGENOMICS 2008



November 3-7, 2008

California Institute for Telecommunications and Information Technology (Calit2)
at the University of California, San Diego

Welcome to



METAGENOMICS2008



**November 3-7, 2008
Atkinson Hall
UC San Diego
La Jolla, California
USA**

This is the third annual Metagenomics conference to be held in San Diego, and the fourth such conference, including Metagenomics 2003, which was held in Germany. At the intersection of many diverse disciplines, metagenomics is engaging an exceptionally wide and interdisciplinary intellectual community. Metagenomics 2008 will provide invaluable opportunities to discuss and accelerate metagenomic research, ranging from microbial ecology, genomics and evolution, bioinformatics and biochemistry, to ecological genomics, population/community genomics and environmental genomics.

Metagenomics 2008 has expanded to five days, including an exciting workshop organized by the Genomic Standards Consortium on Monday. And once again, the conference is hosted by the UC San Diego division of the California Institute for Telecommunications and Information Technology (Calit2). Our thanks to everyone who made this year's conference possible, and we're already looking forward to Metagenomics 2009!

John Wooley
General Conference Chair

Kayo Arima
Co-chair



Cover images by John Wooley; all rights reserved

MONDAY, NOVEMBER 3

[All sessions in Calit2 Auditorium, Atkinson Hall, unless otherwise noted]

Enabling Metagenomics and Genomics: Information Required for Metagenomics and Genomic Standards

Workshop Organized by the Genomic Standards Consortium (GSC)

Organizers:

Dawn Field, Oxford Centre for Ecology and Hydrology
Frank Oliver Glueckner, Max Planck Institute for Marine Microbiology
John Wooley, UC San Diego

MORNING SESSIONS

8:00 On-site Registration at Atkinson Hall, Calit2 at UC San Diego

SESSION I: Genomic and Metagenomic Metadata

- 08:30 Setting the Stage: Mobilizing the Metagenomics Community
John Wooley, UC San Diego
- 08:45 CAMERA Interest in Community Standards for the Capture and Exchange of Metadata
Paul Gilna, CAMERA/UC San Diego
- 09:00 Overview of the GSC and the Minimum Information About a (Meta) Genome Sequence (MIGS/MIMS) Specification
Dawn Field, Oxford Centre for Ecology and Hydrology
- 09:15 The Rapidly Growing Standards Landscape in 'omics
Susanna Sansone, European Bioinformatics Institute
- 09:45 Source Organism and Molecule Information at INSDC and the Trace Archives
Guy Cochrane, EMBL-European Bioinformatics Institute
- 10-10:30 Break

SESSION II: Core GSC Projects

- 10:30 Implementing MIGS/MIMS: The Genomic Contextual Data Markup Language (GCDML)
Renzo Kottmann, Max Planck Institute for Marine Microbiology
- 10:45 StrainInfo and the Linkage of Organisms, Gene and Genomes: The Genomic Rosetta Stone
Peter Dawyndt, University of Ghent
- 11:00 Towards Computer Assisted Markup of Data: Habitat-Lite
Lynette Hirschman, The MITRE Corporation
- 11:15 Towards a Standards Compliant Literature: The GSC eJournal
George Garrity, Michigan State University
- 11:30 Towards Transparency of Computational Analyses: A Central SOP Repository
Owen White, University of Maryland
- 11:45 Questions & Community Comments
- 12:00 Lunch

P
R
O
G
R
A
M
:
D
A
Y
1

MONDAY, NOVEMBER 3

AFTERNOON SESSIONS

SESSION III: Defining the Scope of GCDML

- 01:00 The Importance of Context for the Design and Interpretation of Comparative Metagenomics Studies: The MINIMESS Proposal
Jeroen Raes, European Molecular Biology Laboratory
- 01:15 Should Genomics Contextual Data Fly Coach?
Inigo San Gil, University of New Mexico
- 01:30 Extending MIGS/MIMS to the Description of Ribosomal RNA Sequences
Frank Oliver Gloeckner, Max Planck Institute of Marine Microbiology
- 01:45 Questions & Community Comments

SESSION IV: Ontologies and the Description of Habitat and Geolocation

- 02:00 The Environment Ontology: Linking Environmental Data
Norman Morrison, University of Manchester
- 02:15 Towards an Open Access Gazetteer
Lynn Schriml, University of Maryland
- 02:30 RDP Survey of Habitats Descriptors
James Cole, University of Connecticut
- 02:45 Questions & Community Comments
- 3:00 Break

SESSION V: Metadata Capture: A Key Step for Advancing Understanding

- 03:30 The Genomes Online Database (GOLD): The Value of a Comprehensive Metadata Collection
Nikos Kyrpides, Joint Genome Institute
- 03:45 Metadata Capture in the IMG/IMGm: Getting Scientists to Contribute
Victor Markowitz, Lawrence Berkeley National Lab
- 04:00 The Genome Catalogue: A Future Vision
Lynn Schriml, University of Maryland
- 04:15 Questions & Community Comments
- 04:30-05:30 Panel Discussion: A Community Vision

TUESDAY, NOVEMBER 4

P
R
O
G
R
A
M
:
D
A
Y
2

08:00 On-site Registration at Atkinson Hall, Calit2 at UC San Diego

08:00 Poster Setup/Breakfast

SESSION I: Challenges in Metagenomics for Bioinformatics and Computational Biology

Session Chair: John Wooley, UC San Diego

09:00 Introduction

John Wooley, UC San Diego

09:20 **Components for rational management of genomic and metagenomic information**

Owen White, Univ. of Maryland

10:00 Talk TBA

Adam Godzik, Burnham Institute for Medical Research and UC San Diego

10:40 Break

11:00 Talk TBA

Daniel Huson, Tübingen University

11:40 Efficient Metagenomics Data Processing: Pitfalls and Solutions

Nikos Kyrpides, Joint Genome Institute

12:20 Challenges in Ecological Metagenomics

Patrick Schloss, UMass Amherst

01:00 Lunch

SESSION II: Documenting the Emergence and Opportunities of the Field: Parallel Sessions

02:00 Connecting Metagenomics and Metaproteomics

Chair: Janet Jansson, Lawrence Berkeley Natl Lab

02:00 New Sequence Technologies for Metagenomic Research

Chair: Tim Hunkapiller, Discovery Biosciences and Applied Biosystems

02:00 The Metagenomic Voyage: In Situ to In Silico

Chair: Eric Allen, UC San Diego

02:00 Wellness and Diseases: Implications of Important Microbiota

Chair: Justin L. Sonnenberg, Stanford University

02:00 Soil Options

Chair: Patrick Schloss, UMass Amherst

02:00 GSC Breakouts [details to be announced, to include GCDML and Genome Rosetta Stone]

05:00 - 08:30 Welcome Reception (La Jolla Shores Hotel)

05:30 Welcome Talk: TBA

Mark Ellisman, UC San Diego

05:45 Dinner

07:00 Congress Keynote:

The Evolution of Small Bacterial Genomes in the Ultra-Oligotrophic Ocean

Steve Giovannoni, Oregon State University

WEDNESDAY, NOVEMBER 5

08:00 On-site Registration/Poster Setup/Breakfast - Atkinson Hall, UC San Diego

08:30 Morning Talk: The Importance of Marine Picoeukaryotes and the Search for Lost Time
Alexandra Z. Worden, Monterey Bay Aquarium Research Institute

09:30 Break

SESSION I: Marine Metagenomics

Chair: Stephen J. Giovannoni, Oregon State University

09:40 Photosystem-I Gene Cassettes in Marine Phages
Oded Beja, Technion-Israel Institute of Technology

10:20 Deep-Ocean Metagenomics: Comparative Investigations of Microbes Inhabiting Hydrothermal Vents and the Cold Deep Ocean
Shannon Williamson, J. Craig Venter Institute

11:00 Break

11:10 Metagenomic Analysis of Deep Subsurface Environments
Hideto Takami, JAMSTEC

11:50 Integrative Marine Metagenomics
Elizabeth A. Dinsdale, SDSU/University of Adelaide

12:30 Lunch

SESSION II: New Technologies in Metagenomics

Chair: Paul Gilna, CAMERA/UC San Diego

01:30 Analyzing the Mobilome Using Metagenomics
Julian R. Marchesi, Cardiff University

02:10 Electromicrobiology: Novel approaches for investigating charge transfer and energy transformation in microbial systems
Yuri Gorby, J. Craig Venter Institute

02:50 Accessing the Metatranscriptome for complex marine microbial communities
Jack Gilbert, Plymouth Marine Laboratory

03:30 Break

03:40 Genome Standards Consortium Report
Dawn Field, Oxford Centre for Ecology and Hydrology

SESSION III: Selected Talks by Poster Abstract Authors & Poster Session

04:00 Integrated Information System for Genomic and Metagenomic Data Analysis at NCBI
Anjanette Johnston, NIH

04:20 Targeted Gene Identification from Short Gene Fragments in Metagenome and their Use in Biogeochemical Studies
Robin B. Kodner, University of Washington

04:40 Metabolic Characterization of *Candidatus Accumulibacter Phosphatis* Using Metaproteomic Analysis
Jason Flowers, University of Wisconsin-Madison

05:00 Poster Session

07:00 Dinner

08:00 Dinner Keynote: A Genomic Encyclopedia of Bacteria and Archaea (GEBA) and the Search for the Dark Matter of the Biological Universe
Jonathan Eisen, UC Davis and Joint Genome Institute

P
R
O
G
R
A
M
:
D
A
Y
3

THURSDAY, NOVEMBER 6

P
R
O
G
R
A
M
:
D
A
Y
4

08:00 On-site Registration/Poster Setup/Breakfast

08:30 Morning Talk: Comparative Metagenomics

Peer Bork, European Molecular Biology Laboratory

09:30 Break

SESSION I: Host-Associated Microbes

Chair: Janet Jansson, Lawrence Berkeley National Lab

09:40 Genomic and Genetic Insight into the Gut
Microbiota Function and Manipulation

Justin L. Sonnenburg, Stanford University

10:20 New Insights Into Lignocellulose Conversion by
Termite Gut Microbes

Jared Leadbetter, California Institute of Technology

11:00 Break

11:10 Metaproteomics as a key technology for
characterizing the human microbiome

Nathan C. VerBerkmoes, Oak Ridge National Lab

11:50 Metagenomic Comparison of the Microbial Hindgut
Communities in Drywood- and Grass-Feeding
Termites

Falk Warnecke, DOE Joint Genome Institute

12:10 Screening of the Human Intestinal Microbiota for the
Discovery of Novel Enzymes Efficient for Plant Fiber
Degradation

Gabrielle Veronese, Institut National de la Recherche
Agronomique (France)

12:30 Lunch

SESSION II: Emerging Sequence Technologies

Chair: Nikos Kyrpides, Joint Genome Institute

01:30 Single molecule analyses of DNA in environmental
microbes

Kun Zhang, UC San Diego

02:10 New Sequencing Technologies at JGI and
Applications in Bioenergy Research

Feng Chen, Joint Genome Institute

02:50 Break (move to breakout rooms)

SESSION III: Parallel Tutorials/Demos

03:10 – 05:10 Parallel Tutorials 1

CAMERA Kayo Arima, UC San Diego

RDP-II James Cole, Michigan State University

MEGAN/MetaSim Daniel Huson, Tuebingen University

03:10 - 05:10 Parallel Tutorials 2

IMG, IMG/M Victor Markowitz, LBNL

Greenegenes Todd DeSantis, LBNL

MicrobesOnline Paramvir Dehal, LBNL

03:10 – 05:10 Parallel Tutorials 3

Megx.net: Ivaylo Kostadinov and Melissa Duhaime, Max Planck

ARB & SILVA: Frank Oliver Gloeckner, Max Planck

New Version of MG-RAST Server: Folker Meyer, Argonne National Lab, Univ of Chicago

05:10 Break (move back to Calit2 Auditorium)

05:30 Keynote: International Soil Metagenome Sequencing Project

Timothy M. Vogel, Université de Lyon

06:30 Dinner and After-Dinner Talk: **Exploring Next
Generation Sequencing Today**

Thomas Jarvie, 454 Life Sciences

FRIDAY, NOVEMBER 7

08:00 Breakfast

SESSION I: Connecting Sequence, Structure and Function for Metagenomics

08:30 Introduction	John Wooley, UC San Diego
08:35 The Human Microbiome Project: Key Questions/ Challenges	Claire Fraser, University of Maryland
09:15 Break	
09:30 Diversity Profile of the Human Skin Microbiome in Health and Disease	Elizabeth Grice, NIH
10:10 The Human Oral Microbiome	Floyd Dewhirst, Harvard University
10:50 Break	
11:00 Metabonomic Profiling of the Gut Microbiome: Implications for Human Health	James Kinross, Imperial College, London
11:40 Lunch	

SESSION II: Getting to Molecular and Bacterial Community Function

12:30 Structural Genomics (SG) & Beginning a Research Dialogue with the Metagenomics Community	Ian Wilson, The Scripps Research Institute & Stephen Burley, Eli Lilly & Co.
01:20 Protein Target Selection Challenges and Charge to Discussion Groups	Adam Godzik, Burnham Institute
01:40 Break	
02:00 Discussion Sessions: Role of Structure in Ascertaining Functional Properties of Microbial Communities: How Can These Two Research Domains Interact Most Productively?	
03:15 Break	
03:30 Summary of Subgroup Discussions	
04:00 Closing Remarks	John Wooley, UC San Diego

P
R
O
G
R
A
M
:
D
A
Y
5

MONDAY, NOVEMBER 3

SESSION I: Genomic and Metagenomic Metadata

08:30 **Setting the Stage: Mobilizing the Metagenomics Community**

John Wooley, UC San Diego

08:45 **CAMERA: Interest in Community Standards for the Capture and Exchange of Metadata**

Paul Gilna, CAMERA/UC San Diego

09:00 **Overview of the GSC and the Minimum Information About a (Meta) Genome Sequence (MIGS/MIMS) Specification**

Dawn Field, Oxford Centre for Ecology and Hydrology

09:15 **The Rapidly Growing Standards Landscape in 'omics**

Susanna Sansone, European Bioinformatics Institute

As the size and complexity of scientific datasets and the corresponding information stores grow, reporting standards are playing an increasingly active role. Interoperability among standards, however, becomes pivotal for the development of software applications. Here we present the key synergistic standards activities in the omics domain and the motivation for, an overview of, the BioInvestigation Index infrastructure at the EBI.

The marriage of conventional methods with (meta)genomics, transcriptomics, proteomics and metabol/nomics technologies (hereafter referred as 'omics') has created not only opportunities, but also substantial new informatics challenges. For example, consider the reporting of a complex multi-omics study looking at the effect on a population of worms, investigating the effect of heavy metals by measuring gene and protein expression whole organism (by mass spectrometry and DNA microarrays, respectively), sequencing the genome (by highthroughput) and conducting a series of conventional environmental analysis. It is pivotal that such datasets are reported in a standard manner to enable communication, interpretation and analysis. New approaches are required for describing, formatting, submitting and exchanging both data and metadata (i.e., sample characteristics, study design and execution) from such complex studies.

Many groups are rising to this challenge to this end, including the Genomics Standards Consortium (GSC) [1]. However, standards for data content (minimal information checklists), semantics (ontologies) and syntax (file formats) are being specifically developed to target a particular omics technology or a particular biologically-delineated community. Unfortunately, remaining bounded by a particular discipline, standardisation efforts in gen-

eral remain fragmented and cannot be easily integrated. This result in unnecessary duplication of effort, and more significantly, the development of (arbitrarily) different standards being developed, thereby limiting the scope for data exchange. The result of such 'fragmentation' is also reflected in the implementations. For example, systems such as ArrayExpress and Pride at EBI -built to store microarray-based and proteomics experiments, respectively- employ different submission/exchange formats and terminologies as developed by the standardisation initiatives in their domain. In such scenario, description and submission of multi-omics studies will be difficult if not impossible.

Fortunately, several synergistic activities have begun fostering the harmonization and consolidation of the three kinds of standards being developed. Over 20 projects are registered in the 'Minimum Information about a Bio-medical or Biological' (MIBBI) portal [2,3] set to created orthogonal checklist modules. At present, over 60 groups participate under the OBO Foundry umbrella [4,5] with the objective of developing interoperable ontologies. Several groups participate in the Functional Genomics (FuGE) project [6,7] which underpins the XML-based formats they have developed. Only very recently, another complementary initiative has sprung up from a growing number of communities that work collaboratively on a common tabular framework for presenting the experimental metadata (ISA-TAB) [8,9]. The reuse of common standards and ontologies will ease the task of software developers, vendors, and equipment manufacturers by reducing time and costs for implementing standards-compliant products. In turn, these will be valuable interoperable resources for the system biology community, simplifying the job of data integration. Members of the GSC are actively involved in OBO, MIBBI and the ISA-TAB efforts.

Undoubtedly, the interoperability of reporting standards will ease the task of those developers working to implement standards-compliant systems for complex multi-omic studies, such as the BioInvestigation Index at the EBI [10]. This infrastructure aims to create a common structured representation of the metadata and the sample-data relationship for biological, biomedical and environmental studies employing omics-based technologies along with more conventional methodologies. The BioInvestigation Index infrastructure - along with a first set of publicly available multi-omics datasets- will be launched in Dec 2008.

1. <http://gensc.org>
2. <http://mibbi.sf.net>
3. Taylor, Field, Sansone,....Rocca-Serra,.... et al. (2008) Nat Biotechnol 26(8):889-96.
4. <http://www.obofoundry.org>
5. Smith, Ashburner, Rosse,....Rocca-Serra, Sansone et al. (2007). Nat Biotechnol. 25(11):1251-5.

6. <http://fuge.sf.org>
7. Jones, Miller, Aebersold,...Sansone,...Taylor et al. (2007). *Nat Biotechnol.* 25(10):1127-33.
8. <http://isa-tab.sf.net>
9. Sansone, Rocca-Serra, Brandizi,...Sklyar, Taylor et al. (2008) *OMICS.* 2(2):143-9.
10. <http://www.ebi.ac.uk/bioinvindex>

09:45 **Source Organism and Molecule Information at INSDC and the Trace Archives**

Guy Cochrane, European Nucleotide Archive, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus

The databases of the International Nucleotide Sequence Database Collaboration (INSDC, DDBJ/EMBL/GenBank) and of the Trace and Short Read Archive collaborations between the NCBI and the EBI provide free and comprehensive access to nucleotide sequencing information, from raw sequencing machine output through to functional annotation. As well as offering a direct public portal into nucleotide sequence and annotation, the information embodied in the archives serves, through such projects as UniProt and Ensembl, as foundation for the world's bioinformatics data infrastructure.

Amongst the challenges of ever growing volumes of information is the need to provide sensible data organisation to allow users the simple retrieval of, and computation upon, small and large data sets of interest in the main corpus. Key to this organisation is the systematic capture and structuring of information relating to the biological source organism and molecule that have undergone sequence analysis. While long-established data structures exist for the representation of the more generic elements of this source information, recent community-focused initiatives have provided a number of alternative routes and structures for more specific information.

In the talk, I will outline the services provided by the INSDC and Trace Archives, detail a number of paradigms for the representation of source organism and molecule information and will focus on the emerging strategy for incorporation of MIGS compliance data into INSDC records.

10-10:30 Break

SESSION II: Core GSC Projects

10:30 **Implementing MIGS/MIMS: The Genomic Contextual Data Markup Language (GCDML)**

Renzo Kottmann, Max Planck Institute for Marine Microbiology
Co-authors: Melissa Beth Duhaime, Frank Oliver Glueckner

The MIGS/MIMS checklist developed by the Genomic Standards Consortium (GSC) aims to enrich our ever growing data collection of genomic and metagenomic sequences. Therefore, the checklist defines what minimal list of contextual data attributes has to be added to the sequences in order to conform to the MIGS/MIMS specification. However, MIGS/MIMS itself does not specify how the contextual data should be documented; thus the GSC is developing the Genomic Contextual Data Markup Language (GCDML), which provides the official implementation of MIGS/MIMS. Using XML Schema, GCDML implements a large set of strongly typed contextual meta-data descriptions and so called MIGS-Reports in XML. These reports define in detail how the contextual data is to be documented. The validity and conformance to the MIGS/MIMS specification can then be verified with standard XML parsers. After one year of development and several iterations of refinements, the structure of MIGS-Reports is stabilized and has proven to be easily human-editable. However, the main aim of GCDML is to provide a machine readable representation of contextual data that facilitates the capture, exchange, and comparison of large amount of data in Web Service environments.

More information can be found on the GSC WIKI page: http://gensc.org/gc_wiki/index.php/GCDML

Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* 26:541-547.

