Contents

The 2007 "Finishing in the Future" Organizing Committee:

** Chris Detter, Ph.D., Genome Group / Team Leader, LANL*
** Johar Ali, Ph.D., Staff Scientist / Finishing Group Leader, BCGSC*
** 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., Biotech Core Manager, JCVI*
** Alla Lapidus, Ph.D., Microbial Genomics Group Leader, LBNL-JGI*
** Donna Muzny, M.S., Director of Operations, BCM*

06/18/2007 - Monday				
Time	Type	Abstract #	Title	Speaker
730 - 830am	Breakfast	x	**La Fonda Breakfast Buffet** (French "Texas Toast", Buttermilk pancakes, Poached Eggs Benedict, Grilled breakfast potatoes, sausage links, breads, and fruit, etc.)	x
830 - 845	Intro	x	Welcome Intro	Gary Resnick
845 - 930	**Keynote**	FF0045	Finishing in the New DNA Sequencing Era	George Weinstock
930 - 950	Speaker 1	FF0042	Building Better Genomes	Jeremy Schmutz
950 - 1010	Speaker 2	FF0111	GEBA - A Genomic Encyclopedia of Bacteria and Archaea	Jonathan Eisen
1010 - 1040	Break	x	Beverages and snacks provided	x
1040 - 1100	Speaker 3	FF0080	Manual Sequencing Improvement of Bacterial Genomes	Aye Wollam
1100 - 1120	Speaker 4	FF0087	A High-Throughput cDNA Finishing Pipeline - *Bos Taurus* as a Model	Johar Ali
1120-1220	Panel Discussion	x	**Panel Discussion**	Chair - Johar Ali
1230 - 145pm	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
200 - 220	Speaker 5	FF0088	Assembly and Finishing using 454 Sequencing Data (454 / Roche)	Jim Knight
220 - 240	Speaker 6	FF0110	The Illumina Genome Analyzer System: Cost-effective, High Throughput Genomics Using Solexa DNA Sequencing Technology (illumina)	Gary Schroth
240 - 300	Speaker 7	FF0099	Using a gigabase of short reads: leveraging paired end reads and color base encoding in the AB SOLiD TM sequencing system in finishing and SNP detection (ABI)	Fiona Hyland
300 - 400	Panel Discussion	x	**Panel Discussion**	Chair - Darren Grafham
400 - 430	Break	x	Beverages and snacks provided	x
430 - 600	**Posters - odd #s**	x	**Poster Session**	x
630 - 900pm	**Meet & Greet Party**	x	**Meet & Greet Party - sponsored by Roche --- Food & Drinks**	x

06/19/2007 - Tuesday				
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 - Informatics	Jim Bristow
845 - 930	**Keynote**	FF0011	Automating the finishing process: dreams and realities	Mihai Pop
930 - 950	Speaker 1	FF0018	Celera Assembler: Adapting for the Future	Granger Sutton
950 - 1010	Speaker 2	FF0115	Charting and Sequencing Structural Variation using High-Resolution Paired-End Mapping (HR-PEM)	Jan Korbel
1010 -1040	Break	x	Beverages and snacks provided	x
1040 -1100	Speaker 3	FF0005	Assessment of 454 Sequencing Errors in Microbial Genomes	Stephan Trong
1100 -1120	Speaker 4	FF0026	New Sequencing Technologies and Hybrid Assemblies - A discussion on a shift in finishing paradigm: do we need to analyze each read?	Harindra Arachchi
1120 - 1220	Panel Discussion	x	**Panel Discussion**	Chair - Patrick Chain
1230 - 145pm	Lunch	x	**La Fonda Lunch Buffet** (Pork tenderloin achiote-rubbed and char-grilled with tomatillo-chipotle sauce, Breast of chicken filled with bacon, red onions, green chile, jack and cheddar cheeses, lightly-breaded, flash-fried and baked, accompanied by mild green chile cream sauce, etc.)	x
200-220	Speaker 5	FF0090	Transforming Genomes with New Sequencing Technology	Donna Muzny
220-240	Speaker 6	FF0039	*De novo* Hybrid 454 / Sanger Genome Assembly of *Phytophthora capsici*	Joann Mudge
240 - 300	Speaker 7	FF0032	Incorporating New Sequencing Technologies into Finishing Strategy	Sean Sykes
300 - 400	Panel Discussion	x	**Panel Discussion**	Chair - Donna Muzny
400 - 430	Break	x	Beverages, Wine & Cheese provided - sponsored by IDT, Edge, & Invitrogen	x
430 - 600	**Posters - even #s**	x	**Poster Session with Wine & Cheese - sponsored by IDT, Edge, & Invitrogen**	x
600 - bedtime	on your own	x	Dinner and night on your own - enjoy	x

06/20/2007 - Wednesday				
Time	Type	Abstract #	Title	Speaker
745 - 845am	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
845 - 900	Intro	x	Welcome Back Intro - New Technologies	Paul Richardson
900 - 930	Speaker 1	FF0033d	New Amplification and Cloning Tools for Finishing Genomes	David Mead
930 - 950	Speaker 2	FF0102	TaxSorter: A Solution to Metagenomic Projects	Li Liu
950 -1020	Break	x	Beverages and snacks provided	x
1020 - 1040	Speaker 3	FF0047a	Metagenomic Assembly QC	Alla Lapidus
1040 -1100	Speaker 4	FF0065	Evaluation of New methods and Approaches for Comparative Metagenomic Studies	Emmanuel Mongodin
1100 -1200	Panel Discussion	x	**Panel Discussion** - New Technologies	Chair - Alla Lapidus
1200 - 130pm	Lunch & Close of meeting	x	**La Fiesta Plaza Lunch Buffet - sponsored by illumina** (Cheese enchiladas served with "Christmas" (red and green) chile, 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

06/18/2007 - Monday

Time	Type	Abstract #	Title	Speaker
730 - 830am	Breakfast	x	**La Fonda Breakfast Buffet** (French “Texas Toast”, Buttermilk pancakes, Poached Eggs Benedict, Grilled breakfast potatoes, sausage links, breads, and fruit, etc.)	x
830 - 845	Intro	x	Welcome Intro	Gary Resnick
845 - 930	**Keynote**	FF0045	Finishing in the New DNA Sequencing Era	George Weinstock
930 - 950	Speaker 1	FF0042	Building Better Genomes	Jeremy Schmutz
950 - 1010	Speaker 2	FF0111	GEBA - A Genomic Encyclopedia of Bacteria and Archaea	Jonathan Eisen
1010 - 1040	Break	x	Beverages and snacks provided	x
1040 - 1100	Speaker 3	FF0080	Manual Sequencing Improvement of Bacterial Genomes	Aye Wollam
1100 - 1120	Speaker 4	FF0087	A High-Throughput cDNA Finishing Pipeline - *Bos Taurus* as a Model	Johar Ali
1120-1220	Panel Discussion	x	**Panel Discussion**	Chair - Johar Ali
1230 - 145pm	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
200 - 220	Speaker 5	FF0088	Assembly and Finishing using 454 Sequencing Data (454 / Roche)	Jim Knight
220 - 240	Speaker 6	FF0110	The Illumina Genome Analyzer System: Cost-effective, High Throughput Genomics Using Solexa DNA Sequencing Technology (illumina)	Gary Schroth
240 - 300	Speaker 7	FF0099	Using a gigabase of short reads: leveraging paired end reads and color base encoding in the AB SOLiD TM sequencing system in finishing and SNP detection (ABI)	Fiona Hyland
300 - 400	Panel Discussion	x	**Panel Discussion**	Chair - Darren Grafham
400 - 430	Break	x	Beverages and snacks provided	x
430 - 600	**Posters - odd #s**	x	**Poster Session**	x
630 - 900pm	**Meet & Greet Party**	x	**Meet & Greet Party - sponsored by Roche --- Food & Drinks**	x

Speaker Presentations (June 18th)

Abstracts are in order of presentation according to Agenda

FF0045 – **Keynote**

Finishing in the New DNA Sequencing Era

George Weinstock

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

FF0042

Building Better Genomes

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

Joint Genome Institute – Stanford Human Genome Center
Stanford University School of Medicine, Palo Alto, CA 94304

The Stanford Human Genome Center (SHGC) began a collaboration with the Joint Genome Institute (JGI), concentrating on finishing the DOE portion of the human genome in 1999. 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 DOE mission genomes, we then add further information from directed experiments to produce reference genomic sequences. To date, we have produced reference versions of eight eukaryotic genomes and plan to release improved versions of three difficult genomes this year. Over time, we have developed and improved strategies for working on difficult to sequence genomes, polymorphic genomes and genomes with poor long-range contiguity with the end goal of benefiting the analysis and interpretation of their genomic sequences.

