3rd Annual Finishing in the Future Meeting

Santa Fe, New Mexico

May 28th - 30th, 2008

Los Alamos NATIONAL LABORATORY EST. 1943

Office of Science U.S. DEPARTMENT OF ENERGY

JGI

Contents

Agenda Overview…….. 3

May 28th Agenda……. 5

Speaker Presentations (May 28th)..……. 7

Poster Session (even #s)……..…..... 17

Meet and Greet Party w/ Food & Beverages... 39

May 29th Agenda……. 41

Speaker Presentations (May 29th)..…… 43

Poster Session (odd #s) w/ Wine & Cheese... 57

May 30th Agenda……. 83

Speaker Presentations (May 30th)..…… 85

Close of Meeting Discussion…..…… 90

Attendees…… 95

Map & History of Santa Fe, NM…… 97

The 2008 "Finishing in the Future" Organizing Committee:

* *Chris Detter, Ph.D., JGI-LANL / Genome Sciences Center Director, LANL*
* *Johar Ali, Ph.D., Technology Development Team Leader, OICR*
* *Ruby Archuleta, Genome Administrative Assistant, LANL*
* *Patrick Chain, Finishing Coordinator / Group Leader, LLNL*
* *Michael Fitzgerald, Finishing Manager, Broad Institute*
* *Bob Fulton, M.S., Sequence Improvement Group Leader, WashU*
* *Darren Grafham, Finishing Coordinator, Sanger Institute*
* *Hoda Khouri, M.S., Staff Scientist, NCBI*
* *Alla Lapidus, Ph.D., Microbial Genomics Group Leader, LBNL-JGI*
* *Donna Muzny, M.S., Director of Operations, BCM*

Sponsors: Roche Diagnostics, Applied Biosystems, and Integrated DNA Technologies….Thank You!!!

05/28/2008 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**La Fonda Breakfast Buffet** (French "Texas Toast", Fluffy scramble eggs, Grilled breakfast potatoes, Applewood smoked bacon, breads, and fruit, etc.)	x
830 - 845	Intro	x	Welcome Intro	**Jose Olivares**
845 - 930	**Keynote**	**FF0052**	**Defining the Human Microbiome: Friends or Family?**	**Bruce Birren**
930 - 1000	Speaker 1	FF0090	Sequence Improvement at the Genome Sequencing Center	**Bob Fulton**
1000 - 1030	Break	x	Beverages provided	x
1030 - 1100	Speaker 2	FF0125	Metagenomic Approaches Targeted Toward the Human Microbiome	**Joe Petrosino**
1100 - 1130	Speaker 3	FF0103	Reconstruction of Individual Genomes of Uncultured Microbial Taxa from Complex Communities	**Ramunas Stepanauskas**
1130 - 1200	Speaker 4	FF0110	Targeted genomics of uncultured microbes: from single cells to populations	**Mircea Podar**
1200 - 130pm	Lunch	x	**Coronado Lunch Buffet** (Char-grilled chicken breast with barbecue-chipotle vinaigrette, Pan-seared rainbow trout fillet served with smoked yellow pepper coulis, Roasted garlic mashed potatoes and seasonal vegetables, etc.)	x
130 - 200	Speaker 5	FF0105	New scientific breakthroughs using the 454 Genome Sequencer FLX: 400 base pair reads and more…..Roche 454	**Tim Harkin**
200 - 230	Speaker 6	FF0102	Next Generation Sequencing - Illuminating the Genome - Illumina corp.	**Haley Fiske**
230 - 300	Speaker 7	FF0116	High-resolution Structural Variation Detected with Ultra High-throughput Sequencing of Paired End Libraries…..AB SOLiD	**Michael Rhodes**
300 - 400	Panel Discussion	x	**Panel Discussion** - Next Generation Sequincing Technologies	Chair - Donna Muzny
400 - 430	Break	x	Beverages and snacks provided	x
415 - 530	**Posters - even #s**	x	**Poster Session**	x
530 - 800pm	**Meet & Greet Party**	x	**Meet & Greet Party - sponsored by Roche --- Food & Drinks**	x

05/29/2008 - Thursday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Santa Fe Breakfast Buffet** (Scrambled eggs with a choice of three accompaniments on the side -chilaquiles with green chile and cheese, chorizo sausage and roasted green chile, Grilled breakfast potatoes, applewood-smoked bacon and warm flour tortillas, assorted breads and fruits, etc.)	x
830 - 845	Intro	x	Welcome Back Intro	**Jim Bristow**
845 - 930	**Keynote**	**FF0050**	**The Unfinished Genome**	**Sydney Brenner**
930 - 955	Speaker 1	FF0085	How Close to Finished can an Assembler Get?	**Jim Knight**
955 - 1020	Break	x	Beverages provided	x
1020 -1045	Speaker 2	FF0098	Taking the Next Step: Assembling Large Genomes Using NextGen Sequencing	**Sean Sykes**
1045 -1110	Speaker 3	FF0034	De Novo Assembly of Microbial Genomes from Illumina Whole-Genome Shotgun Data	**Iain MacCallum**
1110 - 1135	Speaker 4	FF0036	Combining Next-Gen Sequence Technologies in Multi-platform Assemblies	**Christian Buhay**
1135 - 1200	Speaker 5	FF0010	The Arcturus – Minerva assembly management system	**Ed Zuiderwijk**
1230 - 130pm	Lunch	x	**New Mexican Lunch Buffet** (Pork tenderloin achiote-rubbed and char-grilled with tomatillo-chipotle sauce, your choice of either Chicken or Cheese enchiladas with red or green chile, etc.)	x
130 -155	Speaker 6	FF0076	Bacterial Genome Finishing on the Turn of New Technology	**Hajnalka Kiss**
155 -220	Speaker 7	FF0088	Towards "Finishing" of Eukaryotic Transcriptomes	**Stephen Kingsmore**
220 - 245	Speaker 8	FF0075	An optimized DNA sequencing pipeline to support drug discovery at Wyeth	**Jan Kieleczawa**
245 - 310	Speaker 9	FF0065	Sequencing the Genome, Transcriptome, Methylome and smRNAome of the Arabidopsis Ecotype, Cape Verde Island	**Ronan O'Malley**
310 - 415	Panel Discussion	x	**Panel Discussion** - Sequencing and Finishing Standards	Chair - Darren Grafham
415 - 630	**Wine & Cheese**	x	**Beverages, Wine & Cheese provided - sponsored by Applied Biosystems**	x
415 - 630	**Posters - odd #s**	x	**Poster Session with Wine & Cheese - sponsored by Applied Biosystems**	x
630 - bedtime	on your own	x	Dinner and night on your own - enjoy	x

05/30/2008 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Healthy Start Breakfast Buffet** (Scrambled Eggs on side tomatoes, scallions and spinach, Turkey sausage links, Assorted chilled fruit juices, Platter of freshly sliced seasonal fruit, Assorted and bran muffins with butter, Granola and oatmeal served with low-fat milk, Individual assorted fruit yogurts, etc.)	x
830 - 845	Intro	x	Welcome Back Intro	**Chris Detter**
845 - 930	**Keynote**	**FF0053**	**Great Expectations: Fulfilling the promises of the Human Genome Project**	**Deanna Church**
930 - 1000	Speaker 1	FF0087	Using Consed and Cross_match in Resequencing Projects	**David Gordon**
1000 -1030	Break	x	Beverages and snacks provided	x
1030 - 1100	Speaker 2	FF0008	1000 Genomes Project Data Management and Analysis	**Hoda Khouri**
1100 -1130	Speaker 3	FF0020	Genome annotation improvement using new massive sequencing technologies.	**François Artiguenave**
1130 -1200	Speaker 4	FF0041	AlpheusTM - a system for nucleotide variant detection and digital gene expression analysis in ultra-high throughput sequencing projects	**Neil Miller**
1200 - 1230	Closing Discussions	x	**Closing Discussions -** discuss next year's plans	Chair - Chris Detter
1230 - 200pm	Lunch & Close of meeting	x	**La Fiesta Plaza Lunch Buffet -** (Chicken and beef fajitas with grilled red onions and bell peppers, Black beans (Vegetarian), Spanish rice (Vegetarian), Pork posole and calabacitas rancheras, Warm flour tortillas and butter, etc.) **End of meeting, enjoy lunch and Santa Fe**	x

05/28/2008 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**La Fonda Breakfast Buffet** (French "Texas Toast", Fluffy scramble eggs, Grilled breakfast potatoes, Applewood smoked bacon, breads, and fruit, etc.)	x
830 - 845	Intro	x	Welcome Intro	**Jose Olivares**
845 - 930	**Keynote**	**FF0052**	**Defining the Human Microbiome: Friends or Family?**	**Bruce Birren**
930 - 1000	Speaker 1	FF0090	Sequence Improvement at the Genome Sequencing Center	**Bob Fulton**
1000 - 1030	Break	x	Beverages provided	x
1030 - 1100	Speaker 2	FF0125	Metagenomic Approaches Targeted Toward the Human Microbiome	**Joe Petrosino**
1100 - 1130	Speaker 3	FF0103	Reconstruction of Individual Genomes of Uncultured Microbial Taxa from Complex Communities	**Ramunas Stepanauskas**
1130 - 1200	Speaker 4	FF0110	Targeted genomics of uncultured microbes: from single cells to populations	**Mircea Podar**
1200 - 130pm	Lunch	x	**Coronado Lunch Buffet** (Char-grilled chicken breast with barbecue-chipotle vinaigrette, Pan-seared rainbow trout fillet served with smoked yellow pepper coulis, Roasted garlic mashed potatoes and seasonal vegetables, etc.)	x
130 - 200	Speaker 5	FF0105	New scientific breakthroughs using the 454 Genome Sequencer FLX: 400 base pair reads and more…..Roche 454	**Tim Harkin**
200 - 230	Speaker 6	FF0102	Next Generation Sequencing - Illuminating the Genome - Illumina corp.	**Haley Fiske**
230 - 300	Speaker 7	FF0116	High-resolution Structural Variation Detected with Ultra High-throughput Sequencing of Paired End Libraries…..AB SOLiD	**Michael Rhodes**
300 - 400	Panel Discussion	x	**Panel Discussion** - Next Generation Sequincing Technologies	Chair - Donna Muzny
400 - 430	Break	x	Beverages and snacks provided	x
415 - 530	**Posters - even #s**	x	**Poster Session**	x
530 - 800pm	**Meet & Greet Party**	x	**Meet & Greet Party - sponsored by Roche --- Food & Drinks**	x

Speaker Presentations (May 28th)

Abstracts are in order of presentation according to Agenda

FF0052 – Keynote

Defining the Human Microbiome: Friends or Family?

Bruce Birren

Broad Institute of MIT & Harvard, Cambridge, MA 02142

FF0090

Sequence Improvement at the Genome Sequencing Center

Robert S. Fulton, Tina Graves, Patrick Minx and Richard Wilson

Washington University School of Medicine, Genome Sequencing Center, St. Louis, MO 63108

The Genome Sequencing Center (GSC) at Washington University School of Medicine has been a world leader in sequence improvement efforts for many years. These works have included finishing efforts on pure clone-based assemblies including *C. elegans*, and human, combined whole genome shotgun and clone based products such as the chimpanzee and mouse genomes, as well as whole genome finishing efforts on many bacterial species. In addition to finishing efforts, the GSC has improved other genomes such as the clone based maize genome, where only "unique" regions of the sequence were improved to "finished" standards, while other more repetitive regions were improved, but not finished. Whole genome improvement initiatives have also been undertaken with primer-directed clone walks totaling well over 750,000 reactions, on genomes such as chicken, chimpanzee, and a variety of *Drosophilia* and nematode species. These efforts have been aided by a production style reaction system leveraging the centers LIMS, production capacity, and software packages such as autofinish and autoedit (David Gordon). Despite this robust system, the introduction of next generation sequencing technology has initiated further refinement of our current approaches as well as new methods of sequence improvement, leveraging the compelling cost and throughput metrics, and minimally biased, robust sequencing dynamics of these next generation platforms. This presentation will highlight some of our successes, as well as outline some of our development processes moving forward.

FF0125

Metagenomic Approaches Targeted Toward the Human Microbiome

Joseph F. Petrosino[1,2], Sarah K. Highlander[1,2], Xiang Qin[1,2], Kim C. Worley[2], Donna M. Muzny[2], and Richard A. Gibbs[2]

[1]Department of Molecular Virology and Microbiology, [2]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030

The microbial communities that comprise the human microbiome have remained largely understudied, however next-generation (next-gen) DNA sequencing and nucleic acid enrichment technologies now enable the comprehensive characterization of the human microbiota and its role in human health and disease. We are implementing a multi-faceted pipeline using next-gen technologies to examine the microbiota of human populations in varying stages of health and disease. Currently, strategies to best assess the diversity and depth of the microbial species found in metagenomic samples from various body sites are being evaluated. Established methods for metagenomic sampling, such as 16S rDNA sequencing, are evolving toward 454-FLX-, Solexa-, and SOLiD-based strategies as these latter approaches provide much deeper data sets at an increasingly lower cost.

We are also proposing a pipeline using existing technologies to enrich for, and determine the sequence of, genomic DNA from unknown/uncultivatable bacteria (UUB). Changes in the populations of these more elusive organisms may be the most significant harbingers of progression/recovery in various human diseases. Only recently have methods been described to detect, isolate, and/or determine the whole genome sequence of uncultivatable bacteria. Our strategy begins with 16S rRNA gene and random DNA sequencing of metagenomic samples to identify sequences from potential UUB that are present in the sample. A Nimblegen high-density microarray incorporating oligos derived from these sequences is used to capture long fragments of DNA from the UUB in the sample. These captured fragments are sequenced and assembled to define longer regions of the UUB genomes. This capture-array strategy is then repeated to produce even longer genomic sequences. Depending on sample complexity, some genomes may be complete at this stage. If desired, whole-organism enrichment using protein-DNA chimeras (called ‘tadpoles’) may be performed to isolate targeted bacteria. DNA can then be isolated from enriched cells and used to determine the genomic sequence of the targeted bacteria. The platforms summarized here will ensure that the “core human microbiome” can be described at high resolution, and that microbiome fluctuations resulting from changes in human health can be detected more easily.

FF0103

RECONSTRUCTION OF INDIVIDUAL GENOMES OF UNCULTURED MICROBIAL TAXA FROM COMPLEX COMMUNITIES

Stepanauskas R, Woyke T, Xie G, Han C, Copeland A, Chatterji S, and Sieracki ME

Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine 04575-0475

Single cell genomics is a novel, transformative approach for studying the uncultured microbial majority, complementing the strengths and limitations of cultivation and metagenomics. We developed protocols for high-speed fluorescence-activated sorting of single cells, their whole genome multiple displacement amplification and subsequent DNA sequencing. We used this approach to generate single amplified genomes (SAGs) from individual prokaryote cells and screened them for phylogenetic markers and for genes encoding biogeochemically significant enzymes. This permitted a robust, culture-independent, high-throughput and cost-effective matching of phylogeny and metabolism. Whole genome sequencing was performed on SAGs of several proteorhodopsin-containing Flavobacteria from the Gulf of Maine. We recovered 60-80% of the genomes in 2-700 kbp contigs, enabling the reconstruction of unique photometabolic and biosynthetic pathways. Comprehensive DNA quality control procedures were developed, indicating significant bias but negligible contamination in the single cell whole genome amplification reactions. Recruitment of fragments from the Global Ocean Sampling database, using the two SAGs as references, resulted in a significant number of >95% identity alignments, all from the Northeast US coast. In contrast, none of the existing isolate genomes, including seven marine Flavobacteria, produced significant fragment recruitment, except for *Pelagibacter, Prochlorococcus, and Synechococcus*. Thus, for the first time, we demonstrate the reconstruction of genomes and metabolic pathways of representative, uncultured microbial taxa from complex environmental communities.

FF0110

Targeted genomics of uncultured microbes: from single cells to populations

Mircea Podar

Oak Ridge National Laboratory

Two main categories of approaches are currently used for genomic sequencing of bacteria and archaea: whole genome sequencing of organisms that can be propagated by clonal expansion in the laboratory and community shotgun sequencing, which uses environmental samples with a varying degree of complexity. Whole community shotgun sequencing ("metagenomics"), while it provides information about the natural sequence variation in populations of related organisms and allows access to the genomes of uncultivated organisms, is limited in terms of coverage depth, assembly options and sequence assignment to specific taxa, especially for highly diverse consortia. A relatively novel technique, whole genome amplification using relatively small populations and even single cells, allows the synthesis of sufficient DNA for genomic sequencing. Such cells or populations can be taxonomically targeted using fluorescence in situ hybridization and specifically isolated from complex consortia using flow cytometry. Application of this concept has resulted in partial genome sequencing for a representative from the uncultured bacterial division TM7. The strengths and current limitations of this approach will be discussed.

FF0105

New scientific breakthroughs using the 454 Genome Sequencer FLX: 400 base pair reads and more

Tim Harkin

454 Life Sciences, A Roche Company, Branford, CT 06405, USA

The Genome Sequencer FLX developed by 454 Life Sciences has been used in a wide variety of applications from the sequencing of complex genomes to transcriptional studies to ultra-deep sequencing for detecting somatic mutations. Later this year, users of the system will be able to obtain sequencing reads in excess of 400 base pairs and have access to generating paired-end reads that span 20 kb of the genome. These improvements are opening even newer opportunities for next generation sequencing. Some recent examples will be presented to demonstrate these improvements.