Renzo Kottmann, Tanya Gray, Sean Murphy, Leonid Kagan, Saul Kravitz, Thierry Lombardot,

Dawn Field, Frank Oliver Glockner, Genomic Standards Consortium. 2008. A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS.* 12(2)

10:45 **StrainInfo and the Linkage of Organisms, Gene and Genomes: The Genomic Rosetta Stone**
Peter Dawyndt, University of Ghent

11:00 **Towards Computer Assisted Markup of Data: Habitat-Lite**

Lynette Hirschman, The MITRE Corporation
Co-authors: Scott Mardis, Cheryl Clark, Kevin Cohen

Habitat-Lite (aka EnvO-Lite) is a set of structured terms designed to facilitate capture of high-level information about habitat and sample source metadata for genomics and metagenomics samples. It is designed to be light-weight and compact. Habitat-Lite terms are

drawn from the full EnvO ontology and are made available by the EnvO Consortium in the EnvO-Lite-GSC OBO file.

Habitat-Lite terms include high level terms (e.g., terrestrial, marine, freshwater, air, organism-associated) and more specific terms (e.g., soil, sediment, hot spring). These terms were initially selected by Dawn Field (NERC, Oxford), in consultation with domain experts. The terms have been evaluated for coverage and ease of use in capturing relevant information from Genbank isolation_source entries, and GOLD HABITAT and ISOLATION entries. We have recently implemented a tool to map automatically from free text phrases into Habitat-Lite terms; we estimate that this tool is capturing 70% of the high level terms correctly, based on comparison to expert manual annotation done by Renzo Kottmann and Pier Buttigieg from the Max Plank Institute, Bremen. Our next steps are to elicit additional use cases from the genomics and metagenomics communities, to develop guidelines for adding terms to Habitat-Lite to capture distinctions critical to these use cases, and to enhance the mapping tool based on these inputs.

*This work has been funded under NSF Small Grant for Exploratory Research IIS-0746650.

11:15 **Towards a Standards Compliant Literature: The GSC eJournal**

George Garrity, Michigan State University

Standards in Genomic Sciences (SIGS) is a new Open Access publication that is being created in cooperation with the Genomic Standards Consortium (GSC). It is intended to fill an unmet need for rapid publication of standards compliant reports on genomes and metagenomes, standard operating procedures relevant to large-scale sequencing initiatives, and various forms of technical reports, policy statements, meeting reports and other content that is pertinent to this rapidly evolving field. In keeping with the spirit of Open Access electronic-only publications, SIGS will be distributed to readers at no cost, with publication costs initially being absorbed through grants from the Michigan State University Foundation and the US Department of Energy. Where the GSC and other standards initiatives have sought to develop interoperable approaches to ensure consistency in semantic and syntactic annotation of genomes and metagenomes, SIGS will apply and extend these standards to published content so as to more tightly integrate the literature with the underlying data. SIGS will also provide a vehicle for rapidly publishing concise, highly structured reports of sequenced genomes and metagenomes so that readers can readily make comparisons across time and taxa in a single place that tightly integrates with past, present and future knowledge existing elsewhere in the literature. The



purpose of this presentation will be to update the community on progress in launching the journal and to solicit feedback.

11:30 **Towards Transparency of Computational Analyses:**

A Central SOP Repository

Owen White, University of Maryland

The methodologies used to generate genome and metagenome annotations are diverse and vary between groups and laboratories. Descriptions of the annotation process are helpful in interpreting genome annotation data. Some groups have produced Standard Operating Procedures (SOPs) that describe the annotation process, but standards are lacking for structure and content of these descriptions. A general structure for SOP documents that are relevant to genome annotation, metagenomics, environmental samples and clinical samples will be presented. In addition, a proposal for a central repository to store and disseminate procedures and protocols for genome annotation will be discussed.

SESSION III: Defining the Scope of GCDML

01:00 **The Importance of Context for the Design and Interpretation of Comparative Metagenomics Studies: The MINIMESS Proposal**

Jeroen Raes, European Molecular Biology Laboratory

Comparative metagenomics is a powerful tool to detect environment-specific adaptation of the communities that inhabit them. However, a wide range of technical and biological interfering factors hamper the correct interpretation of comparative metagenomics results. Here, I will describe some of these factors and pitfalls and propose a minimal metagenome sequence analysis standard (MINIMESS) to cope with these issues. This suggested additional, complimentary layer of reporting provides a standardized description of the metagenome and its inferred community properties and thus facilitates use of these data by the scientific community.

References: Raes et al., *Curr Opin Microbiol* 2007; Raes & Bork, *Nat Rev Microbiol* 2008.

01:15 **Should Genomics Contextual Data Fly Coach?**

Inigo San Gil, University of New Mexico

The Long Term Ecological Research Network (LTER) creates long-term ecological data records in twenty six sites across the US. Five years ago the LTER adopted a metadata specification called the Ecological Metadata Language (EML) to facilitate network data integration and synthesis. A growing number of LTER scientists are conducting genomics research, thus

the need to adopt an integrated genomics and ecological data management system is urgent. Recently, the LTER joined the Genomics Standard Consortium to collaborate in developing Minimum Information about a Genome/ Metagenome Sequence (MIGS/MIMS). In this talk, we discuss some of the differences between the data documenting process of the LTER and GSC and the possible mechanisms to bridge these differences.

01:30 **Extending MIGS/MIMS to the Description of Ribosomal RNA Sequences**

Frank Oliver Gloeckner, Max Planck Institute for Marine Microbiology

Co-authors: Wolfgang Hankeln, Renzo Kottmann, Joerg Peplies

With the MIGS/MIMS specifications the Genomic Standards Consortium has finished the groundwork to enrich our genome and metagenome collections with contextual data. Now, it is time to consider whether these standards could be applied 'as is' in the short-term, and 'with modification/extension' in the longer-term to any genetic marker sequence retrieved from the environment. To move forward and leverage existing interest in the community the proposal for MIENS, the 'Minimum Information about an ENvironmental Sequence', has been accepted as a natural extension to MIGS and MIMS on the 6th GSC meeting in October 2008.

MIENS is meant to be fully compliant with the attributes already accepted by the MIGS/MIMS standards, but it is also intended to identify additional contextual data fields that are needed to enrich our ever growing set of environmental marker gene sequences.

It was decided, to take the ribosomal RNA sequence collections (16S/18S & 23S/28S) as a first use case. By supplementing our sequence collections with more contextual data, it will be possible to retrieve, for example, all 16S sequences related to specific environmental parameters (i.e. location, habitat type, temperature, salinity, oxygen concentration etc.).

A MIENS working group has been established to tackle the following key aspects:

1. Identify which INSDC/MIGS/MIMS contextual data attributes for environmental sequences are most relevant for the community
2. Identification of additional contextual data to be covered by MIENS. This leads to generating a checklist for MIENS indicating the significance of the attributes.
3. Implementation of MIENS: Definition of field names and description of fields and extension of GCDML

4. Collaborate with INSDC to define the modus of sequence & contextual data submission

5. Provide tools for effective sequence & contextual data submission

More information can be found on the GSC WIKI page: http://gensc.org/gc_wiki/index.php/MIENS

Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* 26:541-547.

SESSION IV: Ontologies and the Description of Habitat and Geolocation

02:00 **The Environment Ontology: Linking Environmental Data**

Norman Morrison, University of Manchester

Every biological specimen that is collected or sampled — whether for a museum collection, for epidemiological studies, for population studies, for ecological research, for research into evolution, biodiversity or sustainability, indeed any biological research — comes from a particular habitat where particular physical conditions prevail. However, at present there is no accepted semantic standard for describing the environment from which these biological samples are collected. This is a serious problem for anyone wishing to retrieve and compare environmental data.

The Environment Ontology (EnvO — www.environment-ontology.org) provides an integrated approach to the problem of linking environmental data. The aims of EnvO are to support the semantically consistent description of, and computational reasoning over, environmental information associated with biological data of any organism or biological sample.

In this talk I shall attempt to show how EnvO will help people build services in which you'll be able to run queries like: "retrieve all metagenomic data from samples taken from deep sea thermal vents". Furthermore, at the click of a button, you'll also be able to automatically 'relax' your search space, enabling comparisons against datasets from other marine habitats, such as coral reef atoll or oceanic trench.

02:15 **Towards an Open Access Gazetteer**

Lynn Schriml, University of Maryland

The Gazetteer (Gaz) is a community-based project of the EnvO Consortium for describing instances of organism environments and biological samples supporting consistent annotation of locations and environments. The Gazet-

teer describes places and place names, and the relations between them. The Gaz, with 130,000 plus terms (over 50 % of terms defined) and relations, is being utilized in scientific investigations such as applying Gaz terms to organisms based on free-text from the Encyclopedia of Life dataset for the OntoSpecies project. Additionally, the Gaz is being employed as a controlled vocabulary for the annotation of outbreak and sample collection data of viral and bacterial pathogens in a growing number of resources.

02:30 **RDP Survey of Habitats Descriptors**

James Cole, University of Connecticut

SESSION V: Metadata Capture: A Key Step for Advancing Understanding

03:30 **The Genomes Online Database (GOLD): The Value of a Comprehensive Metadata Collection**

Nikos Kyrpides, Joint Genome Institute

03:45 **Metadata Capture in the IMG/IMGm: Getting Scientists to Contribute**

Victor Markowitz, Lawrence Berkeley National Lab

04:00 **The Genome Catalogue: A Future Vision**

Lynn Schriml, University of Maryland

TUESDAY, NOVEMBER 4

SESSION I: Challenges in Metagenomics for Bioinformatics and Computational Biology

09:00 Introduction

John Wooley, UC San Diego

09:20 Components for rational management of genomic and metagenomic information

Owen White, Univ. of Maryland

10:00 Talk TBA

Adam Godzik, Burnham Institute of Medical Research and UC San Diego

10:40 Break

11:00 Talk TBA

Daniel Huson, Tübingen University

11:40 Efficient Metagenomics Data Processing: Pitfalls and Solutions

Nikos Kyrpides, Joint Genome Institute

Co-authors: Kostas Mavrommatis, Natalia Ivanova

Metagenomics has emerged as a powerful tool for exploration of the functional capabilities of microbial communities regardless of the ability of their members to grow in pure culture. However, little is known for the efficacy of the methods used to process these datasets. Furthermore, sequencing of environmental genomic DNA usually results in large, highly fragmented datasets representing a significant challenge for downstream functional analysis and interpretation due to insufficient sequence coverage preventing assembly of individual reads, higher sequence error rate (frameshifts) resulting in disruption of protein-coding sequences, and unsatisfactory performance of assembly and gene prediction tools. In addition, many metagenomic datasets derived from the communities with one or a few dominant genera exhibit high degree of redundancy due to the presence of very similar, but not identical sequences originating from different strains and species of the same genus. As a result, many of the analyses routinely performed on finished and draft isolate genomes, such as identification of secreted and multi-domain proteins in metagenomic datasets or reconstruction of phylogenetic trees becomes very labor-intensive, if not impossible.

We will discuss the evaluation of methods used for the analysis of metagenomic datasets, as well as a method for post-processing of metagenomic datasets aiming to correcting many of the above errors and reducing the redundancy of metagenomic datasets.

12:20 Challenges in Ecological Metagenomics

Pat Schloss, Department of Microbiology, University of Massachusetts

Since the first metagenomics study was published 17 years ago, the ability to make ecological inferences has been limited by sequencing throughput and the ability to analyze the data. Yet, if the potential of metagenomics to help us decipher the structure and function of microbial ecosystems is to be realized, it is essential that our bioinformatics tools motivate our sequencing efforts. This is analogous to the bioinformatics advances that were made to accelerating the sequencing of the human genome, except that metagenomic projects face perhaps even greater obstacles. Unfortunately, the bioinformatics capabilities for analyzing metagenomic sequencing projects lag our sequencing efforts. Our research group has begun to develop a suite of statistical tools to describe and compare microbial communities at the genomic level using DNA sequence reads obtained via the traditional Sanger sequencing method. These tools have revealed the dominance of protein families with no known functions and the amount of functional overlap between disparate communities. To keep up with our sequencing capacity it is essential that we continue to develop tools that will enable us to analyze genomic and transcript sequences generated from pyrosequencing technologies, proteomics, and microarrays. The end goal is to provide a suite of bioinformatics tools so that we are limited by our sequencing throughput and not our ability to analyze the data.

01:00 Lunch

SESSION II: Documenting the Emergence and Opportunities of the Field: Parallel Discussion Sessions

02:00 Parallel 1: Connecting Metagenomics and Metaproteomics

Chair: Janet Jansson, Lawrence Berkeley National Lab

Current metagenomics sequencing efforts provide enormous amounts of data about the gene composition of microbial communities in different ecosystems, ranging from the ocean to the human gut. This data provides us with unprecedented amounts of information about potential functions of microorganisms in their respective habitats. However, not all genes are expressed under every condition. Also, in metagenomic sequencing efforts the DNA can originate from cells in different physiological states including dormant, inactive states and even dead cells. Therefore, to more accurately determine what genes are functioning, i.e. expressed, in a microbial community, the translated protein products of expressed genes should be directly identified. Recent technological developments

TUESDAY, NOVEMBER 4

T
U
E
S
D
A
Y

A
B
S
T
R
A
C
T
S

in shotgun proteomics approaches enable us to directly identify thousands of expressed proteins by searching genomic and metagenomic sequence datasets. The optimal scenario would be to couple metagenomics and metaproteomics efforts to provide matched datasets for assigning protein identifications to sequence data from the same sample origin, such as demonstrated for the acid mine drainage system, but not yet for more complex communities. Ongoing challenges, technique developments and future goals of combined metagenomics and metaproteomics will be discussed in this session.

02:00 Parallel 2: **New Sequence Technologies for Metagenomic Research**

Chair: Tim Hunkapiller, Discovery Biosciences and Applied Biosystems

02:00 Parallel 3: **The Metagenomic Voyage: In Situ or In Silico**

Chair: Eric Allen, UC San Diego

Metagenomic approaches have the potential to offer unprecedented insight into the form and function of heterogeneous natural communities - microbial, viral, or eukaryotic. However, prior to analyzing the fruits of metagenomic labor - the sequences!- it is necessary to frame the physico-chemical context within which the data is to be interpreted. The initial step in any metagenomic venture requires the collection of samples destined for analysis. The in situ phase of this journey begins with characterizing the sample from multiple vantage points. Ultimately, these parameters constitute the “meta data” intricately tied to a given sample and aid in interpreting the biological significance of the genetic information. The second phase segues from the “field” to the laboratory where ex vivo processing occurs - nucleic acid extraction, library construction, and DNA sequencing. Ultimately, the biological data is made interpretable via in silico analysis where computational power and ingenuity take center stage. The goal of this session is to discuss best practices for initiating and managing metagenomic projects. Our voyage will hit upon the three key phases in the process and identify the needs and approaches best suited for successful metagenomic explorations.

02:00 Parallel 4: **Wellness and Diseases: Implications of Important Microbiota**

Chair: Justin L. Sonnenberg, Stanford University

The microbiota in personalized preventative medicine: how do we get from here to there and what role does metagenomics play? Synopsis—In this session we will (i) identify major questions and hurdles we face in understanding the integration between the human microbiota and human biology (in health and disease) and (ii) determine how these major questions may be addressed

and what tools/technologies are required to overcome current obstacles. We will attempt to identify immediate and long-term goals for addressing these questions, with a particular focus on the role of metagenomics. Finally, we will consider the microbiota in personalized medicine and the potential contribution of metagenomics in facilitating this broader view of human health.

02:00 Parallel 5: **Soil Options**

Chair: Patrick Schloss, University of Massachusetts, Amherst

Metagenomic approaches have the potential to offer unprecedented insight into the form and function of heterogeneous natural communities - microbial, viral, or eukaryotic. However, prior to analyzing the fruits of metagenomic labor – the sequences – it is necessary to frame the physico-chemical context within which the data is to be interpreted. The initial step in any metagenomic venture requires the collection of samples destined for analysis. The in situ phase of this journey begins with characterizing the sample from multiple vantage points. Ultimately, these parameters constitute the “meta data” intricately tied to a given sample and aid in interpreting the biological significance of the genetic information. The second phase segues from the “field” to the laboratory where ex vivo processing occurs - nucleic acid extraction, library construction, and DNA sequencing. Ultimately, the biological data is made interpretable via in silico analysis where computational power and ingenuity take center stage. The goal of this session is to discuss best practices for initiating and managing metagenomic projects. Our voyage will hit upon the three key phases in the process and identify the needs and approaches best suited for successful metagenomic explorations.

04:00 GSC Breakouts

[Details to be announced; to include GCDML and Genome Rosetta Stone]

05:00 – 08:30 Welcome Reception (La Jolla Shores Hotel)

05:30 **Welcome Talk**

Mark Ellisman, UC San Diego

05:45 – 07:00 Dinner

07:00 Congress Keynote:

The Evolution of Small Bacterial Genomes in the Ultra-Oligotrophic Ocean

Stephen J. Giovannoni, Oregon State University
Co-authors: Michael Schwalbach, H. James Tripp and Joshua B. Kitner

The size and coding capacity of a genome are its most fundamental properties. What selective pressures deter-

mine the expansion and contraction of genomes, and how is genome size related to adaptive strategies? Very small marine microbial genomes are providing new insights into these questions. The first reports of genome sequences from the cyanobacterium *Prochlorococcus* and the α -proteobacterium SAR11 (*Pelagibacter*) established that these very abundant marine bacterioplankton clades have unusually small genomes. The genome of *Pelagibacter* (1.31 Mbp) is the smallest reported genome for a free-living heterotrophic cell. *Prochlorococcus* genomes, which range in size from 1.66 to 2.41 Mbp, are the smallest cyanobacterial genomes reported. Recently, the complete genome sequence of HTCC2181, an obligate methylophilic bacterium that is the first cultured member of the OM43 clade of marine bacterioplankton, was determined to be even smaller - 1.30 Mbp. These genomes are the smallest known from free-living cells, but larger than the genomes of some symbionts and parasites that live in close association with eukaryotic hosts. Genome streamlining has been invoked to explain the small genomes of some marine bacterioplankton. The essence of this theory is that selection is most efficient in microbial populations with large effective population sizes, causing the elimination of unnecessary DNA from genomes. Bacterioplankton may be particularly subject to streamlining selection and genome reduction because: 1) they have very large population sizes, 2) they live in a habitat that is frequently limited for the macronutrients N and P, which are stoichiometrically high in nucleic acids, and 3) selection favors high surface to volume ratios in the ocean surface, an adaptation that allows cells to compete effectively for nutrients.

Recently, extraordinary examples of genome reduction in SAR11 have emerged. These cells are deficient in assimilatory sulphate reduction genes, making them dependent on exogenous sources of reduced sulphur, such as 3-dimethylsulphoniopropionate (DMS) or methionine, for growth. We also found that SAR11 cells are glycine auxotrophs that use a unique and elegantly simple glycine-activated riboswitch on malate synthase to control the assimilation of carbon through the TCA cycle into biomass. Studies of ultrastructure and the metaproteomics of cells from an extremely oligotrophic gyre show that these cells have high surface-to-volume ratios and very high ratios of transport proteins, an apparent adaptation to enable efficient replication in ocean "deserts". These observations support the broad conclusion that metabolic versatility has been sacrificed for simplicity and genome reduction in some bacterioplankton, rendering them able to use ambient nutrient resources efficiently but reducing their versatility. The question remains, how does the evolutionary history and ecology of these organisms differ from microbial plankton with genomes of average size?

WEDNESDAY, NOVEMBER 5

08:30 Morning Talk:

The Importance of Marine Picoeukaryotes and the Search for Lost Time

Alexandra Z. Worden, Monterey Bay Aquarium Research Institute (MBARI)

Unicellular eukaryotes are responsible for a massive amount of photosynthetic carbon fixation in marine systems. The smallest among these fall within the "pico" size fraction (<2 micrometers in diameter), are broadly distributed – from coastal to open-ocean environments – and are highly diverse. Picoeukaryotes contribute a significant proportion of the biomass and primary production within this size fraction, often rivaling their cyanobacterial counterparts *Prochlorococcus* and *Synechococcus*. Despite the importance of these eukaryotic phytoplankton, research on their distributions and genetic capabilities has consistently lagged behind that on their bacterial counterparts. This lag persists even now due to their larger genome sizes, tremendous diversity and difficulties in isolating and culturing many of the taxa.

Micromonas pusilla is one of the most widespread picoeukaryotic species. This photosynthetic alga thrives from tropical to polar marine ecosystems and belongs to an anciently diverged clade (prasinophytes) sister to land-plants. We sequenced complete genomes of two strains within this purported 'species,' and found far greater genome variability than anticipated based on their high 18S rRNA gene identities. The *Micromonas* strains have striking differences in terms of gene content, invasion elements and aspects of gene regulation. Genes in each that are mutually exclusive, i.e. 'niche defining genes', have phylogenetic profiles that are dissimilar from 'core' genes, pointing to different selection and acquisition processes. Data from field expeditions show that in addition to prasinophytes, several uncultured lineages are also important components of the photosynthetic picoeukaryotic community. Collectively, these data highlight challenges for metagenomic analyses and reshape how we approach picoeukaryotic populations in situ. Our findings underscore the potential for refined metagenomic approaches to facilitate development of testable hypotheses on population controls of these important primary producers.

