JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Advancing Science with DNA Sequence

JGI's Microbial Genomics and Metagenomics Workshop

JGI's Microbial Genomics and Metagenomics Workshop

a week of cutting edge Science & Technology at the DOE-JGI

- **Workshop Goal:**

The U.S. Department of Energy Joint Genome Institute (DOE JGI) is offering three five-day workshops on Microbial Genomics and Metagenomics during 2008. All three workshops will cover the same material. Each will include two days of intensive seminars and three days of hands-on tutorials. Our goal is to provide you with training in microbial genomic and metagenomic analysis and demonstrate how the cutting-edge science and technology of DOE JGI can enhance your research.

- **Audience:**

Target audience is graduate students, postdocs, and faculty and staff scientists

- **Registration and Number of Participants:**

The course will be free of charge, but students will have to cover their own expenses while here. The maximum number of participants per course will be 40.

- **Workshop Dates:**

January 7 - 11, 2008
May 19 - 23, 2008
September 15 - 19, 2008

Tutorials:

IMG, IMG/M, IMG-ER, IMG-EDU, ARB, VISTA, GREENGENES, CAMERA

Contact Person and Logistics:

Marsha Fenner and David Gilbert
MWFenner@lbl.gov and DEGilbert@lbl.gov

Workshop Agenda:

Monday Introductory seminars (Methods & Technologies)

9.00-9.15 **Nikos Kyrpides**
Welcome and overview of the workshop.

9.15-9.45 **Jim Bristow**
Introduction to the JGI
The powerful high-throughput DNA sequencing technologies catalyzed by the Human Genome Project, which have contributed to dramatic advances in biomedicine, are now being directed to characterizing the genomes of plants and microbes. Leading this effort is the US Department of Energy (DOE) Joint Genome Institute (JGI), a national user facility that unites the expertise of five national laboratories to advance genomics in support of the DOE mission areas of bioenergy, carbon cycling, and bioremediation.

9.45-10.15 **Feng Chen**
New Sequencing Technologies
JGI's future depends on new sequencing technologies. Currently, we are under the process of evaluating, validating, and developing applications for three next-generation sequencing technologies, namely Roche's GS FLX, Illumina's Genome Analyzer, and AB's SOLiD system. Introduction to all three technologies will be given and advantages and disadvantages will be compared and discussed. Examples of applications in genomic research for these new technologies will be presented.

10.15-10.45 **Tanja Woyke**
Single cell genomics
The bulk of finished microbial genomes to date are derived from bacteria and archaea that can be readily grown in culture. However, the vast majority of microorganisms on this planet elude current culturing attempts, severely limiting access to their genomes. While various enrichment methods as well as metagenomic approaches have been successfully applied to aid the genome analysis of such non-cultivable environmental microbes, these methodologies are not suitable for countless community members of interest. Single-cell genomics is a new approach which aims to access the genome from an individual microbial cell. Single cells can be isolated from the community using optical tweezers, micromanipulators, flow-sorting, or serial dilutions. After cell lysis, the microbial genome is amplified by using multiple displacement amplification (MDA), allowing random

genome shotgun sequencing. The advantages as well problems associated with the single-cell genomics approach will be discussed.

10.45-11.00 Break

11.00-11.30 Alla Lapidus

Microbial Genome Assembly and Finishing

The US DOE Joint Genome Institute's mission is to provide the scientific community with high-quality finished genomes. Approximately 400 microbial genomes are currently in the JGI pipeline and to date, 166 have been completed. The value of a totally complete microbial genome was recognized and "appreciated" by scientists. Finished genomes allow, for example, the study of genome-level evolution, while the draft sequences are usually of sufficient quality to determine the basic genetic and metabolic parameters of an organism. Some interesting traits can be lost when only working from draft. Computational and lab approaches will be discussed.