FF0102

Next Generation Sequencing - illuminating the Genome

Haley Fiske

illumina, Inc. 25861 Industrial Blvd, Hayward, CA 94545

The illumina Genome Analyzer, based on the Solexa massively parallel sequencing-by-synthesis technology, is being used for a broad set of functional genomics applications including chromosomal re-arrangements, to single nucleotide variations, variation in DNA methylation, whole transcriptome analysis, small RNA analysis, digital gene expression, DNA-protein, and DNA-RNA interaction analysis. Details on the current state of the technology as well as future technology improvements will be presented.

FF0116

High-resolution Structural Variation Detected with Ultra High-throughput Sequencing of Paired End Libraries

Michael D. Rhodes[1],Heather E. Peckham, Stephen F. McLaughlin, Swati S. Ranade, Christopher R. Clouser, Jonathan M. Manning, Cynthia L. Hendrickson, Lei Zhang, Eileen T. Dimalanta, Tanya D. Sokolsky, Jeffrey K. Ichikawa, Jason B. Warner, Kevin J. McKernan and Joel A. Malek

Applied Biosystems, 500 Cummings Center, Beverly, MA 01915, [1] Applied Biosystems 850 lincoln centre drive, Foster City, CA

The SOLiD™ System next generation of DNA sequencing platform produces short sequencing reads with increased depth of coverage compared to Sanger sequencing. In order to utilize such a system for comprehensive genome resequencing it is essential to use paired end libraries. In a paired end library the library consists of two short tags that were originally separated by a known distance in the target genome. This allows assembly where the target genome has deletions, insertions, duplications, inversions and rearrangements. The use of paired ends also overcomes the problem with placement of short reads on repetitive genomes. We have applied a ligation-based high-throughput sequencing approach to the Yoruban sample NA18507 utilizing eight paired end libraries ranging in size from 600 to 6,000 base pairs. The diverse range of library sizes allows an unprecedented level of resolution of structural variation. We report a high-resolution map of the diploid structural variations present in this Yoruban individual compared to the human genome reference sequence.

Panel Discussion Notes

Panel Discussion Notes

Poster Presentations (Even #s, May 28th)

FF0006

The JGI Contribution to Whole Microbial Genome Sequencing

David Sims[1], Cliff Han[1], David Bruce[1], Tom Brettin[1], Chris Detter[1], and Cheryl Kuske[1]

[1]Los Alamos National Laboratory

The DOE Joint Genome Institute (JGI) was created in 1997 to unite the expertise and resources in genome mapping, DNA sequencing, technology development, and information sciences pioneered at the DOE genome centers at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL).

Since that time, there has been an exponential increase in the number of whole microbial genome sequences that have been made publicly available through the National Center for Biotechnology Information (NCBI) as a result of the work of the JGI. There has been a similar increase in the proportion of the whole genomes that have been submitted by the JGI as compared to all other centers. To date, approximately 200 whole microbial genomes are available at www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1 that were completed by the JGI. This accounts for approximately 1/3 of all of the genomes, and twice the input of the next finishing center.

The JGI continues to strive to be the world leader in its contribution to this resource of basic microbiological knowledge, and to pave the way to next generation genome sequencing techniques and technologies.

LA-UR#081494

FF0010

The Arcturus – Minerva assembly management system

E.J. Zuiderwijk, D. Harper, M-A Rajandream, J. Parkhill, K. Mungall

The Wellcome Trust Sanger Institute Hinxton Cambridge UK

In the past years we have developed an assembly data management system based on the MySQL database engine and Java GUI facilities. The Arcturus back-end consist of a set of Perl-modules and scripts which interface the MySQL database with the the Gap4 tools used by finishers and with the assembly pipeline(s).

Assemblies are represented as mappings between the basic objects (read-contig, contig-contig) which are built during loading operations. Arcturus automatically links newly created contigs to their predecessors. This allows tracking of a contig's history and it enables annotation and tags to be remapped onto the latest version of a contig.

The Minerva front-end enables users to view the contents of assemblies and projects, to move contigs between projects and to import new reads into projects. In addition there are tools for data analysis such as an oligo finder and scaffold analysis.

Recently we have included in the GUI the functionality to import and export data from and to Gap4 databases, which effectively gives the finishers control over all aspects of the system's operation.

FF0018

Optimization of finishing procedures at Genoscope in the frame of the New Sequencing Technologies

Barbe V.[1], Mangenot S. [1], Anthouard V. [1], Aury J.M. [1],Couloux A. [1], Dossat C. [1], Jubin C. [1], Vacherie B. [1], Vallenet D. [1], Scarpelli C. [1], Wincker P. [1] and Weissenbach J. [1]

1-Commissariat à l'Energie Atomique, Institut de Génomique, Genoscope, Evry, France

The appearance of new sequencing technologies has had an important impact on prokaryotic genomes sequencing, assembly and finishing strategies. For about one year, all microbial genomes were sequenced at Genoscope with a mix of ~15X coverage with 454 reads and 4X coverage with Sanger reads obtained from a 10kb low copy plasmid library.

We use the "HybridAssemble" version of Arachne assembler, developed by the Broad institute (www.broad.mit.edu), that combines the 454 contigs and Sanger reads. The assemblies are validated with the Mekano interface, developed at the Genoscope, and based on visualization of clone links inside and between contigs.

The combination of these different sequencing data led to modify the Genoscope "Autofinishing" software tools, usually used for walk and polishing steps.

We will describe new finishing methodologies developed with the aim of completing microbial genomes sequences with a comparable level of quality to those obtained with Sanger-only data.

FF0022

Finishing the 454/Sanger hybrid assembly of the *Phytophthora capsici* genome

Joann Mudge[1], Neil Miller[1], Kurt H. Lamour[2], Sophien Kamoun[3], Paul M. Richardson[4], Darren Platt[4], Igor Grigoriev[4], Alan Kuo[4], Jeremy Schmutz[5], Olga Chertkov[6], Cliff S. Han[6], Chris Detter[6], Greg D. May[1], William D. Beavis[7], Stephen F. Kingsmore[1]

[1]National Center for Genome Resources, Santa Fe, New Mexico, USA, [2]The University of Tennessee, Department of Entomology and Plant Pathology, Knoxville, Tennessee, [3]Sainsbury Laboratory, Norwich, UK, [4]USA Department of Energy Joint Genome Institute, Walnut Creek, California, USA, [5] Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, Palo Alto, California, USA [6]Los Alamos National Laboratory, Los Alamos, New Mexico, USA, [7]Iowa State University, Ames, Iowa, USA

Next generation sequencing technologies have created sequencing opportunities by increasing throughput and decreasing costs. They have also introduced several challenges compared to traditional sequencing technology. In *de novo* assembly, assembly strategies must take into account the large amount of sequencing data as well as the short read lengths and distinctive error profiles of the technology. Hybrid assemblies have been used to mitigate some of these issues by combining next generation sequencing with traditional Sanger sequence and have been particularly effective in *de novo* sequencing of prokaryotic genomes. We have assembled a draft consensus of a eukaryotic genome sequenced using Sanger and 454 sequence. The organism, *Pythophthora capsici,* is an oomycete and devastating pathogen of vegetable crops. Its highly repetitive, 60 Mb genome, with characteristic eukaryotic complexity, as well as the availability of closely related Sanger-sequenced genomes make *P. capsici* an excellent arena for benchmarking hybrid sequencing in eukaryotes. We generated 23X 454 GS20 pyrosequencing singleton reads, 5X Sanger paired reads, and 2M 454 GS20 paired reads. In addition, 20,000 sequencing reactions have been completed to obtain sequence within captured gaps. We assembled these reads using Forge, modified to accommodate the disparate read lengths and pair spacings of the two technologies as well as the unique error profiles of the Sanger and 454 reads. For comparison purposes, a Sanger-only (Arachne) and 454-only (Newbler) assemblies were also generated. Finally, Solexa cDNA libraries from nine life stages have been sequenced to improve gene calls, identify SNPs, and compare gene expression levels. The successes and challenges of doing a 454-based genome sequencing project will be discussed.

FF0024

The CMap Assembly Editor

Faga, B [1], Carmichael, L [2], Belter, E [2], Minx, P [2], Stein, L [1]
1. Cold Spring Harbor Laboratory, Cold Spring Harbor NY
2. Washington University, St. Louis MO

The CMap Assembly Editor (CMAE) is a desktop application being developed to direct editing of large scale sequence assemblies. CMAE allows visualization of multiple sources of genomic data, including optical, physical and genetic maps, together with a sequence assembly to give finishers context with which to make decisions. For example, the editor can display sequence contigs aligned to fingerprinted physical map contigs, which are aligned in turn to genetic maps. Correspondence links between the different objects allow a finisher to see how the assembly agrees with the related data and can highlight possible mis-assemblies. The finisher can then mark mis-assembled contigs to be split, merged, moved, flipped or viewed in a more specialized program. CMAE relies on a highly customizable plug-in system in order to modify data outside of the CMap database.

FF0026

The Zebrafish: Small Stripy Fish, Large Genome Project

Katherine Auger

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

The Wellcome Trust Sanger Institute (WTSI) has employed an integrated pipeline of mapping, sequencing, finishing and annotation in parallel to provide a genome of the highest standard for the Zebrafish (*Danio rerio*) community to use as early as possible. Providing a product of unified quality is important for the community and this is aided by the ability to deal with queries directly from end users through zfish-help. Mapping and sequencing began at WTSI in 2001 and we estimate reaching “essentially complete” by the end of 2009. The reference genome sequence is currently 77.3% complete.

The quality of the sequence and the map is continually improving throughout the project. In mapping core there has been the electronic selection of clones to cover missing cDNAs. This process has utilized the whole genome shotgun (WGS) and has provided a targeted approach to the tilepath alongside extending existing contigs. There is ongoing research into the use of different maps with new marker sets and we are investigating the novel application of new sequencing technologies for mapping purposes.

Following the shotgun process, clones are assessed for redundancy, suitability for sequence improvement and to evaluate sequence overlaps. Clones requiring additional improvement, before entering finishing, are directed through a round of automated sequence improvement. The finishing process, and subsequent quality control, provides gold standard sequence essential for effective manual annotation and ensuing experimental biology.

WGS provides an early representation of the data set and is also utilized during the finishing process, providing an extremely valuable resource. We are able to use contigs with assigned quality values, reflecting the true values of the original Phred scores. We can import reads from the WGS contig for closer interrogation, for example in regions of local repeat. For each public release of the data WGS sequence is replaced by mapped and finished clone sequence where available. The existing assembly, Zv7, can be viewed at http://www.ensembl.org/Danio_rerio/index.html. WTSI are currently integrating finished sequence into a new build, Zv8.

Here we provide an update of the Zebrafish genome sequencing project, detailing the areas that form the continual cycle of improvement to the genome assembly.

FF0032

A SNP Study of *Mycobacterium tuberculosis* using 454 and Solexa/Illumina Sequencing

Danielle Walker and the Pathogen Sequencing Unit

Wellcome Trust Sanger Institute.
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

Tuberculosis is a common and deadly infectious disease caused by mycobacteria, mainly *Mycobacterium tuberculosis*. Tuberculosis most commonly attacks the lungs (as pulmonary TB) but can also affect the central nervous system, the lymphatic system, the circulatory system, the genitourinary system, bones, joints and the skin. *Mycobacterium tuberculosis* H37Rv was sequenced using Sanger capillary sequencing and was published in 1998.
Two NST projects are currently being carried out and have had some interesting developments and hurdles to overcome.

454 Study
Using *M.tuberculosis* H37Rv as the capillary reference, 8 clinical isolates of this organism were taken from 4 patients, one before and one after treatment. Therefore theoretically a sensitive and a resistant strain. They were sequenced using a mix of 454-GS20 and 454-FLX at around 10X coverage. The data was run through a MuMMer and fasta3 comparison, each pair of strains was then compared to each other, the reference, and the other isolates using ssahaSNP. The resulting SNPs were then filtered, using calibrated confidence values and the two comparisons, to extract those SNPs that appeared in both datasets and showed high confidence levels.
These were then viewed in Artemis and the synonymous and non synonymous SNPs allocated to the gene they fell within, for example one SNP fell within the gyrA gene known to be involved in flouroquinolone resistance.

Solexa/Illumina Study
Two *M.tuberculosis* strains, both artificially resistant mutants, were used in this study, the Solexa reads from the Illumina GA platform were viewed in conjunction with the finished sequence of *Mycobacterium tuberculosis* H37Rv in a gap4 interface. An in-house SNP calling pipeline was used to generate a list of SNPs for each strain. These were then verified by manually checking them against the reference sequence, and were also used to correct any errors in the reference sequence. In the future there will be a release of this improved reference sequence. Future work will include looking at 4 more stains as part of this work.

Both of the above studies have given rise to a need for software able to cope with the large amounts of data generated by the NSTs. Software that enables manipulation of this data is also a necessity. The quality values and coverage of the NST data are of great importance in verifying the SNP calls.
Therefore new software developments go hand in hand with working on new finishing projects of this nature, and are the focus for the future development of this work.

FF0042

Applying Project Management Principles to Scientific Programs and Projects within the Joint Genome Institute

David C. Bruce [1], Lynne A. Goodwin [1], Kerrie W. Barry [2], Christa P. Pennacchio [2], Susannah G. Tringe [2], Jim Bristow [2]

1 DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM
2 DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA

The Joint Genome Institute (JGI) made a strategic decision to form a Project Management Office [PMO] in 2006. The initial charge for the PMO was to develop and manage communication paths between external stakeholders (Users) and internal stakeholders (Functional Teams), and facilitate the transit of User projects through JGI processes beginning with project definition and ending with product delivery. In practice, this requires JGI project managers to function as both process analysts and project managers while simultaneously implementing change in the organization. Application of standard project management principles on preexisting processes at the JGI is challenging and requires substantial adaptation of standard project management protocols. Altering people's behavior, process streams and defining progress/performance metrics is partially accomplished by project managers adapting and using analysis tools such as use cases, cross functional maps, sequence diagrams, and data entity relationship models. Durable progress is achieved through repeated process analysis and improvement cycles, and JGI project management practice is slowly converging to standard project management guidelines. Project management implementation techniques developed at the JGI to orchestrate a complex scientific effort is a valuable case study for large scale systems biology enterprises.

This work sponsored by the U.S. Department of Energy under Contracts W-7405-Eng-48, DE-AC02-05CH11231, and W-7405-ENG-36.

FF0044

Cloning with Trace Amount of Unamplified DNA

Julianna Chow, Eileen Dalin, Susan Lucas, Tanja Woyke, and Jan-Fang Cheng

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA

As a user facility, the US Department of Energy's Joint Genome Institute, in collaboration with scientists around the world, are able to generate DNA sequences for a diversity of organisms and environmental samples. Often times, the amount of DNA provided for library construction is limited. It is important to develop a protocol that requires a trace amount of DNA for library construction. In an attempt to test the minimum amount of DNA necessary for library construction, we decided to try a different approach, using AMPure beads instead of agarose gels, to purify DNA. AMPure bead purification, utilizes solid-phase paramagnetic bead technology to purify DNA fragments from contaminates and enzymes, with minimal loss of DNA. We performed the test on a chloroplast genome of a flowering plant Brighamia insignis because of its small genome size (~130Kb) and the nearly finished genome. The concentrations used in this experiment ranges from 100 to 1,000 ng. By using AMPure beads, we eliminate the need for gel separation to purify and size select DNA fragments, which requires 3-5 ug of starting material. The AMPure bead purified DNA was cloned into the pUC18 vector and sequenced to assess the quality of the libraries. The sequencing data from the libraries constructed using AMPure bead are compared with the library constructed using the standard gel purification method to determine the cloning efficiency, insert size distribution, and coverage biases.

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.

FF0048

Gap Closure by PCR Product Transposition

Anna Montmayeur, Daniel Bessette, Tashi Lokyitsang, Tamrat Negash, Thu Nguyen, Michael Fitzgerald
Broad Institute of MIT and Harvard, Cambridge, MA, USA

A multitude of laboratory finishing techniques are available for achieving fine-tuned whole genome sequence when existing clones are available. In cases of uncaptured gaps (physical gaps) or when clones for previously captured gaps (sequence gaps) are no longer available, PCR is often the only available starting point. Commonly, gaps of this nature exist due to either initial cloning bias or depletion of the original clone, all owing to the presence of challenging sequence. Assemblies based on next generation sequencing technology have no underlying clones, presenting the problem on a larger scale. We have selected PCR product transposition to obtain the missing sequence. This process presents a number of advantages over iterative primer walking on PCR amplicons and PCR shatter libraries. We have deployed this process for both clone based and whole genome finishing projects. We will present our production process based on amplicon capture into Stratagene's StrataClone Blunt PCR Cloning system followed by transposition and sequencing along with results.

FF0056

Finishing of Spirochaeta aurantia M1

Alicia Clum[1], Steve Lowry[1], Brian A. Rabkin[1], Hui Sun1, Alla Lapidus[1]

1Lawrence Berkeley National Laboratory
DOE Joint Genome Institute, Walnut Creek, CA 94598

Shotgun sequencing and finishing of an isolate of the Spirochaeta aurantia M1 genome, a free-living nonpathogenic Spirochete, is in process at the Joint Genome Institute. S. aurantia M1 is being sequenced due to its proximity on the phylogenetic tree to bacteria present in the termite hindgut that were partially sequenced during a metagenomic project at the JGI ,and, for improving the understanding of pathogenic Spirocheates through comparative genomic studies. The assembly is currently in 3 contigs and most of the finishing has been done using fosmid templates to close gaps. The shotgun assembly is a combination of fosmid sequencing, a novel Sanger approach, 454 sequencing, and Solexa
data. The novel Sanger approach takes advantage of a bulk cloneless library preparation method involving the amplification of long DNA fragments onto beads, by emulsion PCR, and the subsequent sorting of individual beads by flow cytometry for seeding the Sanger chemistry process.

