DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

DOE Joint Genome Institute

Metagenomics Informatics Challenges Workshop

Ten Thousand Genomes at a Time:
Metagenome Assembly and Beyond

Date

October 12-13, 2011

Location

DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598

Organizers

Patrick Chain, Alex Copeland, Natalia Ivanova, Nikos Kyrpides, Victor Markowitz, Kostas Mavrommatis, Susannah Tringe, Zhong Wang, Tanja Woyke and the DOE JGI Microbial Genomics & Metagenomics Program.

Workshop Goals

Metagenomic studies are an important area for the JGI now and in the near future. In the past two years a handful of large-scale metagenomic studies have emerged, revealing computational bottlenecks in data processing, integration visualization and comparative analysis.

The main goal of this meeting is to explore potential solutions to informatics challenges encountered in the studies of massive amounts of metagenome data.

Workshop Themes

Day 1

Progress and Challenges in Metagenome Assembly: Billions of Reads, Quadrillions of Base Pairs

a. Metagenome-specific Assembly
b. Assembly Quality Evaluation
c. Single Cells and Metagenomes

Day 2

Beyond Metagenome Assembly: Billions of Genes, Quintillions of Blast Hits

d. Metagenome Data Integration, Data Storage and Retrieval
e. Scalability of Comparative Analysis and Novel Algorithms and Tools
f. High-performance Computing

Day 1 Progress and Challenges in Metagenome Assembly: Billions of Reads, Quadrillions of Base Pairs

- ***Shuttle Transport*** - *Walnut Creek Marriott to JGI* ***08:00***

Session 1 09:00-09:30

Introduction to Metagenomics at JGI

Chair – Nikos Kyrpides, DOE Joint Genome Institute

- **Welcome and Goals of the Workshop** **09:00-09:10**
  Nikos Kyrpides, DOE Joint Genome Institute
- **Program Overview and Program Informatics** **09:10-09:30**
  Susannah Tringe, DOE Joint Genome Institute

Session 2 09:30-12:00

Metagenome-specific Assembly

Chairs – Susannah Tringe, DOE Joint Genome Institute, and
Patrick Chain, Los Alamos National Laboratory

Questions:
- Assembly of very large datasets: memory requirements and computation time
- Tackling coverage variation
- Co-assembly of multiple samples
- Co-assembly of data generated with different sequencing technologies
- Co-assembly of data generated with long mate-pair libraries

- **Patrick Chain**, Los Alamos National Laboratory **09:30-09:45**
  JGI Metagenome Assembly

- **Francis Chin**, University of Hong Kong **09:45-10:15**

  Meta-IDBA: A de novo Assembler for Metagenomic Data

  Next-generation sequencing techniques allow us to generate reads from a microbial environment in order to analyze the microbial community. However, assembling of a set of mixed reads from different species to form contigs is a bottleneck of metagenomic research. Although there are many assemblers for assembling reads from a single genome, there are no assemblers for assembling reads in metagenomic data without reference genome sequences. Moreover, the performances of these assemblers on metagenomic data are far from satisfactory because of the existence of common regions in the genomes of subspecies and species which make the assembly problem much more complicated. We introduce the Meta-IDBA algorithm for assembling reads in metagenomic data (http://www.cs.hku.hk/~alse/metaidba) which contain multiple genomes from different species. There are two core steps in Meta-IDBA. It first tries to partition the de Bruijn graph into isolated components of different species based on an important observation. Then, for each component, it captures the slight variants of the genomes of subspecies from the same species by multiple alignments and represents the genome of one species using a consensus sequence. Comparison of the performances of Meta-IDBA and existing assemblers, such as Velvet and Abyss, for different metagenomic datasets shows that Meta-IDBA can reconstruct longer contigs with similar accuracy. Finally, further work to improve the performance of Meta-IDBA will also be discussed.