09:30 Break

SESSION I: Marine Metagenomics

Chair: Stephen J. Giovannoni, Oregon State University

09:40 Photosystem-I Gene Cassettes in Marine Phages

Oded Beja, Technion-Israel Institute of Technology

10:20 Deep-Ocean Metagenomics: Comparative Investigations of Microbes Inhabiting Hydrothermal Vents and the Cold Deep Ocean

Shannon Williamson, J. Craig Venter Institute

The deep-ocean accounts for the majority of the water on our planet. While the cold deep-ocean is the dominant nutrient-challenged ecosystem, hydrothermal vents represent oases along tectonically active areas of the seafloor. Hydrothermal vents and the surrounding cold deep-ocean each support unique communities of microbes and viruses that have become highly adapted to local conditions. Through cycles of infection and host cell lysis, bacteriophages (viruses that specifically infect prokaryotes) have the potential to substantially influence the adaptation and population biology of their hosts. In order to investigate the community dynamics of microbes and viruses inhabiting two very different ecosystems, water samples were collected from a deep-ocean diffuse-flow hydrothermal vent and the surrounding cold bottom waters. Microbial communities were size-fractionated, in situ, by serial filtration onto membrane filters and viruses were concentrated via tangential flow filtration under ambient temperature and pressure. Metagenomic analysis of the organisms captured within the 0.1 μ m-0.8 μ m size fraction indicates that hydrothermal vent and cold deep-ocean ecosystems harbor taxonomically and functionally unique microbial communities; demonstrated by site-specific protein clusters and dominant microbial taxa. Preliminary analysis of sister viral metagenomic libraries indicate that viruses endemic to the deep-ocean contain a high frequency of novel genes and may play an important role in the ad-

aptation of their microbial hosts to the very different, but equally challenging environmental conditions that characterize deep-ocean environments. Together, these observations suggest that microbial diversity and function in the deep-ocean is significantly influenced by the viruses that parasitize them.

11:00 Break

11:10 Metagenomic Analysis of Deep Subsurface Environments

Hideto Takami, Extremobiosphere Research Center, JAMSTEC

As a deep subsurface biosphere is thought to be the biggest biosphere in the earth, it is very interesting to know the phylogenetic and functional diversity in such environments. However, there is a little biological information for them because it is very hard to recover whole microbial community by only culture-base methods. In that sense, metagenomics is one of major useful methods to elucidate the microbial flora in unknown biosphere, which seems to be constructed by mainly unculturable microbes.

We used eight kinds of sediment core sample, which are 9 cm in diameter and 20 cm in length for metagenomic analysis from among all samples recovered by the ocean drilling to 365 meter below seafloor (mbsf) of the D/V Chikyu shakedown cruise in an offing of the Shimokita Peninsula on the northeast Honshu, Japan. DNA isolation directly from each sample was carried out by means of an improved beads beating method with a combination of beads beating and lytic enzymatic treatments. The amount of DNA isolated from the core samples decreased with an increasing of depth from seafloor and only several nanogram of DNA per 5g-core sample was isolated from the deepest layer (348 mbsf) although the total amount of DNA was enough for PCR amplification of 16S rRNA gene. However, since the amount of DNA was too small for construction of metagenomic libraries in some core samples, we attempted to amplify genomic DNA using GenomiPhi. Two thousands archaeal and 1000 bacterial 16S rDNA clones from each sample were sequenced and analyzed phylogenetically to figure out the vertical profile of prokaryotic community in the sediment from the seafloor to 348 mbsf. The patterns of archaeal and bacterial 16S rDNA varied depending on the layer depth and the majority of 16S rDNA were categorized into subsurface group 2 & 3 and α -proteobacteria in the deepest layer, respectively. We constructed the metagenomic shotgun libraries in 5 selected core samples ranging from 0.7 to 107 mbsf and have sequenced both ends of 40,000 clones of each library to investigate a feature of the functional genes in each layer depth.

11:50 **Integrative Marine Metagenomics**

Elizabeth A. Dinsdale, San Diego State University/University of Adelaide

Metagenomics has enabled an unprecedented assessment of the activity of microbial communities within ecosystems. A comparison of almost 15 million sequences from 45 distinct microbiomes and 42 distinct viromes showed that there are strongly discriminatory metabolic profiles across environments. Most of the functional diversity was maintained in all of the communities, but the relative occurrence of metabolisms varied, and the differences between metagenomes predicted the biogeochemical conditions of each environment. The analysis identified when microbial communities within an environment were perturbed and had changed functions. To identify how a perturbed microbial community affect an ecosystem, metagenomic analysis was integrated into the assessments of four coral communities in the central Pacific. The microbes vary dramatically depending on the state of the coral and fish assemblages. On coral reefs with low levels of human activity, in particular no fishing, microbial communities were small and had community structure similar to open oceans. As fishing increased the numbers of microbes increased and functionally the community became more pathogenic. The microbialization was correlated with an increase in unhealthy corals. Both the functional and coral reef analyses identified that viruses are playing a greater role in the function of the ecosystem than anticipated. The motility genes encoded by the viromes were investigated to identify individual viruses carrying the genes. A dinucleotide analysis characterized the signature of the motility genes and compared it to both the microbial and viral signature. The signature suggested viral origins for these genes and that the viruses serve as a repository for storing and sharing genes among their microbial hosts thereby influencing global evolutionary and metabolic processes.

12:30 Lunch

SESSION II: New Technologies in Metagenomics

Chair: Paul Gilna, CAMERA/UC San Diego

01:30 **Analyzing the Mobilome Using Metagenomics**

Julian R Marchesi, Cardiff University

Metagenomic approaches are being used to investigate many diverse ecosystems and reveal many novel functions therein. However these methods can frequently pass over the DNA which we find in the mobile metagenome or mobilome. Much of this DNA is not easily cloned into fosmids or BACs for functional analysis and DNA based metagenomics projects do not always provide sufficient information to determine from where the sequence arose, genome or mobilome. Couple these limitations to

the fact that the current approaches for isolating mobile genetic elements (MGEs), especially plasmids, are either biased to the organisms which are able to grow on laboratory media (endogenous isolation) or by host used to isolate the plasmid (exogenous isolation), it is not surprising that we have a poor appreciation of the diversity of the MGEs in many ecosystems. With this in mind we set out to develop an alternative approach to isolate plasmids from an ecosystem, in this case the human gut, which circumvented the limitations of current methods. Our approach, Transposon Aided Capture (TRACA), was successfully used to isolate plasmids from the distal portion of the human large intestine and we will discuss the method and some of the TRACA plasmids in this presentation.

02:10 **Electromicrobiology: Novel approaches for investigating charge transfer and energy transformation in microbial systems**

Yuri Gorby, J. Craig Venter Institute

Respiratory microorganisms capture energy for growth and maintenance as they transfer electrons from energy sources to appropriate electron acceptors. Controlling the availability of electron donor and acceptor pairs provides opportunities for investigating fundamental aspects of energy transformation and distribution in defined and undefined microbial cultures. For example, we have noted that in all bacteria investigated to date electron acceptor limitation prompts the production of electrically conductive appendages known as bacterial nanowires. These structures apparently serve to connect cells with electron acceptors at distances from tens or hundreds of microns. We have developed novel approaches and cultivation devices to investigate charge transfer in diverse microbial systems with electronic precision. This presentation updates our progress on characterizing the composition and electrical conductive properties of bacterial nanowires in model laboratory microorganisms and how metagenomics will enable a more expansive exploration of nanowires from natural complex microbial communities.

02:50 **Accessing the Metatranscriptome for Complex Marine Microbial Communities**

Jack Gilbert, Plymouth Marine Laboratory

Co-authors: Dawn Field, Weizhong Li, Ying Huang, Rob Edwards, Paul Gilna, Ian Joint

Sequencing the expressed genetic information of an ecosystem (metatranscriptome) can provide information about the response of organisms to varying environmental conditions. Until recently, metatranscriptomics has been limited to microarray technology and random cloning methodologies. The application of high-throughput sequencing technology is now enabling access to both known and previously unknown transcripts in natural communities. We present a study of a complex marine meta-

transcriptome obtained from random whole-community mRNA using the GS-FLX Pyrosequencing technology. Eight samples, four DNA and four mRNA, were processed from two time points in a controlled coastal ocean mesocosm study (Bergen, Norway) involving an induced phytoplankton bloom producing a total of 323,161,989 base pairs. Our study confirms the finding of the first published metatranscriptomic studies of marine and soil environments that metatranscriptomics targets highly expressed sequences which are frequently novel. Our alternative methodology increases the range of experimental options available for conducting such studies and is characterized by an exceptional enrichment of mRNA (99.92%) versus ribosomal RNA. Analysis of corresponding metagenomes confirms much higher levels of assembly in the metatranscriptomic samples and a far higher yield of large gene families with >100 members, ~91% of which were novel. This study provides further evidence that metatranscriptomic studies of natural microbial communities are not only feasible, but when paired with metagenomic data sets, offer an unprecedented opportunity to explore both structure and function of microbial communities – if we can overcome the challenges of elucidating the functions of so many never-seen-before gene families.

03:30 Break

03:40 **Genome Standards Consortium Report**
Dawn Field, Oxford Centre for Ecology and Hydrology

SESSION III: Selected Talks by Abstract Submission

04:00 **Integrated Information System for Genomic and Metagenomic Data Analysis at National Center for Biotechnology Information (NCBI)**

Anjanette Johnston, NCBI, National Library of Medicine, NIH

Co-authors: Azat Badretdin, Stacy Ciufo, Boris Fedorov, Yuri Kapustin, William Klimke, Rich McVeigh, Kathleen O'Neill, Martin Shumway, Leonid Zaslavsky, Tatiana Tatusova, Ilene Mizrahi

Recent advances in biotechnology and bioinformatics has provided a flood of genomic data and tremendous growth in the number of associated data sets. Sequencing projects now include draft assemblies, complete genomes, comparative genomics, and metagenomics where genetic material is sequenced directly from environmental samples.

The NCBI provides integrated systems for data storage, retrieval, and analysis. GenBank, an archival database of DNA sequences, contains consensus sequences assembled from raw sequence reads. The Trace Archive serves as a repository of raw sequence data from a vari-

ety of automated sequencing platforms. In addition, NCBI provides reference sequence collections and specialized tools for sequence analysis and visualization. A novel approach recently developed at NCBI allows the visualization of large phylogenetic trees in an aggregated form with a special representation of subscale details.

Rapid advances in sequencing technologies have created new challenges for information systems. The new NCBI resource, Short Read Archive (SRA) has been designed specifically to handle sequence data from massively parallel sequencing technologies. Specialized metagenomic resources at NCBI include a collection of environmental projects in Entrez Genome Project database, and specialized BLAST databases including Environmental Samples, Whole Genome Shotgun Reads and Trace Archives. The NCBI Metagenomics e-book links together related NCBI resources including sequence data, publications and analysis tools.

The unusual structure of metagenomic data (heterogeneity, fragmentation, redundancy, high error rate, etc.) will require new the development of new analysis tools and visualization techniques. New computational tools for the large-scale analysis of complex metagenomic data (both DNA and predicted proteins) are under development. Ongoing work on 16S rRNA analysis and visualization tools will be presented.

04:20 **Targeted Gene Identification from Short Gene Fragments in Metagenome and their Use in Biogeochemical Studies**

Robin B. Kodner, University of Washington School of Oceanography

Co-authors: Frederick A. Matsen, Ian Hewsen, Jonathan P. Zehr, and E. Virginia Armbrust

Metagenomics allows for a new approach to studying biogeochemical processes in an environment, by giving an approximation of a community's biogeochemical potential through the genes involved in a given process. Lipid biomarkers are well suited to this kind of approach because they are important for a number of biogeochemical questions and some have well studied biosynthetic pathways. One important class of lipid biomarkers, triterpenoids, were investigated in a new metagenomic sample that was collected at station SJ0609.03 in the western tropical Atlantic Ocean within the offshore Amazon River plume (12o15.43'N, 56o8.74'W). We identified proteins involved in triterpenoid biosynthesis that have bacterial and eukaryotic homologs in this metagenome. Though this metagenome sample was 5um prefiltered and then onto a 0.2 um filter for sequencing, the database contains some eukaryotic sequences, along with the target prokaryotic sequences. In this context, genes must be identified with a high degree of confidence, and BLAST searches alone

are not rigorous enough to determine the affinity of short fragments. We test a method for identifying genes for lipid biomarker biosynthesis, using TBLASTN searches coupled to a phylogenetic method for making maximum likelihood trees for distantly related proteins by fitting fragments into a reference tree. By examining the relative likelihood of ML trees with the fragment attached at all points in the reference tree, we gain a measure of confidence in the phylogenetic placement of fragment sequences.

04:40 **Metabolic Characterization of *Candidatus Accumulibacter Phosphatis* Using Metaproteomic Analysis** Jason Flowers, University of Wisconsin-Madison

In many freshwater bodies, phosphorus is a limiting nutrient for bacterial growth. To prevent eutrophication, enhanced biological phosphorus removal (EBPR) of wastewater has been successfully applied at sewage plants for the past thirty years. Previous researchers could only speculate about the biochemical pathways involved in EBPR principally because the organisms responsible for EBPR cannot be isolated in pure culture. Using metagenomic data recently obtained from a *Candidatus Accumulibacter phosphatis* (*Accumulibacter*) enriched bioreactor, we have investigated the metabolism of EBPR through metaproteomic analysis. A total of six protein samples were collected for metaproteomic analysis. Proteins were extracted and either digested with trypsin in a 1D SDS-PAGE and run on a LC-MS/MS or digested in solution with trypsin and run on a 2D LC-MS/MS for protein identification. For all samples, peptide MS data was analyzed using Sequest to construct the original protein sequence by comparison to the previously sequenced metagenome. A total of 964 proteins were matched across all samples. 223 proteins were observed in at least two samples with only 80 identified as originating from *Accumulibacter*. We were able to reconstruct the carbon metabolism for *Accumulibacter* including nearly all proteins for glycolysis, TCA cycle, and the glyoxylate bypass. Additionally, we observed EBPR relevant proteins including polyphosphate kinase and PHA synthase. The overall results confirm the expression of several pathways required for EBPR metabolism, but more work is required to capture the complete proteome that will improve our understanding of the process.

08:00 Dinner Keynote:

A Genomic Encyclopedia of Bacteria and Archaea (GEBA) and the Search for the Dark Matter of the Biological Universe

Jonathan A. Eisen, UC Davis and Joint Genome Institute

There is a glaring gap in microbial genome sequence availability – the currently available genome sequences show a highly biased phylogenetic distribution compared to the extent of microbial diversity known today. This bias has resulted in major limitations in our knowledge of microbial genome complexity and our understanding of the evolution, physiology and metabolic capacity of microbes. Although there have been small efforts in sequencing genomes from across the tree of life for microbes, there are no systematic efforts.

There are many reasons why phylogenetic based sequencing in theory should be of great benefit including: (a) improved identification of protein families and orthology groups across species, which will improve annotation of other microbial genomes (b) improved phylogenetic anchoring of metagenomic data, (c) gene discovery (which tends to be maximized by selecting phylogenetically novel organisms, (d) a better understanding of the processes underlying the evolutionary diversification of microbes (e.g., lateral gene transfer and gene duplication) (e) a better understanding of the classification and evolutionary history of microbial species and (f) improved correlations of phenotype and genotype in microbes.

Based on the potential benefits, we (JGI) have commenced a pilot project to create a Genomic Encyclopedia of Bacteria and Archaea (GEBA). In this pilot, we plan to sequence ~100 genomes selected based on their phylogenetic novelty. This is being done at two phylogenetic scales. About 60 of the genomes are from across the breadth of bacteria and archaea. The remaining 40 genomes are from within the Actinobacteria. By doing this two tiered selection we can test both the value of breadth from across the bacteria and archaea as well as the value of filling in the phylogenetic gaps within a single phyla.

In my talk I will summarize the project and report on the sequencing and analysis of the first 56 genomes. I will discuss how we are using this pilot to test protocols that could be used for a scale up of the GEBA project or for any other large scale microbial sequencing project. In addition I will discuss how collaborations with culture collections can be valuable in such a project. Finally, I will report on the results of tests of the value of phylogenetic based sequencing.

THURSDAY, NOVEMBER 6

T
H
U
R
S
D
A
Y

A
B
S
T
R
A
C
T
S

THURSDAY, NOVEMBER 6

08:30 Morning Talk: **Comparative Metagenomics**
Peer Bork, European Molecular Biology Laboratory

Each metagenome study of a complex habitat is accompanied by lots of unknowns as so far the depth of sequencing is usually insufficient to reveal the entire microbial community of an ecosystem. The probably easiest way of extracting meaningful information is the comparison with existing data, analogous to that of genome comparisons. However, the complexity within the analysis chain of a single metagenome annotation is topped by the difficulties that the integration of heterogeneously analysed metagenomes causes. I will illustrate some of the powers of pitfalls of current basic level analysis tools and will apply comparative strategies to i) the mining of novelty in metagenomes and to ii) a number of diverse habitats such as human and ocean.

09:30 Break

SESSION I: Host-Associated Microbes

Chair: Janet Jansson, Lawrence Berkeley National Lab

09:40 **Genomic and Genetic Insight into Gut Microbiota Function and Manipulation**

Justin L. Sonnenburg, Stanford University School of Medicine

Trillions of microbes live in our digestive tract and influence our biology in profound and diverse ways. Several diseases, including obesity and inflammatory bowel diseases, have been associated with large-scale shifts in microbiota composition. The ability to address basic questions concerning community function and plasticity are fundamental to understanding the extent of causal relationships between host biology and microbiota perturbations, and whether the microbiota is a viable therapeutic target. One of our long-term goals is to achieve a level of functional understanding that, if provided the metagenome of an individual's microbiota, would allow us to accurately predict how the microbial community will functionally adapt to a specific perturbation (e.g., dietary change). To investigate how changes in the intestinal environment alter microbiota function, and how these changes, in turn, influence host biology we have characterized responses of simplified microbiotas living within the gut of gnotobiotic mice to changes in host diet, community membership, and host genotype. These studies have revealed the importance of a finely-tuned system of polysaccharide sensing and utilization in the model symbiont *Bacteroides thetaiotaomicron* (*B. theta*). We are currently using a single polysaccharide utilization locus dedicated to dietary fructan utilization of *B. theta* as a model to understand mechanisms underlying diet-induced changes in microbi-

ota function and composition. Genetic ablation of proteins involved in the multi-step process of sensing, harvesting, degrading, and metabolizing fructans variably cripples *B. theta*'s utilization of fructose-based polysaccharides depending upon which step of consumption is compromised. These findings are consistent with functional differences in fructan utilization between *Bacteroides* species. Together these results set the stage for predicting, based on gene content, how microbiotas respond to changes in the nutrient environment and suggest how metagenomics could facilitate personalized therapeutic manipulation of the microbiota.

10:20 **New Insights Into Lignocellulose Conversion by Termite Gut Microbes**

Jared Leadbetter, California Institute of Technology

Termites and their complex hindgut microbiota are able to convert wood lignocellulose into hydrogen and other products used to fuel their metabolisms. Recent gene and genome and metagenome based analyses on the gut community have revealed novel insights into many bacteria-mediated, important symbiotic functions. The system-wide gene analysis of a microbial community specialized towards plant lignocellulose degradation has both basic and applied implications.

11:00 Break

11:10 **Metaproteomics as a key technology for characterizing the human microbiome**

Nathan C. VerBerkmoes, Oak Ridge National Labs

The human microbiome is a complex system of many microbial communities inhabiting a diversity of environmental niches throughout the human body. With at least an order of magnitude more cells and even greater diversity of genetic potential these microbial communities continually interact with the human host cells in complex but controlled manner that lead to normal human health. Our knowledge of the structure and function of these communities and the interactions with the human host is limited because analyses of microbial physiology and genetics have been largely confined to isolates grown in laboratories. Recent acquisition of genomic data directly from natural samples has begun to reveal the genetic potential of communities (Tyson, *Nature* 2004) and environments (Venter, *Science* 2004). The ability to obtain whole or partial genome sequences from microbial community samples has opened up the door for other system level studies of microbial communities such as community proteomics or metaproteomics (Ram, *Science* 2005, Lo, *Nature* 2007; Wilmes, *ISME* 2008).

The human gut contains a dense, complex, and diverse microbial community. A healthy gut microbiome is clearly

key for human health, thus making this system one of the major target areas of the human microbiome project. Metagenomics has recently revealed the composition of genes in the gut microbiome (Gill, Science 2006), but provides no direct information about which genes are expressed or functioning. Therefore, our goal was to develop a novel approach to directly identify microbial proteins in fecal samples to gain information about what genes were expressed and about key microbial functions in the human gut. We used a non-targeted, shotgun mass spectrometry-based whole community proteomics, or metaproteomics, approach for the first deep proteome measurements of thousands of proteins in human fecal samples from two normal adult twins. The resulting metaproteomes had a skewed distribution relative to the metagenome, with more proteins for translation, energy production, and carbohydrate metabolism compared to what was earlier predicted from metagenomics. Human proteins, including antimicrobial peptides, were also identified, providing a non-targeted glimpse of the host response to the microbiota.