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.

FF0058

Metagenomic Bacterial Finishing at JGI

Stephen Lowry, Alicia Clum, Eugene Goltsman, Hector Garcia Martin, Phil Hugenholtz, Alla Lapidus

DOE Joint Genome Institute, Walnut Creek, CA

The sequencing to completion of uncultured bacterial genomes from mutualistic communities is a demanding process in the best of cases. The complexity of the community, the quantity of genomic DNA available, the fraction of the total DNA collected representing the organism under study are all added to the normal difficulties of establishing a complete sequence. At the JGI we have managed to complete the sequence of three metagenomic organisms, and are investigating a fourth, presenting a considerable range of difficulty.

Candidatus Korarchaeum cryptofilum OPF8, (NC_010482; GI:170289627), is the first of this apparently ancient hyperthermophilic phyletic group to be sequenced (3). The ability to obtain ample DNA of near-monocultural purity and low strain complexity made this the most straightforward sort of metagenomic subject. The target organism constitutes ~40% of the Yellowstone thermal Obsidian Pool community. The community could be maintained in culture, and it was found that *K. cryptofilum* was the most resistant member to SDS lysis, thus allowing DNA purification to better than 90%. Its strain complexity was low as indicated by a SNP rate of ~0.2%.

Some organisms have remote or difficult habitats limiting the availability of source material. This is the case with the thermophile *Desulforudis audaxviator,* (NC_010424; GI:169830219), from fractures in the earth's crust at a depth of 2800 meters in a South African gold mine (4). The bacteria were collected on filters through which large amounts of subterranean water was passed. The surprising fact that this ecosystem contained but one species fortunately meant that the DNA yield of this one-time-only collection was sufficient to complete the genome.

Considerably more complex situations are the rule, as illustrated by the case of *Candidatus Accumulibacter phosphatis Type IIA str. CU-1*. This and closely related species are the principal actors in the sequestration of inorganic phosphate as intracellular polyphosphate in wastewater treatment facilities. Bioreactor sludge derived from a working facility in Wisconsin is a physically unresolvable mixture of organisms, with *A. phosphatis* predominating at about 40%. The entire DNA sample was sequenced and the resulting data then subjected to phylogenetic parsing using the Phytopythia (2) binning technique, greatly reducing the complexity of the subclone libraries. A single 5Mb chromosome was successfully sequenced, along with 3 plasmids of 167, 42 and 38kb.

We are now approaching a still more difficult genome. *Candidatus Endomicrobium trichonymphae* is an intracellular symbiont of a flagellate protist, itself part of the normal hindgut community of a termite host. It is of interest in the pursuit of the efficient breakdown of cellulose and lignin necessary in the hoped-for use of bulk plant materials as CO_2-neutral fuel stocks. Again partitioning with Phytopythia was absolutely necessary. So far this has yielded some 42,000 Sanger reads. Of these about 40% of the assembled contigs are similar to related organisms. Also contigs from 454 pyrosequencing contributed about 750,000 bp of additional coverage (~0.5x). Additional difficulty arises from sample size and strain complexity. There are several closely related organisms with substantial representation. There are several hindguts in the sample, and the lines of descent may be relatively independent even at the protist level.

(1) Martin, H.G., Hugenholtz, P., et al. Nature Biotechnology, v24, no.10, 1263-69. 2006. *Metagenomic Analysis of two Enhanced Biological Phosphorous Removal (EBPR) sludge communities,* (2) McHardy et al. Appl, Environ Microbiol 2000, 66(3):1175-1182, *Accurate phylogenetic classification of variable-length DNA fragments.,* (3) James G. Elkins, et al., PNAS. 2008 (in Press), *The Korarchaeota: Archaeal orphans representing an ancestral lineage of life,* (4) Dylan Chivian, et al., Science 2008 (in Press), *Environmental genomics reveals a single species ecosystem deep within the Earth.*

FF0060

Enhancing Microbial Genome Polishing through Frameshift Targeting

Brian Foster, Eugene Goltsman, Kurt Labutti, Steven Lowry
and Alla Lapidus

Joint Genome Institute Production Genomics Facility, Walnut Creek, CA

The JGI's finishing standards require every nucleotide in the final microbial consensus to be supported by at least two Sanger reads, and to be of at least Q30 quality. Additionally, areas covered by pyrosequence only (454-only) should be inspected for potential errors and re-sequenced.

The polishing process is a very important but time consuming step. Our current polishing strategy developed for Microbial genomes incorporates the use of Solexa data to bring the final consensus quality up to our pre-defined standard (see Kurt LaButti poster). Although this process can eliminate up to 97 percent of our polishing targets automatically, some regions still remain. These regions still require manual analysis based on our traditional (Sanger based) polishing methods.

An addition to our polishing process automation includes selective targeting of the remaining regions. This automation increases efficiency by only polishing targets within proximity to predicted frameshifts (for details of frameshift detection see Andrey Kislyuk poster). This approach automates the remaining polishing process without wasting resources on improving regions which may not be biologically relevant. Our quality improving approach (polishing) will be presented in detail.

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.

FF0066

FINISHING 454 BASED ASSEMBLIES – Mammals and Microbes

Yan Ding, Shannon P. Dugan-Rocha, Christian J. Buhay, Aniko Sabo, Mike E. Holder, Xiang Qin, Guan Chen, Donna Villasana, Huyen Dinh, Christie Kovar, Lynne Nazereth, Donna M. Muzny and Richard Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030

The development and application of new sequencing technologies has brought forth a new era in which traditional Sanger assemblies have been replaced with 454 generated assemblies containing additional data from other sequencing platforms such as Solexa and/or SOLiD. At the BCM-HGSC, these new sequencing technologies have greatly impacted strategies for finishing microbial genomes as well as mammalian BAC-based projects. While high throughput, low cost sequencing without cloning bias are major advantages to these new technologies, template availability, difficult gaps, and base pair accuracy in certain regions remain major issues.

To address these complicated issues, we have implemented direct primer walking on DNA products that have been amplified using the GE TempliPhi Sequence Resolver Kit. Results show success rates of direct BAC primer walking after TempliPhi as high as 98% (average 90%) with Phred20 lengths near 750bp. In addition to its cost effectiveness, this new strategy has been used quite successfully to resolve sequencing problems such as repeats, stops and compressions for many BAC-based projects including rat, macaque and bovine. Approximately 80 mammalian BACs representing 15Mb of data are completed each month utilizing this finishing strategy. For the much larger microbial genomes where DNA availability may be an issue, the GenomiPhi kit has been used to produce microgram quantities of DNA from picogram amounts of starting material. Direct primer walking can then be carried out on these amplified products. To date, five microbes have been completed for a total of 15.5Mb, while seven additional microbes remain in the finishing pipeline.

Furthermore, the roles of 454 and Solexa data in sequencing discrepancies have been investigated. Error rates for 454 and Solexa contigs were calculated by comparison to the finished sequence of 18 rat BACs. Results have shown that 454 data, with an error rate of 8.5×10^{-4}, was prone to indels (90%) in homopolymeric regions, while Solexa sequences with an error rate of 1.8×10^{-4} were dominated by substitutions (90%) in mostly GC rich regions. Combining these two sequencing technologies has thus proven to be an effective and efficient platform for finishing mammalian genomes and can also be applied to microbial finishing. These strategies and parameters will continue to be developed and optimized in preparation for the Human Microbiome Project.

FF0072

New Generation Sequencing and JGI Microbial Genome de novo Sequencing and Finishing

Jeff Froula, Edward Kirton, Paul Winward, Stephan Trong, and Feng Chen

US DOE Joint Genome Institute

2800 Mitchell Dr., B400 Walnut Creek, CA 94598

At JGI, major efforts have been spent on reducing the final cost of finished prokaryote microbial genomes. To achieve this, an optimal ration of traditional Sanger and new sequencing technologies needs to be determined. As part of the goal to figure out how much Sanger sequence coverage is needed for an improved microbial whole genome shotgun sequencing strategy, we looked into the most recent Roche software upgrade systematically to see what the overall quality of the raw reads is and what the coverage it yields. We also looked into the quality score assignment by Roche software for the raw reads. An assessment of the new version of Newbler was also performed to determine any improvements in the assembly parameters which will also impact finishing. Detailed results will be presented.

FF0074

H1 & H2 Haplotype Sequence Assembly and Finishing in the Human 17q21 Region

Amr Abouelleil[1] , Evan Eichler[2] , Gary Gearin[1] , Zhaoshi Jiang[2] , Margaret Priest[1], Mike Zody[1]

1.Department of Computer Finishing, Broad Institute of MIT and Harvard, Cambridge, MA
2.Department of Genome Sciences, Howard Hughes Medical Institute, University of Washington, Seattle, WA

The current reference human genome assembly (build 36) is a mosaic of sequence derived from several different donors. As such, any study interested in examining copy number variation as it pertains to genetic disorders must distinguish between individual chromosomes. Of particular interest is the 17q21 MAPT locus, which has been implicated in several neurodegenerative and cognitive disorders and also contains a large chromosomal inversion. The current human genome build corresponds to the H1 (direct) haplotype despite being derived from various donors. Thus an assembly of the H2 (inverted) haplotype would facilitate further analysis. Through the use of a 79 single nucleotide polymorphism panel we are able to distinguish between the H1 and H2 clones of 17q21 in a unique inverted region of 99.6% sequence similarity. Identity of duplications within one haplotype can reach 99.86%, and some may be copy number variant. To overcome this, we simultaneously constructed both H1 and H2 paths from a single heterozygous individual. Analysis of existing clones and finishing of new clones allowed us to generate two distinct tiling paths specific to the H1 and H2 haplotypes. We will present work used to identify and finish these clone paths.

FF0076

Bacterial Genome Finishing on the Turn of New Technology

Hajnalka Kiss and Cliff Han

Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM

The LANL finishing group is the major group for finishing microbial genomes at the Joint Genome Institute (JGI). In the past, the microbial draft sequences were produced exclusively by Sanger sequencing of random shot-gun libraries prepared with 3, 8 and 40 kb insert size. During the last few years new sequencing technologies have emerged for high-throughput sequencing. Recent microbial draft sequences contain both data from 454 and Solexa/Illumina platforms besides the reduced number of 'old fashioned' Sanger sequences. As the new sequencing technologies become more efficient, we predict, in the near future, that the microbial draft sequences will be prepared without using any Sanger sequencing.

As the drafting strategy changes we have to adjust our finishing strategy. The biggest challenges in finishing are solving the duplications, closing scaffold gaps and going through hard GC stops. We planned and performed several R&D experiments in developing our new finishing approach. Some of the experiments are still ongoing.

Presently, to solve the duplications we use paired end reads information and occasionally we sequence the bridging clone (transposon bomb). Without available templates finishing of highly repetitive bacterial genomes would be impossible. The 454 paired end reads could be a good solution to replace the random shot-gun libraries. We propose a novel mapping method for repeat resolution and gap closure. We used the amplified partial genome of Thricoderma Virens (fungal genome, 39 Mb) to validate the new method. For gap closure we will rely on bridging PCR fragments after performing adapter PCR/pair wise PCR or utilizing the new mapping method. To be able to sequence PCR fragments efficiently we developed an indexing strategy for the Solexa libraries. We also modified the Solexa library preparation to high-throughput rate using a 96-well plate format and vacuum instead of centrifugation. Presently, we use a special chemistry on smaller clones to go through hard GC stops. To evaluate the Solexa technology on this aspect, we sequenced Frankia sp. EAN1pec which contains more than 100 hard GC stops.

As the drafting and finishing strategies change we may have to define new finishing standards to avoid over-finishing.

FF0082

Establishing Optimal Levels of Shotgun Sequence from Different Sequencing Platforms

Eugene Goltsman, Brian Foster, Alex Copeland, Kurt Labutti, Alla Lapidus

DOE-JGI, Walnut Creek, CA

Sequencing-by-synthesis (SBS) technologies have provided new cost efficient ways of obtaining a high quality draft assembly as well as a fully finished genome. The result is also a change in the complex dynamics of shotgun assembly which makes it necessary to re-evaluate the whole process, especially when complete finishing is the goal. As of yet, no single SBS technology has proven itself fully capable of providing all the data necessary for efficient and streamlined finishing, therefore, a combination of SBS and Sanger data is still prefered. JGI currently utilizes both 454/Roche and Solexa/Illumina pyrosequencing in addition to the Sanger big-dye shotgun method. In this study we evaluated how various key assembly features are affected by changes in the levels of sequence from different platforms and attempted to arrive to a combination optimal for cost-efficient finishing. We also looked at how stable and predictable this optimum is in genomes with different GC and repeat composition. The criteria we picked are believed to be the best indicators, available upon the initial draft assembly, of what kind of additional effort, and how much of it, is needed to finish a microbial genome to JGI's quality standards.

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.

FF0086

Finishing Genomes from Single Cell, Photoheterotrophic Flavobacteria

Cliff Han1, Hajnalka Kiss1, Tanja Woyke2, Alex Copeland2, Gary Xie1, Jan-Fang Cheng2, Michael E. Sieracki3 and Ramunas Stepanauskas3

1 The DOE joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87544. 2 The DOE Joint Genome Institute, Production Genome Facility, Walnut Creek, CA 94598, 3 Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine 04575-0475

Sequencing microbial genomes from single cell is one of the emerging technologies for sequencing bacterial that cannot be cultured. Recently, we've been finishing two genomes with DNA amplified from single cell that was flow sorted from sea water. Currently one is about 1.5 Mb and another 2 Mb with estimation of 60 - 80 % coverage of the genomes. The longest contig is more than 600 kb. 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 work, adapter PCR, but also using metagenomic data to closing the gaps in the single cell project.

FF0092

The Genome Reference Consortium

Tina Graves and the Genome Reference Consortium

Washington University School of Medicine, Genome Sequencing Center, St. Louis, MO 63108

The production of a high quality reference sequence of the human genome has produced a step-change in studies of human genetics. It has provided a foundation for genome-wide studies of genome structure, human variation, evolutionary biology and human disease and will continue to provide a platform to further our understanding of human biology. Many of these studies have also revealed, however, that there are regions of the reference human genome sequence that are not represented optimally by the current assembly.

At the time the human reference was initially described it was clear that some regions were recalcitrant to closure with technology available at the time. What was not as clear was the degree to which structural variation affected our ability to produce a truly representative genome sequence at some loci. It is now apparent that some regions of the genome are sufficiently variable that they are best represented by multiple sequences in order to capture all of the sequence potentially available at these loci.

In order to improve the representation of the reference human genome we have formed the Genome Reference Consortium (GRC). The goal of this group is to correct the small number of regions in the reference that are currently misrepresented, to close as many remaining gaps as possible and to produce alternative assemblies of structurally variant loci when necessary. We will provide mechanisms by which the scientific community can report loci in need of further review. In addition, information about loci currently under review and genome assembly production cycles will be made readily available. The human reference assembly is the cornerstone upon which all whole genome studies are based. It is critical to ensure that we have the best possible view of the genome to facilitate continued progress in understanding and improving human health.

FF0096

Community-Involved Genome Annotation and Analysis at JGI-LANL: Facilitating Publication of Genome Papers Through Bioinformatics Support and Training

Jean Challacombe, Gary Xie, Diego Martinez, Ravi Barabote, Monica Misra, Thomas Brettin.

Genome Science/JGI, Los Alamos National Laboratory, Los Alamos, NM.

The role of the genome annotation and analysis team at JGI-LANL is to facilitate publication of JGI genome papers and provide bioinformatics support and training to promote community-involved genome annotation and analysis. Since March 2007, we have hosted 9 JGI collaborators as part of our genome explorer seminar series. In projects where JGI-LANL team members played a leading role in the analysis and preparation of genome papers, 6 genome papers have been published, 1 book chapter is in press and 5 papers are submitted or near submission. In addition to our microbial genome effort, our eukaryotic genome annotation team has worked with the annotation team at JGI-PGF and eukaryotic research community to provide high quality manually curated annotations of eukaryotic genomic sequences. We have hosted 1 off-site annotation Jamboree in the past year to promote community involved analysis for publishing a full analysis and annotation of *Postia placenta.* We also annotated the *Ostreococcus* RCC809, *Trichoderma virens and Trichoderma atroviride* genomes. Work is under way to publish these genome papers.

FF0104

The *Desulfovibrio* "pangenome" as a system for studying the mechanisms of mercury biotransformation (ERSP)

Steven D. Brown, Dwayne A. Elias, Mircea Podar, Amy M. Kucken, Haakrho Yi, Craig C. Brandt, Meghan M. Drake, Lisa A. Fagan, Bruce Roe, Simone Macmil, Graham Wiley, Fares Najar, George Southworth, Igor Jouline, Cynthia C. Gilmour, Judy D. Wall, and Anthony V. Palumbo (PI).