At the Finishing in the Future Meeting, we will present our latest advances in building better genomes. These include: collecting and integrating genomic resources in WGS draft assemblies, creating finishing substrates from polymorphic genomes, merging finished clone information into draft WGS assemblies and a new genome improvement paradigm to address the problem areas of a genome as the WGS sequencing is still in progress in order to produce a better draft assembly. We will also describe the new features in Orchid, our paired-end assembly viewer, including launching external programs from with Orchid and tighter integration with Consed.

FF0111

GEBA - a Genomic Encyclopedia of Bacteria and Archaea

Jonathan Eisen

UC Davis Genome Center, Davis, CA 95616

FF0080

Manual Sequence Improvement of Bacterial Genomes

Aye Wollam, Neha Shah, Robert S. Fulton, and Richard K. Wilson

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

Although base perfect finishing efforts such as that of Human Genome Project remains the best approach in providing in depth information of a genome, it may not always be possible or justifiable when the purpose is to generate and analyze a vast array of genomes for investigative sampling. In order to generate a product, which is qualitatively better than draft assembly, but less costly in terms of time and money than that of finished product, we have developed a strategy, which we termed "manual sequence improvement". Manual sequence improvement entails utilization of the principles of finishing with respect to manual sorting and improvement of the assembly but combined with conscious effort of reducing the cost, time and effort. The specifications of manual improvement are similar to those of comparative-grade finished sequence (Green et al. 2004) with the exception of providing order and orientation between contigs.

The protocol uses the following logistics: 1) the draft assembly must be improved by one round of directed sequencing reactions generated by autofinish program, 2) detectable major artifacts resulting from the sequence –assembly process (e.g. misassemblies) must be resolved, 3) incorrect consensus sequence is manually edited and irrelevant sequence at ends of contigs is trimmed, 4) joins between contigs, if found or confirmed, are manually completed, and 5) order and orientation of contigs (scaffolding information) are provided if possible. To date, we have manually improved twelve bacterial genomes in significantly less time than that of finishing efforts, and the result shows a sharp increase in N50 contig length, as well as increase contiguity in scaffolds, and improved annotation, when compared to the draft assembly.

FF0087

A High Throughput cDNA Finishing Pipeline-*Bos Taurus* as a Model

Johar Ali, Elizabeth Chun, Nancy Liao, Jerry Liu, Diana Palmquist, Peiming Huang, Brian Wynhoven, Robert Kirkpatrick, Robert Holt, Marco Marra, Steven Jones

Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC, Canada V5Z 1L3

The bovine full length cDNA (fl-cDNA) sequencing project provides an example of a high throughput finishing pipeline which will ultimately provide a valuable set of 10,000 high sequence quality full length (fl) cDNA clones for *Bos Taurus*. We have implemented an automated cDNA pipeline developed in-house to process sequenced ESTs, carry out clone selection and finish fl-cDNAs in a high throughput manner. Reference sequence(s) from bovine and other mammals were used to facilitate advanced primer design instead of the usual iterative primer walking approach to allow the finishing of fl-cDNA clones at a faster rate. Each clone is sequenced in both 5' and 3' directions maintaining a PHRED base quality 20 for each base pair with the exception of the 60bp region next to the poly-A which can be sequenced from the 5' end only. The cumulative consensus quality for each base is maintained at PHRED base quality minimum of 30. The sequenced reads are assembled using PHRAP and checked for errors in an automated fashion using in-house developed perl scripts. The automated checks are performed for contiguity, high quality base discrepancies, low quality bases, chimerism, single strand coverage, 5' and 3' single strand adopter primers absence, contamination, DNA quality (DNA purification error and mixed well), open reading frame and other biological problems like frame shifted and truncated. Depending on the nature of the problem(s), fl-cDNA clones are either manually checked, subjected to a further round of finishing by primer walking using Primer3 [1] in an automated fashion to design primers, or eliminated. For most of the eliminated clones, alternate full length clones are selected. By using our automated system we have currently been able to finish 8372 non-redundant fl-cDNA transcripts in a high throughput manner. All of these clones have been submitted to NCBI as part of Mammalian Gene Collection.

References:

Rozen, S. and Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000;132:365-86.

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

FF0088

Assembly and Finishing using 454 Sequence Data

Jim Knight

Roche Diagnostics / 454 Life Sciences, Branford, CT 06405

The characteristics of 454 sequence data and its use in de novo assembly projects (shorter reads, higher depth alignments, coverage across the genome without cloning gaps, lack of clones, different mix of single-ended and paired-end reads, different basecalling error model) impact the strategies and judgments that are made in the finishing process. This talk describes how 454 sequencing data and assembler results can be used for finishing, including the Q2/2007 release of integrated 454/Sanger assembly and full consed folder generation.

FF0110

The Illumina Genome Analyzer System: Cost-effective, High Throughput Genomics Using Solexa DNA Sequencing Technology

Gary P. Schroth

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

This talk will describe the Illumina Genome Analyzer System. Our platform uses Solexa sequencing technology to deliver the most cost-effective, next generation DNA sequencing results available today. We will show examples of how the Genome Analyzer System can be used for DNA sequencing, paired-end sequencing, re-sequencing, gene expression analysis, ChIP-SEQ, small RNA discovery, and full transcriptome analysis. The system routinely delivers in excess of 1 billion bases of high quality sequence information per run and can complete what heretofore were considered enormous DNA sequencing projects in a matter of days.

FF0099

Using a Gigabase of Short Reads: leveraging Paired End Reads and Color Base Encoding in the AB SOLiD TM Sequencing System in Finishing and SNP Detection

Fiona Hyland

Applied Biosystems, Foster City, CA 94404

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

Poster Presentations (odd #s, June 18th)

FF0003

Complete Genome Sequence of *Cyanothece* sp. ATCC 51142

Eric A. Welsh[1], Michelle Liberton[1], Jana Stöckel[1], Thomas Loh[1], Hanayo Sato[1], Jon M. Jacobs[2], Robert S. Fulton[3], Sandra W. Clifton[3], Aye Wollam[3], Richard K. Wilson[3], Richard D. Smith[2], Louis A. Sherman[4], Himadri B. Pakrasi[1]

[1]Department of Biology and [3]Genome Sequencing Center, Washington University, St. Louis, MO, [2]Pacific Northwest National Laboratory, [4]Purdue University, Dept Biological Sciences, West Lafayette, IN 47907

Cyanothece sp. ATCC 51142, isolated from the benthic coastal waters of Port Aransas, Texas, is a Gram-negative unicellular diazotrophic cyanobacterium capable of photoautotrophic as well as heterotrophic growth. In order to temporally separate the oxygen labile nitrogen fixation complexes from oxygenic photosynthesis, *Cyanothece* 51142 exhibits robust circadian rhythms in photosynthetic activity, nitrogenase activity, respiration, and the synthesis and degradation of glycogen and cyanophycin granules, which are stores of carbohydrates and nitrogen, respectively. *Cyanothece* 51142 is the first unicellular diazotrophic cyanobacterium to be fully sequenced, and provides a basis upon which the regulation of such circadian controlled metabolic processes and storage capabilities may be further investigated.

The genome was sequenced at the Washington University Genome Sequencing Center, and the finished assembly independently confirmed using an optical restriction map generated by OpGen, Inc. The 5.5 Mb genome consists of a 4.9 Mb circular chromosome, a 429 kb linear chromosome, and four plasmids ranging in size from 10 kb to 40 kb. Both the genome assembly and the optical map indicate that the 429 kb chromosome is linear. Linear elements have been found previously in other bacteria, as have multiple chromosomes and large plasmids. However, this is the first linear element to be discovered in a phototrophic bacterium. The mechanisms of maintenance and replication of the linear chromosome are unknown.

The majority of the genes with assigned/probable function, and nearly all of the structural RNAs are located on the circular chromosome. Several genes are found on both the circular and linear chromosomes, including a *coxABC* operon, and a cluster on the linear chromosome containing several genes related to glycolysis (*ppk, pyk, pgi, eno, ackA, glgP*). Additionally, within this cluster is the only L-lactate dehydrogenase found in the genome, which is required for the terminal step in lactate fermentation. Other genes unique to the linear chromosome include the integrase-recombinase protein XerC, an *xseA*/*xseB* operon, *metE*, and a *hicA*/*hicB* operon. Analysis of the multi-copy genes does not indicate any conserved synteny between the circular and linear chromosomes. Most of the remaining genes on the linear chromosome are either hypothetical, unknown, or of uncertain function.

A portion of the research was performed as part of an EMSL Scientific Grand Challenge project at the W. R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle. The sequencing was supported by a grant from the Danforth Foundation.

