

Workshop Agenda:

Monday	Introductory seminars (Methods & Technologies)
---------------	---

09.00-09.30 Nikos Kyrpides

Welcome and overview of the workshop.

09.30-10.00 1. 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.

10.00-10.30 2. 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.

10.30-10.45 Break

10.45-11.30 3. 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 4. 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.

12.00-13.00 Lunch

MICROBIAL GENOMICS

13.00-13.30 5. 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.

13.30-14.00 6. 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.

14.00-14.30 7. Igor Grigoriev

Annotation of Eukaryotic Genomes

Over 50 eukaryotic genomes from different taxonomic groups are annotated at JGI using JGI Annotation pipeline. The pipeline integrates several gene prediction, annotation, and analysis tools to annotate a diverse set of genomes in high-throughput but genome-specific manner. To address gene prediction challenges in eukaryotes that often display high repeat content, low gene density, and complex gene structure, we combine different gene predictors with available experimental data and comparative genomics analysis. JGI Eukaryotic Portal provides web-based tools for user communities to enable comprehensive genome analysis and manual curation of predicted genes and functions.

14.30-15.0 Break

METAGENOMICS

15.00-15.30 8. 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 9. 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 10. 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 11. Matthias Hess

Discovery of feedstock-targeted glycosyl hydrolases by Metatranscriptomics

Highly active and stable cellulolytic enzymes are major bottlenecks for the efficient large-scale production of biofuels from lignocellulose. Complex microbial habitats such as the bovine foregut are known to harbor fibrolytic microbes and represent promising sources of novel biocatalysts for lignocellulose degradation. We employed high-throughput pyrosequencing to identify feedstock-targeted enzymes within the transcriptome of bovine rumen microbial communities.

17.00-19.00 JGI Facilities Tour - Poster session and reception

Tuesday	Microbial Genome Analysis & IMG tutorial start
----------------	---

09.00-09.30 12. Amrita Pati

Basic Bioinformatics Tools

Introduction to the concepts behind the most essential tools in computational biology and bioinformatics. These will include blast alignments, hidden Markov models, analysis using sequences, multiple sequence analysis, protein family classifications, and basics of phylogenetics.

09.30-10.00 13. 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.

10.00-10.30 14. 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.

10.30-11.00 Break

11.00-11.30 15. Amrita Pati

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.

11.30-12.00 16. 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.

12.00-13.00 Lunch

13.00-13.30 17. Nikos Kyrpides

Introduction to IMG

13.30-14.00 18. Victor Markowitz

IMG Systems and Design Walk Through

14.00-15.00 Athanasios Lykidis [Live Demo]

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.

15.00-16.30 Users

Hands on IMG (exercises)

16.30-17.00 Athanasios Lykidis

Exercise solutions

TUTORIALS

Wednesday IMG tutorial (annotation and genome analysis)

09.00-09.15 19. 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.

09.15-10.00 20. 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)

10.00-10.15 Break

10.15-11.00 Iain Anderson [Live Demo]

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.

11.00-11.15 Iain Anderson [Live Demo]

MyIMG- configuration environment

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.

11.15-11.45 Kostas Mavrommatis [*Live Demo*]
Gene context analysis on IMG

11.45-12.00 Users
Hands on IMG (exercises)

12.00-13.00 Lunch

13.00-14.00 Users
Hands on IMG (exercises)

14.00-14.30 Iain Anderson
Exercise solutions

14.30-15.00 21. 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 Nikos Kyrpides [*Live Demo*]
IMG-GOLD
Genome Project selection and Metadata curation

15.30-16.00 Kostas Mavrommatis [*Live Demo*]
IMG-ER - submission
Genome Data Submission to IMG-ER

16.00-16.30 Victor Markowitz [*Live Demo*]
IMG-ER - curation environment

16.30-17.00 Users
Hands on IMG-ER

Thursday	IMG/M tutorial (metagenome analysis)
-----------------	--------------------------------------

09.00-10.00 Natalia Ivanova [*Live Demo*]
Metagenome analysis in IMG/M

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.

10.00-10.30 22. Amrita Pati

Statistical analysis of metagenomic datasets

The systematic evaluation of the relative abundances of individual as well as sets of protein functions across various metagenomic datasets, can yield statistically significant deductions about over- and under-representation of protein function(s) and biological pathways in these communities. We can derive statistical methods for comparing the relative abundances of both individual as well as sets of protein families in 2 given metagenomic datasets. Statistical models for modeling individual abundances and methods for identifying protein families whose difference in abundances are statistically significant, will be presented.

10.30-12.00 Users

Hands on IMG (exercises)

12.00-13.00 Lunch

13.00-14.00 Natalia Ivanova

Exercise solutions

14.00-14.30 23. Kostas Mavrommatis

Pre-processing of metagenomic dataset.

The rapid increase of metagenomic projects is leading to an exponential growth of the sequence data, which in turn creates new challenges related to efficient data storage and analysis. This problem is expected to become more prominent as new sequencing technologies are adopted and large scale sequencing projects are carried out (e.g. HMP, GOS). The Genome Biology Group at the DOE-JGI is developing methods to address these challenges in metagenomic projects, which allow efficient compression of the datasets and representation without loss of sequence, contextual and functional information. These methods include the meta-folds and proxy-gene clusters for Sanger and 454 based metagenomic datasets respectively. In many occasions these approaches allow the extraction of information that was previously undetected such as genomic variations and relationships between members of a group.

14.30-15.00 24. Athanasios Lykidis

Metagenome 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 General Discussion & User's Feedback

17.00-17.30 Break

17.30-18.30 Phil Hugenholtz [Live Demo]

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.

Friday CAMERA, Greengenes, and JGI Eukaryotic portal tutorials

09:00-10.00 25. 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 Michael Chiu & Shulei Sun [Live Demo]

CAMERA -II

CAMERA tutorial

12.00-13.00 Lunch

13.00-14.15 26. 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.

14.15-14.30 Break

14.30-15.15 Inna Dubchak [*Live Demo*]

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.

15.15-15.30 Break

15.30-17.00 Andrea Aerts [*Live Demo*]

Eukaryotic Tutorial

Closing Workshop