Oak Ridge National Laboratory, Oak Ridge TN, email: palumboav@ornl.gov

Anaerobic sulfate-reducing bacteria (SRB) are a diverse group of microorganisms that play an important role in the global sulfur and carbon cycles. They are also involved in a wide range of metal ion biotransformations, with implications in the cycling of industrial heavy metal pollutants. Among those, some SRB can transform inorganic mercury into the more toxic methylmercury, which accumulates in the food chain and has become a global environmental problem. The molecular mechanism of methylmercury biosynthesis is unknown. To study this process, we are combining comparative genomics with molecular physiology and are using the *Desulfovibrio* genus as a model system. Some members of this genus are able to methylate mercury whereas others are not. We have sequenced the genomes of two methylating species, *D. africanus* and *D. desulfuricans* ND132, to near completion (~95% based on occurrence of bacterial core genes) using the 454 shotgun approach. The metabolic potentials inferred from the sequences of these two organisms were compared to those of several related *Desulfovibrio* species that were shown to lack the capacity of methylmercury synthesis and for which completed genomes have been publicly released (*D. desulfuricans* G20, *D. vulgaris* DP4 and *D. vulgaris* Hildenborough). The two strains that can methylate mercury have slightly larger genomes (~3.9 - 4.2 Mbp versus ~3.7-3.8 Mbp). Among the ~1900 types of conserved protein domains (based on Pfam classification) that are present in the *Desulfovibrio* "pangenome", we have identified several dozen that are specific to the two strains that methylate mercury.

We are complementing the computational analysis of these specific genes with gene expression microarrays, mutagenesis procedures, and functional complementation to elucidate the mechanism of mercury methylation in *Desulfovibrio*. Preliminary growth studies toward a genetic system have thus far shown antibiotic resistance to only kanamycin and ampicillin of seven antibiotics tested. A complex medium with yeast extract, 0.5% NaCl, lactate as the carbon and electron donor, and sulfate as the electron acceptor yields $1x10^9$ *D. africanus* cells/ml in 48 hours. We are currently determining transformation parameters. Proof-of-principle testing has targeted three candidate loci in *D. africanus* ATCC19997 for potential roles in Hg methylation of which orthologs were either not found in non-methylating *Desulfovibrio* species or did not share synteny with non-methlyators. These studies will provide insight into the molecular mechanisms for methylmercury production by *Desulfovibrio* species and a comprehensive comparative genomic analysis of *Desulfovibrio* will also provide insight into the evolution of metabolic strategies and environmental adaptation in this genus of SRB.

Meet and Greet Party

530pm – 800pm, May 28th

Sponsored by Roche Diagnostics

Enjoy!!!

05/29/2008 - Thursday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Santa Fe Breakfast Buffet** (Scrambled eggs with a choice of three accompaniments on the side -chilaquiles with green chile and cheese, chorizo sausage and roasted green chile, Grilled breakfast potatoes, applewood-smoked bacon and warm flour tortillas, assorted breads and fruits, etc.)	x
830 - 845	Intro	x	Welcome Back Intro	**Jim Bristow**
845 - 930	**Keynote**	**FF0050**	**The Unfinished Genome**	**Sydney Brenner**
930 - 955	Speaker 1	FF0085	How Close to Finished can an Assembler Get?	**Jim Knight**
955 - 1020	Break	x	Beverages provided	x
1020 -1045	Speaker 2	FF0098	Taking the Next Step: Assembling Large Genomes Using NextGen Sequencing	**Sean Sykes**
1045 -1110	Speaker 3	FF0034	De Novo Assembly of Microbial Genomes from Illumina Whole-Genome Shotgun Data	**Iain MacCallum**
1110 - 1135	Speaker 4	FF0036	Combining Next-Gen Sequence Technologies in Multi-platform Assemblies	**Christian Buhay**
1135 - 1200	Speaker 5	FF0010	The Arcturus – Minerva assembly management system	**Ed Zuiderwijk**
1230 - 130pm	Lunch	x	**New Mexican Lunch Buffet** (Pork tenderloin achiote-rubbed and char-grilled with tomatillo-chipotle sauce, your choice of either Chicken or Cheese enchiladas with red or green chile, etc.)	x
130 -155	Speaker 6	FF0076	Bacterial Genome Finishing on the Turn of New Technology	**Hajnalka Kiss**
155 -220	Speaker 7	FF0088	Towards "Finishing" of Eukaryotic Transcriptomes	**Stephen Kingsmore**
220 - 245	Speaker 8	FF0075	An optimized DNA sequencing pipeline to support drug discovery at Wyeth	**Jan Kieleczawa**
245 - 310	Speaker 9	FF0065	Sequencing the Genome, Transcriptome, Methylome and smRNAome of the Arabidopsis Ecotype, Cape Verde Island	**Ronan O'Malley**
310 - 415	Panel Discussion	x	**Panel Discussion** - Sequencing and Finishing Standards	Chair - Darren Grafham
415 - 630	**Wine & Cheese**	x	**Beverages, Wine & Cheese provided - sponsored by Applied Biosystems**	x
415 - 630	**Posters - odd #s**	x	**Poster Session with Wine & Cheese - sponsored by Applied Biosystems**	x
630 - bedtime	on your own	x	Dinner and night on your own - enjoy	x

Speaker Presentations (May 29th)

Abstracts are in order of presentation according to Agenda

FF0050 – Keynote

The Unfinished Genome

Sydney Brenner

The Salk Institute for Biological Studies, San Diego, CA

FF0085

How Close to Finished can an Assembler Get?

James Knight

Roche Diagnostics - 454, Branford, CT 06405, USA

The combination of high sequencing cost and cloning biases involved with Sanger sequencing meant that assemblers were never given sufficient information to be able to reconstruct full genome sequences from their input reads. The advent of next-generation sequencing have reduced the costs and eliminated the cloning biases involved in sequencing, and the development of next-gen, paired-end reads provide the information to span across and through repeats. In particular, generating an 8kb long tag, paired-end library and sequencing it to 20x on the Genome Sequencer FLX or XLR HD platforms should provide enough information content for most bacteria that the assembler should be able to reconstruct the overall genome sequence from those reads. This talk will describe the current state of the GS sequencing technology and the newbler assembler, and how close they are to "finished" assemblies for small genomes, and for large.

FF0098

Taking the Next Step: Assembling Fungal Genomes Using NextGen Sequencing

Sean Sykes, Sarah Young, Theresa Hepburn, Carsten Russ, Chad Nusbaum, Bruce Birren

Broad Institute of MIT & Harvard, Cambridge, MA 02142

454 sequencing technology has been used to generate high-quality bacterial genome assemblies; however, it is still unclear how to maximize its utility for larger genomes such as fungi, that range from 15 to greater than 100 Mb. We have assessed different approaches for using this type of data to produce genome assemblies that are suitable for a range of applications. We have used genomes for which finished references exist to serve as a “truth” data set for evaluating our assemblies. For 454 assemblies produced with a range of sequence coverages from fragment reads and from pairs we measured: 1) the amount of the reference genome that was captured by the assembly; 2) contig size; 3) scaffold size; 4) base accuracy; and 5) assembly errors providing a framework for making decisions about quality and cost trade-offs. We also investigated whether including traditional Sanger sequence of Fosmid clones, or data from other short read technologies can improve the integrity or accuracy of the assembly.

FF0034

De Novo Assembly of Microbial Genomes from Illumina Whole-Genome Shotgun Data

Iain A. MacCallum, Ilya A. Shlyakhter, Dariusz Przybylski, Jonathan Butler, Eric S. Lander, Chad Nusbaum, David B. Jaffe

Broad Institute of MIT and Harvard, Cambridge, MA

For most applications of new short read sequencing technologies a reference genome is required. In contrast to this, another technically challenging application would be to use short reads to sequence 'new' genomes, for which no similar reference is available. This *de novo* assembly problem is difficult, unsolved, and contingent upon high-quality paired-read data.

Recent changes to the instrumentation and chemistry of the Illumina Genome Analyzer have made it possible to generate such data. We did so for five microbes, of size up to 40 Mb, haploid and polymorphic diploid. For each of these microbes we had a reference sequence available to allow assessment of our assemblies.

We developed an algorithm ALLPATHS, applicable to these data. ALLPATHS assemblies are graphs that retain intrinsic ambiguity in the data, from polymorphism, and also from limitations in the data.

Initial results using Illumina data are promising. For example, for *E. coli*, we assembled two lanes of paired 36 bp data from each of two libraries (fragment sizes 200 bp, 4000 bp). Eight edges in the assembly graph cover half the genome, and all but one of these edges match the reference sequence perfectly. The N50 component size is 360 kb. The assembly covers 99% of the genome. These results suggest that the Illumina system could be deployed almost immediately to generate reference sequences of quality approaching that of finished sequence, at least for microbial genomes.

FF0036

Combining Next-Gen Sequence Technologies in Multi-platform Assemblies

Christian J. Buhay, Michael E. Holder, Aniko Sabo, Xiang Qin, Carson Qu, Shannon Dugan-Rocha, Yan Ding, Huyen Dinh, Lynne Nazareth, Christie Kovar-Smith, Yi Han, Jeffrey Reid, Donna M. Muzny and Richard A. Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030

In the last two years, the BCM-HGSC has established a robust pipeline for finishing large-scale genome projects with next-generation platforms. We have implemented three popular commercially available platforms: Roche/454, Illumina/Solexa, and AB SOLiD. All three are either in production or undergoing testing. These sequencing platforms have been applied to a number of sequencing projects including eukaryotes, prokaryotes, and medical resequencing.

The Roche/454 platform is the most established pipeline at the BCM-HGSC. We have generated 586 sequencing runs and produced over 45 Gb of sequence. 454 is utilized in a wide array of projects from eukaryotes such as microbes and rat BAC pools to Nimblegen direct capture sequencing. To date, we have sequenced 65 microbes with 454. Assemblies are comparable to or superior to Sanger sequencing, with N50 contig lengths averaging 51 kb and N50 scaffold lengths averaging greater than 700 kb. We have also started to layer Illumina/Solexa data in a subset of the microbes. Additionally, we developed a robust pipeline utilizing 454 and Solexa platforms for our rat BAC pools. Currently, we have 10 rat BAC pools in different stages of the pipeline. The BAC pools represent 1000 BACs sequenced with 454 and Solexa for genome upgrading and finishing. To date, we have finished 150 BACs to comparative grade and approximately 300 BACs are actively being finished. Layering 454 and Solexa contigs on available Sanger WGS have yielded contig N50 lengths of 40 kb and scaffold N50 lengths of 98 kb. Recently, we have also started to evaluate the SOLiD platform. We have performed an exploratory run with one rat BAC pool (100 BACs). With a pool size of 20 MB, initial results show an estimated coverage of 55X with total contig coverage of 90%. Our current efforts include development of the SOLiD pipeline, and modification to our mapping algorithm.

Additional experiments to asses the sequencing platforms in the same 8 ENCODE regions of 3 organisms are being conducted. A set of 102 BACs from cow, macaque, and rat (34 BACs each) were used to span 8 ENCODE regions: ENm004, ENm014, ENm008, ENr123, ENr131, ENr233, ENr321, and ENr333. Our goal is to establish if sequencing performance varies within platforms for different mammalian genomes.

FF0010

The Arcturus – Minerva assembly management system

E.J. Zuiderwijk, D. Harper, M-A Rajandream, J. Parkhill, K. Mungall

The Wellcome Trust Sanger Institute Hinxton Cambridge UK

In the past years we have developed an assembly data management system based on the MySQL database engine and Java GUI facilities. The Arcturus back-end consist of a set of Perl-modules and scripts which interface the MySQL database with the the Gap4 tools used by finishers and with the assembly pipeline(s).

Assemblies are represented as mappings between the basic objects (read-contig, contig-contig) which are built during loading operations. Arcturus automatically links newly created contigs to their predecessors. This allows tracking of a contig's history and it enables annotation and tags to be remapped onto the latest version of a contig.

The Minerva front-end enables users to view the contents of assemblies and projects, to move contigs between projects and to import new reads into projects. In addition there are tools for data analysis such as an oligo finder and scaffold analysis.

Recently we have included in the GUI the functionality to import and export data from and to Gap4 databases, which effectively gives the finishers control over all aspects of the system's operation.

FF0076

Bacterial Genome Finishing on the Turn of New Technology

Hajnalka Kiss and Cliff Han

Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM

The LANL finishing group is the major group for finishing microbial genomes at the Joint Genome Institute (JGI). In the past, the microbial draft sequences were produced exclusively by Sanger sequencing of random shot-gun libraries prepared with 3, 8 and 40 kb insert size. During the last few years new sequencing technologies have emerged for high-throughput sequencing. Recent microbial draft sequences contain both data from 454 and Solexa/Illumina platforms besides the reduced number of 'old fashioned' Sanger sequences. As the new sequencing technologies become more efficient, we predict, in the near future, that the microbial draft sequences will be prepared without using any Sanger sequencing.

As the drafting strategy changes we have to adjust our finishing strategy. The biggest challenges in finishing are solving the duplications, closing scaffold gaps and going through hard GC stops. We planned and performed several R&D experiments in developing our new finishing approach. Some of the experiments are still ongoing.

Presently, to solve the duplications we use paired end reads information and occasionally we sequence the bridging clone (transposon bomb). Without available templates finishing of highly repetitive bacterial genomes would be impossible. The 454 paired end reads could be a good solution to replace the random shot-gun libraries. We propose a novel mapping method for repeat resolution and gap closure. We used the amplified partial genome of Thricoderma Virens (fungal genome, 39 Mb) to validate the new method. For gap closure we will rely on bridging PCR fragments after performing adapter PCR/pair wise PCR or utilizing the new mapping method. To be able to sequence PCR fragments efficiently we developed an indexing strategy for the Solexa libraries. We also modified the Solexa library preparation to high-throughput rate using a 96-well plate format and vacuum instead of centrifugation. Presently, we use a special chemistry on smaller clones to go through hard GC stops. To evaluate the Solexa technology on this aspect, we sequenced Frankia sp. EAN1pec which contains more than 100 hard GC stops.

As the drafting and finishing strategies change we may have to define new finishing standards to avoid over-finishing.

FF0088

Towards "Finishing" of Eukaryotic Transcriptomes

Stephen Kingsmore

National Center for Genome Resources, Santa Fe, New Mexico, USA

Generation 2 sequencing technologies have powerful functional genomics applications. Illumina GA and GA II, for example, provide more sensitive and quantitative measurement of the abundance of mRNA and small and large non-coding RNA than array hybridization. Given recent findings that the genomes of eukaryotes are almost entirely transcribed, "finishing" of eukaryotic transcriptome analysis appears daunting. Examples of experimental designs, algorithms and discoveries made through Illumina transcriptome analysis will be provided.

FF0075

An optimized DNA sequencing pipeline to support drug discovery at Wyeth

Jan Kieleczawa

Wyeth Research, Cambridge, MA

We have developed very efficient proprietary DNA sequencing support for Wyeth' s growing protein based drug development effort. Custom designed LIMS system is used to electronically submit sequencing requests (in tubes or plate formats). Using the information provided in a request (vector name, concentration, potential difficult regions in a template, reference sequence) lab personnel is able to quickly search database of all available primers and schedule them. If primers are not available LIMS designs new primers and upon receiving conformation that primers are available it schedules them on the next CE run. Currently we are testing robotic setting up of sequencing reactions (primers selected from library of thousands and hundreds of clones) and automated assembly of contigs. Proprietary bioinformatics tool scans reference sequences for potential difficult regions and if needed suggest appropriate chemistry. Upon completion of a sequencing request, sequences are deposited back to the LIMS and email notifies requestor about the availability of data. If all primers are available we are able to fully sequence and edit any clone or number of clones (typically in size range 1-20 kbp) within a day.

FF0065

Sequencing the Genome, Transcriptome, Methylome and smRNAome of the Arabidopsis Ecotype, Cape Verde Island

Ronan O'Malley[12†], Ryan Lister[12†], Jan Korbel[3], Jarrod Chapman[4], Jason Affourtit[5], Zhoutao Chen[5], Brian Desany[5], Srinivasan, Maithreyan[5], Julian Tonti-Filippni[6], Brian Gregory[1], A. Harvey Millar[6], Mark Gerstein[3], Dan Rokhsar[4], Michael Snyder[7], J Michael Egholm[5], Tim Harkins[5], Joseph Ecker[12]

[1] Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [2] Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [3] Molecular Biophysics and Biochemistry Department, Yale University School of Medicine, New Haven, CT 06520. [4] U.S. DoE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [5] 454 Life Sciences, A Roche Company, Branford, CT 06405, USA. [6] ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, WA 6008, Australia. [7] Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, USA. [†] Authors contributed equally.

Trait differences between individuals within a species due to sequence polymorphisms, ultimately result from some combination of changes in expression, splicing, epigenetic, deletions, insertions, and substitutions. With the advent of next-generation sequencing it is now possible to examine all of these biological consequences of polymorphisms genome-wide. Arabidopsis *thaliana* provides an excellent model for such a study: it has a high-quality reference genome whose small size makes it well-suited for current platform capacities, and as a highly-selfing hermaphrodite, it exists in the wild as a collection of naturally occurring inbred populations, or ecotypes, with a wide range of trait differences. Using a combination of 454 paired-end and XLR reads, we have carried out de novo sequencing and assembly of the Cape Verde Island (Cvi-0) ecotype and identified SNPs, small indels, and larger structural variations that collectively represent over a 2% sequence content difference to the Columbia reference genome. Utilizing this high-quality Cvi-0 genome as a reference, we have mapped transcriptome, bisulfite, and smRNA reads to create a comprehensive view of the consequences of polymorphisms within a species.