FF0009

A Novel Approach to High Throughput Microbial Genome Finishing: Incorporation of 454 Sequence Data for Gap Closure in Low Quality Sanger Data

Olga Chertkov, Avinash Kewalramani, Riley Arnaudville, Cliff Han and Thomas Brettin

Los Alamos National Laboratory, Los Alamos, New Mexico, USA

The microbial genome finishing effort at Los Alamos takes an assembly line approach to finishing. We have unfinished genomes which are comprised of sequencing reads generated on the Sanger sequencer but also consisting of gaps. Gap closure is achieved by picking primers belonging to a wide array of techniques, like primer walking to close captured gaps, pcr walks to close uncaptured gaps, transposon bombing to close gaps associated with repeats. However most recently with the advent of new sequencing technologies like the 454 it has become necessary to evaluate the sequencing data produced by these machines against Sanger reads. This evaluative study helps us in understanding how the 454 data can be used, if at all, to close the gaps in Sanger data. The coverage depth of Sanger data can be reduced to as low as 5X and this will give rise to lot of gaps. If we can incorporate 454 data and close these gaps, then it will lead to less primer picking for gap closure and hence lesser reactions to be performed in the lab, thus increasing our efficiency in finishing microbial genomes with high quality data from different sequencing technologies with lower costs. Though not completely rosy our approach still provides the community and us with an opportunity to understand the new sequencing technology data and incorporate it into our finishing processes.

FF0019

Automated Finishing System at JCVI

Hoda Khouri, Luke Tallon, Heather Forberger, Brent Bradley, and Dan Kosack

J. Craig Venter Institute, Rockville, MD, USA

In a continuing effort to further automate finishing, the 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, adds the correct primer to each sample, and performs sequencing reactions to generate the finishing reads. The finishing reads are detected by an automated tracking system and incorporated into the targeted contigs and scaffolds using an assembly stitching process. Results are reported automatically to finishing project managers.

In this study, we analyze genomes sequenced to 8X coverage with traditional Sanger methods and genomes sequenced to 5X Sanger plus 454 LS pyrosequencing. The reads were assembled into contigs using the Celera assembler and then further grouped into scaffolds. Intrascaffold gaps and low coverage regions were targeted using autoCloser. The resulting finishing reads were incorporated into the project upon reassembly. An average of 60% of targeted gaps and 70% of low coverage areas were resolved after a single iteration of autoCloser.

In this poster we report the autoCloser data, compare the genomes studied, and evaluate the cost and the efficiency of the system. AutoCloser proves useful for both complete or partial finishing of both Sanger only and hybrid (Sanger/454) sequenced genomes.

FF0023

Discovery and Resolution of Repetitive Regions in Microbial Genomes

Brad Toms, Yasmin Mohamoud, Diana Radune, Heather Forberger, and Nadia Fedorova

J Craig Venter Institute, Rockville, MD, USA

There are clear theoretical reasons and many well-documented examples which show that repetitive regions are essential for genome function. Repetitive signals are necessary to regulate expression of coding sequences and to organize functions essential for accurate genome replication. Additionally, repeat elements have evolutionarily significant role in genome architecture and reorganization. Studies done on prokaryote genomes that contain direct repeats suggest that recombination between direct repeats is a widely conserved mechanism to promote genome diversification. Due to the functional importance of these regions to the organism, the JCVI finishing group dedicates a lot of resources to accurately assembling of repetitive areas in the genome. During the microbial random shotgun process, thousands of random sequences are assembled into contigs using Celera Assembler program. The assembler considers the sequence similarities and the clone size constraints; however in many instances it has difficulty resolving large repeats. This may lead to gaps, misassembled contigs, and/or collapsed repeats, which leaves these regions to be resolved and finished manually. Even the novel Pyrosequencing technology using (GS 20) 454 machines does not help in the resolution of repetitive regions because the generated reads are too short. In this poster, we will discuss the process by which the JCVI microbial finishing group identifies, confirms, and sequences repeats. We will also display many interesting examples of large difficult repeats, which we have successfully resolved.

FF0027

Whole Genome Assembly Assessment and Pre-Finishing

Brown, Adam, Labutti, K, Young, S, Zimmer, A and FitzGerald, M

Broad Institute of MIT and Harvard, Cambridge, MA

While informatics groups have improved the overall quality of genome assemblies as well as provided several useful tools for judging the integrity of such assemblies, visual inspection of these large assemblies is still necessary. During the inspection process several aspects of the assembly and genome itself can be noted, aiding the finishing process. Additionally, software tools may also be suggested as a result of this review process. An ARACHNE assembled database of the pathogenic fungus *Coccidioides immitis* was recently analyzed before the finishing process was started in order to determine misassembled and low quality regions that were missed by software tools, in addition to the difficulty of remaining gaps. By determining the signatures of regions that are missed or not counted by current software tools, new tools or revisions of old tools can be made to more fully automate the assessment process, and therefore further simplifying the finishing process as a whole.

FF0031

The Finishing Laboratory at the Broad Institute

Daniel Bessette, Nicole Allen, Mostafa Benamara, Chelsea Dunbar, Xiaohong Liu, Tashi Lokyitsang, Rakela Lubonja, Charles Matthews, Anna Montmayeur, Tamrat Negash, Thu Nguyen, Michael FitzGerald

The Broad Institute, Cambridge, MA, USA

Using a combination of high-throughput automation and small-scale custom techniques, the Laboratory Finishing group at the Broad Institute improves rough draft genomes by providing new sequencing data to fill in gaps and resolve conflicts. We work in collaboration with the Computer Finishing and Production Sequencing teams to fulfill work orders that include:

- Primer Walking (Fosmid and Plasmid)
- Transposition (Fosmid and Plasmid)
- Resequencing (Fosmid and Plasmid)
- Custom PCR
- Shatter Library creation (PCR products)

Built into these general workflows are special chemistries to address difficult-to-sequence regions. Recently the Shatter Library workflow was revamped and now utilizes a topoisomerase-based cloning system for ease of use and cost reduction. This modest-sized group has the capacity to work on at least one mammal and several fungal- and bacterial-sized genomes simultaneously. The structure of the group also allows it to have significant flexibility and scale-up capability: our ability to address clone tiling path based projects (*e.g*. mouse), pure whole genome finishing *(e.g*. Mycobacterium tuberculosis) and a mix of the two (*e.g*. dog) gives us flexibility, our workflow (rather than project) based tasks allows for increased efficiency, personnel cross-training allows us to shuffle personnel to various workflows according to need, and a mix of automated and manual protocols as well as a seasoned LIMS enable us to tend to large or small orders very efficiently.

FF0033a

Bias Free Linear Vector for Cloning Recalcitrant DNA and Accelerating Sequence Finishing

Ronald Godiska[1], Rebecaa Hochstein[1], Sarah Vande Zande[1], Nikolai Ravin[2], Attila Karsi[3], David A. Mead[1]

1Lucigen Corporation, Middleton, WI 53562
2Centre Bioengineering, Russian Academy of Science, Moscow, Russia
3Mississippi State University

We have developed a novel linear vector for unbiased cloning of 10-30 kb inserts in *E. coli*. This vector, termed “pJAZZ”, shows unprecedented ability to maintain large inserts from very AT-rich genomes. The otherwise difficult to clone genome from *Flavobacterium columnare* (70% AT) was sequenced to seven fold coverage using the pJAZZ vector, resulting in less than 10 clone gaps left in this 3.2 Mb genome. The linear vector was able to maintain 20-30 kb fragments from *Lactobacillus helveticus* (65% AT) and 2-4 kb inserts from Piromyces species (up to 96% AT), which were unclonable in conventional plasmids. Unlike fosmid cloning, the construction of large-insert libraries (10-20 kb) in pJAZZ is simple and robust, using standard methods of transformation and plasmid purification. We are evaluating the use of a single pJAZZ shotgun library to eliminate the need for multiple libraries, making finishing easier and more cost effective. Enhanced stability of inserts in the pJAZZ vector is attributed to both the lack of supercoiling and the lack of transcriptional interference. Torsional strain inherent to supercoiled plasmids can induce localized melting and generate secondary structures, which are substrates for deletion or rearrangement by resolvases and replication enzymes. For example, the instability of tandem repeats and palindromic sequences is presumably due to cleavage of hairpin structures or to replication slippage across the secondary structures. Most conventional plasmid vectors also induce strong transcription and translation of inserted fragments, and they allow transcription from cloned promoters to interfere with plasmid stability. As a result, certain DNA sequences are deleterious or highly unstable, leading to sequence “stacking”, clone gaps, or a complete inability to construct libraries, especially from AT-rich genomes or toxic cDNAs. The transcription-free, linear pJAZZ vector also minimizes “sequence gaps” caused by secondary structures, as shown by its stable cloning of inverted repeats and di- and tri-nucleotide repeats.