Similar integrated genomic, transcriptomic and proteomic studies are on-going in germ free gnotobiotic mice. In this system the mice can be directly colonized with known sequenced human gut-derived type strains of *B. theta* or *E. rectale* or even more complex communities. Furthermore, in this mouse system, the cecum can be removed and the microbial communities directly assayed thus providing a perfect model system for the development of omic techniques for studying host-microbial interactions.

11:50 Metagenomic Comparison of the Microbial Hindgut Communities in Drywood- and Grass-Feeding Termites

Falk Warnecke, Microbial Ecology Program, Joint Genome Institute, DOE

Co-authors: Natalia Ivanova, Martin Allgaier, Nikos Kyrpides, Rudolf Scheffrahn, and Phil Hugenholz

Termites are highly efficient in degrading lignocellulosic biomass. In a recent metagenomic study we showed for the first time that microbes inhabiting the termite hindgut encode hundreds of carbohydrate-active enzymes, e.g. glycosyl hydrolases (GHs) and tens of carbohydrate-binding modules (CBMs) and by implication are responsible for the bulk of the lignocellulose deconstruction (1). In this initial study we analyzed the metagenome of the microbial hindgut community of the drywood-feeding higher termite species *Nasutitermes corniger* collected from a nest in a pristine tropical rainforest in Costa Rica. Here we present results from a similar study however focusing on termites exposed to different food sources. We analyzed the hindgut of laboratory-kept *N. corniger*, as well as two samples of the grass-feeding termite species *Amitermes wheeleri*.

The lab-kept termites have been maintained in captivity for several years in the laboratory of Prof. R. Scheffrahn. A key difference between the wild and lab-kept *Nasutitermes* is diet; the Costa Rican termites are foragers and most likely fed on a wide variety of plant species while the lab counterparts were fed almost exclusively on Brazilian pepper wood. The two samples of grass-feeding termites were collected in Texas tentatively representing different feeding regimes as well: the first was collected from pristine grassland, while the second from within pieces of cow dung on pasture land. In summary this collection of samples allows several interesting comparisons: (i) drywood-feeding versus grass-feeding termite hindgut microbial communities, (ii) wild versus lab-kept drywood feeders, and (iii) pristine grassland versus cow dung-feeding *Amitermes*.

The GH and CBM profiles from the same termite species are very similar whereas the profiles show pronounced differences between species. In addition to the metagenomic sequence data we prepared a 16S rRNA PCR library to analyze the microbial community composition. We discover

12:10 Screening of the Human Intestinal Microbiota for the Discovery of Novel Enzymes Efficient for Plant Fiber Degradation

Gabrielle Veronese, Institut National de la Recherche Agronomique (INRA)

Co-authors: Lena Tasse, Sandra Pizzut-Serin, Sophie Bozonnet, Juliette Bercovici, Patrick Robe, Julien Tap, Marion Leclerc, Pierre Monsan, and Magali Remaud-Simeon

The human intestinal microbiota plays a key role in the metabolism of non-digestible food components (dietary fibres). In the frame of the project Alimintest*, a metagenomic DNA library has been prepared from fecal samples of a man volunteer submitted to a fiber-rich vegetarian diet, to evaluate the functional potential of the human gut, and to explore this biodiversity as a source of novel enzymes responsible for plant fiber hydrolysis. The library consists of 156 000 fosmids, each clone comprising a 40 kb insert, covering a total of 6 gigabases. We have developed various efficient high-throughput functional screening methods for the discovery of glycoside-hydrolases specific for resistant polysaccharide degradation: alpha-glycosidases, beta-glycosidases (cellulases, hemicellulases, beta-glycanases), pectinases... A screening rate of 200 000 clones per week and per enzyme activity is achieved using our HTS facility, enabling an average of 0.15 % hits. So far 50 clones have been selected for their catalytic efficiency and the originality of their specificity. To identify the genes responsible for the screened functions, pyrosequencing is performed on the corresponding fosmids. The most original genes will then be sub-cloned in the view of protein purification for enzyme characterisa-

THURSDAY, NOVEMBER 6

T
H
U
R
S
D
A
Y

A
B
S
T
R
A
C
T
S

tion and 3D-structure resolution.

*Project 200206-01-01, funded by of the French National Research Agency. Consortium combining the teams of LibraGen SA., Timothy Vogel (UMR CNRS 5005 Lyon), Joël Doré (UEPSD/UGM INRA Jouy-en-Josas), Stéphanie Passot (INRA INAPG), Eric Fontaine (CNRH Grenoble), François Pompanon (UMR CNRS Grenoble) and Pierre Monsan (LISBP INRA-CNRS-INSA Toulouse).

12:30 Lunch

SESSION II: Emerging Sequence Technologies

Chair: Nikos Kyrpides, Joint Genome Institute

01:30 Single molecule analyses of DNA in environmental microbes

Kun Zhang, UC San Diego

Comprehensive characterization of genomic composition in environmental microbial samples has been challenging, because the majority of microorganisms are difficult to culture. Additional challenges include a high degree of genetic diversity between and within species, various level of relative abundance and ubiquitous presence of cell-free DNA. Single cells or single molecules genomic assays hold great promise for tackling these challenges. Recent progresses in the developments of these technologies, including single cell genome sequencing, multiplex polony-based analysis of single cells and microbial cell sorting with lab-on-chip devices will be discussed.

02:10 New Sequencing Technologies at JGI and Applications in Bioenergy Research

Feng Chen, Joint Genome Institute

The US DOE Joint Genome Institute (JGI) is a high-throughput sequencing and genomic research center involved in a myriad of sequencing projects. JGI's major effort is the sequencing of genomes and transcriptomes of plants, microbes and environmental metagenomic samples of relevance to the DOE missions of carbon sequestration, bioremediation and energy production.

Roche/454's platform utilizes emulsion PCR for template amplification and pyrosequencing technology on high well-density picotiter plate. Illumina/Solexa's platform uses bridge amplification on glass surface for template preparation and reverse terminator technology for sequencing. Both platforms provide high throughput and high quality sequencing production at low cost, however, the read length and error distribution differ between these two significantly.

Based on the different characteristics of the sequencing data, we developed, implemented and optimized vari-

ous applications to best utilize the sequencing data from these two platforms. In the presentation, I am going to give examples of some of these applications such as whole genome shotgun sequencing of microbial genomes, bacterial community 16S diversity study, transcriptome sequencing (RNA-Seq), and metatranscriptome sequencing.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

02:50 Break (move to breakout rooms)

SESSION III: Parallel Tutorials/Demos

03:10 Parallel Tutorials 1

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

Kayo Arima, UC San Diego

The aim of CAMERA (<http://camera.calit2.net/>) is to serve the metagenomic and microbial ecology research community by providing annotated data and metadata for metagenomic sequences and will increasingly include relevant software tools focused on the requirements for studying microbial communities and the specific attributes of metagenomic datasets.

The distributed architecture of the CAMERA compute engine is based on the OptIPuter, which will allow up to a hundred-fold increase in bandwidth to end-users accessing our facility over the new National Lambda Rail cyberinfrastructure backplane. Tiled display walls, used as termination devices for the OptIPuter backplane, will enable researchers to share, analyze and visualize scientific data interactively on very high resolution (multi-megapixel) scales.

CAMERA sustains highest priority for acquisition of environmental metagenomics data from marine and other environments, along with a community-defined level of metadata, and next will add biological metagenomics or microbiome data sets and reference genomes. This growing resource will enable the community, as end-users, to conduct rigorous analyses of the metagenomic data in order to obtain deeper insight into microbial ecology and other challenges. This tutorial will include an overview of the CAMERA project and will focus on the hands-on use of the portal.

MEGAN and MetaSim

Daniel Huson, Tuebingen University

MEGAN is a tool for metagenome analysis that we have developed that seems to be quite popular. We plan to officially release version 2.0 for the conference, which can handle very large datasets (Blast files of 300-700 GB) and can be used to compare such datasets. MetaSim is a metagenome simulator that is in press and can be used to simulate complex metagenomes and sequencing technologies for sequencing them.

RDP-II

James Cole, Michigan State University

03:10 Parallel Tutorials 2

IMG Family of Systems for Comparative Analysis and Annotation of Microbial Genomes and Metagenomes

Victor Markowitz, Lawrence Berkeley National Lab

Co-authors: Amrita Pati and Nikos Kyrpides

The IMG family of systems provide support for comparative analysis and annotation of microbial genomes and metagenomes in a comprehensive integrated context, and consist of: the Integrated Microbial Genomes (IMG) data analysis system (<http://img.jgi.doe.gov>), the IMG Expert Review (IMG/ER) genome data curation system (<http://img.jgi.doe.gov/er>), the IMG Metagenome (IMG/M) metagenome data analysis system (<http://img.jgi.doe.gov/m>), and the IMG/M Expert Review (IMG/MER) metagenome data curation system (<http://img.jgi.doe.gov/mer>). Scientists can analyze and curate their genome and metagenome datasets in the integrated context of IMG ER or IMG/M ER by submitting their datasets via the IMG Submission Site (<http://img.jgi.doe.gov/submit>) which is also used for collecting genome and metagenome specific metadata.

We will overview the purpose, content, and analytical tools of the IMG systems and then walkthrough each system. We will also discuss some of the challenges of processing and comparing metagenome datasets and show how these challenges are addressed within IMG/M and IMG/MER.

Greengenes Services for 16S rRNA Gene Analysis

Todd DeSantis, Lawrence Berkeley National Lab

Greengenes (<http://greengenes.lbl.gov>) is a web application assisting molecular ecologists with data analysis. Aligning 16S rRNA gene sequences, removing chimeras, and classifying the members of a microbial community against all of the five dominant bacterial and archaeal taxonomies will be covered. Two advanced methods will also be discussed: integration of PhyloChip community analysis with sequencing data and how to import your

Greengenes pre-processed data into ARB for visualization. Participants may preview the online tutorial from the Greengenes website.

MicrobesOnline

Paramvir Dehal, Lawrence Berkeley National Lab

03:10 Parallel Tutorials 3

Megx.net, ARB & SILVA: Integrated Diversity and (Meta)genomic Analysis in the Marine Environment

Frank Oliver Glöckner, Ivaylo Kostadinov and Melissa Duhaime, Max Planck Institute for Marine Microbiology
Co-authors: Renzo Kottmann, Wolfgang Hankeln, Elmar Prüsse, Christian Quast, Karin Dietrich, and Jörg Peplies

Sequencing ribosomal RNA (rRNA) genes is currently the method of choice for phylogenetic reconstruction, nucleic acid based detection and quantification of microbial diversity. To cope with the deluge of data, the SILVA system (from Latin silva, forest) was implemented to provide a central, comprehensive web resource for up to date, quality controlled databases of aligned small- and large subunit rRNA sequences from the Bacteria, Archaea and Eukarya domains. All sequences are checked for anomalies and carry a rich set of process information. An intuitive ranking system allows the user to get a rapid overview about the quality of the sequences. SILVA is designed as a central comprehensive resource by integrating multiple taxonomic classifications and the latest validly described nomenclature as well as additional information, such as if a sequence was derived from a cultivated organism, a type strain, or belongs to a genome project. All databases are fully compatible with the widely used ARB software package for phylogenetic inference and probe design.

Megx.net, a Portal for Marine Ecological GenomiX, provides access to specialized georeferenced databases and tools for the analysis of marine bacterial, archaeal and viral genomes and metagenomes including the Global Ocean Sampling (GOS) reads. The key element of Megx.net is the 'Genes Mapservers', which facilitates the interpretation of sequence data in its environmental context. However, this only works as long as the environmental parameters are clearly referenced with longitude, latitude, time and depth/altitude (x, y, z) coordinates. In this respect, the unique strength of the 'Genes Mapservers' is its Global Information System (GIS), which integrates environmental data layers extracted from the World Ocean Atlas (WOA). Currently, the 'on the fly' interpolation covers temperature, salinity, dissolved oxygen, apparent oxygen utilization, percent oxygen saturation, phosphate, silicate, and nitrate at standard depths, averaged over annual, seasonal, and monthly periods for any location in the marine system.

By linking SILVA with the 'Genes Mapservers', phylogenetic

THURSDAY, NOVEMBER 6

T
H
U
R
S
D
A
Y

A
B
S
T
R
A
C
T
S

diversity and marine (meta)genomics will be married in light of on site habitat parameters. The SILVA databases and Megx.net portal can be found at www.arb-silva.de & www.megx.net. Please visit the tutorial session for Megx.net and SILVA for a practical introduction to the websites and tools.

03:10 **The New Version of the MG-RAST Server**
Folker Meyer, Argonne National Lab; Univ of Chicago

As of October 2008 more than 45 gigabases of metagenome data have been analyzed by MG-RAST. More than 1300 data sets have been submitted by several hundred users from more than 30 countries.

URL: <http://metagenomics.nmpdr.org>

MG-RAST is a free, open-access, online resource that allows upload and analysis of metagenome data sets (Sanger and 454 is supported). The server computes metabolic and organismal reconstructions using SEED data structures and existing ribosomal RNA databases. The results are available via a user friendly web interface and download.

The version 2.0 of MG-RAST offers adjustable parameters for metabolic and community reconstruction and a greatly enhanced web user interface. In addition to the databases already in the system (SEED NR, SEED Subsystems, RDP-II, Greengenes) for analysis, we have included the new Silva ribosomal RNA database. New comparative tools have been created (e.g. recruitment plots), as well as several more download options, making all data we compute available in a variety of formats. Using the new "invite a friend" feature the server now allows data sharing between groups of users. Owners of data sets can also release data as public, effectively publishing the data sets. In the tutorial we will walk users through the analysis of a sample data set.

05:10 Break (move back to Calit2 Auditorium)

05:30 Keynote: **International Soil Metagenome Sequencing Project**

Timothy M. Vogel, Université de Lyon

Soil is often considered to be one of the main reservoirs of microbial diversity on the planet. This diversity could provide a range of information about the origins of microbial functional diversity as well as novel genetic resources. However, our historical inability to cultivate the majority of soil bacteria has hampered our understanding and exploitation of this large diverse community (more than 95% are often considered inaccessible through traditional culture techniques). Over the last 20 years, new methods have been developed to overcome the limitations due to culture

techniques and access a significantly greater soil microbial diversity through the direct extraction and exploration of DNA from soil. This metagenomic toolbox allows accessing, storing, and analyzing the DNA extracted from the quasi total microbial community and thus can provide an otherwise hard-to-attain insight into the biology and evolution of environmental micro-organisms. The limits and benefits of this approach need to be discussed before attempting to completely sequence a soil metagenome.

The complete sequencing of a soil metagenome, (i.e. the genome of all bacteria inhabiting the soil environment) is now an imaginable objective that would require a strong international collaboration including soil microbiologists, geneticists, bioinformaticians etc. The establishment of a working public international consortium for the complete sequencing of the metagenome of a reference soil would provide focus for an increased effort on the barriers in accessing the entire soil metagenome. These barriers include imperfect methods of soil sampling, DNA extraction, DNA purification, cloning and sequencing. The soil system proposed for investigation, Park Grass, Rothamsted (UK), is a charismatic, internationally recognized unique resource that includes ongoing experiments that have been running for over 140 years. This unique long term ecological site (LTER) provides a history of soil biology and chemistry, as well as an archive of soil samples representing detailed studies of the impact of manipulations and its metagenome sequence could constitute the "reference" to which other soils could be compared.

06:30 Dinner

07:30 Dinner Talk: **Exploring Next-Generation Sequencing Today**

Thomas Jarvie, 454 Life Sciences

FRIDAY, NOVEMBER 7

FRIDAY, NOVEMBER 7

SESSION I: Connecting Sequence, Structure and Function for Metagenomics

08:30 Introduction

John Wooley, UC San Diego

08:35 The Human Microbiome Project: Key Questions and Challenges

Claire Fraser, University of Maryland

09:30 Diversity Profile of the Human Skin Microbiome in Health and Disease

Elizabeth A. Grice, National Institutes of Health
Co-authors: Heidi H. Kong, Sean P. Conlan, Alice C. Young, NISC Comparative Sequencing Program, Gerard G. Bouffard, Robert W. Blakesley, Maria L. Turner, Julia A. Segre

The concept that the human body is host to trillions of microbes is revolutionizing our view of the human genome while underscoring the role of the gene-environment interface in complex disorders. One such disorder, with a known genetic component, is the very common inflammatory skin disorder atopic dermatitis (AD; eczema) whose incidence has tripled in the past 30 years. Our previous 16S phylo-type diversity survey of the most commonly affected human site in AD, the antecubital fossae (inner elbow), demonstrated a unique skin core microbiome dominated by Janthinobacteria and Pseudomonas (both Proteobacteria) with less representation from five other bacterial divisions. Skin provides an unprecedented opportunity to sample multiple sites from the same individual, many with underlying left-right symmetry. Skin sub-sites have unique environmental niches: moist/dry, haired/non-haired, acid/basic, sebaceous (oily)/non-sebaceous. Associated with these specific areas are stereotyped human disorders; e.g. psoriasis affects outer elbow and AD affects the inner elbow. We are currently ascertaining 21 skin sub-sites from healthy humans to comprehensively survey the resident microbiota and address the fundamental question of whether there is a baseline cutaneous microbiome. This data is a foundation for our studies investigating alterations of skin microbiota in a disease state, specifically AD. Our long term goal is to elucidate the contribution of the cutaneous microbiome to complex skin disorders and translate this into novel pharmacological treatments.

10:10 The Human Oral Microbiome

Floyd Dewhirst, Harvard University

The human oral cavity is a diverse habitat that contains approximately bacterial 600 predominant species. The oral microbiome is comprised of 44% named species, 12% isolates representing unnamed species, and 44% phylo-types known only from 16S rRNA based cloning studies.

Species from 11 phyla have been identified: Firmicutes (211), Bacteroidetes (106), Proteobacteria (99), Actinobacteria (64), Spirochaetes (49), Fusobacteria (29), TM7 (12), Synergistetes (10), Chlamydiae (1), Chloroflexi (1) and SR1(1). Full and survey sequences have been obtained for over 30 oral species, and in the course of the Human Microbiome Project over 300 essentially complete genome sequences should be determined. An Oral Microbiome Project is in progress and data from this project should be available soon. The talk will discuss the diversity of the oral microbiome, the Human Oral Microbiome Database (a resource for exploring the Oral Microbiome), and efforts to connect the oral metagenome with the oral metaproteomics and structural metagenomics.

11:00 Metabonomic profiling of the gut microbiome: Implications for human health

James Kinross, Imperial College, London

Microbial-mammalian metabolic cooperation is defined as the human 'metabonome'. Transgenomic co-metabolic interactions within the metabonome greatly increase system complexity and this presents a significant challenge for elucidating the mechanisms by which the intestinal microbiome influences the host phenotype. By measuring and mathematically modelling changes in the levels of products of metabolism found in mammalian biological fluids and tissues, metabonomics offers fresh insight into the effects of the gut microbiome on human health. This talk explores current concepts in this rapidly evolving field, and describes how metabonomics may be used as part of a systems approach for studying the intestinal microbiome.

SESSION II: Getting to Molecular and Bacterial Community Function

12:30 Structural Genomics (SG) & Beginning a Research Dialogue with the Metagenomics Community

Ian Wilson, The Scripps Research Institute & Stephen Burley, Eli Lilly & Co.

01:20 Protein Target Selection Challenges and Charge to Discussion Groups

Adam Godzik, Burnham Institute of Medical Research and UC San Diego

02:00 Discussion Sessions: Role of Structure in Ascertaining Functional Properties of Microbial Communities: How Can These Two Research Domains Interact Most Productively?

03:30 Summary of Subgroup Discussions

04:00 Closing Remarks

John Wooley, UC San Diego



Metagenomics 2008 Poster Abstracts

Comparative genomics of *Bacillus* sp. from a desiccation lagoon in Cuatro Ciénegas, Mexico

Luis David Alcaraz, Departamento de Ingeniería Genética
Co-authors: Valeria Souza, Luis Herrera-Estrella and Gabriela Olmedo

Bacillus coahuilensis was recently described as a new species and its genome has been published. *B. coahuilensis* was isolated from the water column of a desiccation lagoon in the Chihuahuan desert. This oligotrophic water system is rich in magnesium and sulfate, but its extremely low phosphorus (P) level (>0.5mM). This bacterium seems to have overcome phosphorus limitation by the replacement of phospholipids by sulfolipids and by a reduction of its genome size. The genes required for the sulfolipid synthesis were acquired by means of Horizontal Gene Transfer (HGT) from Cyanobacteria, as was the case for a sensory rhodopsin. We decided to go further and so we sequenced the genome of another isolate from the same environment, *Bacillus* m3-13. Based on the 16S rRNA, it seems to be very distant from *B. coahuilensis*, and closer to *B. horikoshii*. Interestingly, *B. coahuilensis* and *B. sp.* m3-13 share more genes than expected. An explanation to this could be that they share the same environmental conditions, constrains, gene pool and we suppose that HGT is shaping this relatedness. Differences are remarkable. *B. sp.* m3-13 has a larger genome (4.3 Gb) than *B. coahuilensis* (3.35 Gb), and this is reflected in its metabolic capabilities; whereas the first one is robust and could play a generalistic role, *B. coahuilensis* seems to be bound to primary producers. Additionally, *B. sp.* m3-13 seems to overcome P limitation not by sulfolipid but by phosphate uptake thus reflecting different gene strategies to the same limitation.

Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions

Linda Amaral-Zettler, Marine Biological Laboratory, Woods Hole
Co-authors: Phillip Neal, Susan Huse, and Mitchell L. Sogin

ICoMM is one of 14 ocean realm projects of the Census of Marine Life Program (CoML) that seeks to determine the diversity, distribution and abundance of microbes in the ocean. The ICoMM Secretariat hosts the website <http://icomm.mbl.edu> and the MICROBIS database. It has sponsored meetings for the five primary working groups (Benthic, Open Ocean and Coastal Systems, Technology, Informatics and Data Management and Microbial Eukaryotes) and annual meetings of its Scientific Advisory Council. In collaboration with the international community

of marine microbiologists, ICoMM has forged a large-scale effort to characterize microbial diversity in the sea through massively-parallel, 454-based sequencing of hypervariable regions of the SSU rRNA genes of bacteria, archaea and microbial eukaryotes. Sequencing is underway on 52 separate projects from a diversity of environments including deep and shallow hydrothermal vent systems, polar regions, coastal and estuarine environments, the open ocean, the deep biosphere, oxygen minimum zones, corals and 8 of the 13 aquatic US Long Term Ecological Research Sites (MIRADA project). ICoMM is simultaneously collecting data on environmental parameters that characterize all sampling sites and making these available through its affiliated website VAMPS (<http://vamps.mbl.edu>) that provides the ICoMM community with tools for comparing similarities and differences in the composition of microbial populations. To date, ICoMM has completed sequencing for roughly one third of its projects and has generated over 10 million tags and counting. Analyses are underway integrating diversity data with contextual information that should inform us about the interplay between microbial mediated activities and oceanic processes.

Comparative analysis of human gut microbiota by barcoded pyrosequencing

Anders F. Andersson, Uppsala University
Co-authors: Hedvig Jakobsson, Johan Dicksved, Ben Willing, Mathilda Lindberg, Jonas Halfvarson, Thomas Abrahamsson, Janet Jansson, Lars Engstrand

The human gut houses complex microbial ecosystem that have co-evolved with its host and are likely to play important roles in the maintenance of health and in the etiology and outcome of various disease states. Finding microbes associated with higher or lower risk of developing diseases thus holds promise for developing therapeutic and preventive agents. For this reason we are comparing the gut microbiota of groups of patients and healthy controls, employing barcoded 454 pyrosequencing of a hypervariable region of the bacterial 16S ribosomal RNA gene. We have developed a bioinformatics pipeline for grouping sequences into OTUs and performing phylogenetic annotations. Using this approach we are currently investigating the links between gut microbiota, allergy, probiotics and mode of delivery; the role of microbiota in Inflammatory Bowel Disease (using monozygotic twin-pairs discordant for IBD); and consequences of antibiotic treatment on the gut microbiota.

A statistical approach for a better assessment of community taxonomy and average genome size

Florent Angly, San Diego State University

Metagenomic-based analysis of carbon management in diverse soils

Dionysios A. Antonopoulos, Argonne National Laboratory
Co-authors: Areej Ammar, Marc H. Domanus, Rob Edwards, Julie Jastrow, Kenneth Kemner, R. Michael Miller, Kelly A. Skinner-Nemec, Rick Stevens, Jared Wilkening, Yizhong Zhang, Folker Meyer

Understanding nutrient cycling in soil as it relates to questions of carbon sequestration and general carbon management requires not only an understanding of the impact of the plant community but also of the underlying microbially-mediated processes. Microbial concentrations can reach 2×10^9 organisms per gram of soil in the top meter, and 1×10^8 per gram at 1-8 m depths. As part of an effort to better catalog the processes involved in carbon and nutrient cycling mediated by these organisms, we have begun a metagenomic-driven initiative for describing the metabolic potential of the organisms present. Soil samples collected at the Fermilab National Environmental Research Park were used as the source of metagenomic DNA for construction of 454 shotgun sequencing. The largest fraction of identifiable sequences annotated and analyzed using the metagenomics RAST (Rapid Annotation using Subsystems Technology) server were classified within the carbohydrate metabolism subsystem category. Additional soil samples from sites within the National Ecological Observatory Network (NEON) have also been collected for metagenomic sequencing. These sites represent a variety of terrestrial environments from throughout the United States providing context for the Fermilab-derived samples.

Putting biodiversity to work: New approaches to functional enzyme screening in *E. coli*

Meike Ballschmiter, University of Leipzig
Co-authors: Antje Eichler, Anja Kunert, Thomas Greiner-Stoeffele

Over the last years it has become established, that a nearly endless number of microorganisms can be found in the various habitats of our planet. This biodiversity brings about a seemingly unlimited number of potentially novel enzyme activities. Yet accessing this biodiversity and effective activity-based screening for industrial relevant enzymes in nature's toolbox is a bottleneck. We have developed a 2-component vector system, matched to the proprietary cluster screening system (Greiner-Stoeffele, T., Struhalla, M., WO2004002386) of our industrial partner that enables us to functionally screen in *E. coli* with a higher throughput than the conventional screening approach. Using a metagenomic library from sheep rumen and a lipase/esterase screen as our test system we show that with our combination of vector and screening system we can detect one enzyme as initial hit per 9.7 Mbp. With a conventional pUC library from the same metagenomic DNA we can only detect one target enzyme per 92 Mbp. Although *Pseudomonas* species are scarce in our library, two out of three characterized hit enzymes are putative *Pseudomonas* autotransporter lipases/esterases. This casts a spotlight on the known problem of the strong expression bias of the heterologous host *E. coli*. To circumvent the host bias alternative screening systems in addition to the optimized *E. coli* screen are necessary. Therefore we are developing an in vitro functional screening system based on *E. coli* extracts, which has yielded first results. We are also able to detect proteases from *Bacillus* in a complex library screening approach.

Metagenomic Analysis of the Epiphytic Bacterial Community of the Green Macroalga *Ulva Australis*

Catherine Burke, Centre for Marine Bio-innovation
Co-authors: Nicky Yeo, Aaron Halpern, Staffan Kjelleberg, Torsten Thomas

Epiphytic microbial communities from algae are thought to protect their hosts from biofouling via the production of bioactive metabolites, be involved in nutrient exchange and acquisition and in the case of green algae have been shown to influence the morphology of their host. In order to better understand these processes, we have investigated the phylogenetic and functional diversity of the bacterial community associated with the green macroalga *Ulva australis*. A novel method was developed that specifically extracts DNA from the bacterial community of algal surfaces and this DNA was used to generate extensive 16S rRNA gene libraries (>1000 clones) and DGGE analyses. This has provided detailed phylogenetic information on the differences between the surface community and the community in the surrounding seawater. Furthermore a number of fosmid metagenomic libraries (totalling more than 80 000 clones) were created and screened for antibacterial, antifungal and LuxR-inducing activities.

We have recovered a number of positive clones with antibiotic and LuxR-inducing activities. Sequencing analysis and in vitro transposon mutagenesis revealed no similarities to known antibiotic genes or pathways and the genes responsible for LuxR-induction showed no homology to known LuxR inducers. In conclusion, the epiphytic bacterial community of *U. australis* is specific to this living surface and contains species which produce novel bioactive compounds, and may communicate via novel quorum sensing mechanisms.

Metaproteome and metagenome analysis of microbial communities in the phyllosphere

Samuel Chaffron, University of Zurich
Co-authors: Nathanaël Delmotte, Claudia Knief, Gerd Innernebner, Christian von Mering and Julia Vorholt

The aerial parts of plants (phyllosphere) are the greatest organic surface area on the planet ($\pm 1.109 \text{ km}^2$) and therefore represent an important environment, ecologically and economically. The phyllosphere constitutes a particular interesting habitat, host of numerous microorganisms. To date little is known about their identity, the mechanisms involved in niche colonization and the complex processes maintaining the equilibrium of this specific ecosystem. In order to perform a functional in situ analysis of a natural phyllosphere population, a whole community proteome analysis has been undertaken for different naturally or agriculturally grown plant species. Moreover, this study has been complemented with a metagenomic approach using shotgun pyrosequencing to facilitate and improve protein identification and phylogenetic assignments. This work allowed us to identify specific microorganisms and to detect protein families that are highly expressed. From the identification of these proteins, we can learn how protein

functions are partitioned among the principal members of the phyllosphere as a first step to unravel their importance in this habitat. Finally, this study shows how powerful metaproteogenomic (metaproteomics plus metagenomics) approaches are to better characterize complex environmental communities of microorganisms.

High-resolution metagenomics: understanding metabolism and ecological function of individual members of complex microbial communities

L. Chistoserdova, University of Washington
Co-authors: M.G. Kalyuzhnaya, G. Bosch, A. Lapidus, N. Ivanova, A. McHardy, M. Hackett, M.E. Lidstrom

Most microbes on this planet remain uncultured and unknown. Whole genome shotgun (WGS) sequencing of environmental DNA (Metagenomics) has become a powerful tool for uncovering genetic and metabolic potentials of microbial communities in the absence of cultivation. However, when applied to complex communities, the resolution of metagenomics remains insufficient for linking phylogenetic identity of microbes to their ecological function. We aimed at increasing the resolution of metagenomics by focusing the sequencing effort on specific functional types, by the way of specific enrichments. We employed stable isotope probing (SIP) to label DNA of target populations, in this case those oxidizing different single-carbon (C1) compounds in Lake Washington (Seattle, USA), then applied WGS sequencing to the labeled DNA. A total of 255 Mb of high quality sequence were obtained, followed by assembly and annotation. Phylogenetic marker analysis and compositional binning of assembled contigs demonstrated specific sequence enrichments in response to different C1 substrates, indicating the presence of populations with distinct nutritional preferences. A nearly complete genome of the major methylamine degrader, *Methylotenera mobilis* was extracted from the methylamine labeled sample, enabling genomic reconstruction and genome-wide analyses, as well as the downstream transcriptional and proteomic analyses. Partial genomes of other C1 utilizers, including yet uncultivated phyla, were recovered, providing insights into their metabolic peculiarities. Our results demonstrate that use of a novel variation on the standard metagenomics approach, employing function-specific enrichment, allows high-resolution genomic analysis of ecologically relevant species and has the potential to be used in a wide variety of ecosystems.

Analysis of methods to extract, quantify, and identify the metaproteome of soil and groundwater samples

Adina S. Chuang, University of Iowa
Co-author: Timothy E. Mattes

Increased availability of metagenomic bacterial sequences has encouraged the use of metaproteomic techniques to study microbial populations in their natural habitats. Proteins are promising biomarkers as they can relate key functions to specific members of a bacterial community and also indicate microbial diversity within an ecosystem. The development of protein biomarkers for environmental

samples relies on efficient protein extraction and quantification. Protein extraction from the environment is challenging because of the complexity of a metaproteome and large variations between different ecosystems. In our research, we have evaluated several protein extraction methods to extract proteins from soil and groundwater samples. These methods vary in the types of buffers used, physical techniques to lyse cells, direct and indirect extraction of cells from samples, and means of concentrating protein extractions.

Furthermore, various quantification techniques for protein concentrations were studied. Specifically, proteins from environmental samples spiked with a model bacterium, *Nocardioides* sp. JS614, were extracted, quantified, and identified. Strain JS614 aerobically degrades and assimilates vinyl-chloride, a common groundwater pollutant and known human carcinogen. Vinyl-chloride degraders are widespread in the environment and could play a major role in the natural attenuation of vinyl-chloride. A protein biomarker for the presence and function of vinyl-chloride degraders within a contaminated ecosystem would be a valuable tool for bioremediation of environmental sites.

Determining bacterial community by pyrosequencing of SSU rRNA and functional genes

Jim Cole, Michigan State University
Co-authors: Woo Jun Sul, Qiong Wang, Shoko Iwai, Eder-son Da Conceicao Jesus, James M. Tiedje

Determining bacterial community composition of environmental samples is accomplished not only by SSU rRNA genes but also by functional genes of interest. New sequencing technologies such as multiplex pyrosequencing of mixed rRNA amplicons allow in-depth analysis of bacterial rRNA composition to be carried out rapidly and inexpensively. We developed an analysis pipeline to automate the data processing of such large sequencing libraries starting with raw sequences. Multiplex runs containing multiple samples are sorted into individual samples and filtered for quality. Sequences are assigned to taxa using the RDP Classifier. In parallel, sequences are aligned using a fast aligner that incorporates rRNA secondary structure information, clustered using the furthest neighbor method, and several common ecological metrics calculated. The processed data are available in formats suitable for common ecological and statistical packages including Spade, EstimateS, and R. We carried out SSU rRNA genes pyrosequencing of more than 50 samples from soils, rhizosphere, seawater, marine sediment, bioreactor, permafrost, and animal's feces. The bacterial communities were distinct by habitat-type and sample origin. Pyrosequencing of functional genes also shows vast diversity and potentials in the environments. Using a similar strategy, we obtained 3000 sequences targeted by biphenyl dioxygenase primers. These sequences were highly diverse but contained residues conserved in aryl dioxygenases, indicating the range of related genes is much larger than predicted from known sequences.

Pfam Annotations of Environmental and Metagenomic Datasets

Robert D. Finn, Wellcome Trust Sanger Institute
Co-authors: Penny C Coghill, Miao He, Jaina Mistry, John Tate, Alex Bateman

Pfam, the protein families databases, contains over 10,000 families based on the sequences found in the UniProt sequence database. Since the beginning of 2007, we have provided Pfam annotations for additional protein sequence datasets: in GenPept and Metaseq. Metaseq is a collection of proteins from 22 different environmental and metagenomic samples and exceeds 6.6 million sequences. As of Pfam release 23.0 there are 3.7 million domain annotations, matching 46% of sequences in Metaseq and 34% of residues (compared with 73% and 51% for UniProt, respectively). The metagenomic sequence and Pfam match data are available from <http://pfam.sanger.ac.uk> and <http://pfam.janelia.org/> via FTP as flatfiles. These annotations are consistent across the different samples found in Metaseq and maintained as new families are added to Pfam. Having the Pfam domain annotations in one place, enables the domain composition of different samples to be compared. We found enrichment of uncharacterised family DUF1291 in the acid mine environmental sample. Further investigation led us to hypothesise its function in sulfur oxidation. Our sequence alignments also allow the intra- and inter- sequence variation with the samples to be analysed. Furthermore, these alignments provided an excellent resource for phylogenetic inference. The absence of Pfam annotations allows the identification of sequences that are potential novel and may form new families.

Targeting the V1-V2 region of the 16S rRNA gene yields improved measures of microbial community composition

Anthony Fodor, UNC Charlotte
Co-authors: Timothy J. Hamp, W. Joe Jones

The 16S rRNA gene has long been used as a "barcode" to characterize diversity in complex microbial communities. Recent advances in sequencing technology have decreased the cost and effort involved in generating sequences from PCR experiments targeting the 16S rRNA gene. In this study, we explored how the choice of primers affects estimates of microbial diversity. Using a sample taken from the aerobic basin of the activated sludge of a North Carolina wastewater treatment plant, we performed pyrosequencing reactions on PCR products generated with sets of primers targeting the V1-V2, V6 and V6-V7 regions of the 16S rRNA gene. We compared these sequences to 16S rRNA gene sequences found in a whole-genome shotgun pyrosequencing run performed on the same sample. Sequences generated from primers targeting the V1-V2 variable region had the best match to the whole-genome shotgun reaction across a range of taxonomic classifications from phylum to family. These

results demonstrate that by comparing different kinds of pyrosequencing runs, it is possible to find primers that minimize the bias introduced by the initial PCR targeting the 16S rRNA gene.

Comparative Metagenomics of the Human Lung: Characterization of Bacteriophage Communities in Cystic Fibrosis versus Non-Cystic Fibrosis Individuals

Dana Hall, San Diego State University
Co-authors: Mike Furlan, Matt Haynes, Florent Angly, Robert Schmieder, Douglas Conrad, and Forest Rohwer

The usual cause of death for Cystic Fibrosis (CF) patients is a persistent, chronic microbial infection of the lungs which eventually leads to suffocation. Bacteriophage have dramatic effects on the distribution and abundance of the microbes they infect. As phage have been shown to be present in the CF lung, they may have important effects on the dynamics and microbial ecology of the diseased lung. Here, we compare phage communities isolated from five individuals with CF and five control subjects. Viral DNA was sequenced using the Roche/454 Life Sciences GS FLX platform, which produced sequence reads with an average length of 225 base pairs. Metagenomic libraries from CF subjects contained on average more unknown sequences (38%) than those from controls (18%) as determined by BLASTX comparison to the SEED database. CF viral libraries were significantly enriched in the virulence and membrane transport metabolic pathways, and overall exhibited differing metabolic potential from controls. Relative abundances of bacteriophage were dissimilar in CF versus control libraries, as were the distributions of potential phage hosts, with CF libraries representing significantly more phage of *Staphylococcus* and *Streptococcus* species among others. Biodiversity measures including the Shannon Index indicated that phage communities in controls were less variable than those in CF individuals. Based on these results, the bacteriophage community in the CF lung appears to have different taxonomic and functional composition. Future studies will utilize more subjects to further elucidate the unique characteristics of the CF lung bacteriophage community and to determine potential clinical implications.

Finishing Genomes from Single Cell, Photo-heterotrophic Flavobacteria

Cliff S. Han, Joint Genome Institute, DOE
Co-authors: Hajnalka Kiss, Tanja Woyke, Alex Copeland, Gary Xie¹, Jan-Fang Cheng, Michael E. Sieracki and Ramunas Stepanauskas

Sequencing microbial genomes from single cell is one of the emerging technologies for sequencing bacterial that cannot be cultured. We developed protocols for fluorescence activated cell sorting to separate single bacterial cell, amplification of its DNA with whole genome amplification kits from Qiagen, and sequencing the genomes. Recently, we've been finishing two genomes with DNA amplified from single cell that was flow sorted from seawater near Bigelow Laboratory in Maine. Currently one is

about 1.5 Mb and another 2 Mb with estimation of 60 - 80 % coverage of the genomes. The longest contig is close to 700 kb. The sequencing results show the procedures for cell isolation, DNA amplification, and sequencing are robust with minimal contamination. A third genome containing 100% identical rRNA sequence has been amplified and will be sequenced later. Specific problem in finishing genomes from single cell amplification are significant coverage bias in the amplified genomic DNA, higher chimeric rate due to amplification. We applied not only the conventional finishing methods such as the primer walk, adapter PCR, but also using metagenomic data from the Global Ocean Sampling to closing the gaps in the single cell projects. It is found that genomic sequences from the two single cells are abundant in the data sets of samples collected from the Northeast US coast.

MetaBar: A Tool for Consistent Contextual Data Acquisition for Ecological Genomics

Wolfgang Hankeln, Max Planck Institute for Marine Microbiology

In molecular microbial ecology, a higher level of data analysis is facilitated by complementing sequence data with the physical-chemical and biological parameters (contextual data) describing the sample. A consistent and integrated dataset covering the diversity and genetic potential of organisms in their environmental context makes it possible to tackle questions such as “Who is out there?”, “How many of what kind are there?” and “What are they doing?” or even “Where, and under which environmental conditions, can certain genes be found?”. This goal can be achieved if investigators are able to efficiently capture, merge and share sequence and contextual data gathered in the field. The MetaBar (<http://www.megx.net/metabar>) tool has been developed to accomplish Consistent Contextual Data Acquisition (CCDA) on a routine basis. As a multiuser web application it is designed to assist researchers worldwide in consistently storing contextual data, such as longitude, latitude, depth/altitude and time (x, y, z, t), and other habitat parameters. MetaBar assigns unique identifiers to samples, generates barcode labels, and provides uploadings of contextual data via spreadsheets. Tracking of samples is simplified by standardized barcode labeling with global unique identifiers. MetaBar is part of the megx.net data portal (www.megx.net), allowing visualization of the sample data on the Genes Mapserver (www.megx.net/gms). It is compliant with the contextual data standards for genomes (MIGS) and metagenomes (MIMS) proposed by the Genomic Standards Consortium (<http://gensc.org>).

Molecular characterisation of the bacterial flora in the lung of healthy, asthmatic and chronic obstructive pulmonary diseased (COPD) subjects

Markus Hilty, Imperial College, London

Co-authors: Conor Burke, Len Poulter, Miriam F. Moffatt and William O.C. Cookson

Microbiological cultures have often been used to describe bacterial infections as the causes of exacerbations

in chronic obstructive pulmonary disease (COPD) and Asthma. However, current culture technologies might be limited as unculturable bacteria are neglected and possible causative agents overseen. In our study, we applied a culture-independent molecular approach, based on 16S small-subunit ribosomal RNA sequencing, to protected specimen brushes (PSB) samples of 24 adult subjects with Asthma (11), COPD (5) and no diseases (8). After collecting the PSBs, DNA was extracted, quantified and adjusted. Different primers were evaluated and 16S clone libraries produced. For subsequent DNA sequencing, 1188 clones were picked of which 1070 resulted with a high quality sequence of exact the same length.