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

Wine & Cheese Poster Session

415pm – 630pm, May 29th

Sponsored by Applied Biosystems

Enjoy!!!

Poster Presentations (Odd #s, May 29th)

FF0005

Assembly of Sanger and 454 genomic sequence data from the microsporidian Enterocytozoon bieneusi

Hilary G. Morrison (MBL) and Donna E. Akiyoshi (Tufts Cummings School of Veterinary Medicine)

Marine Biological Laboratory, Woods Hole, MA

Single celled organisms that are essentially uncultivable present a special challenge for genomic scale studies. This category includes most members of the microsporidia, protists that are obligate intracellular parasites believed to be degenerate relatives of fungi. A notable exception is Encephalitozoon cuniculi, which can be grown in mammalian cell culture and was one of the first eukaryotic genomes completed (2.9 MB; Katinka et al. 2001). Research on other members of this evolutionarily diverse group of parasites, with its wide host range (fish, insects, mammals) and extreme genome compaction is limited by the difficulty in obtaining adequate amounts of clean, high molecular weight DNA. Here we report on the progress of a genomic sequence survey of the AIDS-associated human microporidial parasite, Enterocytozoon bieneusi, using microgram amounts of DNA extracted from spores purified from fecal material. Sanger sequencing (Agencourt) required a genomic DNA amplification approach. In contrast, an equivalent amount of sequence data was produced from a 454 library constructed using fewer than 3 ug of genomic DNA. The datasets were combined using the Roche Newbler assembler.

FF0009

XRpro Technology for Characterization of the Metallome and Biomarkers of Response to Heavy Metals: a Gene-Environment Initiative (GEI) Project

Birnbaum ER[1], Peterson LJ[1], Harris MN[1], Miller RLE[1], Touchet N[2], Warner BP[1]

[1] Caldera Pharmaceuticals, Inc. Los Alamos, NM 87544
[2] Harvard College of Medicine, Boston, MA 02108

NHGRI is a leader of the Genes, Environment and Health Initiative (GEI), a branch of the Human Genome Project (HGP). The GEI is a trans-NIH research effort to combine comprehensive genetic analysis and environmental technology development to understand disease etiology. Caldera Pharmaceuticals uses their XRpro technology to identify biomarkers of response to environmental stressors by measuring differential proteomic properties related to toxic metal exposures. Arsenic, lead, and mercury, as well as essential micronutrients like zinc, iron, copper, and selenium have affinity for similar protein binding sites in the human body. XRpro simultaneously monitors response to environmental toxins and essential elements in proteins and biospecimens. Concentrations of DNA and the degree of post-translational modifications to proteins (i.e., phosphorylation, sulfonation, ubiquitilation) may also be assessed. XRpro identifies differences in the metallome that result from exposure to various forms of metals, such as metals in *different valence states* or metals with *different chemical ligands*. This is performed with *no a priori knowledge* of which proteins to investigate and with absolutely *no labeling or tethering* of proteins, ligands, drugs, co-factors, or other small molecules (labeling introduces experimental artifacts). XRpro will provide information that will allow identification of variations in multiple genes and environmental factors in interconnected biological pathways or networks, like zinc homeostasis or arsenic elimination. Correlation of environmental exposures with genotype and phenotype information will help to identify susceptible populations and individuals. Characterization of individual metallomes can be correlated with HGP data to improve individualized medicines, diagnosis/early detection, and response to treatments. Our initial research focuses on environmentally-encountered toxic metals involved in Autism Spectrum Disorder (ASD), which has been shown on the basis of inheritance patterns to have the highest heritability of any psychiatric disorder and as many as 10 potentially related genes.

FF0017a

Improving Efficiency Through Automation: A New Model for the Viral Finishing Pipeline

Erin Hine, Katie Proudfoot, Nadia Fedorova, Jeff Sitz, Danny Katzel, Maria Sarmiento, Larry Overton, Tamara Tsitrin, Jessica Hostetler, David Spiro

J. Craig Venter Institute, Rockville, MD U.S.A.

The high throughput pipeline currently used by the Viral Genomics and Finishing groups at JCVI has been able to sequence and close over 2800 Influenza A and B genome from the beginning of 2005 to present (http://msc.tigr.org/influenza/index.shtml). The current pipeline employs amplicon-based PCR sequencing, and contigs are then built through an assembly software suite called Elvira. A combination of standard and custom primers are then used to close each genome, while every sample is thoroughly tracked using the Closure Task Manager (CTM), a tracking web interface. Despite the current pipeline's success, there is much more that can be done to improve and automate the finishing processes. This will increase the efficiency of the viral pipeline further still. The new Viral Finishing pipeline proposes to automate the initial assembly, editing and task assignment steps, which are currently done manually. This will be accomplished using existing tools, like Elvira and the CTM, as well as new software, such as a Contig Checker Program. All of these programs will comprise one large suite of processes that Influenza samples will initially pass through without manual interaction. Subsequently, the manual work by closure personnel will only be required at later stages of the pipeline for problem areas. An automated Viral Finishing pipeline that operates in such a manner will significantly increase production, affectively decreasing costs and the time a sample spends in finishing.

FF0017b

Automated Finishing System at JCVI

Jessica Hostetler, Brent Bradley, Heather Forberger, Hoda Khouri, Luke Tallon, Dan Kosack and Danny Katzel

J. Craig Venter Institute, Rockville, MD U.S.A.

In a continuing effort to further automate finishing, autoCloser was developed and implemented at JCVI. Software tools analyze assembly results, identify finishing targets, design primers, select clones, and choose laboratory reactions to resolve each target. Finishing features targeted by autoCloser include intra-scaffold gaps, low coverage regions, repeat elements, and scaffold ends (physical gaps). Laboratory reactions are grouped by type and clustered to form work orders. A separate LIMS element, called Clover, processes the work orders into laboratory consumable instructions including barcoding, primer orders, clone locations within the template blocks (source plates) and new clone locations in re-arrayed plates (destination plates). A Hamilton® MicroLabSTAR robotic system reads the files generated by Clover, re-arrays the templates and adds the correct primer to each sample. The plates are then tracked through the JTC's high throughput sequencing pipeline to generate the finishing reads. The finishing reads are incorporated into the targeted contigs and scaffolds using the Celera Assembler. Results are reported automatically to finishing project managers. The autoClosure pipeline is currently being updated to work with JLIMS tracked data resulting in an upgrade of Clover, soon to be Grover. Planning for automated finishing of genomes sequenced primarily with next generation technology is actively underway.

FF0019

Bacillus subtilis ten years later: will we learn something new?

Stéphane Cruveiller[1], Valérie Barbe[1], Eric Pelletier[1], Jérôme Lesaux[1], Patricia Lenoble[1], Claudine Médigue[1, 2] and Antoine Danchin[3]

1- Commissariat à l'Energie Atomique – Institut de Génomique –
Genoscope, Evry – France
2- CNRS UMR8030 "Génomique Métabolique", Laboratoire de Génomique
Comparative, CEA - Institut de Génomique – Genoscope, Evry – France
3- Institut Pasteur, Paris – FRANCE

In 1997, the shotgun sequencing strategy permitted the first comparison between model organisms [1, 2]. At the beginning of 2008, we are witnessing around 3600 genome sequencing projects and more than 700 genomes are now completely deciphered (source GOLD database [3]).

Biochemical experiments, although of fundamental importance, are slow. Hence it is essential to complete those with in silico analyses instead of bench experimental tools. These analyses rest heavily on the quality of the data, and the saying "garbage in, garbage out" is unfortunately demonstrated daily. This therefore raises questions about the quality of published sequences. The B. subtilis sequence currently available in public databases has taken ten years, at the early times of sequencing technology and is the result of the work of a consortium of more than 30 laboratories [2]. It was therefore extremely important of resequencing it with the new approaches now available, while reannotating entirely the sequence with the literature that has kept unfolding in the ten past years.

Sequencing technologies improved drastically these last years providing huge amounts of data in a single run. Because the perspective of completing a genome sequence in a single shot is somewhat exciting, many forget that these new approaches may have drawbacks.

This project was designed as a pilot study aiming at testing the Roche 454-FLX™ technology on the Bacillus subtilis str. 168 genome, reciprocally validating the reference sequence and potentially revising the original annotation set released ten years ago. Interestingly, preliminary analyses revealed some differences between the current assembly compared to the reference: a high frequency of point mutations that are not uniformly distributed along the genome and not necessarily correlated to the presence of a homopolymer in the nearest neighboring. This raised the new challenging task of discriminating between true sequencing errors and mutations that have occurred naturally during ten years of evolution.

References

1. Blattner, F.R., et al., The complete genome sequence of Escherichia coli K-12. Science, 1997. 277(5331): p. 1453-74.
2. Kunst, F., et al., The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature, 1997. 390(6657): p. 249-56.
3. Liolios, K., et al., The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res,2008. 36(Database issue): p. D475-9.

FF0025

The Implications of New Technologies on the Construction of DNA Libraries

Giselle Kerry

The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA

WTSI is a world leader in genome sequencing. For the past 12 years our sequencing production has been achieved using capillary sequencing with the Sanger dideoxy sequencing method. To achieve our sequencing output we have generated 119,356 sequencing libraries (equivalent to 9946 per year, or 200 a week). However, with sequencing strategies shifting from capillary sequencing to new technologies, involving massively parallel DNA sequencing, such as Illumina and 454, changes to the current processes are required.

High quality DNA and sequencing library construction is a prerequisite to successful sequencing projects; the library quality having a critical effect on the resulting assembly. Capillary sequencing libraries are typically constructed by a multi-step process involving physical shearing, end repair to make it blunt-ended, size selection by gel electrophoresis, ligation to a sequencing vector and transforming into *E.coli*.

The new sequencing technologies have different library construction protocols and different levels of throughput are required to utilize them to maximum capacity. Attention to quality in generating these libraries will also be crucial, poorly constructed libraries will yield chimearic reads and biased sequence coverage.

The WTSI has a number of projects that are being sequenced on the capillary platform and traditional libraries are required for these. In parallel with this there is a demand for approximately 100 libraries per week for the Illumina platform and 2-4 libraries per week for the 454. The resource allocated to these activities will therefore need to evolve in line with future demands.

Presented here is an overview of Sanger versus Illumina and 454 library construction their differences, issues encountered and the implications on possible future methodology.

FF0027

New Sequencing Technologies – A New Era in Finishing?

Siobhan Whitehead

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

The Wellcome Trust Sanger Institute (WTSI) is a world leader in genomic sequencing. Most notably it is responsible for the completion of one third of the sequence of the human genome, as well as the genomes of model organisms such as Mouse and Zebrafish and more than 90 pathogen genomes. To date, these genomes have been finished to the highest quality possible using predominantly standard BigDye terminator sequencing and ABI capillary sequencers, with the data assembled by Phrap (Phil Green). The finishers at WTSI then visualise and manipulate the data using Gap4 (Staden).

The new sequencing technologies were first introduced to WTSI in 2006, consisting initially of one 454 GS20 machine and one Illumina machine. This has now increased to two 454 FLX machines and 27 Illumina GA instruments, in operation alongside 54 ABI capillary sequencers. The advent of this focus on the new sequencing technologies, however, brings with it a new era in finishing.

The WTSI has an ongoing commitment to sequence the Parasitic Helminths using a strategy of combining sequencing technologies. To date, involvement with the new technology data has been in improving the sequence of genomes used for comparative studies and in assisting the finishing of *de novo* genomes. The experience of these finishers has helped to formulate ideas as to how the finishing department as a whole will use new technology data and which software tools will be needed. A fundamental software requirement is a tool, similar to Gap4, which can both visualise and manipulate the data.

Presented here is a comparison of both new sequencing technologies (454 and Solexa) with an aim to discuss how using data from these platforms, either individually or in combination with capillary sequence data, affects the finishing process. Also presented is a look at the application of the new sequencing technologies to the finishing of future *de novo* genomes, using Helminth as an example, and the progress being made on a new visualisation and manipulation tool; Gap5.

FF0033

Genome Assembly Visualization Tool for Short Read Data

Aijazuddin Syed, Stephan Trong, Harris Shapiro, Eugene Goltsman, Kurt LaButti, Alla Lapidus, and Anthony Kosky.

US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94597.

Usually it is a challenging job for genome analysts to accurately debug, troubleshoot, and validate genome assembly results, including identifying such problems as mis-assemblies, low-quality regions, and repeats. Genome analysts rely on visualization tools to help with these tasks. Short read data generated by the new generation of high-throughput sequencing technologies add further complexity and make it extremely challenging for the visualization tools to scale and to view all needed assembly information. As a result, there is a need for new visualization tools that can scale to display assembly data from the new sequencing technologies.

We present GAViT, an assembly visualization tool developed at the DOE Joint Genome Institute (JGI), which can support data from new sequencing technologies and addresses the aforementioned concerns. GAViT incorporates numerous data handling, organization, and display optimizations, which make it interactive and scalable. In GAViT data is accessible at various levels of resolution, including scaffold, contig, read, and consensus. GAViT displays bird's eye views of assembly, inconsistently placed reads in the assembly, clone/read depth and GC graphs, quality scores, annotations, and read pair information by read or library type. A link viewer in GAViT provides a graphical view of how the contigs in a scaffold are arranged, and the linking reads within contigs are colored and oriented by the read type. Also, in GAViT we can edit library information on fly, group reads by type/library/insert/sequencing technology, generate various assembly analysis reports, and it has the facilities to add user annotations and user specific color schemes, and to view multiple contigs in either strand.

We have completed development and are currently testing GAViT with short-read data sets. The initial tests with traditional whole genome shotgun assembly data are very promising, scaling to view assemblies with 1.8 GB of genome sequence, and indicate that GAViT will be a scalable tool for visualizing large genome assembly data.

FF0043

Ligation-mediated PCR Amplification as a Tool to Finish Microbial Genomes

Hope N. Tice, Eileen M. Dalin, Ze Peng, Susan Lucas, and Jan-Fang Cheng

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA

As Sanger sequencing is being replaced by higher throughput and lower cost of next generation sequencing, finishing microbial genomes will face two major challenges. First, the technology will need to be fast enough to handle many more drafted genomes. Second, it will have to incorporate a clone-free approach to fill gaps. We have been testing a method that utilizes a universal “bubble-tag” to perform primer walking and gap closure in a clone-free condition. The “bubble-tag” method was first described by Doug Smith (PCR Methods Appl. 2: 21-27, 1992) to amplify and sequence lambda DNA. There is no evidence however, that this approach will work for the more complex microbial genome.

Here we describe the experimentation of this approach in primer walking of the E. coli genome. Genomic DNA was sheared, blunt-end repaired, and ligated to the bubble adaptors. Site specific primers were used together with the universal bubble primer to amplify the regions of interest. We applied the Ampure beads binding and washing step to reduce the amount of small amplified fragments. The remaining large amplified DNA appears to be suitable for sequencing. Different bead-to-DNA ratios were tested in order to generate long amplified templates. This approach enables the primer walking and gap closure in a clone-free sequencing process with the new sequencing platforms. More importantly, the uniformity of this approach is amenable for an automated finishing process.

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.

FF0055

Machine Learning Techniques for Filtering Low Quality Pyrosequencing Reads

Dibyendu Kumar, Yijun Sun, Li Liu and William Farmerie

Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL-32607

High throughput pyrosequencers such as the 454 Life Sciences GS FLX have certainly revolutionized our approach to DNA sequencing. 454 technology is used extensively not only for genome sequencing but also transcriptome analysis, ultra-deep sequencing for detecting rare sequence variants, 16S ribosomal RNA profiling, and metagenomics. Pyrosequencing has a different sequence error profile when compared with conventional Sanger sequencing. Generally, in genome sequencing projects, building consensus through multifold coverage masks individual base call errors. In metagenomics or transcriptomics projects, where high coverage depth and consensus assembly is difficult or practically impossible, low read quality may have a significant impact on downstream analysis. Therefore, evaluation of individual read quality is an important step to prevent biased or incorrect interpretation due to sequencing errors. While 454 Life Sciences is continuously trying to make its base by base quality scoring method consistent with PHRED scores, we decided to approach the question of overall sequence accuracy by looking at sequence signatures. We built a mathematical model based on supervised machine learning techniques to predict read quality. This model includes and weights various sequence features, such as read length, GC ratio, homopolymer repeats, and presence of ambiguous bases. We use this evaluation method as a filter to remove low quality sequences prior to downstream analysis. Numerical experiments are presented that demonstrate the effectiveness of the newly proposed model.

FF0057

POLISHER: a tool for using ultra short reads in microbial genome finishing

Kurt LaButti[1], Brian Foster[1], Steve Lowry[1], Stephan Trong[2], Eugene Goltsman[1], and Alla Lapidus[1]

1Lawrence Berkeley National Laboratory, 2Lawrence Livermore National Laboratory DOE Joint Genome Institute, Walnut Creek, CA 94598

Polishing is one of the major steps of genome finishing at the JGI (Joint Genome Institute). Along with repeat resolution and gap closure, it is required to produce a fully sequenced high quality genome. Polishing consists of consensus error correction and quality improvement such that the resulting consensus meets a pre-defined standard. This has traditionally been done through targeted Sanger clone based sequencing, making it the most time consuming and resource intensive stage in finishing. Our pilot experiments, conducted using Illumina data produced by the JGI for several microbes, demonstrated that aligning ultra short reads against unpolished contigs help correct a significant amount of consensus errors and greatly reduce the amount of quality improvement necessary to produce a finished genome. A prototype tool named the “Polisher” was developed in order to automate this process. It facilitates polishing and error correction of a subject assembly (acefile), typically a draft or closed assembly, using Illumina read data.