11.30-12:00 Nikos Kyrpides

Microbial Genomics

Since the release of the first completely sequenced microbial genome, more than a decade ago, the genomics world has been changing rapidly as large amounts of microbial sequencing data have been accumulating at an exponential rate. Microbial genomics, fueled by recent advancements in sequencing technology, is now playing a central role in medicine and biotechnology and has greatly expanded our understanding of the available phylogenetic and metabolic complexity. Where are we going next? The past, present, and future of microbial genomics will be discussed.

12.00-13.15 Lunch & JGI Facilities Tour

JGI Science

13.15-13.45 Iain Anderson:

Archaeal Genomics

Archaea are the least well characterized organisms of the three domains of life. Yet, they share many important features with eukaryotes and are the key in understanding the origins and nature of the last common ancestor. JGI has a strong interest in archaea because of their broad biotechnological applications as well as their relevance in energy production, and therefore a large number of archaeal sequencing projects are currently under way. The analysis of two crenarchaeal genomes that have been completely sequenced will be presented. Examples will be shown of how unique genes and genes uniquely missing from these genomes can be identified and characterized.

13.45-14.15 Cheryl Kerfeld

Discovering Bacterial Organelles with IMG and Structural Biology

Bacterial microcompartments, polyhedral bodies composed entirely of protein, were first discovered by electron microscopy more than 40 years ago. Until recently, they were thought to be confined to a few species of bacteria. Now genomic sequence data is revealing the widespread occurrence of these bacterial organelles and providing clues about their functional diversity.

14.15-14.45 Rotem Sorek

Genome-wide experimental determination of barriers to horizontal gene transfer

Investigations of horizontal gene transfer in microbial genomes have generally been limited to computational sequence analyses, while experimental studies are largely lacking. We show that cloning gaps in sequenced microbial genomes are indicative of lack of horizontal transfer. By analyzing the attempted experimental transfer of 246,045 genes from 79 prokaryotic genomes to E. coli, we identified sets of genes that consistently fail to transfer, and form phylum-independent barriers for gene flow between prokaryotes.

14.45-15.00 Break

15.00-15.30 Phil Hugenholtz:

Introduction to Metagenomics

Metagenomics, the application of high-throughput sequencing to environmental samples, is an emerging field that is rapidly advancing our understanding of how microbial communities function and evolve. This introductory talk will trace the roots of metagenomics and its current practice and speculate on future developments in the field.

15.30-16.00 Susannah Tringe:

Metagenomics projects at JGI

Metagenomics, the sequencing of DNA from uncultivated microbial communities, offers us the opportunity to study organisms that cannot be domesticated in the lab. In the past several years, advances in DNA sequencing techniques, throughput, and analysis have allowed valuable glimpses into this uncharted genomic space. The DOE Joint Genome Institute has taken the lead in several key metagenomic projects and is currently involved in the sequence-based study of dozens of environmental and symbiotic microbial communities. I will discuss metagenome-specific processes for sequencing, assembling, annotating, and analyzing metagenomic data, and scientific insights gained through the application of these processes to a variety of communities, both simple and complex.

16.00-16.30 Victor Kunin:
Metagenomics of Hypersaline mats
The Guerrero Negro hypersaline microbial mat in Baja California is one of the most complex and diverse microbial communities yet described. We have generated shotgun sequence of 10 successive layers of a ~6-cm-thick mat core for comparative analysis. Millimeter-scale functional gradients were inferred from gene and pathway frequency distributions that often tracked with the physicochemical profile of the mat. The environment and the results of the metagenome analysis will be presented and discussed.

16.30-17.00 Falk Warnecke:
Termite gut metagenomics
Termites efficiently decompose plant biomass (i.e., lignocellulose). By using the random shotgun sequencing approach, this project tried to elucidate the role that microorganisms inhabiting the termite's hindgut play in this complex process. In the future, the discovered hydrolytic enzymes may be used to convert waste biomass or energy crops into transportation fuel.

17.00-19.00 Poster session and reception

Tuesday Microbial Genome Analysis

09.00-09.30 Natalia Ivanova
Finding the genes in microbial genomes
Annotation of microbial genomes usually starts with finding the genes coding for stable RNAs (rRNA and tRNA) and protein-coding genes (CDSs). The principles underlying gene prediction in microbial genomes, as well as different implementations of these algorithms and most popular gene finding tools will be discussed.