- **Break** **10:15-10:30**

- **Yasubumi Sakakibara**, Keio University **10:30-11:00**

  MetaVelvet: An Extension of Velvet Assembler to *de novo* Metagenome Assembly from Short Sequence Reads

  Motivation: An important step of "metagenomics" analysis is the assembly of multiple genomes from mixed sequence reads of multiple species in a microbial community. Most conventional pipelines employ a single-genome assembler with carefully optimized parameters and post-process the resulting scaffolds to correct assembly errors. Limitations of the use of a single-genome assembler for de novo metagenome assembly are that highly conserved sequences shared between different species often causes chimera contigs, and sequences of highly abundant species are likely mis-identified as repeats in a single genome, resulting in a number of small fragmented scaffolds. The metagenome assembly problem becomes harder when assembling from very short sequence reads.

  Method: We modified and extended a single-genome and de Bruijn-graph based assembler, known as "Velvet" (Zerbino and Birney, 2008), for short reads to metagenome assembly, called "MetaVelvet", for mixed short reads of multiple species. Our fundamental ideas are first decomposing de Bruijn graph constructed from mixed short reads into individual sub-graphs and second building scaffolds based on every decomposed de Bruijn sub-graph as isolate species genome. We make use of two features, graph connectivity and coverage (abundance) difference, for the decomposition of de Bruijn graph.

  Results: On simulated datasets, MetaVelvet succeeded to generate higher N50 scores and smaller chimeric scaffolds than any compared single-genome assemblers, produce high-quality scaffolds as well as the single assembly using Velvet from single species sequence reads, and MetaVelvet reconstructed even relatively low-coverage genome sequences as scaffolds. On a real dataset of Human Gut microbial read data, MetaVelvet produced longer scaffolds, increased the number of predicted genes, and improved the assignments of a phylum-level taxonomy in the sense that the number of predicted genes that cannot be assigned to any tanoxomy is reduced.

- **Jared Simpson**, Sanger Institute **11:00-11:30**
  Memory efficient sequence analysis using compressed data structures
  The analysis of large collections of sequence data in the absence of a reference genome is computationally challenging. I will present our approach to this problem which uses the FM-index, a data structure that allows efficient queries to be performed over a compressed representation of text. To illustrate our approach, I will discuss the application of the FM-index to error correction and de novo genome assembly.

- **Sergey Koren**, University of Maryland **11:30-12:00**
  Assembly Strategies for Metagenomics
  Metagenomic datasets present significant challenges to traditional genome assemblers. Direct application of these methods on metagenomic datasets can result in highly fragmented assemblies or egregious misassemblies, or both. Recently, several novel assemblers have appeared that are designed specifically for metagenomic datasets. The presentation will discuss how assembly can be/is being tailored to metagenomic datasets. I will also discuss my experience to date with metagenomic assembly and also present results using an end-to-end metagenomic pipeline, MetAMOS. I will conclude with a discussion of future directions of research.

- **Lunch and Discussion** **12:00-13:00**

Session 3 13:00-14:55

Assembly Quality Evaluation

Chair – Alex Copeland, DOE Joint Genome Institute

Questions:
- Informative assembly quality metrics
- Validation of assemblies in the absence of reference genomes
- Quality of assembly in relation to population abundance

- **Alex Copeland**, DOE Joint Genome Institute **13:00-13:10**
  JGI Quality Metrics

- **C. Titus Brown**, Michigan State University **13:10-13:40**
  Approaches to Scaling and Improving Metagenome Assembly
  We have developed a number of technical approaches to metagenome assembly that have enabled us to tackle large data sets from the Great Prairie Grand Challenge. In addition to scaling assembly and improving the quality of basic assemblies from short reads, we have found some apparent artifacts in large Illumina data sets that challenge assemblers. We have also been working on ways to compare assemblies quantitatively.

- **Mihai Pop**, University of Maryland **13:40-14:10**
  Genome Assembly Forensics: Metrics for Assessing Assembly Correctness
  I will provide an overview of approaches/metrics for assessing the correctness of genome assemblies when a reference genome is not available, and highlight ways in which assemblers can 'cheat' these metrics, thereby appearing to perform better than they actually do.