The Polisher corrects consensus errors and supports correctly called bases that would normally be targeted for polishing due to their below standard quality by analyzing Illumina alignment data and automatically modifying the consensus. The Illumina read data in Fasta format is aligned to the subject sequence and simultaneously parsed for the best hit based on percent identity. The best hit alignment results are then used to determine coverage and discrepancy information per base in the subject. A list of errors is generated where there is overwhelming evidence that the consensus base is wrong and needs to be changed. Such areas represent mismatches, deletions, and insertion. Mismatches and deletions are automatically corrected in the acefile while insertions, for now, remain as tags for manual inspection. In addition to the previously described error correction, other areas of the genome that would normally be targeted for polishing are inspected. These areas currently represent low quality, single subclone, and 454 only regions and exist as tags in the acefile. If there is overwhelming evidence that a particular base targeted for traditional polishing is correct, then that base is termed Supported and retagged. If there is not enough evidence for support, then the original polishing tag remains for traditional manual polishing.

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.

LLNL-ABS-402508

FF0063

Drafting with 454 without Sanger

Olga Chertkov, D. Sims, L. Meincke, C. Han

Los Alamos National Laboratory, Los Alamos NM

JGI is moving forward with replacing Sanger drafting strategy with 454 pyrosequencing. It will result in a very fast movement of microbial projects through JGI pipeline from drafting to finishing. It will also reduce significantly cost of drafting (by eliminating the cost of Sanger library construction). But will it deliver comparable sequence quality, since 454 draft is based on very short reads?

We have looked at the improvement of new data generated by 454 /FLEXA and assembled with new newbler assembler. We can see that with longer reads (250 bp compared to 100 bases before) and new base calling, error rate dropped significantly and majority of the errors are now small insertion/deletions. We have plans to implement frameshift detection software to identify possible errors in 454 data in the absence of Sanger reads.

Since 454 data now is more reliable, we do not need to confirm single clone coverage of 454 data with Sanger. It will reduce significantly number of finishing reactions picked for each project. We present new way of introducing 454 data to the assembly by creating overlapping fake reads going in both directions and assigning higher phred score. It reduces number of primers picked by 30%.

There is a need for new screening tool to identify contaminations in 454 data. Since reads are short (250 bp) regular blast may not be very efficient tool. Probably feedback needed to Roche about identifying small chimeric reads. And some reads are still discrepant with Sanger.

We are also looking forward to get new Solexa data as complementary tool for finishing. Currently we are in the process of implementing Polisher (from PGF-JGI) for use with Solexa data.

FF0067

Multi-Platform Joint Assemblies of Mammalian BACs

Michael E. Holder, Christian J. Buhay, Aniko Sabo, Xiang Qin, Carson Qu, Shannon Dugan-Rocha, Yan Ding, Huyen Dinh, Lynne Nazareth, Christie Kovar-Smith, Yi Han, Jeffrey Reid, Donna M. Muzny and Richard A. Gibbs

The BCM-HGSC developed a process to finish selected gene rich regions exploiting the strengths of multiple sequencing technologies and using the whole genome shotgun assembly of Rat genome as a starting draft. The original rat WGS assembly (Rnor 3.1) was built from Sanger sequence data and achieved an average coverage of 6X which left many gaps and low quality areas throughout the assembly. By incorporating 454, Solexa, and SOLiD we are able to both close gaps and improve the low quality regions. We use the BAC as the unit of sequence and construct a tiling path of BACs which spans the genome in order to determine the start and stop positions of each. Additionally, the regions to finish are mapped to the assembly in order to determine which BAC(s) span these regions. Once the BACs containing the regions to finish are determined, we use the BCM BAC fisher to isolate the set of whole genome Sanger reads which are required to build an initial Phrap assembly for each BAC

To effectively utilize the new sequencing technologies, such as 454, Solexa, and Solid, the DNA corresponding to the previously selected BACs are combined into pools of 100. Shotgun libraries for each pool are prepared and sequenced. Once sequencing is complete, the data is processed to yield a set of reads which represent the sample. Subsequently we use 2 approaches to identify which reads belong to each BAC in the pool. For the data generated by the 454 platform, we perform a de novo assembly using the 454 Newbler assembler. The Sanger reads from each BAC are mapped using BLAST onto the 454 assembly contigs in order to pick contigs that contain one or more entire Sanger read(s). The selected contigs are formatted into artificial PHD files which are added to the initial Phrap assembly to produce a combined 454 and Sanger assembly of the given BAC. This step offers the possibility of closing gaps since the 454 platform doesn't have the same sequencing bias as the Sanger data. For data generated by the Solexa and Solid platforms we use Mosaik to map reads to the portions of the genome assembly that have been identified to belong to BACs in the given pool. Once these reads are mapped, Mosaik is able to produce a set of contigs which are transformed into artificial PHD reads and added to the 454 plus Sanger BAC assemblies. This portion of the process aims to improve low sequence quality in the original Sanger data as well as provide a means to correct homo-polymer issues that may have been present in the 454 data.

By combining these different types of sequencing data, we have been able to close gaps and improve low quality areas in the targeted finishing regions. In areas where gaps have been closed by 454 data, PCR is used to check and correct home-polymer issues that may arise. In those areas where quality has been improved, questionable base calls are biased towards a consensus of the data provided by the new platforms.

FF0069

Optimization of Direct BAC Walk and GenomiPhi for Finishing

Guan Chen, Qiaoyan Wang, Shannon Dugan-Rocha, Yan Ding, Zhangwan Li, Donna Villasana, Shalini N. Jhangiani, Christian J. Buhay, Aniko Sabo, Mike E. Holder, Donna M. Muzny and Richard A Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030

New sequencing technologies including 454 and Solexa have recently been applied at the BCM –HGSC. These new methods have greatly changed the finishing platform from traditional Sanger assemblies into combined assemblies with various levels of 454, Solexa and Sanger coverage. Although these new technologies offer great advantages for finishing such as high throughput and cost effective sequencing without cloning bias, problems such as template availability and highly repetitive regions remain issues. Therefore, we have implemented and optimized finishing methods including direct genomic sequencing using the GenomiPhi kit and direct BAC walk with the TempliPhi Sequence Resolver kit to deal with these challenges.

Direct genomic sequencing has been designed not only for closing or extending gap regions where templates are not available, but also for bridging contigs without order and orientation. Although results may vary depending on the size and complexity of the target sequence, recent results have shown success rates of 60%-88% with an average Phred20 of 520bp. Direct primer walking is now routinely done on mammalian DNA that has been amplified and is sequenced at $1/16^{th}$ cocktail concentration in a 96 well format.

Direct BAC Walk with the TempliPhi Sequence Resolver kit was initially tested at BCM-HGSC and used quite successfully with small plasmids to resolve sequencing problems such as repeats, sequencing stops and compressions. Earlier success with these smaller plasmids prompted experiments with larger BAC templates. Recently, direct BAC walks on BACs amplified with the TempliPhi kit have been successfully completed for various species including Rat, Pea Aphid, Rhesus Macaque, Sea Urchin, and Bovine. This new technique has been successful in many kinds of difficult sequencing situations such as gaps without order and orientation, tandem repeats, small duplications, sequence compressions, GC-rich regions and regions of complicated secondary structure. Various tests were conducted to optimize this kit for use with mammalian BAC DNA. Initial testing of optimal DNA starting concentration was completed and further testing was done to increase product yield. These tests included varying the annealing time and temperature as well as the initial denaturization. Although these tests did not produce significant differences in product concentration, altering the incubation time from 18 hours to 20 hours proved very effective. With this change in incubation time, the product yield increased from ~100ng/ul to ~175ng/ul. Although results may vary depending on the size and complexity of the target sequence, recent results have shown success rates of 95%-98% with an average Phred20 of 550bp. Protocols and results of these strategies will be presented.

FF0077

Modification of the 454 LT Paired-end Library Protocol for Constructing Longer Insert Size Libraries

Ze Peng[1], Matthew Hamilton[1], Sara Ting[2], Hank Tu[1], Eugene Goltsman[1], Alla Lapidus[1], Susan Lucas[2], and Jan-Fang Cheng[1]

1Lawrence Berkeley National Laboratory, 2Lawrence Livermore National Laboratory US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA

Paired-end library sequencing has been proven useful in scaffold construction during *de novo* assembly of genomic sequences. The ability of generating mate pairs with 8 Kb or greater insert sizes is especially important for genomes containing long repeats. While the current 454 GS LT Paired-end library preparation protocol can successfully construct libraries with 3 Kb insert size, it fails to generate longer insert sizes because the protocol is optimized to purify shorter fragments. We have made several changes in the protocol in order to increase the fragment length. These changes include the use of Promega column to increase the yield of large size DNA fragments, two gel purification steps to remove contaminated short fragments, and a large reaction volume in the circularization step to decrease the formation of chimeras. We have also made additional changes in the protocol to increase the overall quality of the libraries. The quality of the libraries are measured by a set of metrics, which include levels of redundant reads, linker positive, linker negative, half linker reads, and driver DNA contamination, and read length distribution, were used to measure the primary quality of these libraries. We have also assessed the quality of the resulted mate pairs including levels of chimera, distribution of insert sizes, and genome coverage after the assemblies are completed. Our data indicated that all these changes have improved the quality of the longer insert size libraries.

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.

LLNL-ABS-402237

FF0079

Augmenting via Fishing of 454 and Sanger-454 Integrated Assemblies

Harindra M. Arachchi, Sarah Young, Margaret Priest, Michael G. Fitzgerald

The Broad Institute, Cambridge, MA, USA

Next generation sequencing technology has dramatically reduced the cost of whole genome shotgun data. The same cannot be said for genome finishing. The efficiency of generating targeted, and frequently difficult sequence will never compare with that for shotgun data, but nevertheless we are investigating various methods in which to reduce finishing cost. One approach is manual addition of shotgun data not incorporated into the WGA. Next-gen assemblies typically provide a high density of sequence data, not all of which gets utilized by assemblers. We frequently observe assemblies including roughly 80% of the available shotgun data. We will present data on the composition and utility of the unincorporated read sets. We will attempt to add these reads to our assembly via a technique known as fishing, which entails searching of unique “bait” sequence from a gap edge against the unincorporated read set known as our “pond”. The fished in reads will have their authenticity verified via paired end or other analysis.

FF0081

Development of a Genome Sequencing Pipeline for Single Stranded RNA Viruses at the Broad Institute

Elizabeth Ryan[1], Niall Lennon[1], Scott Anderson[1], Rachel Erlich[1], Lisa Green[1], Qing Yu[1], Kamran Rizzolo[1], Yama Thoulutsang[1], Toby Bloom[1], Keenan Ross[1], Matthew Henn[2], Bruce Birren[2], Robert Nicol[1], Jennifer Baldwin[1]

[1]Genome Sequencing Platform and [2]Genome Sequencing and Analysis Program
Broad Institute of MIT & Harvard, 7 Cambridge Center, Cambridge, MA USA.

As part of the NIAID-funded Microbial Sequencing Center at the Broad Institute we have created infrastructure to support sequencing large numbers of isolates of Flaviviruses, single stranded RNA viruses. Current projects include Hepatitis C Virus (HCV), Dengue Virus and Human Immunodeficiency Virus (HIV). Armed with this information, scientists will be better able to understand how viruses evolve, spread and cause disease, as well as how they respond to and evade host immune pressures. The data may also yield information about potential targets for new therapies, vaccines and diagnostic research.

Single stranded RNA viruses pose several significant challenges to traditional sequencing pipelines. First, the samples consist of highly diverse mixtures from complex populations of virus. Generally, individual molecules in samples are over ninety percent identical and yet diverse enough to make PCR primer design challenging; second, the RNA must be converted to cDNA, and then prepared by PCR for sequencing; third, because of the small genome size (typically on the order of 10kb) and large volume of isolates to be sequenced, substantial development was required to build a production process to track samples from sample receipt to assembly and annotation submission.

We present here the design, building and implementation of a robust, high-throughput Sanger-chemistry-based sequencing pipeline for RNA viruses. The current capacity is >200 genomes per week, but this can easily be scaled to generate several thousand per week. This system arose from an extensive collaborative development process with our colleagues in the operations, informatics, assembly and annotation teams and using methods co-opted from industrial process design and optimization. Ongoing continuous improvement of this process will increase capacity and decrease costs.

FF0097

SEQUENCING AND DEFINING DIVERSITY OF RARE HCV GENOTYPES USING 454

Aaron Berlin[1], Thomas Kuntzen[2], Niall Lennon[1], Sarah Young[1], Rachel Erlich[1], Sante Gnerre[1], Carsten Russ[1], Chad Nusbaum[1], Todd Allen[2], Bruce Birren[1], Matthew R. Henn[1]

1) Broad Institute of MIT & Harvard, Cambridge, MA 02142
2) Partners AIDS Research Center, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129

The Hepatitis C Virus contains 6 genotypes, each consisting of numerous subtypes. Defining HCV genotypes has clinical implications as treatment outcome has been linked to viral genotype. However, for most rare genotypes full-length reference sequences are unavailable, making it difficult to characterize these by conventional amplification and sequencing approaches. We have established a robust long range PCR protocol to amplify the entire HCV open reading frame in two 4.5 kb amplicons, and used this in combination with 454 pyrosequencing to produce full genome sequences, and to examine the population variability within individual serum samples. Due to the lack of a proofreading function of the RNA dependent RNA polymerase HCV is present in multiple quasispecies within a single patient, and also shows considerable sequence differences even between patients who share the same HCV subtype. This high degree of sequence variability prevented the use of current 454 assemblers including reference-assisted assembly. An in-house assembler was built on the Arachne platform to assemble the 454 reads *de novo*. Exploiting the non-repetitive nature of the virus, the assembler iteratively merges the diverse contigs into a more contiguous assembly. Once a consensus is achieved, reads are aligned back to the reference to evaluate overall diversity within a single serum sample. Nucleotide frequencies at each base are determined from aligned reads using a Neighborhood Quality Score algorithm. Consensus sequences from these methods allow us to better define the HCV genotype phylogeny. In addition, high-resolution information about HCV diversity within a single patient provides a novel opportunity to study evolution of the virus within its host.

FF0101

Genome Improvement and Sequencing Activities at JGI-SHGC

Schmutz, J[1], Grimwood, J[1], JGI-SHGC Group Members[1], and R.M. Myers[1]

[1]Joint Genome Institute – Stanford Human Genome Center Stanford University School of Medicine 975 California Avenue, Palo Alto, CA 94304

Since the completion of the sequencing of the human genome, the JGI has rapidly expanded its scientific goals in several DOE mission-relevant areas. At the JGI-SHGC, we have kept pace with this rapid expansion of projects with our focus on assessing, assembling, improving and finishing eukaryotic whole genome shotgun (WGS) projects for which the shotgun sequence is generated at the Production Genomic Facility (JGI-PGF). We follow this by combining the draft WGS with genomic resources generated at JGI-SHGC or in collaborator laboratories (including BAC end sequences, genetic maps and FLcDNA sequences) to produce an improved draft sequence. For eukaryotic genomes important to the DOE mission, we then add further information from directed experiments to produce reference genomic sequences that are publicly available for any scientific researcher. Also, we have continued our program for producing BAC-based finished sequence, both for adding information to JGI genome projects and for small BAC-based sequencing projects proposed through any of the JGI sequencing programs. We have now built our computational expertise in WGS assembly and analysis and have moved eukaryotic genome assembly from the JGI-PGF to JGI-SHGC. We have concentrated our assembly development work on large plant genomes and complex fungal and algal genomes. In the winter of 2008, we will be moving our facility from our current home at Stanford University to the newly constructed non-profit HudsonAlpha Biotechnology Institute in Hunstville, Alabama where we expect to be able to increase our impact on JGI genome projects.

FF0107

Single cell genome reconstruction of an uncultured, proteorhodopsin-containing Flavobacterium

Tanja Woyke, Alex Copeland, Gary Xie, Cliff Han, Jan-Fang Cheng, Hajnalka Kiss, Jimmy Saw, Pavel Senin, Michael E. Sieracki and Ramunas Stepanauskas

DOE, Joint Genome Institute, LBL, Walnut Creek, CA

Determining the genetic makeup of predominant microbial taxa with specific metabolic capabilities remains one the major challenges in microbial ecology and bioprospecting, due to the limitations of current cell culturing and metagenomic methods. The complexity of microbial communities and intraspecies variations hinders the assembly of individual genomes from metagenomic shotgun libraries. Here we report the use of single cell genomics to access the genome of a proteorhodopsin-encoding flavobacteria from Gulf of Maine bacterioplankton. We use high throughput fluorescence-activated sorting of single cells, whole genome amplification via multiple displacement amplification, PCR-screening and subsequent shotgun sequencing of this single amplified genome (SAG), allowing the genomic analysis of its novel photometabolic system and the sequence comparison to environmental marine sequence data.