FF0033b

Chimera Free Cloning of Single DNA Inserts Using “GC Cloning”

David Mead, Rebecaa Hochstein, Keynttisha Jefferson, Ronald Godiska, Spencer Hermanson

Lucigen Corp., Middleton, WI 53562

The efficiency of shotgun DNA sequencing depends to a great extent on the quality of the random libraries used. It is important to minimize chimeric inserts to facilitate accurate sequence assembly. Single-insert clones are usually ensured by ligating asymmetric linkers, typically with a BstXI recognition site, to the insert DNA. Subsequently, the excess linkers must be completely removed from the insert DNA before ligation to a vector containing complementary BstXI ends. We have developed a novel "linker free" cloning strategy to eliminate the need for linker addition and removal in constructing high-quality shotgun libraries. It is based on a new GC cloning technology and a unique DNA end blocking chemistry developed at Lucigen. The pSMARTGC vector contains a single 3’-C overhang, which is compatible with the single 3’-G overhang added to blunt ended DNA using PyroPhage DNA polymerase or many other non-proofreading polymerases. The unique combination of a C tailed vector and G tailed insert blocks the ligation of multiple fragments. This protocol is robust and showed five to ten-fold higher yields of clones compared to previous protocols, and is significantly faster than TA cloning. The level of chimerism is ~ 1% in the library, and the background of clones without an insert was <1%. Another important benefit is the ability to construct complex libraries using 10-100 ng insert without compromising the sequence coverage of level of empty background. The procedure is very rapid, as libraries were completely processed in a day. High copy, low copy, single copy and linear vector versions of the GC cloning vectors with chimera free capabilities have been constructed

FF0033c

Random-Shear BAC Library Construction and Efficient Genome Gap Closing

Chengcang Wu, Sarah Vande Zande, Rebecaa Hochstein, David Mead, Ronald Godiska

Lucigen Corporation, Middleton, WI 53562

Bacterial artificial chromosome (BAC) libraries are indispensable for physical mapping, positional cloning, genetic analysis, and sequencing of large genomes. BAC libraries have been created from many species, including Arabidopsis, Drosophila, rice, mouse, and human. A significant limitation of the current methods is the use of partial restriction digestion to generate genomic DNA fragments of 100-300 kb. The inherently skewed genome distribution of restriction sites causes at least 10-fold under- or over-representation of particular sequences, with some regions being entirely absent from the BAC libraries. Another drawback is the instability of inserts in current cloning vectors due to transcription and secondary structure formation. As a result, existing BAC libraries built with conventional methods and vectors are biased, and numerous gaps exist in all of the physical and sequencing maps of eukaryotic multi-cellulargenomes. To circumvent these problems we have successfully developed techniques to construct unbiased, randomly-sheared BAC libraries (>100 kb inserts). We have demonstrated that a single 5X random shear BAC library covers various genome regions uniformly and closes several gaps in the *Arabidopsis thaliana* genome. We believe it will be possible to finish the physical mapping and sequencing of Drosophila, Arabidopsis, rice, mouse, and human with this approach, closing all of the existing genomic gaps, including centromeres. We have also developed transcription-free BAC vectors. These vectors show much higher stability of inserts containing AT-rich sequences, direct and inverted repeats, and other deleterious DNAs. It is thus possible for the first time to construct unbiased BAC libraries to achieve complete closure of a large complex genome.

FF0035

454-Sanger Joint Assemblies in Mammalian BAC Pools

Christian J. Buhay, Michael E. Holder, Aniko Sabo, Xiang Qin, Huyen H. Dinh, Peter R. Blyth, Sandra L. Lee, Lynne V. Nazareth, Christie L. Kovar-Smith, Huaiyang Jiang, Erica Sodergren, Donna M. Muzny, George M. Weinstock and Richard A. Gibbs

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

There is an ongoing need to investigate time and cost effective methods of sequencing and finishing large-scale genome projects. A robust and versatile 454 sequencing pipeline has been established at the BCM-HGSC. Advancements include optimization of the BCM pipeline to achieve 100+ Mb per GS-FLX run, and the application of amplicon and paired-end sequencing methods supported by 454 Life Sciences. The 454 sequencing methods have been applied to numerous sequencing projects over the past year including 37 microbial genomes. We are developing strategies to integrate 454 assemblies of mammalian BAC pools with Sanger reads.

Microbial 454 sequencing and assemblies have resulted in assemblies comparable to, or better than that of Sanger sequencing, with N-50 contig sizes of 25kb. To evaluate how well the 454 sequencing technology would perform on mammalian genomes, we sequenced a pool of 100 macaque BAC clones using 454 technologies. Sequence data generated for the BAC pools were repeat masked and assembled using the 454 assembly tools. The assembled 454 contigs were then aligned to finished BAC sequence to assess contiguity and accuracy. We found that on average, 64% of finished BAC sequence was covered by 454 contigs in a 4x assembly; however, the N-50 contig length was low at roughly 3kb. Further investigations of strategies to combine whole genome shotgun Sanger reads with 454 data are in progress. Sanger 4x assemblies were created to simulate whole genome shotgun assemblies, and contigs from the 454 assembly were mapped to each Sanger assembly. We evaluated ways to combine Sanger reads and 454 contigs and the influence of progressively increasing the 454 coverage (4x – 8x – 12x) on the quality of the combined assembly. Initial results indicate that the 454 contigs are closing some gaps, with N50 contig lengths increasing to 12kb or better. Conventional finishing tools such as Autofinish can then be used to close the remaining gaps. These 454/Sanger finishing techniques have direct applications in microbial finishing as well as upgrading mammalian genome regions.

FF0047b

New Technology Strategies for Microbial Assembly and Finishing

Alla L.Lapidus, Eugene Goltsman, Stephan Trong, Kurt M. LaButti, Brian Foster, Ed Kirton, Feng Chen, Paul Richardson

DOE Joint Genome Institute, Walnut Creek, CA

FF0051

Finishing Large Tandem Repeats in the Zebrafish Genome

Alan Tracey

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

As part of ongoing involvement with large genomes the Wellcome Trust Sanger Institute (WTSI) usesD a BAC by BAC approach to finish clones to HTGS phase 3. The majority of clones are successfully resolved with our normal strategies including autoprefinishing, custom oligos and alternative chemistries.

Some of the clones which typically evade successful resolution by the normal methods include large tandem repeats of >10kb which can be very difficult to construct using automated assembly programs such as Phrap (Green) in isolation and may even be resistant to preliminary manual finishing efforts. Such clones may warrant specialist attention in order to resolve the assembly problems. The WTSI uses the specialist skills of the most experienced finishers to tackle these clones and assess the likelihood of resolution before committing a significant amount of resources.

In one of our current genomes, the Zebrafish, large tandem repeats occur in 3.3% of clones seen in finishing to date and due to time constraints we have found it necessary in some cases to employ alternative strategies to resolve some of these regions. Preanalysis is undertaken to select those repeats which warrant additional effort. Non-coding repeats are constructed to agree with the repeat pattern and restriction digest data.

FF0053

The PGP Viewer and its Application in Finishing the Zebrafish Genome

Kerry Giselle

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

The assembly and finishing of genomes by combining sequence generated from mapped clones and whole genome shotgun approaches has led to the development of bioinformatics tools at the Wellcome Trust Sanger Institute (WTSI) that allow alignment of multiple sets of sequence data to be viewed, together with clone map data. One of the recent developments, the PGP (Pseudo Golden Path) viewer is being used to assess the assemblies of zebrafish chromosomes as we proceed with the finishing of the *Danio rerio* genome. The reference genome sequence is currently 67% finished and we aim to reach “essential completion” of the genome by the end of 2008.

The PGP viewer is built on the ENSEMBL platform and shows alignments of the latest build of the reference sequence, which combines finished and unfinished clones sequences with whole genome shotgun contigs, individual clones or uploaded contigs, ESTs, cDNAs, BAC or fosmid end sequences and markers. The viewer has been used extensively to identify and resolve regions of the genome that are represented by multiple haplotypes.

The PGP viewer is very versatile and continues to evolve as the project moves forward. It is our aim that it will be extended and transferred within the WTSI to help with other sequencing projects in the future. As the PGP viewer is based on ENSEMBL it could be made available for use by other sequencing centres. We will describe examples of the utility of the PGP viewer for the genome finishing process.

FF0055

Microbial Finishing in the New 454/Sanger Hybrid Era

Stephanie Malfatti, Lisa Vergez, Maria Shin, Mari Christensen, Jeff Elliott, Dorothy Lang and Patrick Chain

Lawrence Livermore National Laboratory, Livermore, CA;
DOE Joint Genome Institute, Walnut Creek, CA

The Joint Genome Institute (JGI) has contributed ~180 finished microbial genomes to public databases. In order to accomplish this task, the JGI has in the past employed a whole genome shotgun strategy which consists of 3kb and 8kb plasmid libraries, along with a 40kb fosmid library. More recently, with the advent of new sequencing technologies such as Roche's 454 pyro-sequencing, we have had to adapt our finishing strategy to incorporate this novel data type and are anticipating other changes in the near future, as several other platforms mature, such as Illumina's Solexa Sequencing Technology and Applied Biosystem's SOLiD. With the newest genomes coming through the JGI pipeline, we have omitted the creation of 3kb libraries and have replaced it with 454 sequencing coverage. In the future, we anticipate this hybrid Sanger/454 strategy will evolve to 454 sequencing runs plus 3-5 fold coverage of only one (8kb or 40kb) Sanger sequencing library. Within the JGI-LLNL finishing pipeline, we currently have 15 Sanger/454 hybrid genome projects. We will present the advantages of this new strategy, as well as some of the main challenges that remain in finishing microbial genomes in the 454/Sanger hybrid era.