- **Alex Sczyrba**, DOE Joint Genome Institute **14:10-14:40**
  Evaluation of the Cow Rumen Metagenome
  Assembly by Single Copy Gene Analysis and Single Cell Genome Assemblies
  The paucity of enzymes able to degrade lignocellulosic biomass efficiently represents a major bottleneck in the industrial production of biofuels. To identify and characterize biomass degrading genes and genomes of the cow rumen, we sequenced and analyzed about half a terabyte of metagenomic DNA isolated directly from microbes adherent to rumen incubated switchgrass. Despite the complexity of the microbial community, ultra-deep sequencing technology enabled the de novo assembly of numerous draft genomes from uncultured biomass degrading microbes. The completeness and authenticity of individual assembled genomes was validated by complementary methods such as single copy and core gene analysis as well as sequencing of single cells isolated from the same sample.

- **Break** **14:40-14:55**

Session 4 14:55-17:30

Single Cells and Metagenomes

Chair – Tanja Woyke, DOE Joint Genome Institute

Questions:
- Strategies of metagenome/single cell coassembly
- Leveraging single cell data for metagenome assembly and analysis

- **Tanja Woyke**, DOE Joint Genome Institute **14:55-15:05**
  JGI Introduction

- **Ramunas Stepanauskas**, Bigelow Laboratory **15:05-15:15**
  Single Cell and Metagenomic Assemblies: Biology Drives Technical Choices and Goals
  The value of research tools ultimately depends on their power to answer previously unaddressable questions. I will provide a brief "user perspective" of how technical advances in single cell and metagenomic assembly translate into new opportunities for biology research. I will also discuss purely biological factors, such as DNA organization within and among cells, that define the relevance of single cell and metagenomic assemblies to address specific research questions.

- **Steve Quake**, Stanford University **15:15-15:45**
  Sequencing Single Cell Microbial Genomes with Microfluidic Amplification Tools
  I will discuss the recent results in my group using microfluidic devices to assist in single cell genome amplifications. We have been able to generate draft genome sequences for a variety of uncultured microbes using this approach.

- **Doug Rusch**, J. Craig Venter Institute **15:45-16:15**

  Marriage or Civil Unions for Single Cells and Metagenomes

  Single cell sequencing techniques hold great promise for dissecting microbial communities at the cellular level. However technical limitations mean that single cell approaches can only be applied to those cells that can be easily lysed and it is the exceptional case where a single cell project results in a complete genome. Metagenomics can be applied to all the members of a microbial community but the diversity and complexity of the genomic data often precludes assembly of the data into informative units. Clearly these approaches are distinct and complementary but can they be synergistic? Can the union of metagenomic and single cell data result in the better single cell assemblies? Here we will explore the merits and pitfalls associated with building a better single cell genome using metagenomics.

- **Phil Hugenholtz**, University of Queensland **16:15-16:45**

  Comparison of Normalized and Unnormalized Single Cell and Population Assemblies

  We assembled Illumina sequence data obtained from an uncultured archaeal species flow-sorted from a methane-driven denitrifying bioreactor. Five datasets were assessed; i) unnormalized multiple displacement amplification (MDA) of a single cell, ii) single cell MDA normalized using duplex-specific nuclease, iii) population (10e4 cells) unnormalized MDA, iv) population normalized MDA and v) combined data using two short read assemblers; Velvet and Sassy. Sassy was developed by Mike Imelfort at the University of Queensland and differs from most other assemblers in that it does not rely on coverage information and does not classify contigs as repetitive or non-repetitive, instead iteratively extending the longest contigs using paired read and overlap information. This means that Sassy is less affected by highly variable coverage as produced by unnormalized MDA. Both Velvet and Sassy produced more fragmented assemblies for single vs population cell datasets and unnormalized vs. normalized datasets. In particular, the most fragmented assembly from the unnormalized single cell dataset had a substantial number of zero coverage regions relative to the least fragmented assembly (combined dataset). Sassy consistently produced less fragmented assemblies than Velvet. For example with the combined dataset, SaSSY produced 286 contigs with an n50 of 89 Kbp, a largest contig of 366 Kbp and a total assembly size of 3.48 Mbp, whereas Velvet produced 747 contigs with an n50 of 22 Kbp, a largest contig of 127 Kbp and a total assembly size of 3.12 Mbp. We suggest that Sassy is well suited to assembling MDA products and that where possible multiple normalized MDAs should be used to produce draft genomes of uncultured species.