FF0109

Error Detection in Microbial Genomic Sequences

Andrey Kislyuk[1], Alexandre Lomsadze[2], Alla Lapidus[3]
and Mark Borodovsky[2*]

[1]Georgia Institute of Technology, Atlanta, Georgia 30332
[2]Georgia Institute of Technology and Emory University, Atlanta, Georgia 30332
[3]Department of Energy, Joint Genome Institute (DOE-JGI), 2800 Mitchell Drive, Walnut Creek, California 94598
*borodovsky@gatech.edu

The advent of sequencing technologies, such as 454 Pyrosequencing, accelerates genome research by an order of magnitude. At the same time, the new techniques carry a larger volume of assembling operations and the risk for more frequent sequence errors. In this context, computational sequence error detection tools will make a higher impact on the final sequence quality and have to be included in the QC protocols, especially as the gene finding and annotation of genes in novel sequences is concerned. Moreover, frameshift detection at early sequencing stages can further reduce the overall cost of microbial genome assembly finishing. We have implemented an ab initio statistical algorithm for frameshift detection with immediate verification by protein sequence alignment. The algorithm focus is on discrimination of frameshifts caused by sequencing errors from gene overlaps that may naturally occur in the same DNA strand. We present an evaluation of the method accuracy for test sets of known genomes with artificially inserted errors as well as for test sets of genomes subjected to 454 Pyrosequencing with sequence errors recovered by the Sanger re-sequencing. We propose a mathematical model of 454 Pyrosequencing induced errors and compute the theoretically attainable error detection accuracy based on this model.

FF0111

Benchmarking and fragment recruitment of Flavobacteria sequences from Global Ocean Sampling Expedition data

Jimmy H. Saw, Pavel V. Senin, Ramunas Stepanauskas, Gary Xie

DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM

Significance:
Fragment recruitment is a useful method in metagenomic studies of diverse environments for a number of reasons: novel sequence data present in the environmental samples can be identified through comparison with existing genomic data in public databases, and useful data from closely related species can be applied to complement the analysis of draft genomic sequences.

Methods:
Using various alignment tools: MUMMER, BLAST, BLAT, and BLASTZ, we have benchmarked performance of these alignment tools using simulated metagenomic data. Next, through fragment recruitment, we have identified from the Global Ocean Sampling (GOS) metagenomic data, the sequences from closely related species of Flavobacteria that was recently sequenced through single-cell amplification method. We also compared the fragment recruitment of various complete and draft Flavobacteria genomic sequences available publicly.

Results:
We determined that for species level sequence identity, MUMMER alignment tool is adequate and was chosen due to its speed in performing the alignments. Fragment recruitment results of Flavobacteria were plotted using the plotting tools we have developed. Visualization tool we have developed clearly shows the sequences with high identity to the draft Flavobacteria genome we have sequenced and we are able to extract these sequences for further analysis. Also, through analysis of sequence identities from various sampling sites, we have determined that the close neighbor of the Flavobacteria we have sequenced is predominantly found in the New England coastal waters.

Conclusions:
Fragment recruitment method enabled us to identify from the GOS data the sequence reads of close neighbors to the draft Flavobacteria species we are interested in. These recruited reads may help us in further work involving closing of gap regions in this draft genome.

FF0113

Use of Paired 454 Reads to Scaffold a Bacterial Sanger Assembly

[1]Michele L. Williams, [2]Jeremy B. Zaitshik, [2]Dave W. Dyer, [2]Allison F. Gillaspy, [1]Mark L. Lawrence

[1]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762; [2]Oklahoma University Health Sciences Center, Oklahoma City, OK 73104

The greatest obstacle confronting the *Edwardsiella ictaluri* genome sequencing project is the large number of repeat elements leading to unscaffolded contigs. Following high-throughput shotgun assembly, we utilized traditional closure methods including primer walking on gap-spanning fosmid clones, primer walking of gap-spanning amplicons identified from paired-end reads, and 19-fold whole genome coverage using GS20 pyrosequencing reads (711,603 reads). These methods brought the contig number down to 75 contigs (42 contigs <2 kb in size). The remaining gaps were unscaffolded. Paired-end 454 sequencing was conducted to 10.7-fold sequencing coverage, which resulted from 539,666 paired reads that were assembled by Newbler into 702 large contigs comprising 63 scaffolds. Utilizing information from the paired-end 454 scaffolds, we were able to construct primers for directed PCRs followed by DNA sequencing of the amplicons to close our Sanger assembly. As a result, only six unscaffolded contigs remain in the assembly, making the use of combinatorial PCR to close remaining gaps feasible.

Notes

Notes

05/30/2008 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Healthy Start Breakfast Buffet** (Scrambled Eggs on side tomatoes, scallions and spinach, Turkey sausage links, Assorted chilled fruit juices, Platter of freshly sliced seasonal fruit, Assorted and bran muffins with butter, Granola and oatmeal served with low-fat milk, Individual assorted fruit yogurts, etc.)	x
830 - 845	Intro	x	Welcome Back Intro	**Chris Detter**
845 - 930	**Keynote**	**FF0053**	**Great Expectations: Fulfilling the promises of the Human Genome Project**	**Deanna Church**
930 - 1000	Speaker 1	FF0087	Using Consed and Cross_match in Resequencing Projects	**David Gordon**
1000 -1030	Break	x	Beverages and snacks provided	x
1030 - 1100	Speaker 2	FF0008	1000 Genomes Project Data Management and Analysis	**Hoda Khouri**
1100 -1130	Speaker 3	FF0020	Genome annotation improvement using new massive sequencing technologies.	**François Artiguenave**
1130 -1200	Speaker 4	FF0041	AlpheusTM - a system for nucleotide variant detection and digital gene expression analysis in ultra-high throughput sequencing projects	**Neil Miller**
1200 - 1230	Closing Discussions	x	**Closing Discussions -** discuss next year's plans	Chair - Chris Detter
1230 - 200pm	Lunch & Close of meeting	x	**La Fiesta Plaza Lunch Buffet -** (Chicken and beef fajitas with grilled red onions and bell peppers, Black beans (Vegetarian), Spanish rice (Vegetarian), Pork posole and calabacitas rancheras, Warm flour tortillas and butter, etc.) **End of meeting, enjoy lunch and Santa Fe**	x

Speaker Presentations (May 30th)

Abstracts are in order of presentation according to Agenda

FF0053 – Keynote

Great Expectations: Fulfilling the promises of the Human Genome Project

Deanna Church

National Center for Biotechnology Information (NCBI), Bethesda, MD 20894

FF0087

Using Consed and Cross_match in Resequencing Projects

David Gordon and Phil Green

Genome Sciences Dept, Univ of Washington and Howard Hughes Medical Institute

We are currently adapting the sequence editor Consed and Phrap package for use in resequencing projects with large numbers of short reads. Cross_match (the sequence comparison program distributed with Phrap) can now find gapped alignments for two million 36 bp Solexa reads against a 100 Mb reference genome in less than 5 minutes (3.6 Ghz CPU), and has been given a number of new features useful in resequencing applications. Consed now has the capability to read in and display Solexa and 454 reads (and its "traces"), and can handle resequencing assemblies with ~5 million aligned reads with response times similar to those for small assemblies. Additional features have been added to Consed, such as filters and navigators, making it useful (and not overwhelming) to view such assemblies. These and other improvements will be presented.

FF0008

1000 Genomes Project Data Management and Analysis

Hoda Khouri, Martin Shumway and Stephen Sherry

National Center for Biotechnology Information (NCBI), Bethesda, MD

The 1000 Genomes Project is an international research effort to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups within the next three years, using newly developed faster and less expensive sequencing technologies. By doing so the researchers aim to first discover >95 % of the sequence variants (e.g. single nucleotide polymorphisms (SNPs), copy number variants (CNVs), and insertion/deletions (indels) with minor allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions, and then identify the haplotype backgrounds and the linkage disequilibrium (LD) patterns of variant alleles. Furthermore, the project aims to improve the human reference sequence, support studies of variation in multiple populations, refine knowledge about the process of mutation and recombination, and develop better SNP and probe selection for genotyping platforms.

In the first phase of the 1000 Genomes Project, researchers will conduct three pilots. The results of these pilots will be used to decide how to most efficiently and cost effectively produce the project's detailed map of human genetic variation.

- Pilot study 1 will sequence the genomes of 180 people at low coverage (2-4X). This will test the ability to use low coverage data from new sequencing platforms to identify sequence variants and to put them in their genomic context.

- Pilot 2 will sequence the genomes of two nuclear families (both parents and an adult child) at deep coverage (20X) of each genome. This will provide a comprehensive dataset from six people to be a basis for comparison for other parts of the effort.

- Pilot 3 will sequence exons of about 1,000 genes in about 1,000 people. This is aimed at exploring how best to obtain a more detailed catalog in the approximately 2 percent of the genome that is comprised of protein-coding genes.

The data generated by the 1000 Genomes Project will be held by and distributed from The European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). There will also be a mirror site for data access at Beijing Genomics Institute (BGI). The computer and data storage requirements are considerable. The sequencing centers will submit the primary data to the Data Coordinating Center (DCC) in Short Read Format (SRF) files. The DCC will accession the data into the Short Read Archives (SRA) and make it publicly available in SRF files and fastQ files to the Analysis groups. The DCC will also store the alignment files and other analysis results for the researchers to evaluate and use in comparison studies. http://www.1000genomes.org/

FF0020

Genome Annotation Improvement Using New Massive Sequencing Technologies

François Artiguenave [1,2], Jean-Marc Aury [1,2], Benjamin Noël [1,2], Odile Rogier[1], France Denoeud[1,2], Claude Scarpelli[1], Patrick Wincker[1,2], and Olivier Jaillon[1,2]

1 Genoscope (CEA), 2 rue Gaston Crémieux CP5706, 91057 Evry, France
2 CNRS UMR 8030, 2 rue Gaston Crémieux CP5706, 91057 Evry, France

Automatic gene structure prediction in eukaryotic genome requires multiple, independent, complementary or contradictory analysis methods. These methods include alignment or *ab initio* gene prediction softwares. Usually, the different gene evidences obtained are combined, automatically or manually, to propose full-transcript structures. In this process, it has been well documented that the best evidences to build gene structures are produced by sequence comparison with collections of ESTs or FL cDNA sequences. In this context, new sequencing instruments enable rapid and inexpensive DNA sequence data production. Hence, we have now access to deep sequencing coverage of genes with transcript sequences. We will present new developments on structural annotation using data obtained with Illumina's Solexa and 454 Roche sequencers. We will show how these technologies allow us to annotate the near-full set of genes of a large eukaryote genome.

FF0041

AlpheusTM - a system for nucleotide variant detection and digital gene expression analysis in ultra-high throughput sequencing projects

Neil Miller, Joann Mudge, Andrew Farmer, Lar Mader, Selene Virk, C. Forrest Black, M. Kathy Myers, Dan Weems, Melodie Rice, Terri Gomez, Linda Julien, Sharon Lewis, Steven Day, Kamal Gajendran, Susan M. Baxter, Faye Schilkey, Gregory D. May, Stephen F. Kingsmore

National Center For Genome Resources (NCGR), Santa Fe, NM USA

Massively parallel sequencing technologies present researchers with the ability to generate sequence data at a rate and cost that was previously unimaginable. However, with this opportunity comes the challenge of analyzing tremendous amounts of characteristically short reads that are problematic for conventional assembly and analysis pipelines. We have developed the Alpheus system specifically for analysis, queries and visualization of gigabase-scale results from high-throughput technologies such as the Illumina Genome Analyzer, Roche-454 and Applied Biosystems SOLiD instruments. Designed for resequencing projects, Alpheus enables the identification and evaluation of nucleotide variants including synonymous and non-synonymous substitutions, micro-insertions and deletions and provides dynamic filters that allow researchers to dramatically reduce the number of false positive variant calls. In addition, Alpheus provides tools for comparing gene expression between samples using read abundance as a digital gene expression metric. Alpheus consists of a computational analysis pipeline, a results database and a web-based query and visualization interface.