This work was performed under the auspices of the U.S. DOE by the University of California, LLNL under Contract No. W-7405-Eng-48.

FF0067

Finishing Sanger/454 Hybrid Sequenced Genomes at JCVI

Mary Kim, Nadia Fedorova, Yasmin Mohamoud, Daniela Puiu, Luke J. Tallon

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850

As new and promising genome-finishing techniques emerge on the scene, they must undergo thorough analysis to determine their full efficacy. At JCVI we strive to test and incorporate new methods with our own to achieve optimal results in finishing genomes. Currently we generate genome assembly sets from a combination of 454 sequence data and Sanger sequence data (i.e. hybrid sequenced genomes). These genomes proceed through the finishing process to close remaining gaps and resolve misassemblies, hard stops and low quality regions. Both generating a viable assembly set and manipulating the hybrid genome through our database and software tools present an array of challenges and difficulties. Furthermore, modification of our quality standards for finished hybrid sequenced genomes is a need that must be addressed. Closing such genomes will give us a better understanding of how to work in tandem with new approaches such as 454 technology, and thus help to advance genome-finishing techniques.

FF0069

Finishing at the Arizona Genomics Institute

Currie, J., Mueller, T., Yu, Y., Angelova, A., Rajasekar, S., Ko, A., Wing, R.

Plant Sciences Department, University of Arizona, Tucson, Arizona, USA

Attempts to make finishing an automated process have been unsuccessful in developing accurate final sequence. These failures have proven that the process of finishing is a valuable endeavor. A combination of automated and manual editing continues to be the most efficient and necessary method to confirm the assembly of the shotgun sequence. For most projects, standard finishing protocols will allow the Finisher to achieve contiguity, delineate insert/vector junctions, resolve sequence discrepancies and raise each base to a standard of quality designated by the scientific community. There are a few instances where more intensive procedures are required to meet these standards because of secondary structure and repetitive sequence. Through the use of sorting, large insert libraries, and G/C homopolymeric oligos, we have been successful in resolving difficult sequence regions. In areas where the number or size of repeats makes sorting misassemblies nearly impossible, construction of a large insert library may be the best choice. Using the pCNS vector and BstXI adaptor we have been successful in making 10kb libraries and utilizing *in vitro* transposons to fill large gaps and confirm assemblies. Sequence with mononucleotide runs of Gs or Cs are extremely difficult to resolve with current techniques. We have had success sequencing through these areas with the use of a specialized oligo. The oligo is a heptamer or tetradecamer of Cs. This homopolymeric oligo blocks the template from re-annealing allowing the polymerase to pass through this usually difficult if not impossible sequencing region.

FF0077

A Pipeline for Viral Genome Closure

Jennifer M. Zaborsky, Vik Subbu, Jeffrey D. Sparenborg, Naomi Sangamalay, Torrey L. Gallagher, Larry J. Overton, Xinyue Liu, Jeff Sitz, Kristine Jones, Luke J. Tallon, David Spiro

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850

The Viral Genomics and Closure group at JCVI have developed a high throughput pipeline to sequence the influenza A genome. The extensive genetic variation of influenza and the large number of viral samples sequenced has led the Closure team to develop new finishing techniques and software to match the unique needs of this pipeline. Closure uses a software suite, Elvira, to build assemblies from amplicon-based PCR reads. Assemblies are tracked through the pipeline in the Closure Task Manager where laboratory procedures are assigned until the assembly is finished. A combination of standard and custom primers is used in the finishing process to handle the highly varied influenza genome and maintain the efficiency of the pipeline. To date, JCVI's pipeline has published over 2000 complete Influenza genomes in Genbank. The pipeline has recently been adapted to sequence and finish influenza B, coronavirus, and rhinovirus samples. The evolution of the pipeline reflects the growth of interest in viral genomics and the need for finishing techniques that efficiently close highly varied genomes.

FF0081

Maize Genome Sequence Improvement

Laura Courtney

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

The Maize Genome Sequencing Consortium, led by the Genome Sequencing Center at Washington University School of Medicine (GSC) in collaboration with Arizona Genome Institute (AGI), Cold Spring Harbor Laboratory (CSHL) , and Iowa State University (ISU) and funded by a three year grant from NSF, undertook improving the B73 maize genome. Sequencing and improvement efforts on the maize genome provided particular challenges due to the fact that the genome is comprised of 70% repeat sequence. These challenges spurred the development of an array of new programs and techniques in an effort to increase speed and efficiency, while reducing overall costs. The consortium employed a BAC based approach that utilizes a relatively low coverage of approximately 4-6X. An average draft assembly contains about 26 contigs and multiple unoriented gaps. To aid in the order and orientation of the draft assembly, programs are used to integrate paired end fosmid reads. Projects are screened against a repeat database using a version of Vmatch, a program used to solve large-scale sequence matching tasks. Any areas found to match known repeat sequence in the genome are not targeted for improvement efforts. The first stage of improvement efforts, involve the use of the program autofinish (Gordon, D. University of Washington) to target gaps and low quality areas within the unique regions. Following the automated improvement efforts, high-Cot and methyl-filtration reads along with mRNA data are incorporated into the assembly both to provide scaffolding information as well as support existing BAC data. The second stage of improvement involves manual resolution of misassemblies as well as a limited amount of directed sequencing attempts targeting the remaining oriented gaps and low quality areas within the unique regions. Direct BAC sequencing has been somewhat successful in addressing the remaining unoriented gaps containing at least some unique sequence. Quality assurance checks are made to insure the integrity of the assembly and sequence in accordance with the standards set by the Maize Consortium. Upon completion of the above-mentioned tasks, the sequence is submitted to Genebank as phaseI – HTGS improved.

FF0089

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

With the advent of low-coverage sequencing and exotic sequencing technologies it is becoming increasingly necessary to use all available genome mapping data, including optical, physical and genetic maps, during the assembly and finishing phases of a sequencing project. The CMap Assembly Editor (CMAE), a desktop application, is being developed to assist in visualizing and editing large scale sequence assemblies for the maize sequencing project. Using the CMap comparative mapping database, CMAE allows sequence assemblies to be superimposed on top of diverse other types of mapping data,, allowing the finisher to view assemblies in the context of a cascade of mapping data at a variety of resolutions. For example, the editor can show sequence contigs aligned to fingerprinted physical map contigs, which are aligned in turn to genetic maps. Correspondence links between the different objects indicate the quality of the assembly and highlight possible mis-assemblies. The editor will then allow mis-assembled contigs to be split, merged or moved, or the troubled contigs can be exported to a more specialized program.

FF0113

The Azotobacter Vinelandii Sequencing Project

Jing Lu[1,] Phil Latreille[1], Nancy Miller[1], Brad Goodner[2], Dennis Dean[3], Derek Wood[4], Steve Slater[5], and Barry Goldman[1]

[1]Monsanto Company, [2]Hiram College, [3]Virginia Tech, [4]Arizona State University, [5]Seattle Pacific University

Azotobacter vinelandii is an aerobic, free-living, nitrogen-fixing bacterium and a member of gamma proteobacteria that contains many genes associated with energy consumption, nitrogen metabolism, and carbon sequestration. Unlike most diazotrophic bacteria, *Azotobacter* can fix nitrogen when grown in atmospheric oxygen (20%). Like most eubacteria, *A. vinelandii* contains a single circular chromosome, however, the copy number of this chromosome is dramatically variable. During exponential growth phase, the number of chromosomes per cell is low, however, when cultures reach stationary phase, the number of chromosomes can increase to 50-100 per cell. The genomic sequence of *A. vinelandii* will help researchers dissect the genetic and biological basis for these remarkable capabilities.

The Joint Genome Institute (JGI) originally sequenced the genome to 8X coverage in 2002 using one fosmid and two plasmid libraries. The 50-contig assembly has been available to the public through the JGI and NCBI websites. To finish the genome, we combined traditional finishing technologies with the optical mapping technology available from Opgen, Inc. Optical mapping provided a sequence-independent methodology to resolve misassembles and aided in overall completion of the project. Using these we brought the assembly to five contigs and one scaffold without sequencing a single read.

Meet and Greet Party

630pm – 900pm, June 18th

Sponsored by Roche Diagnostics

Enjoy!!!