- **Discussion** **16:45-17:30**

- ***Shuttle Transport*** *- JGI to Dinner Offsite* ***17:30***

Workshop Dinner *(By Invitation Only)* 18:00-20:00

Il Fornaio Restaurant
1430 Mt Diablo Blvd
Walnut Creek, CA 94596
(925) 296-0100

- ***Shuttle Transport*** *- Il Fornaio to WC Marriott* ***20:00***

Day 2 Beyond Metagenome Assembly: Billions of Genes, Quintillions of Blast Hits

- ***Shuttle Transport*** - *Walnut Creek Marriott to JGI* ***08:00***

Session 3, continued 09:00-09:30
Assembly Quality Evaluation

Chair – Alex Copeland, DOE Joint Genome Institute

- **Jill Banfield**, University of California, Berkeley **09:00-09:30**

  Recovery of Innumerable Near-complete Genomic Datasets from Community Metagenomic Projects

  Almost one decade ago, JGI began sequencing community genomic DNA from an acid mine drainage system biofilm. The collaboration led to the first reconstruction of near-complete genomes from a natural sample (Tyson et al. 2004). This early project was notable in that the biofilm was dominated by a handful of relatively abundant organisms. For years the question remained: can the approach be scaled to achieve comparable results from much more complex samples? Only recently has it has become feasible to increase sequencing allocations to the tens of Gb scale and to deal with the challenges of short read datasets. The question has been answered, and the answer is yes. Major challenges for a community metagenomics study center on the linked tasks of assembly, assembly curation, and binning. Recent results indicate that good and accurate assemblies can be generated for several Gb-scale sequencing efforts applied to complex community samples. However, additional bioinformatic steps are needed to remove inevitable assembly errors (which tend to occur within a single genome). New approaches that were developed by Sharon et al. (in prep.) for quality evaluation and improvement of assemblies overcome problems resulting from the use of de-bruijn graph based single genome assemblers. The methods have been validated manually and by comparison of de novo assemblies with closely related (thus syntenous) isolate genomes. To date, our approaches have yielded almost 100 genomes of bacteria and phage from samples collected in the IFRC in Rifle, Colorado, in which most organisms are present at the < 1 % abundance level (a total of ~ 25 Gb of Illumina DNA sequence spread across three samples). Given current sequence generation, computational and bioinformatic capacity, our group is enthusiastic about the prospect of "ten thousand microbial genomes at a time. Specifically, we anticipate ttargeting background sediment for development of genomically-informed bioremedation strategies.

Session 5 09:30-11:00

Scalability of Comparative Analysis, Novel Algorithms and Tools

Chair – Kostas Mavrommatis, DOE Joint Genome Institute

Questions:
- Fast novel methods of data processing (alignment-based, alignment-free, etc.)

- **Kostas Mavrommatis**, DOE Joint Genome Institute **09:30-09:45**
  Overview of IMG Challenges

- **Weizhong Li**, San Diego Supercomputer Center **09:45-10:10**
  Effective Analysis of Large Amounts of NGS Metagenomic Data with Ultra-fast Clustering Algorithms
  Tremendous computational analyses are required to process large and complex NGS metagenomic data and to deal with NGS-specific challenges such as sequence errors and artifacts. A fast clustering algorithm provides an effective way to solve many computational problems by identifying and analyzing clusters of duplicates, reads, genes, ORFs, contigs or genomes. Clustering methods can quickly identify sequence errors with very high accuracy. Sophisticated clustering analyses also significantly improve the performance of sequence assembly, diversity analysis, gene prediction and annotation.