Discussion Notes

Discussion Notes

Discussion Notes

Discussion Notes

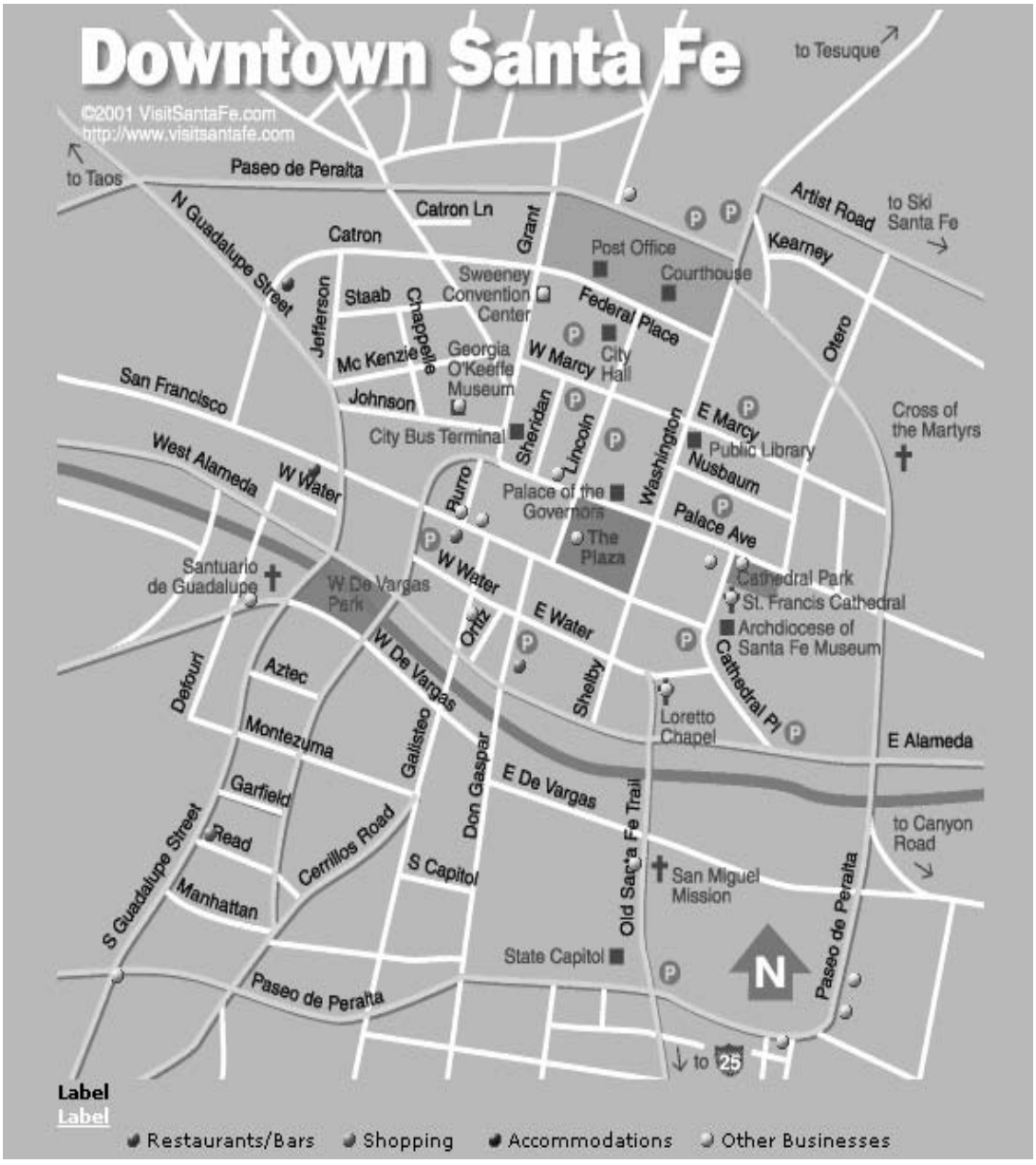
Discussion Notes

2008 Attendee List

FF #	Name	Affiliation	email
1	**Chris Detter**	Los Alamos National Laboratory - JGI	cdetter@lanl.gov
2	**Chad Geringer**	Illumina, Inc.	cgeringer@illumina.com
3	**Lori Court**	Caldera Pharmaceuticals, Inc	court@CPsci.com
4	**Eva Birnbaum**	Caldera Pharmaceuticals, Inc	eva@cpsci.com
5	**Hilary Morrison**	Marine Biological Laboratory	morrison@mbl.edu
6	**David Sims**	Los Alamos National Laboratory - JGI	dsims@lanl.gov
7	**Nicole Touchet**	Caldera Pharmaceuticals, Inc	touchet@cpsci.com
8	**Hoda Khouri**	National Center for Biotechnology Information (NCBI)	khourih@ncbi.nlm.nih.gov
9	**Rebecca Miller**	Caldera Pharmaceuticals, Inc	miller@cpsci.com
10	**Ed Zuiderwijk**	The Wellcome Trust Sanger Institute	ejz@sanger.ac.uk
11	**Tara Bennink**	Edge BioSystems	TBennink@edgebio.com
12	**Omayma Al-Awar**	Edge BioSystems	oalawar@edgebio.com
13	**Robert Blakesley**	NIH Intramural Sequencing Center (NISC)	rblakesl@nhgri.nih.gov
14	**Jyoti Gupta**	NIH Intramural Sequencing Center (NISC)	jyotig@mail.nih.gov
15	**Holly Coleman**	NIH Intramural Sequencing Center (NISC)	hcoleman@mail.nih.gov
16	**Darren Grafham**	The Wellcome Trust Sanger Institute	dg1@sanger.ac.uk
17	**Jessica Hostetler**	J Craig Venter Institute (JCVI)	Jessicah@jcvi.org
18	**Valérie BARBE**	GENOSCOPE	vbarbe@genoscope.cns.fr
19	**Stéphane CRUVEILLER**	GENOSCOPE	scruveil@genoscope.cns.fr
20	**François ARTIGUENAVE**	GENOSCOPE	artigue@genoscope.cns.fr
21	**Arnaud COULOUX**	GENOSCOPE	acouloux@genoscope.cns.fr
22	**Joann Mudge**	National Center for Genome Resources (NCGR)	jm@ncgr.org
23	**Jennifer Heddens**	Integrated DNA Technologies	jheddens@idtdna.com
24	**Ben Faga**	Cold Spring Harbor Laboratory	faga.cshl@gmail.com
25	**Giselle Kerry**	The Wellcome Trust Sanger Institute	gh2@sanger.ac.uk
26	**Katherine Auguer**	The Wellcome Trust Sanger Institute	kaa@sanger.ac.uk
27	**Siobhan Whithead**	The Wellcome Trust Sanger Institute	slw@sanger.ac.uk
28	**Jim Bristow**	Joint Genome Institute - LBL	JBristow@lbl.gov
29	**John Havens**	Integrated DNA Technologies	jhavens@idtdna.com
30	**Brad Thomas**	Seqwright, Incorporated	kbthomas@seqwright.com
31	**Ken Taylor**	Integrated DNA Technologies	ktaylor@idtdna.com
32	**Danielle Walker**	The Wellcome Trust Sanger Institute	dw2@sanger.ac.uk
33	**Aijazuddin Syed**	Joint Genome Institute - LBL	ASyed@lbl.gov
34	**Iain MacCallum**	Broad Institute of MIT	iainm@broad.mit.edu
35	**Scott Sammons**	Center for Disease Control (CDC)	zno6@CDC.GOV
36	**Christian Buhay**	Baylor College of Medicine	cbuhay@bcm.tmc.edu
37	**Justin Johnson**	J Craig Venter Institute (JCVI)	jjohnson@jcvi.org
38	**Thomas Brettin**	Los Alamos National Laboratory - JGI	brettin@lanl.gov
39	**Isaac Meek**	Caliper Life Sciences	Isaac.Meek@caliperls.com
40	**Savita Shanker**	University of Florida	ss@biotech.ufl.edu
41	**Neil Miller**	National Center for Genome Resources (NCGR)	nam@ncgr.org
42	**David Bruce**	Los Alamos National Laboratory - JGI	dbruce@lanl.gov
43	**Hope Tice**	Joint Genome Institute - LLNL	tice1@llnl.gov
44	**Julianna Chow**	Joint Genome Institute - LBL	JChow@lbl.gov
45	**Martha Trela**	Pacific Biosciences	mtrela@pacificbiosciences.com
46	**Jose Olivares**	Los Alamos National Laboratory	olivares@lanl.gov
47	**Stephanie Malfatti**	Joint Genome Institute - LLNL	malfatti3@llnl.gov
48	**Anna Montmayeur**	The Broad Institute	annamont@broad.mit.edu
49	**Maqsudul Alam**	University of Hawaii at Manoa	alam@hawaii.edu
50	**Sydney Brenner**	Salk Institute	backhill@hotmail.co.uk
51	**Teri Mueller**	Roche Diagnostics	teri.mueller@roche.com
52	**Bruce Birren**	Broad Institute of MIT	bwb@broad.mit.edu
53	**Deanna Church**	National Center for Biotechnology Information (NCBI)	church@ncbi.nlm.nih.gov
54	**Peter Welch**	Invitrogen Corporation	Peter.Welch@invitrogen.com
55	**Dibyendu Kumar**	University Of Florida	dkumar@ufl.edu
56	**Alicia Clum**	Joint Genome Institute - LBL	aclum@lbl.gov
57	**Kurt LaButti**	Joint Genome Institute - LBL	klabutti@lbl.gov
58	**Steve Lowry**	Joint Genome Institute - LBL	slowry@lbl.gov
59	**Lynne Goodwin**	Los Alamos National Laboratory - JGI	lynneg@lanl.gov
60	**Brian Foster**	Joint Genome Institute - LBL	bfoster@lbl.gov
61	**Alla Lapidus**	Joint Genome Institute - LBL	alapidus@lbl.gov

62	**Chris Munk**	Los Alamos National Laboratory - JGI	cmunk@lanl.gov
63	**Olga Chertkov**	Los Alamos National Laboratory - JGI	ochrtkv@lanl.gov
64	**Linda Meincke**	Los Alamos National Laboratory - JGI	meincke@lanl.gov
65	**Ronan O'Malley**	The Salk Institute for Biological Research	omalley@salk.edu
66	**Yan Ding**	Baylor College of Medicine	yding@bcm.tmc.edu
67	**Michael Holder**	Baylor College of Medicine	mholder@bcm.tmc.edu
68	**Qiaoyan Wang**	Baylor College of Medicine	qiaoyanw@bcm.tmc.edu
69	**Guan Chen**	Baylor College of Medicine	guanc@bcm.tmc.edu
70	**Xiaohong Liu**	Broad Institute of MIT	xliu@broad.mit.edu
71	**Hui Sun**	Joint Genome Institute - LBL	HSun@lbl.gov
72	**Feng Chen**	Joint Genome Institute - LBL	fchen@lbl.gov
73	**Mike Fitzgerald**	Broad Institute of MIT	fitz@broad.mit.edu
74	**Amr Abouelleil**	Broad Institute of MIT	amr@broad.mit.edu
75	**Jan Kieleczawa**	Wyeth Research	JKieleczawa@wyeth.com
76	**Hajnalka Kiss**	Los Alamos National Laboratory - JGI	hajkis@lanl.gov
77	**Ze Peng**	Joint Genome Institute - LBL	zpeng@lbl.gov
78	**Jan-Fang Cheng**	Joint Genome Institute - LBL	jfcheng@lbl.gov
79	**Harindra Arachchi**	Broad Institute of MIT	harindra@broad.mit.edu
80	**Daniel Bessette**	Broad Institute of MIT	danielb@broad.mit.edu
81	**Elizabeth Ryan**	Broad Institute of MIT	eryan@broad.mit.edu
82	**Eugene Goltsman**	Joint Genome Institute - LBL	egoltsman@lbl.gov
83	**Liz Saunders**	Los Alamos National Laboratory - JGI	ehs@lanl.gov
84	**Andrew Bradbury**	Los Alamos National Laboratory	amb@lanl.gov
85	**Jim Knight**	Roche Diagnostics - 454	james.knight@roche.com
86	**Cliff Han**	Los Alamos National Laboratory - JGI	han_cliff@lanl.gov
87	**David Gordon**	Howard Hughes Medical Institute at University of Washington	dgordon@u.washington.edu
88	**Stephen Kingsmore**	National Center for Genome Resources (NCGR)	sfk@ncgr.org
89	**Greg May**	National Center for Genome Resources (NCGR)	gdm@ncgr.org
90	**Bob Fulton**	Washington University in St. Louis	bfulton@watson.wustl.edu
91	**Patrick Minx**	Washington University in St. Louis	pminx@watson.wustl.edu
92	**Tina Graves**	Washington University in St. Louis	tgraves@watson.wustl.edu
93	**My-Hanh Nguyen**	Roche Diagnostics	my-hanh.nguyen@roche.com
94	**Brian Desany**	Roche Diagnostics, 454 Life Sciences	brian.desany@roche.com
95	**David Gilbert**	Joint Genome Institute - LBL	degilbert@lbl.gov
96	**Jean Challacombe**	Los Alamos National Laboratory - JGI	jchalla@lanl.gov
97	**Aaron Berlin**	Broad Institute of MIT	amberlin@broad.mit.edu
98	**Sean Sykes**	Broad Institute of MIT	ssykes@broad.mit.edu
99	**Stephan Trong**	Joint Genome Institute - LLNL	trong1@llnl.gov
100	**Jane Hutchinson**	Roche Applied Science	jane.hutchinson@roche.com
101	**Jeremy Schmutz**	Stanford Genome Center	jeremy@shgc.stanford.edu
102	**Haley Fiske**	Illumina, Inc.	hfiske@illumina.com
103	**Ramunas Stepanauskas**	Bigelow Laboratory for Ocean Sciences	rstepanauskas@bigelow.org
104	**Steve Brown**	Oakridge National Laboratory	brownsd@ornl.gov
105	**Tim Harkin**	Roche Diagnostics	tim.harkins@roche.com
106	**Marvin Stodolsky**	Dept. of Energy, OBER	Marvin.Stodolsky@science.doe.gov
107	**Tanja Woyke**	DOE Joint Genome Institute - LBNL	TWoyke@lbl.gov
108	**Johar Ali**	Ontario Institute for Cancer Research (OICR)	johar.ali@oicr.on.ca
109	**Andrey Kislyuk**	Georgia Institute of Technology	kislyuk@gatech.edu
110	**Mircea Podar**	Oakridge National Laboratory	podarm@ornl.gov
111	**Jimmy Saw**	University of Hawaii and JGI-LANL	jsaw@lanl.gov
112	**Pavel Senin**	University of Hawaii and JGI-LANL	pvs@lanl.gov
113	**Michele Williams**	Mississippi State University	MWilliams@cvm.msstate.edu
114	**Craig Pierson**	Illumina, Inc.	cpierson@illumina.com
115	**Jon Sorenson**	Pacific Biosciences	JSorenson@pacificbiosciences.com
116	**Michael Rhodes**	Applied Biosystems	rhodesmd@appliedbiosystems.com
117	**Tim Hunkapiller**	Applied Biosystems	tim.hunkapiller@appliedbiosystems.com
118	**Patrick Chain**	DOE Joint Genome Institute - LLNL	chain2@llnl.gov
119	**Donna Muzny**	Baylor College of Medicine	donnam@bcm.tmc.edu
120	**Lance Green**	DOE Joint Genome Institute - LANL	ldgreen@lanl.gov
121	**Anne Foster-Kloepple**	Applied Biosystems	fosteram@appliedbiosystems.com
122	**Cody Cain**	Applied Biosystems	Cody.Cain@appliedbiosystems.com
123	**Brandon Blakey**	Applied Biosystems	BlakeyBM@appliedbiosystems.com
124	**Eric Mathur**	Synthetic Genomics, Inc.	EMathur@SyntheticGenomics.com
125	**Joe Petrosino**	Baylor College of Medicine	jpetrosi@bcm.tmc.edu
126	**Sarah Highlander**	Baylor College of Medicine	sarahh@bcm.tmc.edu

Map of Santa Fe, NM



History of Santa Fe, NM

Thirteen years before Plymouth Colony was settled by the Mayflower Pilgrims, Santa Fe, New Mexico, was established with a small cluster of European type dwellings. It would soon become the seat of power for the Spanish Empire north of the Rio Grande. Santa Fe is the oldest capital city in North America and the oldest European community west of the Mississippi.

While Santa Fe was inhabited on a very small scale in 1607, it was truly settled by the conquistador Don Pedro de Peralta in 1609-1610. Santa Fe is the site of both the oldest public building in America, the Palace of the Governors and the nation's oldest community celebration, the Santa Fe Fiesta, established in 1712 to commemorate the Spanish reconquest of New Mexico in the summer of 1692. Peralta and his men laid out the plan for Santa Fe at the base of the Sangre de Cristo Mountains on the site of the ancient Pueblo Indian ruin of Kaupoge, or "place of shell beads near the water."

The city has been the capital for the Spanish "Kingdom of New Mexico," the Mexican province of Nuevo Mejico, the American territory of New Mexico (which contained what is today Arizona and New Mexico) and since 1912 the state of New Mexico. Santa Fe, in fact, was the first foreign capital over taken by the United States, when in 1846 General Stephen Watts Kearny captured it during the Mexican-American War.

Santa Fe's history may be divided into six periods:

Preconquest and Founding
(circa 1050 to 1607)

Santa Fe's site was originally occupied by a number of Pueblo Indian villages with founding dates from between 1050 to 1150. Most archaeologists agree that these sites were abandoned 200 years before the Spanish arrived. There is little evidence of their remains in Santa Fe today.

The "Kingdom of New Mexico" was first claimed for the Spanish Crown by the conquistador Don Francisco Vasques de Coronado in 1540, 67 years before the founding of Santa Fe. Coronado and his men also discovered the Grand Canyon and the Great Plains on their New Mexico expedition.

Don Juan de Onate became the first Governor-General of New Mexico and established his capital in 1598 at San Juan Pueblo, 25 miles north of Santa Fe. When Onate retired, Don Pedro de Peralta was appointed Governor-General in 1609. One year later, he had moved the capital to present day Santa Fe.

Settlement Revolt & Reconquest
(1607 to 1692)

For a period of 70 years beginning the early 17th century, Spanish soldiers and officials, as well as Franciscan missionaries, sought to subjugate and convert the Pueblo Indians of the region. The indigenous population at the time was close to 100,000 people, who spoke nine basic languages and lived in an estimated 70 multi-storied adobe towns (pueblos), many of which exist today. In 1680, Pueblo Indians revolted against the estimated 2,500 Spanish colonists in New Mexico, killing 400 of them and driving the rest back into Mexico. The conquering Pueblos sacked Santa Fe and burned most of the buildings, except the Palace of the Governors. Pueblo Indians occupied Santa Fe until 1692, when Don Diego de Vargas reconquered the region and entered the capital city after a bloodless siege.

Established Spanish Empire
(1692 to 1821)

Santa Fe grew and prospered as a city. Spanish authorities and missionaries - under pressure from constant raids by nomadic Indians and often bloody wars with the Comanches, Apaches and Navajos-formed an alliance with Pueblo Indians and maintained a successful religious and civil policy of peaceful coexistence. The Spanish policy of closed empire also heavily influenced the lives of most Santa Feans during these years as trade was restricted to Americans, British and French.

The Mexican Period
(1821 to 1846)

When Mexico gained its independence from Spain, Santa Fe became the capital of the province of New Mexico. The Spanish policy of closed empire ended, and American trappers and traders moved into the region. William Becknell opened the l,000-mile-long Santa Fe Trail, leaving from Arrow Rock, Missouri, with 21 men and a pack train of goods. In those days, aggressive Yankeetraders used Santa Fe's Plaza as a stock corral. Americans found Santa Fe and New Mexico not as exotic as they'd thought. One traveler called the region the "Siberia of the Mexican Republic."

For a brief period in 1837, northern New Mexico farmers rebelled against Mexican rule, killed the provincial governor in what has been called the Chimayó Rebellion (named after a village north of Santa Fe) and occupied the capital. The insurrectionists were soon defeated, however, and three years later, Santa Fe was peaceful enough to see the first planting of cottonwood trees around the Plaza.

Territorial Period
(1846 to 1912)

On August 18, 1846, in the early period of the Mexican American War, an American army general, Stephen Watts Kearny, took Santa Fe and raised the American flag over the Plaza. Two years later, Mexico signed the Treaty of Guadalupe Hidalgo, ceding New Mexico and California to the United States.

In 1851, Jean B. Lamy, arrived in Santa Fe. Eighteen years later, he began construction of the Saint Francis Cathedral. Archbishop Lamy is the model for the leading character in Willa Cather's book, "Death Comes for the Archbishop."

For a few days in March 1863, the Confederate flag of General Henry Sibley flew over Santa Fe, until he was defeated by Union troops. With the arrival of the telegraph in 1868 and the coming of the Atchison, Topeka and the Santa Fe Railroad in 1880, Santa Fe and New Mexico underwent an economic revolution. Corruption in government, however, accompanied the growth, and President Rutherford B. Hayes appointed Lew Wallace as a territorial governor to "clean up New Mexico." Wallace did such a good job that Billy the Kid threatened to come up to Santa Fe and kill him. Thankfully, Billy failed and Wallace went on to finish his novel, "Ben Hur," while territorial Governor.

Statehood
(1912 to present)

When New Mexico gained statehood in 1912, many people were drawn to Santa Fe's dry climate as a cure for tuberculosis. The Museum of New Mexico had opened in 1909, and by 1917, its Museum of Fine Arts was built. The state museum's emphasis on local history and native culture did much to reinforce Santa Fe's image as an "exotic" city.

Throughout Santa Fe's long and varied history of conquest and frontier violence, the town has also been the region's seat of culture and civilization. Inhabitants have left a legacy of architecture and city planning that today makes Santa Fe the most significant historic city in the American West.

In 1926, the Old Santa Fe Association was established, in the words of its bylaws, "to preserve and maintain the ancient landmarks, historical structures and traditions of Old Santa Fe, to guide its growth and development in such a way as to sacrifice as little as possible of that unique charm born of age, tradition and environment, which are the priceless assets and heritage of Old Santa Fe."

Today, Santa Fe is recognized as one of the most intriguing urban environments in the nation, due largely to the city's preservation of historic buildings and a modern zoning code, passed in 1958, that mandates the city's distinctive Spanish-Pueblo style of architecture, based on the adobe (mud and straw) and wood construction of the past. Also preserved are the traditions of the city's rich cultural heritage which helps make Santa Fe one of the country's most diverse and fascinating places to visit.

SameDay Oligos®

Order online by 2:00 P.M. ET and receive your SameDay® Oligos the next business morning via priority shipping.

- 2-OD guarantee (sufficient for > 250 PCR reactions)
- Deprotected & desalted
- Lyophilized
- Available within the U.S. and Canada
- 15-45 bases

Learn more at www.idtdna.com/sameday

IDT®
INTEGRATED DNA TECHNOLOGIES

INNOVATION AND PRECISION IN NUCLEIC ACID SYNTHESIS

800-328-2661 (US & Canada)
+1-319-626-8400 (Outside US & Canada)
+32 (0)16 28 22 60 (Europe)
www.idtdna.com

ISO 9001:2000
FM88954