06/19/2007 - Tuesday				
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 - Informatics	Jim Bristow
845 - 930	**Keynote**	FF0011	Automating the finishing process: dreams and realities	Mihai Pop
930 - 950	Speaker 1	FF0018	Celera Assembler: Adapting for the Future	Granger Sutton
950 - 1010	Speaker 2	FF0115	Charting and Sequencing Structural Variation using High-Resolution Paired-End Mapping (HR-PEM)	Jan Korbel
1010 -1040	Break	x	Beverages and snacks provided	x
1040 -1100	Speaker 3	FF0005	Assessment of 454 Sequencing Errors in Microbial Genomes	Stephan Trong
1100 -1120	Speaker 4	FF0026	New Sequencing Technologies and Hybrid Assemblies - A discussion on a shift in finishing paradigm: do we need to analyze each read?	Harindra Arachchi
1120 - 1220	Panel Discussion	x	**Panel Discussion**	Chair - Patrick Chain
1230 - 145pm	Lunch	x	**La Fonda Lunch Buffet** (Pork tenderloin achiote-rubbed and char-grilled with tomatillo-chipotle sauce, Breast of chicken filled with bacon, red onions, green chile, jack and cheddar cheeses, lightly-breaded, flash-fried and baked, accompanied by mild green chile cream sauce, etc.)	x
200-220	Speaker 5	FF0090	Transforming Genomes with New Sequencing Technology	Donna Muzny
220-240	Speaker 6	FF0039	*De novo* Hybrid 454 / Sanger Genome Assembly of *Phytophthora capsici*	Joann Mudge
240 - 300	Speaker 7	FF0032	Incorporating New Sequencing Technologies into Finishing Strategy	Sean Sykes
300 - 400	Panel Discussion	x	**Panel Discussion**	Chair - Donna Muzny
400 - 430	Break	x	Beverages, Wine & Cheese provided - sponsored by IDT, Edge, & Invitrogen	x
430 - 600	**Posters - even #s**	x	**Poster Session with Wine & Cheese - sponsored by IDT, Edge, & Invitrogen**	x
600 - bedtime	on your own	x	Dinner and night on your own - enjoy	x

Speaker Presentations (June 19th)

Abstracts are in order of presentation according to Agenda

FF0011 – **Keynote**

Automating the Finishing Process: Dreams and Realities

Mihai Pop

University of Maryland, Computer Science Department and Center for
Bioinformatics and Computational Biology, College Park, MD 20742

FF0018

Celera Assembler: Adapting for the Future

Granger Sutton[1] (presenting), Brian Walenz[1], Art Delcher[2], Eli Venter[1], Aaron Halpern[1], Gennady Denisov[1], Jason Miller[1], Justin Johnson[1]

1. J. Craig Venter Institute, Rockville, MD
2. Center for Bioinformatics and Computational Biology, University of Maryland, College Park

The Celera Assembler is an open source (https://sourceforge.net/projects/wgs-assembler/) shotgun fragment assembly software package. The Celera Assembler is under active development and is in production use for large scale genome assembly. A current focus of our work is to effectively incorporate 454 sequencing reads in conjunction with Sanger sequencing reads into the Celera Assembler. Modifications are necessary for two reasons: deeper sequence coverage with shorter reads, and a distinctly different error model for homopolymer runs for 454 sequencing. The deeper coverage with shorter reads decreases the spacing on the target sequence between the starts of shotgun fragments impacting accurate detection of branching in the overlap graph. The homopolymer error model impacts multiple levels: initial shared k-mer seed detection for overlapping, overlap detection, fragment error correction, allele separation, and consensus generation. We will present algorithmic modifications to address all of these hurdles. We will present known problems with the Celera Assembler, possible current work arounds, and our proposed solutions. Finally, we will discuss future enhancements to directly support rounds of directed finishing.

FF0115

Charting and Sequencing Structural Variation using High-Resolution Paired-End Mapping (HR-PEM)

Korbel, Jan; Urban, Alexander Eckehart; Affourtit, Jason; Grubert, Fabian; Kim, Philip; Du, Lei; Carriero, Nicholas; Godwin, Brian; Turcotte, Cynthia; He, Wen; Taillon, Bruce; Simons, Jan; Kidd, Kenneth, Carter, Nigel; Hurles, Matthew; Weissman, Sherman; Harkins, Tim; Gerstein, Mark; Egholm, Michael; Snyder, Michael

Yale University, New Haven, CT; 454 Life Sciences, Branford, CT; The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; Roche Applied Science, Indianapolis, IN

Structural variants (SV), i.e. deletions, duplications, insertions, and inversions involving kilo- to Megabases of genomic DNA, were recently suggested to be responsible for a considerable amount of phenotype variation, and possibly, disease in humans (1-6). However, to date, most methods for identifying structural variants have resolutions in the order of 50–75 kb (7), and thus do not precisely identify the boundary sequences (i.e. *breakpoints*) of SVs. Furthermore, the majority of approaches used so far for cataloging SVs in the human genome do not detect copy-number neutral variation events such as inversions and balanced translocations.

We present a novel approach, High-Resolution Paired-End Mapping (HR-PEM), which makes use of 454/Roche sequencing technology, and combines computational analysis, high-throughput PCR assays, and amplicon-cocktail-sequencing to rapidly identify SVs at high resolution, and subsequently sequence across the breakpoints associated with these variants. The approach involves sequencing the ends of circularized 3 kb genomic fragments and mapping them onto the human genome reference sequence. The resolution of breakpoint assignments is ≤3 kb and thus well-suited for PCR validation. We have used HR-PEM to map and sequence SVs – i.e. simple deletions, insertions, and inversions, as well as more complex structural rearrangements – in two individuals in order to generate a precise map of SVs and their associated breakpoints. From 21 million and 10 million paired-end sequences, respectively, from each individual, several hundred SVs have been predicted so far, ranging from 2 kb to several Mb in size. A first pass PCR analysis indicates that at least 60% of the predicted SVs can be amplified in a single PCR band and analyzed using DNA sequencing. Our results reveal as yet unexplored aspects of structural variation in the human genome, and suggest mechanisms by which this layer of genomic variation has arisen.

References

1. Sebat *et al.* (2004) *Science* **305,** 525-8.
2. Iafrate *et al. Nat Genet* **36,** 949-51.
3. Tuzun *et al. Nat Genet* **37,** 727-32.
4. Redon *et al.* (2006) *Nature* **444,** 444-54.
5. Gonzalez *et al.* (2005) *Science* **307,** 1434-40.
6. Stranger *et al.* (2007) *Science* **315,** 848-53.
7. Coe *et al.* (2007) *Genomics* **89,** 647-53.

FF0005

Assessment of 454 Sequencing Errors in Microbial Genomes

Stephan Trong, Patrick Chain, Ed Kirton, Eugene Goltsman, Brian Foster, Paul Richardson and Alla Lapidus

U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA

The introduction of pyrosequencing-based sequencing platforms such as 454-sequencing along with the Sanger technology for whole-genome shotgun sequencing has provided an alternative way to produce cost-effective sequence data. For microbial genome finishing, the 454/Sanger hybrid approach has made a considerable impact in reducing the time required to close a genome. However, errors within the 454 sequencing data remain to be addressed to make sure that final consensus represents high quality sequence.

JGI is working on further cost reduction of sequencing and finishing processes and is planning significant cut of Sanger sequencing. This will increase the portion of the genome covered with pyrosequence only. Thus the analysis of errors produced by this new technology becomes very important.

We conducted a study to assess the quality of the 454 sequence data in order to determine its impact on error rate, and to devise a strategy to correct or reduce the errors. The study analyzed 29 microbial genomes containing both Sanger and 454 only assemblies to look for mismatches in the 454 sequence data. We examined both homopolymer tracts and non-homopolymer regions with respect to read depth and quality assignment. An overview of our findings and strategy for reducing the 454 errors will be presented.

FF0026

New Sequencing Technologies and Hybrid Assemblies – A Discussion on a Shift in Finishing Paradigm: do we need to analyze each read?

Harindra M. Arachchi, Manuel Garber, Chad Nusbaum, Mike Zody, Sarah Young, Michael FitzGerald.

Broad Institute of MIT and Harvard, Cambridge, MA

New sequencing technologies such as 454 and Solexa have ushered in a new era in genome technology, while creating a myriad of questions on their usage and analysis. A shared aspect of the new technologies is the ability to bypass bacterial cloning procedures. This could be used to sequence across previously uncaptured material.

In order to circumvent suspected cloning bias we applied a PCR approach to a human chromosome 15 gap. The amplicon was sequenced with both standard Sanger and 454 technology. Data were assembled and analyzed with a variety of methods, including use of Arachne (Broad) and Newbler (454) assemblers. We will present the results of this analysis and information supporting our current process. Our assembly begins with Sanger data before independently assembled Newbler contigs and reads are joined in. No individual 454 trace data are reviewed in Gap 4. This is driven by the volume of trace data, typically 20X, and to the specific nature of pyrosequencing data.