- **Paramvir Dehal**, Berkeley Lab **10:10-10:35**
  Managing and Storing Large Datasets in MicrobesOnline, metaMicrobesOnline and the DOE Knowledgebase
  In order to keep up with the rapidly increasing rate of new sequence and functional genomics data for both isolates and metagenomes, we have implemented new strategies for computationally efficient analysis and data storage. These methods eliminate computationally expensive all-against-all analysis while allowing analysis to take place in a more accurate phylogenetic context and significantly reducing database storage needs. We have added the ability to analyze metagenomic datasets to the MicrobesOnline database (http://meta.MicrobesOnline.org), which currently also holds over 1500 microbial genomes. Comparison of environmental sequence data with isolate genomes allows for phylogenetic classification and functional characterization of the members of communities. We have developed the FastTree program, which allows us to create phylogenetic trees for all gene families, even those with over 100,000 members, so that genes can be studied within an evolutionary framework. This permits high-resolution analysis beyond simple gene-family counting, allowing an understanding of the evolutionary processes taking place in a community such as gene family expansions, as linking which gene families play the key roles in given communities tells us their particular ecological demands. Additionally, we have determined Microbial Clade-Oriented Sequence Markers (MicroCOSM), gene-families that are phylogenetically informative for given clades, such as at the phylum level, allowing for increased coverage of reliable phylogenetic classification of metagenomic sequence data. This coupling of metagenomic and isolate genomic data, population structure analysis, sequence classification, and gene-subfamily targeted analysis allows the user to identify the important processes taking place in an environment and the roles played by each member of a community.

- **Discussion** **10:35-10:45**

- **Break** **10:45-11:00**

Session 6 11:00-12:30

Metagenome Data Integration, Data Storage and Retrieval

Chair – Victor Markowitz

Questions:
- Methods of efficient storage and fast retrieval of massive amounts of integrated data without losing the ability to run sophisticated queries

- **Victor Markowitz**, DOE Joint Genome Institute **11:00-11:15**
  Overview of IMG Challenges

- **Johannes Goll**, J. Craig Venter Institute **11:15-11:40**
  Using Solr/Lucene for Large-scale Metagenomics Data Retrieval and Analysis
  We developed JCVI Metagenomics Reports (METAREP), an open source web based tool for high-performance comparative metagenomics (www.jcvi.org/metarep). The software utilizes Solr/Lucene as efficient and flexible data retrieval component. As of today, we have indexed more than 400 million entries from environmental and human microbiome metagenomics samples. We will highlight our approach and summarize lessons learned. Topics include:
  i. brief overview of METAREP and Solr/Lucene
  ii. data import; schema and indexing; index organization
  iii. data retrieval; query language and faceting; statistical component; hierarchical data
  iv. performance considerations; load balancing and distributed searches
  v. website integration; METAREP screenshots
  vi. summary advantages/disadvantages

- **Guy Cochrane,** *EMBL-EBI* **11:40-12:05**
  A Collaborative Ecosystem Model for Metagenomics Data Preservation
  For the last three decades, the International Nucleotide Sequence Database Collaboration has provided a celebrated example of global scientific data sharing. The European Molecular Biology Laboratory (EMBL) has been the European driver of the collaboration and the European resource, now known as the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) and hosted at the EMBL European Bioinformatics Institute, brings together and integrates a spectrum of databases and services, covering raw data, assembly information, read alignments and submitted functional annotation.

  Ongoing technological advances in nucleic acid sequencing and dramatic falls in sequencing costs have driven a relentless growth in data volumes and broad expansion of the range of applications to which sequencing can be put. Indeed, sequencing has become a near-ubiquitous assay platform across and beyond the life sciences. Despite the common misconception that growing data volumes provide the most threatening difficulties in operating large sequence databases, it is rather the broadening of applications of sequencing that brings the less bounded and tractable challenges; a generic data resource such as ENA is mandated to capture and present data across all sequencing applications, but it is clear that accessing the appropriate expertise to be able to model experimental configurations, curate incoming data sets and provide appropriate data access services across the breadth the resource's content raises issues of scalability.