Preliminary results revealed 118 operational taxonomic units (OTUs) at the level of 99% similarity. The phyla Proteobacteria, Bacteroidetes and Firmicutes dominated the microbiological lung flora. Certain single bacterial types were characteristic in a few but not all asthmatic and COPD patients while disease control cases contained normally a higher bacterial diversity. In conclusion, molecular techniques provide a much better picture of the bacterial flora in the lung than more traditional techniques and offer a wide number of possibilities for their investigations in health and disease.

Predicting genes in short metagenomic sequencing reads with high specificity

K. J. Hoff, University of Göttingen

Co-authors: M. Tech, P. Meinicke

The prediction of protein coding genes is an important step in analysing sequenced metagenomes. In contrast to genomic sequence data, a large proportion of metagenomic sequencing reads cannot be assembled reliably into contigs and remains as single short fragments. Conventional gene prediction methods are not suitable for predicting genes in phylogenetically undetermined, short metagenomic sequencing reads. We recently introduced a machine learning approach based on neural networks that achieves accurate and fast gene predictions in anonymous Sanger read-length fragments (BMC Bioinformatics 9:217). Linear discriminants, trained on 140 annotated prokaryotic genomes, were used to extract features from open reading frames (ORFs). An artificial neural network combines these features with ORF length and fragment GC-content to compute a posterior gene probability. The neural network was trained on randomly excised 700 bp fragments from the above prokaryotic genomes. The overall performance of our method is similar to that of MetaGene (Noguchi et. al, 2006, NAR 34(19):5623-5630). In detail, the neural network showed a higher prediction specificity while MetaGene is more sensitive. A high specificity is advantageous if large-scale metagenomic studies do not allow manual validation of the predictions. However, on fragments shorter than 500 bp, we observed a significant performance loss.

Here, we show that our method performs well on DNA fragments as short as 300 bp if the neural network is trained on shorter fragments. This finding indicates that

our method could be valuable for gene prediction in pyrosequenced metagenomes.

Integration of Phenotypic Metadata and Protein Similarity in Archaea Using a Soft Clustering Approach

Sean D Hooper, Joint Genome Institute

Co-authors: Iain J Anderson, Amrita Pati, Daniel Dalevi, Kostantinos Mavromatis, Nikos C Kyrpides

In order to simplify and meaningfully categorize large sets of protein sequence data, it is commonplace to cluster proteins based on the similarity of those sequences. However, it quickly becomes clear that the sequence flexibility allowed a given protein varies significantly among different protein families. The degree to which sequences are conserved not only differs for each protein family, but also is affected by the phylogenetic divergence of the source organisms. Clustering techniques that use the same pre-defined similarity threshold value for all protein families cannot allow for these variations and thus cannot be confidently used for applications such as automated annotation and phylogenetic profiling.

In this work, we present a novel protein clustering method which provides meaningful groupings by taking into account both sequence flexibility and the effects of phylogenetic divergence. This approach was applied to all proteins from 46 archaeal genomes. Comparisons between different taxonomic levels allowed us to study the effects of phylogenetic distances on cluster structure. Likewise, by associating functional annotations and phenotypic metadata with each protein, we could compare our protein similarity clusters with both protein function and associated phenotype.

Integrating our soft-clustering method with functional annotations, phylogeny, and associated phenotype adds further value to the protein clusters generated. For example, this enabled us to identify protein clusters that are independent of phylogenetic distance due to strong sequence conservation, clusters that are specific to a particular phenotype, and clusters with a probable history of lateral gene transfer. We are also offering an intuitive, graphical online tool (at <http://coal.jgi-psf.org:8180/coal>) where these clusters can be analyzed and explored further.

Inter-individual variation in the human metabolic phenotype of equol production is associated with variation in gut microbial communities

Meredith A J Hullar, Fred Hutchinson Cancer Research Center

Co-authors: Charlotte Atkinson, Fei Li, Johanna W Lampe.

Inter-individual differences in human intestinal bacterial metabolism of dietary components may influence human health. For example, 30-50% of people harbor intestinal bacteria capable of converting the soy isoflavone, daidzein, to equol. Equol-producer status has been associated inversely with the risk of several hormone-dependent cancers. We examined the microbial community composition

(MCC) in equol producers (EP; n=19) and equol non-producers (ENP; n=15) using terminal restriction length polymorphism (tRFLP). Genomic DNA was extracted from frozen fecal samples and, the 16S rRNA gene was amplified using primers 27F-FAM and 1492R, digested with ALU I, and analyzed by tRFLP. To identify microbiota represented by peaks in the tRFLP traces, 16S rRNA genes were cloned from fecal samples from EP and ENP, sequenced, and compared using neighbor-joining analysis. Non-metric multidimensional scaling (NMS) of tRFLP patterns resolved the MCC from EP versus ENP and explained 80 % of the variation in the fecal data. Multi-response permutation procedures (MRPP) showed that the MCC was significantly different between the EP and ENP ($P=0.028$). In addition, within EP, there were significantly different MCC associated with the same range of urinary equol concentration. Using a high throughput fingerprinting technique, we have identified unique gut microbial communities associated with the human metabolic phenotype, equol production.

Functional and phylogenetic characterisation of the gut of the asian elephant with a metagenomic approach

Nele Ilmberger, University of Hamburg

Co-authors: Julia Pottkaemper, Christel Schmeisser, Wolfgang R. Streit

Metagenomics offers the possibility to explore the uncultivated bacteria for new biocatalysts and to advance our knowledge on the microbial ecology of complex communities. We have initiated work to characterise the metagenome of the gastrointestinal tract of the asian elephant (*Elephas maximus indicus*) via 16S rDNA analysis and construction of a large insert library. As faecal microflora is very diverse and the elephant lives on substrates rich in biopolymers, this habitat is supposed to comprise microbes with hydrolytic enzymes in high quantity and diversity. The 16S rDNA analysis supports the assumption that the community of the gastrointestinal tract of the asian elephant is highly diverse. About 50% of the sequences show highest homology to uncultivated organisms.

To further analyse this microbial community and to screen for biocatalytic genes we constructed a metagenomic library containing 6,800 clones in the mobilizable cosmid vector pLAFR3. The library was conjugated via triparental mating from *E. coli* to the alpha-Proteobacterium *Rhizobium* sp. NGR234. Both libraries were screened for cellulolytic clones. Three putative positive clones could be identified in both hosts. Six additional putative positive clones were identified in *E. coli*, seven in *Rhizobium*. Sequencing of putative genes is underway. To further advance the knowledge of the enzymes present in the gastrointestinal tract, we are constructing a cDNA library, subsequently analysed by 454 sequencing. Furthermore, we initiated work for the establishment of a screening system which enhances the detection potentiality of enzymes produced by gram-positive bacteria which can hardly be detected in common screening systems.

Life in deep bedrock aquifers – metagenomics of deep borehole biosphere

Merja Itävaara, VTT Technical Research Centre
Co-authors: Mari Nyysönen, Aura Nousiainen, Lasse Ahonen, Petri Auvinen, Ilmo Kukkonen

Deep biosphere of Outokumpu deep borehole (2500 m) located in eastern Finland is under intensive research. Fractures of the Early Proterozoic (about 1900 million years old) crystalline bedrock of the site contain saline waters; total salinity up to tens of grams/L, main components being Ca-Na-Cl). The deep saline waters have typical anoxic methanic characteristics; no measurable oxygen, sulphate below detection limit and dissolved methane in abundance. Microbial sampling has revealed the existence of microorganisms in depths of more than two kilometers. Characterization of the microbial communities is underway. The microbial community composition in the bore hole is investigated by direct pyrosequencing of Bacterial and Archaeal 16S rRNA gene fragments PCR amplified from different depths.

Genome sequence analysis of *Bifidobacterium bifidum* and *B. animalis* subsp. *lactis*

Haeyoung Jeong, Korea Research Institute of Bioscience and Biotechnology (KRIBB)
Co-authors: Dong-Su Yu, Sang-Haeng Choi, Dae-Won Kim, Myeong-Soo Park, Geun Eog Ji, Hong-Seog Park, Tae Kwang Oh, and Jihyun F. Kim

Microbes living in the human gut are considered to play a pivotal role for supporting human health by preventing pathogen colonization, degrading non-digestible compounds, producing vitamin K, facilitating absorption of ions, and enhancing gut mucosal immunity. Bifidobacterial species are natural inhabitants of human and animal gut forming a phylogenetically distinct group with characteristic biochemical and probiotic properties.

To elucidate their health-promoting effects and host-microbe interaction mechanisms at the genome level, we determined the complete genome sequences of two bifidobacterial species isolated from human fecal samples. *Bifidobacterium bifidum* BGN4, exhibiting a prominent adhesive capacity for intestinal epithelial cells, has been demonstrated to have high immunomodulatory activities in the mouse model. We applied 454 pyrosequencing technology for genome sequencing and gaps were closed by multiplex PCR methods without any conventional plasmid/fosmid clone-derived end sequencing. For genome sequencing of *B. animalis* subsp. *lactis*, a probiotic bacterium that also shows high immunomodulatory activities, a standard Sanger whole-genome shotgun method was carried out. Both bifidobacterial genomes contain a circular compact chromosome with no plasmid (ca. 2 Mb). Comparative genome analysis shows a high proportion of genes that are conserved among bifidobacteria, but a high level of genome rearrangement was observed with a limited syntenic relationship between species. Detailed genome analysis is underway to identify genes involved

in their probiotic effects such as interactions with the host, biosynthesis of chiro-inositol, and metabolism of bifidogenic factors.

Metagenomic Analysis of Airborne Microorganisms in Yellow Sand in Korea

GwangPyo Ko, Seoul National University
Co-authors: S. Lee, B. Choi, S. M. Yi and G. Ko

Yellow Sand is a seasonal meteorological phenomena affecting East Asia in early spring. It has been receiving increased attention. Since large amount of dust are moving from Desert in China to Korea and further to Japan. There are fair amount of chance to include various microorganisms in yellow storm dust for long range travel. However, even though there was a fair amount of chemical analysis, very little has been microbiologically studied for analyzing yellow storm dust. Hence, we collected 30 microbiological air samples by PM2.5 cyclone sampler in Seoul, Korea from April, 2007 to March, 2008 and analyzed the microbiological air samples. Six samples of yellow dust samples and 24 samples of non-yellow dust samples were analyzed. In addition to chemical analysis by Terminal-optical transmission (TOT), total nucleic acids in collected samples were extracted, and 16S rDNA was amplified by PCR and analyzed by denaturing gradient gel electrophoresis (DGGE). Banding pattern of DGGE were analyzed by Bionumerics® software, and the major band of DGGE were subsequently sequenced and analyzed by BLAST. Dendrogram based on DGGE indicated that 6 samples of yellow storm were clustered each other. Microorganisms identified in yellow sand including *Aquabacteriu* sp., *Dietzia* sp., and *Bacillus* sp., etc whereas microorganisms in non-yellow sand included *Propionibacterium* sp., *Denitratisoma* sp., *Hymenobacter* sp., *Sphingomonas* sp., and *Acinetobacter* sp. etc. Our study demonstrated that different microorganisms in yellow storm were exposed to human in this region and could be important factor for affecting human health in this region.

Environmental Transcriptomics Of Southern Ocean Phytoplankton Assemblages

Asuncion Lago-Leston, Universidade do Algarve
Co-authors: Ester Serrao, Gareth Pearson

The Southern Ocean is one of the worlds major high-nutrient low-chlorophyll oceanic regions, considered to have had an enormous effect on global climate over geological time. Due to their role in macronutrient fluxes and carbon export, understanding phytoplankton community function and regulation are critical for modeling global biogeochemical cycles. The objective of the project will be to characterize community-level transcriptomes in natural Southern Ocean phytoplanktonic assemblages in response to experimental manipulation of major bottom-up drivers of ecosystem function (Fe/Si/light, UVR), focusing particularly on diatom-dominated assemblages. This will be achieved with deep EST coverage and/or SAGE profiling using high-throughput pyrosequencing technology. By preserving and including community-level com-

plexity in our experiments, the identification/annotation of differentially expressed transcripts will provide insights into genotype-environment interactions that include those arising from biotic interactions, in Southern Ocean planktonic ecosystems.

Metagenomic Finishing at JGI

Alla Lapidus, Joint Genome Institute

Co-authors: Alicia Clum, Eugene Goltsman, Phil Hugenholtz, Stephen Lowry, Susan Lucas, Hector Garcia Martin

Normally, complete genomes are obtained by growing the organism of interest in pure culture, generating the shotgun sequencing and closing the gaps. In case of metagenomic samples, it is difficult to expect a completed, ungapped genome, considering the organism is being sequenced directly from the environment. Despite this fact, finishing of three dominant populations within the metagenome datasets of different complexity that had draft level coverage was successfully performed. This was possible due to 1) the enrichment of the target organism in the population; 2) generation of draft sequence using traditional metagenomics; 3) computational identification of sequences derived from the target organism; 4) gap closure; 5) use of pyrosequencing. The complexity of the community, the quantity of genomic DNA available as well as the size of the fraction of the total DNA, which represented the organism under study all added on to the normal difficulty level of having to establish a complete sequence. At the JGI we have managed to complete the sequence of:

- *Candidatus Korarchaeum cryptofilum* OPF8. - low complexity case; single contig 1.59Mb in length with a GC content of 49%;
- *Desulforudis audaxviator* - very limited quantity of source material; 454 pyrosequencing to compensate for the cloning bias of Sanger libraries
- *Candidatus Accumulibacter phosphatis* Type IIA str. CU-1. A. *phosphatis* - complex project. Binning/reassembly approach was developed and used to complete this genome (5Mb chromosome and with 3 plasmids of 167, 42 and 38kb).

Performed finishing and projects in progress will be presented in more details.

Discovery of novel mosquito-associated viruses using metagenomics

Yan Wei Lim, Genome Institute of Singapore

Co-authors: Christina Nilsson, Dana Hall, Robert Schmieder, Forest Rohwer, Chia Lin Wei, Yijun Ruan

Mosquitoes are vectors of viral diseases such as Dengue fever, Yellow fever, and Japanese encephalitis. Mosquitoes may be associated with a much larger spectrum of potential viral pathogens, the majority of which have not yet been identified. Traditional methods of identifying and characterizing viruses rely largely on culture-dependent or PCR-based techniques which, in addition to being time and labor intensive, are biased and provide only minimal opportunities for viral discovery. Here, we have used a metagenomic approach that includes isolation of the full

cohort of mosquito-associated viruses coupled with high-throughput DNA sequencing. Our goal was to identify all the viruses carried by mosquitoes. We have collected mosquitoes from two distinct and geographically distant locations. Three metagenomic libraries were generated from mosquitoes caught in San Diego, California, and two metagenomic libraries were generated from mosquitoes caught in Singapore, which were sequenced using the Roche GSFLX platform. Similarity-based sequence analysis indicated that all libraries contained a variety of viruses associated with humans, insects, plants, and protozoans, indicating successful isolation of total DNA from viral particles. Metagenomic comparisons between the Singapore and San Diego samples demonstrated different viral communities, reflecting the different biogeographic conditions of the two environments.

UniMES: a repository for Metagenomic and Environmental Sequences

Maria Martin, European Bioinformatics Institute

Co-author: UniProt Consortium

The Universal Protein Resource (UniProt) provides a stable, comprehensive, freely accessible, central resource on protein sequences and functional annotation. One of the UniProt components is the UniProt Metagenomic and Environmental Sequence database (UniMES). UniMES provides the metagenomics community with a repository for public deposition of their data allowing the enrichment of protein information by well-established analytical tools. UniMES currently contains data from the Global Ocean Sampling Expedition (GOS) which was originally submitted to the International Nucleotide Sequence Databases (INSDC). The initial GOS dataset is composed of 25 million DNA sequences primarily from oceanic microbes and predicts nearly 6 million proteins. UniMES provides free access to metagenomic data combining the predicted protein sequences with automatic classification by InterPro, the integrated resource for protein families, domains and functional sites. Additionally, the UniProt automatic annotation system for functional annotation of newly predicted protein sequences is being extended to UniMES and will be presented in this poster. UniProt also provides UniMES clusters of sequences at two resolutions (100% and >90%) using the CD-HIT algorithm (Li W., Jaroszewski L., and Gofzik A, *Bioinformatics*, 17:282-283, 2001). UniMES is available in the UniProt FTP (<ftp://ftp.uniprot.org>) site in FASTA format with a UniMES matches to InterPro methods file.

Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles

Ana-Belén Martín-Cuadrado, Universidad Miguel Hernández

Co-authors: Elena Ivars-Martínez, Giuseppe D'Auria, Francisco Rodríguez-Valera

Alteromonas macleodii is a widespread marine heterotrophic bacterium. Its main niche is particulate material

in the water column where it can grow relatively fast due to the high, localized nutrient concentrations. We describe the genome of the strain *A. macleodii* AltDE isolated from 1000 m deep in the Mediterranean and compare its genome with that of the type strain ATCC27126, a surface isolate. Both genomes have also been compared to the Global Ocean Survey (GOS) database. The association of this species to particles and relatively warm waters was confirmed and AltDE was proportionately more represented in cooler waters.

The genomes indicate that AltDE is better adjusted to live in more crowded conditions with much more evidence of phage predation, better adhesion and resistance to toxic heavy metals. Metabolically AltDE seems better suited for degradation of recalcitrant compounds and urea, and for being exposed to microaerophilic conditions by specialized respiratory chains and nitrate respiration. On the other hand, the surface isolate had more capabilities for regulation and motility, consistent with a more heterogeneous environment and degraded more sugars and amino acids. A large part of the differential gene content is found in islands of over 20 Kbp that generally also recruit poorly in the GOS database. Overall the image depicted by both genomes indicate that ATCC27126 represents lineages specialized in marine surface, nutrient-rich environments with better chances to survive in the free-living phase, while AltDE would display more chances in larger particles that sink rapidly to meso and bathypelagic depths.

Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions

Adam C. Martiny, UC Irvine

Co-authors: Ying Huang and Weizhong Li

The cyanobacterium *Prochlorococcus* is the numerically dominant phototroph in oligotrophic parts of the oceans. Recently, it was shown that the distribution of phosphate acquisition genes did not match the 16S rRNA phylogeny among isolates from this group but rather appeared related to phosphate availability where the strains had been isolated.

To further understand adaptation to phosphate limitation in *Prochlorococcus*, the distribution of phosphate acquisition genes was investigated in different ocean regions and related to local ortho-phosphate concentration. In regions characterized by less than 0.1 μM phosphate, most *Prochlorococcus* cells contain genes involved in phosphate uptake, regulation and utilization of organic phosphates. In contrast, most of these genes are absent in regions with more than 0.1 μM phosphate with the exception of genes involved in transport of phosphate (*phoE* and *pstABCS*) and three genes of unknown function. This pattern of phosphate acquisition genes showed no significant correspondence to the distribution of rRNA phylotypes. In addition, it was demonstrated that several genes in a separate genomic island were commonly present in low P sites while absent in high P sites. Overall, this study further demonstrates a linkage between envi-

ronmental conditions in the ocean and genome content of *Prochlorococcus*.

Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites: MIRADA (LTERS)

Elizabeth McCliment, Marine Biological Laboratory, Woods Hole

Co-authors: Susan Huse, Linda Amaral-Zettler

Understanding microbial ecosystem processes requires information about the population structures of microbial communities including estimates of richness and relative abundance (evenness), and the relationship of this diversity to the underlying physical and chemical environment. In 2007, we established a biodiversity survey focused on Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research sites (MIRADA-LTERS) in the NSF US LTER program with the long-term objective of documenting and describing baseline diversity of microbial communities across each of the 13 aquatic sites. Our project employs high-throughput pyrosequencing technology based on the comparison of hypervariable regions in the small subunit ribosomal RNA gene, enabling the discovery of novel microbial diversity in the bacterial, archaeal and eukaryal domains, and the creation of high-resolution microbial population structure profiles under different ecological and biogeographical constraints. The scope of this project initially necessitated the development of variations on currently used methodologies to retrieve and characterize the broadest possible cross-section of diversity across all three domains. Our preliminary results reveal substantial diversity within microbial populations of our first LTER test sites, confirming the presence of low-abundance taxa comprising the 'rare biosphere' in a community otherwise dominated by relatively few major groups. The data resulting from this project will directly benefit each of the participating LTERs and allow for broad cross-site comparisons of microbial diversity, as well as valuable baseline data for integrating microbial population structure with ecosystem change.

Hydrogenobaculum Population Genomics: Linking Phylogeny, Genetics, and Ecologic Function

Timothy R. McDermott and Scott Clingenpeel, Montana State University

Background: *Hydrogenobaculum* inhabiting the Yellowstone Geothermal complex are the subjects of a population genomics study currently underway. Several pure culture isolates that are phylogenetically 100% identical (full-length 16S rDNA sequence and ITS sequence), but that differ in their ability to utilize H₂ and/or H₂S for growth are being genome sequenced. This effort is in parallel with a metagenomic thrust of the mat community that these organisms dominate. Based on extensive chemical analysis, we know these phenotypes are all relevant to the environment these organisms inhabit, which is chemostat-like in nature with respect to temperature, pH, and flux of electron donors. Importantly, however, their habitat is also

comprised of overlapping temperature and geochemical gradients that provide a continuum of niche opportunities that theoretically could support the emergence and maintenance of various genetic alterations that might account for the above-described physiologic diversity.