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

FF0090

Transforming Genomes with New Sequencing Technology

Donna M. Muzny, Xiang Qin, Christian Buhay, Mike Holder, Aniko Sabo, Huaiyang Jiang, Shannon Dugan-Rocha, Yan Ding, Huyen H. Dinh, Christie L. Kovar-Smith, Sandra L. Lee, Lynne V. Nazareth, Erica Sodergren, Kim Worley, Sarah K. Highlander, George M. Weinstock and Richard A. Gibbs

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

The advent of new DNA sequencing platforms offers cost-effective and technically manageable opportunities for finishing and genome refinement. Challenges of using these new systems include optimizing the use of shorter read lengths with different error models. Applying these techniques to sequencing of microbial genomes is now well established, but their use for more complex genomes requires development. To date, we have performed 200 runs generating over 8.5Gb of sequence using the GS-20 and GS-FLX sequencing instruments. We have sequenced over fifty different prokaryotic genomes ranging in size from 1.1Kb to 12 Kb. When FLX data was used (>200bp), contig length N50 increased e.g. from 29kb to 108kb for *Staphylococcus aureus*. In other cases, paired end data dramatically improved contiguity – for example in *Enterococcus faecalis* scaffold number was reduced from 128 to 8, even less than with Sanger data. Of these genomes, *Staphylococcus aureus USA-300* has been finished and five additional genomes (*Pantoea stewartii, Bacillus pumilus, Treponema paraluiscuniculi*, *Moraxella bovis* and *Enterococcus faecalis OG1RF*) are now in the finishing pipeline with variable Sanger sequence coverage.

Development of techniques and protocols for finishing these mixed 454/Sanger assemblies include direct genomic sequencing, multiplex PCR and long range PCR. Informatics tools were also developed to utilize the 454 paired end reads within CONSED and Autofinish parameters were adjusted for primer appropriate primer coverage.

For BAC-based genomes, pooling of BACs for 454 sequencing allows targeted shotgun sequencing of the regions of the genome represented by the BAC clones. We have performed 454 sequencing of BACs from human, rat, mouse, macaque, and bovine genomes. Pools containing from 9 to 151 mammalian BACs have been successfully sequenced and assembled. Reads from both GS20 and FLX machines yield only short contigs (<5kb), bounded by Alus and other repeats, regardless of the depth of coverage. However, addition of paired ends yielded scaffolds with N50s of 35 kb (GS20) and 54 kb (FLX). Assembly quality is high, with an overall error of 5x10-4, with indels representing about 80% of the errors. Strategies to combine whole genome shotgun Sanger reads with 454 data are in progress. Here 4X Sanger read coverage was combined with increasing 454 read coverage (4-12X) of the BAC pools. Initial results indicate that the 454 contigs are closing gaps, with N50 contig lengths of 12-31Kb. Current efforts in mixed assemblies involve utilizing the 454 mapping software for correct placement of 454 pooled BAC reads before assembly with Newbler.

FF0039

De novo Hybrid 454/Sanger Genome Assembly of *Phytophthora capsici*

Joann Mudge[1], Stephen F. Kingsmore[1], Neil Miller[1], Sophien Kamoun[2], Kurt H. Lamour[3], Paul M. Richardson[4], Darren Platt[4], Igor Grigoriev[4], Alan Kuo[4], Greg D. May[1], William D. Beavis[1]

[1]National Center for Genome Resources, Santa Fe, New Mexico, USA Department of Plant Pathology, [2]Ohio State University, Wooster, Ohio, USA
[3]The University of Tennessee, Department of Entomology and Plant Pathology, Knoxville, Tennessee, [4]USA Department of Energy Joint Genome Institute, Walnut Creek, California, 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 looked at the feasibility of using a hybrid sequencing approach for *de novo* eukaryotic genome sequencing by creating a draft genome sequence for *Pythophthora capsici,* 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. With funding from DOE-CSP, USDA, and NSF, we generated 23X 454 GS20 pyrosequencing singleton reads, 5X Sanger paired reads, and 2M 454 GS20 paired reads. 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. The assembly results will be discussed. In addition, I will demonstrate the use of the Alpheus^{TM} software system for compiling a searchable SNP and indel database of *P. capsici*.

FF0032

Incorporating New Sequencing Technologies into Finishing Strategy

Sean M. Sykes, Sarah Young, Chandri Yandava, Chinnappa Kodira, David Jaffe, Bruce Birren and Chad Nusbaum

The *Neurospora crassa* genome contains numerous regions of high AT content that are refractory to cloning thus difficult to finish by traditional clone-based methods. We have analyzed the utility of 454 sequencing to capture these unclonable regions.

We generated ~25-fold coverage of the *Neurospora* genome in 454 data, and assembled the data using 454's assembler. Alignment of the 454 assembly to the finished contigs revealed a significant amount of new sequence. Specifically, 1.5 Mb of the approximately 40 Mb genome were represented in the 454 assembly but not in the assembly of clone-based shotgun data. The base composition of this new sequence was 26% GC, whereas the rest of the genome had a GC content of 49%. 454 data spanned 36 gaps between scaffolds, and extended and additional 200 ends of scaffolds. In the regions covered by the 454 assembly only, we found very low coverage (<0.5X) in ABI reads.

We have also used 454 to sequence the genomes of 18 isolates of the bacterium *Listeria monocytogenes*. We had previously observed that as much of 25% of these genomes was not obtained from shotgun clone libraries, making *Listeria* an excellent candidate for a 454 approach. Indeed, 454 sequence assemblies of *Listeria* showed near complete coverage of the genomes, and dramatically larger contigs. The unclonable regions in *Listeria* differ from those of *Neurospora* in that they are not related to GC content.

Annotation of these *Listeria* genomes revealed a number of genes with apparent frame shifts associated insertion/deletion errors in the consensus. The rate of apparent frame shifts was substantially higher in assemblies with less than 20-fold coverage. However, even for assemblies with high 454 coverage, 2-15% of gene models were disrupted by indels. These indels correspond to regions containing mononucleotide runs.

We are examining ways to combine a variety of different data types to most efficiently achieve an accurate finished product.

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

Wine & Cheese Poster Session

400pm – 600pm, June 19th

Sponsored by IDT, Edge, & Invitrogen

Enjoy!!!

Innovation & Precision
In Nucleic Acid Synthesis
IDT®
INTEGRATED DNA TECHNOLOGIES

Edge BioSystems
get the edge®

invitrogen™

Poster Presentations (even #s, June 19th)

FF0004

Finishing by Comparative Genome Comparisons

Haibao Tang, Alex Feltus, Andrew H. Paterson
Department of Plant Biology, University of Georgia

As more genomes are being sequenced at major sequencing centers, there are cases where closely related genomes are sequenced one after another, e.g. human and chimpanzee, *Arabidopsis thaliana* and *Arabidopsis lyrata,* etc. In those cases, the later-sequenced genome can utilize the closely related yet more "finished" genome to help in various stages of sequencing, including map construction, reads assembly and finishing. Using a reference genome can reduce coverage requirement to finish a genome and as a result, reduce cost associated with sequencing reactions and computer hardware. However, it is non-trivial to show that reference genome structure can be used in an un-biased way. We describe here a novel algorithm that automatically picks finishing primers by comparing to reference genome and scan the assembly to enforce finishing standard. We have applied this algorithm in finishing a large segment (>1Mb) of *Sorghum propiquum* genome by referencing a close relative, *Sorghum bicolor*. The software "*cmp_finish*" is written in C++ and available online.

FF0006

Use of Near-Neighbor PCR to Close Scaffold Gaps in Microbial Genomes

A. Christine Munk, Yan Xu, Avinash Kewalramani, Riley Arnaudville, Roxanne Tapia, Thomas S. Brettin, Cliff S. Han

Los Alamos National Laboratory (JGI), Los Alamos, NM 87545

One of the challenges of finishing microbial genomes is large numbers of uncloned regions (scaffold gaps). Gaps with no clone links require many expensive pcr reactions to connect scaffolds. Ordering and orienting scaffolds with no clone links is difficult unless a closely-related genome is available to identify possible links. This abstract describes near-neighbor PCR (nnPCR), a method of closing these 'scaffold' gaps with a minimum of pcr reactions using closely related finished genomes.

In the manual process of near-neighbor PCR, scaffold ends are identified visually using the Consed program's Assembly view and map of scaffolds. Blast is performed using as query a fasta file containing alternatively 1) the sequence of all unordered contigs <2kb and >10 reads, or 2) the final 10kb of each scaffold. A closely-related-genome downloaded from Genbank is used as subject. The blast output is parsed and a tab-delimited file is produced which can be opened as a spreadsheet. Links between contigs are identified and pcr primers are chosen and paired according to the observed links. PCR is performed and products are end-sequenced and assembled to close scaffold gaps. If PCR products are >1500 bp, they are shattered, subcloned and subclones are end-sequenced and assembled to make a consensus sequence to close scaffold gaps.