  Not only is it economically impossible to grow expertise across all sequencing applications in house at the EMBL European Bioinformatics Institute, but it is inefficient as it fails to recognise and take advantage of pre-existing expertise in the user community. The approach taken by the ENA is one in which a central core of generic cross-application technologies, including submission, storage and data access systems, are

developed at EBI and made useful by close collaboration with a large number of external partners in a collaborative ecosystem. Specific user groups with expertise in particular sequencing applications or disciplines within biology collaborate with ENA to provide specialist submission and data presentation services around the generic core. With echoes of large infrastructure development initiatives in operation under the European Commission's ELIXIR project, the ENA collaborative ecosystem model promises long-term sustainability and increased utility for its part of the scientific record.

In the talk, I will briefly introduce the ENA, describe the collaborative ecosystem model, taking existing collaborations as example, and describe the opportunity to develop the model in the context of environmental genomics data. I will outline programmatic submission and data access interfaces and will provide an update on ongoing pilot studies to establish metagenomics data brokering to ENA under the collaborative ecosystem model.

- **Owen White**, University of Maryland **12:05-12:30**

  Metagenome Data Integration, Data Storage and Retrieval

  We are developing an Open Data Framework (OSDF) to provide genomic data sets and analysis resources to the scientific community. OSDF consists of a database, a data exchange format, and an Application Program Interface (API) to support data retrieval and submissions from the community. Unlike conventional public archives such as GenBank, EMBL or DDBJ that utilize centralized databases, file systems and web servers to deliver data to the user, the OSDF will utilize cloud computing and cross-site replication technologies to support a distributed data access and storage model. The OSDF is specifically intended to break the paradigm of conventional web-based genome resources (e.g., the UCSC Genome Browser, FlyBase, WormBase) typically used by model organism databases that are confined to monolithic web servers with centralized data. This infrastructure will significantly more scalable and able to accommodate rapidly growing genomic metagenomic datasets derived from next generation sequencing and promote data sharing. OSDF will support access through an API and pre-defined virtual machines, including Qiime (http://qiime.sourceforge.net/) and CloVR (http://clovr.org/). The API is designed for use in both analytical pipelines and online web resources. OSDF also takes advantage of virtual machines to access and analyze the data. In previous efforts we have shown that users are able to launch ~1000 virtual machines on cloud based systems and represents unprecedented access to remote, dynamically scalable computational power to the average user.

- **Lunch and Discussion** **12:30-13:30**

Session 7 13:30-15:30

High-performance Computing

Chair – Zhong Wang, DOE Joint Genome Institute

Questions:
- Porting of available software onto HPC infrastructure
- Use of GPU

- **Shane Canon**, Berkeley Lab **13:30-13:45**

  Exploiting HPC Platforms for Metagenomics: Challenges and Opportunities

  The rapid growth in data from Next-Generation Sequencing has created an analysis challenge that threatens to limit the potential impact from these new platforms. Consequently, scientists are looking at a variety of computational platforms to address this demand. HPC centers provide an immense scale of computing resource and can potentially play an important role in meeting this demand. However, these platforms also introduce new challenges. We will discuss some of these challenges along with approaches that have been used to overcome these limitations. We will also discuss some of the still untapped potential of these systems and how they can further play a role in meeting future computing needs.

- **Narayan Desai**, Argonne National Lab **13:45-14:15**

  Scaling MG-RAST to Terabases

  MG-RAST is a large scale analysis resource for metagenomic data. In order to support the analysis of many large scale datasets, we have adapted MG-RAST to perform its computations across an ensemble of HPC systems. In this talk, I will describe the overall architecture for MG-RAST's computational architecture, as well as the tactics we have employed to ensure good scalability and efficiency.

- **Brooklin Gore**, Morgridge Institute for Research **14:15-14:45**

  HTC in Support of Genomics

  For more than two decades we have been serving the High Throughput Computing (HTC) needs of a broad range of science domains. This includes scientists using HTC to extract biological insight form genomics data. The broad adoption of our tools and technologies by these scientists are a clear display of the value our approach brings to the ever-growing computational challenges of genomics research. HTC allows researchers to harness not only the power of their local resources, but also campus, national and cloud resources. We will present the general capabilities of HTC and the requirements for performing genomic science with HTC. Examples from projects in genotype-to-phenotype mapping, optical mapping, and next-generation sequencing will be used and support for analysis tools like R and MatLab will be discussed. An overview of the Center for High Throughput Computing at the University of Wisconsin-Madison and the Open Science Grid will be provided. The talk will conclude with insights into future work integrating HTC with interactive portals like Galaxy and emerging computing resources like GPUs.