This project is in its early stages. Draft genome coverage of one isolate is being examined for specific functions of interest and initial available metagenome reads are also being assessed. Genome sequence revealed novel sequences coding for arsenite oxidases and that have been used for expression and diversity analysis. Metagenome sequencing uncovered significantly more *Hydrogenobaculum* phylotypes than previously observed with traditional PCR-based cloning. These and other current developments will be summarized. This work has been supported by the NSF Microbial Observatories and DOE-JGI-Community Sequencing programs, and the NASA-supported Thermal Biology Institute.

An Uncultivated Crenarchaeota Contains Functional Bacteriochlorophyll a Synthase

Jun Meng, Xiamen University
Co-authors: Fengping Wang

A fosmid clone 37F10 containing an archaeal 16S rRNA gene was screened out from a metagenomic library of Pearl River sediment, southern China. Sequence analysis of the 35 kbp inserted fragment of 37F10 found that it contains a single 16S rRNA gene belonging to Miscellaneous Crenarchaeotal Group (MCG) and 36 open reading frames (ORFs). One ORF (orf11) encodes putative bacteriochlorophyll a synthase (bchG) gene. Bacteriochlorophyll a synthase gene has never been reported in a member of the domain Archaea, in accordance with the fact that no (bacterio)-chlorophyll has ever been detected in any cultivated archaea. The putative archaeal bchG (name as ar-bchG) was cloned and heterologously expressed in *Escherichia coli*. The protein was found to be capable of synthesizing bacteriochlorophyll a by esterification of bacteriochlorophyllide a with phytyl diphosphate or geranylgeranyl diphosphate. Furthermore, phylogenetic analysis clearly indicates that the ar-bchG diverges before the bacterial bchGs. Our results for the first time demonstrated that a key and functional enzyme for bacteriochlorophyll biosynthesis does exist in Archaea.

Marine medicine, learning from complex barriers and microbiota in the ocean

Rebekka Metzger, Christian-Albrechts-Universität Kiel
Co-authors: Diana Meske, Nancy Weiland, Ruth Schmitz

Marine microbial communities are highly diverse and have evolved during extended evolutionary processes of physiological adaptations under the influence of a variety of ecological conditions and selection pressures. They harbor an enormous diversity of microbes with still unknown and probably new physiological characteristics and are thus rich sources for isolating novel bioactive compounds and genes. The surfaces of marine organisms are typically

covered by a consortium of epibiotic bacteria and act as barriers where diverse interactions take place. These interactions between microorganisms and hosts can be disadvantageous as well as beneficial resulting in specifically associated microorganisms on the host tissue. Therefore, the microbial consortia on marine multicellular host tissues are attractive model systems to understand the complex interplay between microbes and their host cells that may be also relevant to the human barrier organs and its microbiota. Thus, insights into ancient mechanisms of host/microbial interactions may allow understanding human barrier disorders and may provide insight into the development of diseases in humans and identify new drug targets. First results studying microbial consortia on different marine multicellular host tissues including a metagenomic approach will be presented and discussed.

Comparative Metagenomics with MEGAN 2.0

Suparna Mitra, University of Tuebingen
Co-authors: Daniel C. Richter, Alexander F. Auch, Stephan C. Schuster and Daniel H. Huson

Metagenomics is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation & evolution in different environments. A main promise of metagenomics is that it will accelerate drug discovery and biotechnology by providing new genes with novel functions. The recent development of ultra-high throughput sequencing technologies, which do not require cloning or PCR amplification & can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects. Increasingly, there is a need for new ways of comparing multiple metagenomics datasets & for fast and user-friendly implementations of such approaches. Methods & results: Here we describe some new features of our application MEGAN[1,2] which has been developed to fit large metagenomic datasets onto a taxonomical tree, thus leading to a visualization of the biodiversity of a given sample. MEGAN provides many options for fine-tuning thresholds for high-scoring segment pair selection and taxon matching. Some new features such as 'rate of discovery', 'functional assessment' & 'meta-data analysis' has enhanced these analyses. In addition to existing features of MEGAN the new version 2.0 allows the comparative analysis of different datasets that can be brought together and compared for taxonomic and functional content. Conclusion: A main goal is to provide a tool that can be used to perform large analysis efficiently on a laptop. We have developed an interactive and fully customizable chart viewer for MEGAN2.0 that allows one to extract a number of different comparisons directly from the multiple comparison tree view.

[1] Daniel H Huson, Alexander F Auch, Ji Qi & Stephan C Schuster. Megan analysis of metagenomic data. *Genome Res*, 17(3):377–386, Mar2007.

[2] Daniel H Huson, Daniel C Richter, Suparna

Mitra, Alexander F Auch & Stephan C Schuster. Methods for Comparative Metagenomics. Submitted to APBC 2009.

Exploring fiber degradation by rumen microbial communities

Christina D. Moon, AgResearch, Ltd.
Co-authors: Dong Li, Carrie Sang, Eric Altermann and Graeme T. Attwood

The rumen is the fermentative forestomach of ruminant animals and is densely populated with a diverse community of anaerobic microbes. Collectively, the rumen microbiota is responsible for degrading forage material and providing simpler substrates for use by the host. To do this, these microbes encode various enzymatic activities to break down complex structural polysaccharides, such as those found in plant cell walls. We seek to gain a more comprehensive understanding of the variety of mechanisms employed by rumen microbes to mediate plant biomass conversion in the rumen. However, the vast majority of rumen microbes have not been cultured, and so, metagenomic analyses provide a powerful means to explore the rumen microbiota. The rumen contents of two fistulated dairy cows grazing a ryegrass and clover diet were fractionated to obtain the liquid, plant-associated and plant-adherent microbial communities. Metagenomic DNA was isolated from the plant-associated and plant-adherent fractions and used to construct large-insert fosmid libraries. Over 11,500 clones were arrayed and screened on artificial substrates, revealing α -1,4-endoglucanase activities (expressed by 0.24% of clones), α -1,4-endoxylanase activities (0.41%), cellobiohydrolase activities (0.47%) and α -glucanase activities (0.63%). These activities were more abundant in the plant-adherent library compared to the plant-associated. The identification of active clones will enable the genes and enzymes conferring these activities to be identified and further characterized. In addition, we are currently examining rumen metagenomic DNA sequence data, which, together with functional data, will further expand our knowledge of the mechanisms and diversity of enzymes involved in ruminal fiber degradation.

Functional variation analysis of human gut metagenomic data

Kozo Nishida, Nara Institute of Science and Technology
Co-authors: Hiroshi Mori, Hideki Noguchi, Takeaki Taniguchi, Yoshitoshi Ogura, Masahira Hattori, Tomomi Kuwahara, Takehiko Itoh, Tetsuya Hayashi, Shigehiko Kanaya and Ken Kurokawa

One of the goals of metagenomics is to declare functional characterizations of microbial communities in each environment. Although recent studies have tried to achieve the purpose, still we have little knowledge about it. This is probably because of the methodological difficulties of detailed analysis of each member from metagenomic data. Here, we present the clear method to estimate the functional variations among microbial communities. Two American and thirteen Japanese gut metagenomic data

were used for our analysis. Predicted genes using MetaGene were assigned to the KEGG pathway and the SEED databases using BLASTP. In order to identify specific metabolic characters of each sample, we compared the mapped results among pathways or subsystems. We will present the differences of metabolic characters among microbial communities.

Metagenomic analysis of plasmids in coastal marine cyanobacterial populations

Brian Palenik, Scripps Institution of Oceanography, UC San Diego
Co-authors: E. Latham, B. Brahmsha, V. Tai, Q. Ren, I.T. Paulsen

Because they do not require culturing, metagenomic approaches have the potential to reveal the genetic diversity of the microbes actually present in an environment, including the contribution to diversity of mobile elements such as plasmids. From coastal California seawater, a complex and diverse environment, the marine cyanobacteria of the genus *Synechococcus* were enriched by flow cytometry-based sorting and the population metagenome was analyzed with 454 sequencing technology. Interestingly, at least three distinct mobile DNA elements (plasmids) not found in model *Synechococcus* strain genomes were detected in the assembled contigs. Two of these had some similarity to plasmids from other cyanobacteria, suggesting for the first time the likely importance of plasmids in marine cyanobacterial populations and their possible role in horizontal gene transfer. Further analysis of natural samples confirmed the presence of plasmids and is shedding light on their diversity.

Statistical modeling of the relative abundance of protein functions in metagenomic datasets

Amrita Pati, Joint Genome Institute
Co-authors: Daniel Dalevi, Natalia Ivanona, Ernest Szeto, Victor Markowitz, Phil Hugenholtz, and Nikos Kyrpides

Microbial populations evolve complex biological processes to cope with specific impacts of their natural habitats. Large-scale environmental metagenomic datasets with functional annotations provide significant contextual insights to fill gaps in the understanding of relationships between environmental factors, gene functions, and phenotypes of microbial communities. Members of a protein family typically have high sequence similarity and similar functions. In order to compare metagenomic datasets generated from different microbial communities, the systematic evaluation of the relative abundances of individual or groups of protein families can yield statistically significant deductions about the over- and under-representation of the protein function(s) and their corresponding biological pathways within these communities.

In this work, we present novel statistical methods for comparing the relative abundances of protein families in two metagenomic datasets. First, we propose a statistical model for modeling the abundance of a given protein fam-

ily and develop a method for identifying protein families with statistically significant abundance variation in the two datasets. Then, we extend this method to groups of protein families. Finally, we illustrate the application of these methods in the comparison of metagenomic datasets within the Integrated Microbial Genomes with microbiomes (IMG/M) system.

The Identification of Genomic Islands Harboring Phage and Secondary Metabolites in the Genomes of Obligate Marine Actinobacteria and a New Mechanism of Marine Adaptation

Kevin Penn, Scripps Institution of Oceanography, UC San Diego

Co-authors: Caroline Jenkins, Brad Moore, Dan Udwarý, Alla Lapidus, Sheila Podell, Eric Allen, Paul R Jensen

A pattern that has emerged from the comparative analysis of closely related bacterial genomes is the presence of genomic islands that house genes associated with ecological adaptation. In this study comparative genomics has been used to infer differences between the obligate marine actinobacteria *Salinispora tropica* (ST) and *S. arenicola* (SA) both isolated from marine sediments. SA and ST contain 1320 and 936 unique genes, respectively, and the majority of these (79 and 73%, respectively) reside in 21 (>20 kb) genomic islands. Phage and secondary metabolite encoding genes represent a major component of the islands and are likely associated with ecological differentiation of the two species. The spacers from Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) in SA have exact matches to genes from prophage in ST suggesting that SA has been exposed to and developed immunity to phage that infect ST. Both genomes contain 30 duplicated polymorphic membrane proteins (Pmp) that have no significant similarity to other Actinobacteria genes and may provide a new mechanism of marine adaptation. Pmps likely act as autotransporters that form pores in the cell membrane preventing *Salinispora* from osmoregulating in low osmotic environments. Fragment recruitment from CAMERA shows metagenomic matches to *Salinispora* originate mostly from open water sites while *Streptomyces* largely recruit metagenomes from terrestrial influenced ocean environments.

Diel Environmental Metatranscriptomics at the Hawaii Ocean Time-Series and the Sapelo Island Microbial Observatory

Rachel S. Poretsky, University of Georgia

Co-authors: Ian Hewson, Shulei Sun, Jonathan Zehr, Mary Ann Moran

Metatranscriptomic analyses of microbial assemblages from surface water at the Hawaiian Ocean Time-Series (HOT) and the Sapelo Island Microbial Observatory (SIMO) revealed community-wide metabolic activities and diel patterns of differential gene expression. Following isolation of total RNA and removal of rRNA, mRNA was amplified linearly by in vitro transcription and then converted to double-stranded cDNA for sequencing by 454

pyrosequencing. Sequencing of HOT and SIMO mRNA resulted in more than 100,000 reads from both night and day samples, 35- 50% of which was rRNA-derived. Experimental metatranscriptomics at SIMO using dissolved organic carbon (DOC) amendments identified pathways used by coastal bacterial communities in transforming vascular plant- and algal-derived DOC in seawater. Taxonomic binning of the sequences showed a dominance of genes expressed by Proteobacteria and Cyanobacteria. Expressed gene bins representing several taxa of heterotrophic bacteria including the Roseobacter and SAR11 groups indicated a significant representation of predicted highly expressed genes, as identified by genomic analysis of cultured relatives, in the metatranscriptomic libraries. Statistical comparisons of the day vs. night libraries revealed differences in relative representation of COGs and KEGG pathways, reflecting the conditions at the time of sampling as well as DOC amendments. Our results demonstrate the application of environmental metatranscriptomics to the examination of dominant and differential gene expression in marine microbial assemblages.

Annotation, Classification and Comparative Analyses of Membrane Transporters from Environmental Metagenomic Samples

Qinghu Ren, J. Craig Venter Institute

Co-authors: Ian T. Paulsen

In recent years, metagenome-based approaches have led to the accumulation of an increasing number of gene sequences from the uncultured microbes from different environments. Using the metagenome sequences to fully understand how complex microbial communities function and how microbes interact within these niches represents a major challenge for microbiologists today. A large percent of predicted metagenomic genes are expected to be transporters, which play crucial roles in fundamental cellular processes and functions in all prokaryotic and eukaryotic organisms. Systematic annotation and analyses of these transporter genes are of great value to increase our knowledge on microbial ecology and physiology of these microbes.

The transporter annotation of metagenomic sequence data is challenging. The source organism of a sequence is typically unknown. Due to the "cosmopolitan" nature of metagenomic samples, a large fraction of the protein sequences predicted in the metagenomic data will be fragmentary. We significantly improved our transporter annotation pipeline (TransAAP), a web-based transporter annotation tool, to handle the volume, complexity, heterogeneity and fragmentary nature of the metagenomic data. We employed various sequence-based and domain-based bioinformatics searches to maximize the possibility of predicting all the potential candidates of transporter genes out of metagenomic samples. We plan to include information such as transporter type and family classification, substrate prediction, the confidence levels, Pfam/TIGRfam hits and values, COG hits and value, the top hit to the non-redundant protein database and the top hit species,

as well as a pre-computed phylogenetic tree with known transporters in the same family. A comparative analysis of the transporter profiles of metagenomic samples from the different environment will illustrate the relationship between the transporter features and organism's physiology and living environment.

MetaSim: A Sequencing Simulator for Genomics and Metagenomics

D.C. Richter, University of Tuebingen
Co-authors: F. Ott, A.F. Auch, D.H. Huson

The new research field of metagenomics is providing exciting insights into various, previously unclassified ecological systems. Next-generation sequencing technologies are producing a rapid increase of environmental data in public databases. There is great need for specialized software solutions and statistical methods for dealing with complex metagenome data sets. To facilitate the development and improvement of metagenomic tools and the planning of metagenomic projects, we introduce a sequencing simulator called MetaSim. Our software can be used to generate collections of synthetic reads that reflect the diverse taxonomical composition of typical metagenome data sets. Based on a database of given genomes, the program allows the user to design a metagenome by specifying the number of genomes present at different levels of the NCBI taxonomy, and then to collect reads from the metagenome using a simulation of a number of different sequencing technologies. A population sampler optionally produces evolved sequences based on source genomes and a given evolutionary tree.

Using MetaSim, the user is able to simulate individual read datasets that can be used as standardized test scenarios for planning sequencing projects or for benchmarking metagenomic software.

Classifiers for Metagenome Fragment Annotation

Gail Rosen, Drexel University

With the advent of next-generation sequencers, environmental samples can be quickly sequenced by fragmenting DNA into very short segments (around 25 bp). Therefore, whole 16S rRNA segments cannot be sequenced for taxonomic analysis. We show that a fundamental classifier can train on previously known genomes, in order to classify fragments to the nearest strain, species, and genus. Different scoring mechanisms can also be used to discern whether the fragment of interest is new and not from the training set, and we investigate sensitivity and specificity for detecting such fragments.

The performance of a Naive Bayes classifier (NBC) using a training set of 635 genomes (all microbes available as of Feb. 2008) and tested on fragments of size 500bp, 100bp, and 25bp. The advantage is that it does not rely solely on rRNA sequences and can use ANY fragment. Also, the classifier obtains reasonable accuracy with small 25-bp fragments, which are common read-lengths from fast next-gen sequencing. The results are comparable to BLAST,

but do not have the ambiguity that BLAST causes with highly similar fragments. Also, the classifier obtains an accuracy of 92% for 500bp fragments and 77% for 25bp fragments for genus-level accuracy, which outperforms the state-of-the-art RDP classifier by Wang et. al (with a genus-level accuracy of 89% for 400bp sequences and 52% for 50-bp sequences). Finally, we investigate scoring schemes and predicting unknown genomes.

Metagenomic analysis of an epilithic microbial biofilm in a Hawaiian lava cave

Jimmy H. Saw, University of Hawaii at Manoa; Los Alamos National Laboratory

Co-authors: Mark V. Brown, Gayle Philip, Durrell D. Kapan, Jamie S. Foster, Alexis S. Templeton, John Berest-ecky, Stuart P. Donachie

While lava caves in terrestrial habitats are features amenable to detailed microbial characterization, the phylogenetic structure and metabolic functions of microbial communities on recently erupted (<100 years) volcanic rocks are poorly characterized. Indeed, studies of how microbes interact with subsurface basalts have largely focused on deep-sea hydrothermal vents, ridge systems, and deep aquifers and bore holes. We have, however, begun to elucidate the microbial characteristics of an epilithic biofilm on the shaded entrance wall of Big Ell, a lava cave in the 1919 flow in volcanically active Kilauea Caldera, Hawaii. Conditions around the biofilm include temperatures above 100°C and humidity over 100%. We pyrosequenced community genomic DNA from this biofilm through 386,217 'reads' of 247bp average length (approximately 95 million nucleotides in total). BLASTX comparisons of these data with the SEED non-redundant database classified 33% of sequences as unknown. The remainder belonged in 27 functional categories, as predicted by SEED comparisons. Sequence assembly yielded 39,258 contigs and 153,962 singletons. Only 4 contigs exceed 10,000bp. An analysis with MEGAN showed the biofilm community comprises diverse Bacteria phyla, including Acidobacteria, Actinobacteria, Bacteroidetes, Cyanobacteria, Chloroflexi, Firmicutes, Planctomycetes, and Proteobacteria. A preliminary survey of microbial diversity through a 16S rDNA clone library revealed a similar community structure. This epilithic biofilm contains a phylogenetically diverse and functionally complex community, one we aim to describe through further metagenomic analyses of microbially-mediated processes in this environment.

ADAPTdb/ADAPT - A Framework for the Analysis of ARISA Data Sets

Robert Schmieder, San Diego State University

Co-authors: Matthew Haynes, Elizabeth Dinsdale, Forest Rohwer, and Robert Edwards

The characterization of natural microbial assemblages in community profiling projects introduces the major scientific challenges of understanding and predicting the function and response to environmental changes of microbes of an ecosystem. Here, we present a system for the auto-

matic analysis of ARISA data sets. ARISA is a method for analyzing the composition of microbial communities, which performs faster and at a much lower cost than other community profiling techniques. ARISA relies on the analysis of intergenic regions called internal transcribed spacer (ITS), which are located between the 16S and 23S rRNA genes. The database ADAPTdb was created to store and maintain ITS regions along with information about their source organisms. The data stored in ADAPTdb is retrieved from different data resources, such as the Entrez sequence databases. The program ADAPT was developed to taxonomically characterize ARISA data sets using ADAPTdb. The additional organism information for each ITS region in the ADAPTdb database is used by ADAPT for pathogenic and autotrophic/heterotrophic comparisons of organisms among different ARISA samples. The program is publicly available through a user-friendly web interface, which allows onsite analysis of ARISA data sets and computation of the output. The interactive web interface facilitates navigation through the output and export functionality for subsequent analysis.

Frequent occurrence of clones encoding for lactonase family proteins in metagenomes

Wolfgang, R. Streit, University of Hamburg
Co-authors: W. R. Streit, C. Schipper, N. Weiland, C. Horning, P. Bijtenhoorn, M. Quitschau, S. Grond, R. Schmitz

Life in microbial consortia depends on bacterial cell-cell communication. To further advance our knowledge on processes linked to bacterial communication in complex communities we have screened several metagenome libraries for the presence of autoinducer-I (Ail) and autoinducer-II (AII) degrading genes and enzymes. Here we report on the isolation of nine metagenome-derived clones that interfere with bacterial quorum sensing and affect the stability of the Ail molecule 3-oxo-C8-acetyl-homoserine lactone. The identified ORFs were designate bpiB01-bpiB09. While bpi07 was weakly similar to a lactonase no significant similarities were observed for all other bpi genes and the deduced aa sequences. A further biochemical characterization including a HPLC-MS analysis of the hydrolytic products released by the recombinant Bpi proteins suggested that most of them encode for novel lactonases and therefore can be grouped in novel lactone hydrolyzing protein families (EC 3.1.1.--.) The high frequency of the identification of these novel lactonases in soil communities suggests that the lactone hydrolyzing enzymes may play an important role in establishment of the soil bacterial communities. Most striking at least two of the isolated proteins were able to hydrolyze autoinducer I and II molecules.