09.30-10.00 Athanasios Lykidis
Gene models Quality Control
Accurate gene prediction is an indispensable step for correct subsequent genome analysis. All currently available tools for automatic gene-finding have a 10-15% error rate in their accuracy. A methodology for gene model validation and manual curation will be presented.

10.00-10.30 Sean Hooper
Sequence space Gene Clustering
One of the first steps following (meta-)genome assembly is to organize and sort large numbers of potential open reading frames.

We will look at some approaches that can be used to categorize DNA sequences into groups based on sequence similarities to known or unknown sequences.

10.30-10.45 Break

10.45-11.15 **Athanasios Lykidis**
Annotation: function prediction & metabolic reconstruction
In this section we will discuss methodologies for assigning functions to gene products. Methods based on homology, common motif occurrence, and chromosomal context will be presented. The steps necessary to reconstruct the metabolic network of an organism will be presented.

11.15-11.45 **Natalia Ivanova**
IMG Terms and Pathways
Description of the Control Vocabularies for the annotations in IMG (IMG Terms) and the curation of the IMG pathway database (IMG pathways)

11.45-13.00 Lunch

13.00-13.45 **Pilar Francino:**
Phylogenomics
Clarifying the relationships among bacterial lineages is important to provide a phylogenetic framework on which to trace the evolution of bacterial diversity. The large number of complete genomes now available from numerous bacterial lineages has greatly augmented the power of phylogenetic analyses. Large-scale protein alignments and other approaches are revealing the topology of many areas of the bacterial tree, although others still remain controversial. We'll discuss how to use genomic information to investigate bacterial phylogeny, in spite of horizontal transfer and other phenomena that entangle phylogenetic reconstruction, and how the phylogenies obtained can advance our understanding of genome evolution in bacteria.

13.45-14.15 **Kostas Mavrommatis**
Data Sources
Genome analysis and gene function prediction depends on the comparison of sequences to the existing information stored in databases. They can either be simple repositories of nucleotide or protein sequence, or contain curated information related to the function of the genetic elements. Used in combination, bioinformatics databases constitute the most powerful method for gene function prediction. In this presentation, databases commonly used for genome analysis will be discussed.

14.15-14.45 **Nikos Kyrpides**

Introduction to Microbial Genome Annotation

Microbial genome annotation generally refers to the process of interpreting the raw sequence data with respect to the biological properties of an organism, by identifying protein-coding sequences and other genome features and determining their physiological functions. The identification of the complete set of functions of any organism provides the foundation upon which our understanding of the biology of that organism rests. In essence, it forms the basic framework that any genome project targets, and from which any biological interpretation originates. Pipelines and methodologies for microbial genome annotation will be presented and discussed.

14.45-15.15 Break

15.15-16.15 **Victor Markowitz**

*Roadmap of the IMG Systems and Desig*n **(IMG-ER & IMG-EDU)**

IMG-Expert Review (IMG-ER) and IMG-Educational (IMG-EDU) are two recently released systems of the IMG-family. IMG-ER allows users to upload their genome under secure access and manually curate and analyze it before its public release. IMG-EDU provides support for undergraduate and graduate level courses in microbial genome analysis and annotation. The goals and specific functionality of both systems will be discussed

16.15-17.00 **Victor Markowitz**

Anatomy of IMG Data Integration

Effective genome comparative analysis relies on coherent and consistent data integration. We will review the main steps of genome data integration in IMG, including selection of genomic and functional data sources, data collection and review, and functional annotation. The challenges of integrating different types of genomes and of coping with continuously evolving data will be also discussed.

TUTORIALS

Wednesday IMG tutorial (annotation and genome analysis)

09.00-10.00 Athanasios Lykidis
IMG Genes & Genomes
Microbial genome data analysis in IMG is set in the comparative context of multiple microbial genomes. IMG allows navigating the microbial genome data space along three key dimensions: genomes (organisms), functions (terms and pathways), and genes. In this section, IMG-based comparative analysis of gene families and genomes will be presented. Tools that will be discussed include phylogenetic profiles and occurrences, homology-based and chromosomal context analysis, VISTA, abundance profiles, and genome clustering.