- **Karan Bhatia**, Amazon **14:45-15:15**

  AWS for Scientific Computing

  Cloud computing is not just used for building highly scalable websites -- the same technologies can be used to build highly elastic analytic frameworks that are flexible and cost effective. In this talk, I'll describe some of the new service offerings that would be appropriate for scientific computing.

- **Break** **15:15-15:30**

Session 8 15:30-16:30

Break-out Discussions on Workshop Themes

- **Day I: Progress and Challenges in Metagenome Assembly**
  Chairs – Alex Copeland, DOE Joint Genome Institute
  Patrick Chain, Los Alamos National Laboratory
  - Metagenome-specific Assembly
  - Assembly Quality Evaluation
  - Single Cells and Metagenomes

- **Day II: Beyond Metagenome Assembly**
  Chairs – Victor Markowitz and Kostas Mavrommatis, DOE Joint Genome Institute
  - Metagenome Data Integration, Data Storage and Retrieval
  - Scalability of Comparative Analysis and Novel Algorithms and Tools
  - High-performance Computing

Session 9 16:30-17:00

General Discussion & Workshop Conclusion

Chairs – Nikos Kyrpides, Susannah Tringe, DOE Joint Genome Institute

- ***Shuttle Transport*** *- JGI to Walnut Creek Marriott* ***17:00***

Attendance

Confirmed Speakers

1.	**Jill Banfield**	University of CA, Berkeley	jbanfield@berkeley.edu
2.	**Karan Bhatia**	Amazon Web Services	karanb@amazon.com
3.	**C. Titus Brown**	Michigan State University	ctb@msu.edu
4.	**R. Shane Canon**	Berkeley Lab	scanon@lbl.gov
5.	**Francis Chin**	University of Hong Kong	chin@cs.hku.hk
6.	**Guy Cochrane**	EMBL-EBI	cochrane@ebi.ac.uk
7.	**Paramvir Dehal**	Berkeley Lab	psdehal@lbl.gov
8.	**Narayan Desai**	Argonne National Lab	desai@mcs.anl.gov
9.	**Johannes Goll**	J. Craig Venter Institute	jgoll@jcvi.org
10.	**Brooklin Gore**	Morgridge Inst. for Research	BGore@morgridgeinstitute.org
11.	**Phil Hugenholtz**	University of Queensland	Phugenholtz@gmail.com
12.	**Sergey Koren**	University of Maryland	sergek@umd.edu
13.	**Weizhong Li**	San Diego Supercomputer Center	liwz@sdsc.edu
14.	**Mihai Pop**	University of Maryland	mpop@umiacs.umd.edu
15.	**Steve Quake**	Stanford University	quake@stanford.edu
16.	**Doug Rusch**	J. Craig Venter Institute	drusch@jcvi.org
17.	**Yasubumi Sakakibara**	Keio University	yasu@bio.keio.ac.jp
18.	**Alex Sczyrba**	DOE Joint Genome Institute	ASczyrba@lbl.gov
19.	**Jared Simpson**	Sanger Institute	js18@sanger.ac.uk
20.	**Ramunas Stepanauskas**	Bigelow Laboratory	rstepanauskas@bigelow.org
21.	**Owen White**	University of Maryland	owhite@som.umaryland.edu

Invited Participants

1.	**Kostas Billis**	DOE Joint Genome Institute	KBillis@lbl.gov
2.	**Jim Bristow**	DOE Joint Genome Institute	JBristow@lbl.gov
3.	**Patrick Chain**	Los Alamos National Lab	PChain@lanl.gov
4.	**Amy Chen**	DOE Joint Genome Institute	IMAChen@lbl.gov
5.	**Dylan Chivian**	Berkeley Lab	DCChivian@lbl.gov
6.	**Scott Clingenpeel**	DOE Joint Genome Institute	SRClingenpeel@lbl.gov
7.	**Alex Copeland**	DOE Joint Genome Institute	ACCopeland@lbl.gov
8.	**Bob Cottingham**	Oak Ridge National Lab	cottinghamrw@ornl.gov
9.	**Rob Egan**	DOE Joint Genome Institute	RSEgan@lbl.gov
10.	**Marsha Fenner**	DOE Joint Genome Institute	MWFenner@lbl.gov