A Metagenomic Analysis of Fungus Garden Communities in Leaf-Cutter Ants

Garret Suen, University of Wisconsin-Madison
Co-authors: Jarrod Scott, Sandye Adams, Kerrie Barry, Sussanah G. Tringe, Cameron R. Currie

For ~50 million years, fungus-growing ants have been farming fungus for food. This evolution has culminated in the conspicuous leaf-cutting ants, which forage leaves as substrate for their fungus gardens. This ant-fungus symbiosis is one of the most complex in nature, and is composed of six known members, including four mutualists and two pathogens. All members of this symbiosis are located within the fungus gardens of the leaf-cutters, and it is thought that the microbial community within the gardens changes constantly since new leaf material is being constantly incorporated. To determine if this community fluctuates temporally, we performed 16S and 18S metagenomic analyses on fungus garden material sampled from colonies of the leaf-cutter, *Atta columbica*, collected in Panama in January and June of 2008. Preliminary analysis indicates that there is a radical shift of some groups of microbes; however, there remains a stable and persistent core of microbes within the gardens. This suggests that these core microbes may play a specific role within the garden as part of the ant-fungus symbiosis. We also measured the abiotic, enzymatic, and chemical conditions across different layers within the garden, and, coupled with further metagenomic analysis across these layers, attempt to define specific roles for each microbial community.

Horizontal gene transfer in coastal marine cyanobacterial populations

Vera Tai, Scripps Institution of Oceanography, UC San Diego
Co-authors: B. Palenik, Q. Ren, I.T. Paulsen

The extent by which cultured strains represent the genetic diversity of a group of microorganisms is poorly understood. Without the need for culturing, metagenomic approaches have the potential to reveal the genetic diversity of the microbes actually present in an environment at a given time. From coastal seawater, a complex and diverse environment, we have tailored this metagenomic study to specifically analyze marine cyanobacteria populations from the genus *Synechococcus*. The *Synechococcus* population was isolated by flow cytometry sorting and the metagenome was analyzed with 454 sequencing technology. The sequence data were compared to model *Synechococcus* genomes including those of two strains isolated from the same or similar coastal environment and metagenomes were compared to each other. Here we report that the natural population had high homology to most genes from the coastal model strains, but diverged greatly from these genomes in regions of atypical trinucleotide content.

These results are explained by extensive horizontal gene transfer with presumably large differences in the genetic material transferred. Thus, significant aspects of the

genomic diversity and physiological potential of natural populations of *Synechococcus* are not represented by cultured isolates. The ecological significance of the genes transferred, the mechanisms of horizontal gene transfer, and the implications for microbial evolution and adaptation are topics that will be further investigated with these data.

The distribution of prokaryotes in natural and artificial environments supports everything is everywhere

Javier Tamames, University of Valencia

Co-authors: Juan Jose Abellan, Miguel Pignatelli, Andres Moya

In this study, we have performed a thorough study of the relationships between individual prokaryotic taxa and different environments. We have compiled all the available samplings for prokaryotes that have been deposited in GenBank database, taxonomically assigned the sequences found, and classified the samples in one of forty different environments. Our results show that many taxa show clear environmental preferences, that is, they tend to appear in some environments more than in others. This advocates for the statement that “everything is everywhere, but environment selects”. On the other hand, clear-cut environmental specificity is rare. The presence of several ubiquitous taxa is also detected. The conclusions stand for different taxonomic ranks, from families to species. The environmental distributions detected for taxa can be used to explore the characteristics and similarities of the environments themselves. The results discriminates clearly five environmental groups. According to their selectiveness, these main groups are: host-associated, thermal, salines, terrestrial and aquatic. This conforms, as far as we know, the most comprehensive assessment of the worldwide distribution of prokaryotic taxa and its relationships to different environments.

Using a metagenomic approach to determine the microbial ecology of hypersaline mats

Rion G. Taylor, University of South Carolina

Co-author: R. Sean Norman

Photosynthetic microbial mats form complete self-sustaining ecosystems at the millimeter scale and impact environmental processes on a planetary scale. They have also played a crucial role in the evolution of life on Earth and are often considered analogs of the extensive Precambrian microbial communities. Among microbial mats, those forming in hypersaline environments have been the focus of numerous studies. Within these extreme environments, the lack of multi-cellular organisms allow very specialized microbial interspecific interactions and make them attractive models for examining the links between community dynamics and ecosystem function. We have been examining the molecular ecology of microbial mats forming in a hypersaline (45-100 psu) pond located at San Salvador, Bahamas. Since greater than 99% of bacteria cannot be readily cultured, we have taken a metagenomic approach using next generation DNA sequencing to shotgun sequence more than 100 million bases of the

mat metagenome. From these sequence data, we are reconstructing the phylogenetic and metabolic diversity of the bacteria thriving in this extreme environment to understand (1) what bacteria can survive in this environment (2) what interspecific interactions are occurring in the mat and (3) what genetic pathways are allowing bacteria to survive under extreme hypersaline conditions.

Bayesian Markov Chain Monte Carlo Approach to Binning of Short Metagenomic Reads

Joshua Weitz, Georgia Tech

Co-authors: Andrey Kislyuk, Srijak Bhatnagar, Jonathan Dushoff

The development of an effective environmental shotgun sequence binning method is a key problem in algorithmic analysis of metagenomic data. While previous methods have focused primarily on supervised learning involving extrinsic data, a first principles statistical model combined with a self-training fitting method offers distinct advantages and can enhance the overall resolving power of a binner based on a combination of methods. We present a mathematical formalism suitable for clustering short sequences by their taxonomic origin on the basis of their k-mer distributions. We find that binning accuracies obtained via a Bayesian MCMC approach are competitive with those obtained using supervised approaches when tested on synthetic constructions of low and medium complexity communities.

Statistical methods for detecting differentially abundant features in metagenomic samples

James Robert White, University of Maryland–College Park

Co-author: Mihai Pop

Numerous studies are currently underway to characterize the microbial communities inhabiting our world. These studies will dramatically expand our understanding of the microbial biosphere and, more importantly, will reveal the secrets of the complex symbiotic relationship between us and our commensal bacterial communities. An important prerequisite for such discoveries are computational tools able to rapidly and accurately compare large datasets generated from complex bacterial communities.

Results: We describe a statistical method for detecting differentially abundant features between two populations using count data (e.g. 16S rRNA surveys to find differentially abundant taxa). In high-complexity environments, our method employs the false discovery rate to improve specificity and properly handles low abundance taxa. We demonstrate the use of our tool on several publicly available datasets: 16S rRNA surveys of human and mouse gut microbiomes, and metabolic subsystem data from 85 microbial and viral metagenomes.

These methods provide a statistical approach for analyzing frequency data to detect differentially abundant categories between two populations, specifically targeted at clinical studies comprising large numbers of samples. A web server implementation of our methods and freely

available source code are available at <http://www.cbcb.umd.edu/~whitej/metastats/detection.shtml>.

Fragment Recruitment : Benchmarking and its application on Metagenomics Data

Gary Xie, Los Alamos National Lab

Co-authors: Jimmy H. Saw, Pavel V. Senin, Ramunas Stepanauskas, Nick Hengartner

Fragment recruitment is a powerful method that uses environmental metagenomic data to explore variation in structure, geography, and sequence relative to a reference sequence or genome. Using various sequence similarity tools: MUMmer, BLAST, BLAT, and BLASTZ for recruiting simulated metagenome sequences to the given reference sequence, we have benchmarked performance of these alignment tools. Meanwhile, well-spaced representative genomes have been chosen according to their phylogenetic distance to test the boundary/limitation of fragment recruitment. These studies help us gain insight into the impact of rRNA distance on recruiting efficiency and sensitivity. Due to the speed of performance, we have chosen MUMMER as the tool for sequence alignment.

e performed benchmarking of sequence similarity search parameters on simulated metagenomic data and Global Ocean Sampling (GOS) expedition data to understand the capabilities of this method. Fragment recruitment method allowed us to identify sequences of Flavobacterial origin from the GOS data that are closely related in sequence similarity to the draft *Flavobacteri* draft genome (Woyke et al, in press). Realizing the importance of visualization tools for metagenomic data, we have implemented a software pipeline for performing fragment recruitment analysis and visualization of results through web interface. Our website will allow users to submit either reference genome sequence (draft or complete), and also their own metagenomic data to compare against existing data.

Phylogenetic screening of the microbial metagenomic library using homing endonuclease restriction and marker insertion (HERMI)

Pui Yi Yung, Centre for Marine Bio-innovation, University of New South Wales

Co-authors: Catherine Burke, Staffan Kjelleberg, Torsten Thomas

One major aim of metagenomic studies is to link phylogeny with function to better understand the ecological role certain groups of organism might play in the environment. This link is often established by screening and sequencing fosmid libraries for clones containing phylogenetic anchors (such as the 16S rRNA gene) or by assigning fragments from shotgun-sequencing projects to taxonomic groups based on nucleotide patterns (e.g. tetra-nucleotide frequency). Here we present a rapid and general strategy to screen pooled fosmid libraries for clones containing a 23S rRNA gene using homing endonuclease restriction and marker insertion (HERMI). We applied this method to a pooled metagenome fosmid library of the marine sponge

Cymbastela concentrica (6500 fosmid clones covering ~260 Mb of genomic data) and identified twelve unique fosmid clones containing 23S rRNA genes. Sequencing of 16S and 23S rRNA genes of these clones unambiguously assigned them to phylogenetic clades within the Piscirickettsiaceae, Bradyrhizobiales and alphaproteobacteria. Identical sequences were also found in 16S rRNA gene PCR libraries directly generated from sponge DNA, which indicates that the HERMI strategy can recover fosmids of relevant phylotypes. Furthermore we can link these fosmids to previously unassigned shotgun-sequencing fragments by nucleotide patterns matching and similarity searches. This allows us to substantially improve the linkage between phylotypes and functional genes in the sponge metagenome project.

Bacterial metagenome from a Neanderthal bone sample

Katarzyna Zaremb, Uppsala University

Co-authors: Siv Andersson, Jennifer Ast

A metagenomic approach was used to analyze DNA extracted from a Neanderthal bone fossil (data from Svante Pääbo, Max Planck Institute of Evolutionary Anthropology, Germany). Although most of the DNA was of unknown origin, preliminary analysis suggested that bacteria constitute a majority of the sequences that could be identified [Green et al 2006]. We refined the analysis using ribosomal RNA sequence searches against the Silva Ref rRNA database. Procedures were tested and error rates were estimated using 454-sequence reads of similar lengths from a bacterial genome project. Searches were limited to high-quality data to limit the number of false positives. The results suggest that the taxonomic diversity of the sample is rather limited, with about 80% of the identified sequences being bacterial rRNA sequences. The dominating group is Actinobacteria, specifically Actinomycetales, which accounts for 50% of the rRNA hits. Other bacterial phyla represented in the sample with a few percent each include Proteobacteria, Firmicutes and Acidobacteria. Circa 8% of the hits were to Metazoa (putative Neanderthal sequences), 4% to other eukaryotes and 1% to Archaea. Overall, we estimate that the sample contained more than 1 Gb of bacterial DNA and 0.7 Gb of Neanderthal (Metazoa) DNA, out of the 7 Gb in total. In contrast, similar analysis of DNA extracted from a mammoth fossil gave estimates that summed up to the size of the sample, suggesting that the DNA from the Neanderthal fossil is more degraded and therefore more difficult to identify.

POSTERS



Author	Email	Affiliation
Luis David Alcaraz	ldalcaraz@gmail.com	Instituto de Ecología, UNAM
Linda Amaral-Zettler	amaral@mbl.edu	Marine Biological Laboratory Woods Hole
Anders Andersson	doubleanders@gmail.com	Uppsala University
Florent Angly	florent.angly@gmail.com	San Diego State University
Dionysios A. Antonopoulos	dion@anl.gov	Argonne National Laboratory
Meike Ballschmiter	ballschmiter@uni-leipzig.de	University of Leipzig
Catherine Burke	c.burke@student.unsw.edu.au	Centre for Marine Bio-innovation
Samuel Chaffron	samuel.chaffron@molbio.uzh.ch	University of Zurich
L. Chistoserdova	milachis@u.washington.edu	University of Washington
Adina S. Chuang	adina-chuang@uiowa.edu	University of Iowa
Jim Cole	sulwoo@msu.edu	Michigan State University
Robert D. Finn	rdf@sanger.ac.uk	Wellcome Trust Sanger Institute
Anthony Fodor	anthony.fodor@gmail.com	UNC Charlotte
Dana Hall	halld@rohan.sdsu.edu	San Diego State University
Cliff S. Han	han_cliff@lanl.gov	Joint Genome Institute
Wolfgang Hankeln	whankeln@mpi-bremen.de	Max Planck
Markus Hilty	m.hilty@imperial.ac.uk	Imperial College, London
K. J. Hoff	khoff@gwdg.de	University of Göttingen
Sean D Hooper	SHooper@lbl.gov	Joint Genome Institute
Meredith A J Hullar	mhullar@fhcrc.org	Fred Hutchinson Cancer Research Center
Nele Ilmberger	Nele.Ilmberger@uni-hamburg.de	University of Hamburg
Merja Itävaara	merja.itavaara@vtt.fi	VTT Technical Research Centre, Finland
Haeyoung Jeong	hvieong@kribb.re.kr	Korea Research Institute of Bioscience and Biotechnology
GwangPyo Ko	gko@snu.ac.kr	Seoul National University
Asuncion Lago-Leston	alago@ualg.pt	Universidade do Algarve
Alla Lapidus	alapidus@lbl.gov	Joint Genome Institute
Yan Wei Lim	limyw@gis.a-star.edu.sg	Genome Institute of Singapore
Maria Martin	martin@ebi.ac.uk	The European Bioinformatics Institute
Ana-Belén Martín-Cuadrado	amartin@umh.es	Universidad Miguel Hernández
Adam C. Martiny	amartiny@uci.edu	University of California, Irvine
Elizabeth McCliment	lmcccliment@mbl.edu	Marine Biological Laboratory, Woods Hole
Tim McDermott	timmcder@montana.edu	Montana State University
Jun meng	mengjun_quo@hotmail.com	Xiamen University
Rebekka Metzger	rmetzger@ifam.uni-kiel.de	Christian-Albrechts-Universität Kiel
Suparna Mitra	mitra@informatik.uni-tuebingen.de	University of Tuebingen
Christina D. Moon	christina.moon@agresearch.co.nz	AgResearch Ltd
Kozo Nishida	kozo.nishida@gmail.com	Mitsubishi Research Institute
Brian Palenik	bpalenik@ucsd.edu	University of California, San Diego
Amrita Pati	apati@lbl.gov	Joint Genome Institute
Kevin Penn	kpenn@ucsd.edu	Scripps Institution of Oceanography
Rachel S. Poretsky	poretsky@uga.edu	University of Georgia
Qinghu Ren	qren@icvi.org	J. Craig Venter Institute
D.C. Richter	drichter@informatik.uni-tuebingen.de	University of Tuebingen
Gail Rosen	gailr@ece.drexel.edu	Drexel University
Jimmy H. Saw	jimmy@hawaii.edu	University of Hawaii
Robert Schmieder	rschmieder@gmail.com	San Diego State University
Wolfgang, R. Streit	wolfgang.streit@uni-hamburg.de	University of Hamburg
Garret Suen	gsuen@wisc.edu	Michigan State University
Vera Tai	vtai@ucsd.edu	University of California, San Diego
Javier Tamames	javier.tamames@uv.es	University of Valencia
Rion G. Taylor	rtaylor@qwm.sc.edu	University of South Carolina
Joshua Weitz	jsweitz@gatech.edu	Georgia Institute of Technology
James Robert White	whitej@umd.edu	University of Maryland - College Park
Gary Xie	xie@lanl.gov	Los Alamos National Lab
Pui Yi Yung	mariayungpy@gmail.com	University of New South Wales
Katarzyna Zaremba	Katarzyna.Zaremba@ebc.uu.se	Uppsala University

Poster

- Comparative genomics of *Bacillus* sp. from a desiccation lagoon in Cuatro Ciénegas, Mexico
- International Census of Marine Microbes (ICoMM): Unveiling the Ocean's Hidden Majority Through Community 454 Tag Pyrosequencing
- Comparative analysis of human gut microbiota by \square barcoded pyrosequencing
- A statistical approach for a better assessment of community taxonomy and average genome size
- Metagenomic-based analysis of carbon management in diverse soils
- Putting biodiversity to work: New approaches to functional enzyme screening in *E. coli*
- Metagenomic Analysis of the Epiphytic Bacterial Community of the green macroalga *Ulva australis*
- Metaproteome and metagenome analysis of microbial communities in the phyllosphere
- High-resolution metagenomics: understanding metabolism and ecological function of individual members of complex microbial communities
- Analysis of methods to extract, quantify, and identify the metaproteome of soil and groundwater samples
- Determining bacterial community by pyrosequencing of SSU rRNA and functional genes
- Pfam Annotations of Environmental and Metagenomic Datasets
- Targeting the V1-V2 region of the 16S rRNA gene yields improved measures of microbial community composition
- Comparative Metagenomics of the Human Lung: Characterization of Bacteriophage Communities in Cystic Fibrosis versus Non-Cystic Fibrosis Individuals
- Finishing Genomes from Single Cell, Photoheterotrophic Flavobacteria
- MetaBar: A tool for consistent contextual data acquisition for ecological genomics
- Molecular characterisation of the bacterial flora in the lung of healthy, asthmatic and chronic obstructive pulmonary diseased (COPD) subjects
- Predicting genes in short metagenomic sequencing reads with high \square specificity
- Integration of Phenotypic Metadata and Protein Similarity in Archaea Using a Soft Clustering Approach
- Inter-individual variation in the human metabolic phenotype of equal production is associated with variation in gut microbial communities.
- Functional and phylogenetic characterisation of the gut of the asian elephant with a metagenomic approach
- Life in deep bedrock aquifers – metagenomics of deep borehole biosphere
- Genome sequence analysis of *Bifidobacterium bifidum* and *B. animalis* subsp. *Lactis*
- Metagenomic Analysis of Airborne Microorganisms in Yellow Sand in Korea
- Environmental Transcriptomics Of Southern Ocean Phytoplankton Assemblages
- Metagenomic Finishing at JGI
- Discovery of novel mosquito-associated viruses using metagenomics
- UniMES: a repository for Metagenomic and Environmental Sequences
- Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles
- Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions
- Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites: MIRADA (LTERS)
- Hydrogenobaculum Population Genomics: Linking Phylogeny, Genetics, and Ecologic Function
- An Uncultivated Crenarchaeota Contains Functional Bacteriochlorophyll a Synthase \square
- Marine medicine, learning from complex barriers and microbiota in the ocean
- Comparative Metagenomics with MEGAN 2.0
- Exploring fiber degradation by rumen microbial communities \square
- Functional variation analysis of human gut metagenomic data
- Metagenomic analysis of plasmids in coastal marine cyanobacterial populations
- Statistical modeling of the relative abundance of protein functions in metagenomic datasets
- The Identification of Genomic Islands Harboring Phage and Secondary Metabolites in the Genomes of Obligate Marine Actinobacteria and a New Mechanism of Marine Adaptation
- Diel Environmental Metatranscriptomics at the Hawaii Ocean Time-Series and the Sapelo Island Microbial Observatory
- Metagenomic transporter annotation
- MetaSim – A Sequencing Simulator for Genomics and Metagenomics
- Classifiers for Metagenome Fragment Annotation
- Metagenomic analysis of an epilithic microbial biofilm in a Hawaiian lava cave
- ADAPTdb/ADAPT - A Framework for the Analysis of ARISA Data Sets
- Frequent occurrence of clones encoding for lactonase family proteins in metagenomes
- A Metagenomic Analysis of Fungus Garden Communities in Leaf-Cutter Ants
- Horizontal gene transfer in coastal marine cyanobacterial populations
- The distribution of prokaryotes in natural and artificial environments supports everything is everywhere
- Using a metagenomic approach to determine the microbial ecology of hypersaline mats
- Bayesian Markov Chain Monte Carlo Approach to Binning of Short Metagenomic Reads
- Statistical methods for detecting differentially abundant features in metagenomic samples
- Fragment Recruitment : Benchmarking and its application on Metagenomics Data
- Phylogenetic screening of the microbial metagenomic library using homing endonuclease restriction and marker insertion (HERMI)
- Bacterial metagenome from a Neanderthal bone sample



'Class Photo' from Metagenomics 2007

Sponsors



Industry Sponsors



Silver Sponsor



<http://metagenomics.calit2.net>

**University of California, San Diego
Metagenomics 2008**

**Cover images by
John Wooley**