10.00-11.15 Users
Hands on IMG (exercises)

11.15-12.00 Athanasios Lykidis
Exercise solutions

12.00-13.00 Lunch

13.00-13.45 Iain Anderson
IMG Functions & Pathways
IMG has several ways for users to interact with protein functions and pathways, including Clusters of Orthologous Groups (COGs) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. In addition, JGI is developing a controlled vocabulary for the representation of functions and pathways known as IMG Terms and Pathways. The use of the various Functional Groups and their Pathways and their importance in comparative genome analysis will be presented and discussed.

13.45-14.15 Iain Anderson
MyIMG
The functional annotation for individual genes can be modified using the MyIMG Annotations features of MyIMG. In addition to curation of functional annotations, MyIMG provides support for uploading user genome selections that have been saved earlier from the Genome Browser or Genome Statistics and for setting systemwide user preferences. The use and functionality of MyIMG features will be discussed.

14.15-14.45 Kostas Mavrommatis
Gene context analysis on IMG

14.45-16.00 Users
Hands on IMG (exercises)

16.00-17.00 Iain Anderson
Exercise solutions

Thursday IMG/M tutorial (metagenome analysis)

09.00-10.00 Natalia Ivanova
Metagenome analysis in IMG/M – Part I
A snapshot of microbial community structure can be derived from analysis of metagenomic data. IMG/M methods and tools for establishing the taxonomic identity of community members will be presented along with tools for determining the fine population structure, genetic variation, and genome dynamics of the dominant populations. Methods for assessing the diversity and abundance of microbial communities will be discussed.

10.00-11.15 Users
Hands on IMG/M (exercises)

11.15-12.00 Natalia Ivanova
Exercise solutions

12.00-13.15 Lunch

13.15-13.45 Sean Hooper
COAL
A web-based system which is using a novel soft protein clustering method that allows the correlation of information from genetic, phenotypic and phylogenetic data.

13.45-15.00 Kostas Mavrommatis
A Genome Analysis test case
The methodology and steps to analyze a genome in IMG will be presented with a user case

15.00-15.30 Break

15.30-17.00 Athanasios Lykidis
Metagenome analysis in IMG/M – Part II

The methodology and steps to analyze a genome in IMG will be presented with a user case

Friday VISTA, ARB, Greengenes, CAMERA tutorials

09.00-10.00 **Paul Gilna**
CAMERA -I
CAMERA stands for Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis. The aim of this project is to serve the needs of the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis.

10.00-10.15 Break

10.15-12.00 **Kayo Arima**
CAMERA -II
CAMERA tutorial

12.00-13.00 Lunch

13.00-14.15 **Inna Dubchak**
VISTA
The VISTA portal (http://genome.lbl.gov/vista) is a comprehensive comparative genomics resource that provides scientists with a single unified framework to generate and download multiple sequence alignments, visualize the results in the context of existing annotations, and analyze comparative results in the search for important sequence signals in alignments. Among the servers for user-submitted sequences are GenomeVISTA, for aligning a sequence (draft or finished) against whole genome assemblies; mVISTA and wgVISTA, for globally aligning sequences of different species up to 10 Mb long; rVISTA, which uses conservation among species to improve prediction of transcription factor binding sites; and Phylo-VISTA, for visualization of multiple alignments with a phylogenetic tree.

14.15-10.30 Break

14.30-15.45 **Todd DeSantis**
Greengenes
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.

15.45-16.00 Break

16.00-17.00 **Falk Warnecke**

ARB

ARB is a software package designed to allow the efficient analysis of ribosomal RNA sequences. It incorporates tools for database management, automatic and manual sequence alignment, phylogenetic tree calculation, and the design of discriminatory oligonucleotides used as probes (e.g., for fluorescence in-situ hybridization) and primers.

Closing Workshop