11. **Tracey Freitas**	Los Alamos National Lab	tracey.freitas@gmail.com
12. **Jeff Froula**	DOE Joint Genome Institute	JLFroula@lbl.gov
13. **Igor Grigoriev**	DOE Joint Genome Institute	IVGrigoriev@lbl.gov
14. **Steve Hallam**	Univ. of British Columbia	shallam@interchange.ubc.ca
15. **Shaomei He**	DOE Joint Genome Institute	SHe@lbl.gov
16. **Matthias Hess**	Washington State University	matthias.hess@tricity.wsu.edu
17. **Adina Chuang Howe**	Michigan State University	adina.chuang@gmail.com
18. **Marcel Huntemann**	DOE Joint Genome Institute	MHuntemann@lbl.gov
19. **Natalia Ivanova**	DOE Joint Genome Institute	NNIvanova@lbl.gov
20. **Janet Jansson**	Berkeley Lab	JRJansson@lbl.gov
21. **Denis Kaznadzey**	DOE Joint Genome Institute	DKaznadzey@lbl.gov
22. **Ed Kirton**	DOE Joint Genome Institute	ESKirton@lbl.gov
23. **Nikos C. Kyrpides**	DOE Joint Genome Institute	NCKyrpides@lbl.gov
24. **Dino Liolios**	DOE Joint Genome Institute	KLiolios@lbl.gov
25. **Stephanie Malfatt**i	DOE Joint Genome Institute	SAMalfatti@lbl.gov
26. **Victor Markowitz**	DOE Joint Genome Institute	VMMarkowitz@lbl.gov
27. **Kostas Mavrommatis**	DOE Joint Genome Institute	KMavrommatis@lbl.gov
28. **Brett McMillion**	Amazon Web Services	mcmillen@amazon.com
29. **Henrik Nordberg**	DOE Joint Genome Institute	HNordberg@lbl.gov
30. **Krishna Palaniappan**	DOE Joint Genome Institute	KPalaniappan@lbl.gov
31. **Amrita Pati**	DOE Joint Genome Institute	APati@lbl.gov
32. **Abhishek Pratap**	DOE Joint Genome Institute	APratap@lbl.gov
33. **Chris Rinke**	DOE Joint Genome Institute	CRinke@lbl.gov
34. **Dan Rokhsar**	DOE Joint Genome Institute	DSRokhsar@gmail.com
35. **Eddy Rubin**	DOE Joint Genome Institute	EMRubin@lbl.gov
36. **Matt Scholtz**	Los Alamos National Lab	MScholz@lanl.gov
37. **Itai Sharon**	University of CA, Berkeley	itai.sharon@gmail.com
38. **Weibing Shi**	DOE Joint Genome Institute	WShi@lbl.gov
39. **Shannon Steinfadt**	Los Alamos National Laboratory	Shannon@lanl.gov
40. **Ernest Szeto**	DOE Joint Genome Institute	ESzeto@lbl.gov
41. **Brian Thomas**	University of CA, Berkeley	bcthomas@berkeley.edu
42. **Julien Tremblay**	DOE Joint Genome Institute	JTremblay@lbl.gov
43. **Susannah Tringe**	DOE Joint Genome Institute	SGTringe@lbl.gov
44. **Kai Wang**	DOE Joint Genome Institute	KaiWang@lbl.gov
45. **Zhong Wang**	DOE Joint Genome Institute	ZhongWang@lbl.gov
46. **Tanja Woyke**	DOE Joint Genome Institute	TWoyke@lbl.gov
47. **Fangfang Xia**	Argonne National Lab	Fangfang.Xia@gmail.com
48. **S. M. Yiu**	University of Hong Kong	smyiu@cs.hku.hk