In one microbial finishing project with 14 uncaptured gaps considered, 9 out of 14 pcr reactions were successful, and 9 gaps were closed. In another project with 23 uncaptured gaps, 8 out of 19 pcr reactions were successful and 5 gaps were closed. This can be compared to combinatorial PCR, which would require 378 and 703 PCR reactions respectively. Software to automate and improve this procedure has been developed and is currently being tested. This nnPCR software uses the Consed autoreport function to generate a file which contains a map of contigs in scaffolds, and creates a fasta file for the blast query with sequence from the contigs at scaffold ends. nnPCR uses megablast to search all finished genomes currently in Genbank for hits within the same organism within a 20 kb range. The software chooses primers and pairs them, and submits a work order directly to the lab. For the project with 14 uncaptured gaps considered, 12 out of 14 pcr reactions chosen by the nnPCR software were successful. For the project with 23 uncaptured gaps, 3 of 6 pcr reactions were successful. Several microbial genomes have been significantly improved using near-neighbor PCR. This method has significantly reduced the number of PCR reactions that would be required if combinatorial PCR were necessary. Possible improvements to the software are being tested and results will be included in the poster.

FF0008

Comparative Sequencing of Eukaryote Genomes: Experiences with Three Leishmania Genomes

David Harris, Kathy Seeger, Lee Murphy, Chris Peacock and the Pathogen Sequencing Unit

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

We have sequenced the genomes of two Leishmania species (*Leishmania infantum* and *L. braziliensis*) for comparison with the high quality finished sequence of *L. major*. This was done by selecting parts of the process we use to finish eukaryote sequences with the aim of producing useful data in a highly cost effective way. For each genome we produced shotgun sequence to 5-fold coverage and then performed one round of automated pre-finishing to close gaps. The resulting contigs were matched against *L. major* chromosome sequences and were assigned to assembly splits on the basis of their similarity. This gave manageable sets of contigs that were further improved by manual correction of mis-assemblies and by finishing selected regions. These projects were successful in producing useful comparative data and gave us experience in genome sequence improvement. A number of problems were identified. For the finisher, these included difficulties with working on low coverage shotguns (much of the software had been developed for high coverage data) and the initial lack of clearly defined criteria for the final data quality. Annotators and finishers worked closely to define the project's objectives and to decide what was possible with the lower than usual data coverage. The analysis of the final fragmented data also proved to be difficult and we had problems in describing in measurable, objective terms the final state of the projects when we came to write up the comparisons. A description of these projects and how we intend to perform similar ones in the future will be presented.

FF0010

Life-cycle of a JGI Microbial Finishing Project – LANL

David Sims, Olga Chertkov, Chris Munk, Hajnalka Kiss, Liz Saunders, Sue Thompson, Linda Meincke and Cliff Han

Los Alamos National Laboratory (JGI), Los Alamos, NM 87545

The Joint Genome Institute (JGI) provides a network of services for the finishing of whole genome sequences for microbes. The customer for these services is a collaborator who submits a proposal and genomic DNA to the JGI for sequencing. JGI-Walnut Creek prepares Sanger and 454 Libraries creating a draft shotgun sequence of the genome. The draft sequence is annotated at JGI-Oak Ridge and assigned for finishing at one of three JGI finishing teams: Walnut Creek, Laurence Livermore National Laboartory or Los Alamos National Laboratory (LANL).

Each finishing team approaches the finishing process utilizing a variety of strategies.
JGI-LANL uses computer programs (greca and dupFinisher) lab techniques (transposon bombs) to resolve duplications within the genome. Primer walks on Sanger libraries and PCR products are utilized to improve coverage and quality of reads, while specialized chemistries are used to resolve hard G/C stops. To date, over 100 complete genomes have been sequenced at JGI-LANL.

When a genome sequence is completed, a quality validation step is conducted at Stanford University. Then JGI-Oak Ridge annotates the finished genome. JGI-Walnut Creek verifies the annotation and provides the collaborator with the complete sequence with annotation. After a period of time when the collaborator has exclusive access to the annotated genome, it is made public through Genbank and is available to the public at NCBI. The collaborator and/or JGI staff then write publications based on the data generated through this process, thus completing the “life-cycle” from genomic DNA to published research.

LA-UR 07-1235

FF0014

Recent Developments in Finishing Strategies for BAC-based and Medical Sequencing

J. Gupta[1], B. Barnabas[1], S. Y. Brooks[1], A. Young[1], N. F. Hansen[2], NISC Comparative Sequencing Program[1,2], G. G. Bouffard[1,2], E. D. Green[1,2], and R. W. Blakesley[1,2]

[1]NIH Intramural Sequencing Center (NISC) and [2]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

We continue to refine our sequence-finishing procedures in order to improve efficiency and expand capabilities. First, we investigated potential modifications of sequence-assembly programs to reduce the amount of manual correction required with initial Phrap assemblies of BAC sequences. In an examination of 102 BAC sequences with misassemblies, alternative assembly-program routines completely resolved the problems in 29 BACs, while another 38 BAC sequences showed a reduction in the severity of the misassembly. Investigation of alternative assembly routines is now being applied to any Phrap-based assembly showing misassembled sequences. Second, we have found that the use of a plasmid copy-control strain of *E. coli* improves the uniformity of sequence-read distribution across assembled BAC sequences. Furthermore, the frequency of uncaptured sequencing gaps due to 'unclonable' DNA fragments is dramatically reduced when such a strain is used. This has proven especially helpful for sequencing BACs that yield assemblies with higher numbers of gaps, such as those derived from platypus, owl monkey and hedgehog. Finally, as our scientific portfolio expands to include major initiatives in medical sequencing, we are refining our capabilities in sequence finishing. Towards that end, we are adapting our traditional tools to improve the sequence data being generated in this context.

FF0020

Assembling and Analysing a Bacterial Genome Using a Combination of Sanger Capillary Data and 454 Shotgun Data

Mandy Sanders and the Pathogen Sequencing Unit

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

Streptococcus suis SC84 is the first bacterial genome at the Wellcome Trust Sanger Institute to be co-assembled using Sanger capillary shotgun data and 454 shotgun data. We have previously finished the sequence of *Streptococcus suis* P1/7 using fluorescent di-deoxy DNA sequencing and subsequently used *Streptococcus suis* P1/7 as a model to test new sequencing technologies. Streptococcal genomes are a favoured organism to test new technologies on due to a severe cloning bias in conventional fluorescent di-deoxy DNA sequencing. The shotgun of *Streptococcus suis* SC84 has been achieved using 19 x 454 shotgun coverage combined with 5 x shotgun Sanger capillary data. After repeat tagging and contig ordering, using *Streptococcus suis* P1/7 as a reference strain, two different assemblies have been compared. We will report on differences found in assembling the 2 different shotgun sequencing technologies based on which technology is used as the backbone for the assembly. We will also describe how we aim to contiguate the genome and analyse the differences in sequence between *Streptococcus suis* SC84 and the reference strain.

FF0024

Challenges in Entire Genome Finishing: TB Haarlem

Priest, M, Shea, T, Labutti, K, Young, S, Zimmer, A, FitzGerald, M.

Broad Institute of MIT and Harvard, Cambridge, MA

Since the completion of the Human and Mouse genomes our finishing efforts have shifted from BAC tiling to whole genome shotgun assemblies, focusing less on mammals and more on pathogens. We are finishing multiple isolates of pathogenic bacteria as part of an effort to further understand issues like pathogenicity and drug resistance. The resurgent pathogen Mycobacterium tuberculosis is part of this program. There are three currently finished M. tuberculosis genome sequences available, H37Rv (lab strain, Sanger), CDC1551 (TIGR) and F11 (Broad). The latest strain the Broad Institute is finishing is TB Haarlem. Here we will report on efforts to finish Haarlem (source material?) and compare that to the process and results for F11. Issues including effect of shotgun quality and necessity of new informatic tools will be reviewed.

FF0028

Use of Multi-BAC Assemblies to Improve the Dog Genome

Abouelleil, A., Aftuck, L., Arachchi, H., Berlin, A., Brown, A., Fitzgerald, M., Gearin, G., Johnson, J., Lui, A., Macdonald, P., Pirun, M., Priest, M., Shea, T., Zimmer, A.

Broad Institute of MIT and Harvard, Cambridge, MA

The dog genome is important for comparative analysis of mammalian genome biology and evolution. CanFam 2.0, the current dog draft assembly, covers approximately 99% of the euchromatic portion of the genome. Through the use of targeted finishing, the accuracy and integrity of this draft assembly can be significantly improved in an efficient and cost effective manner.

The use of multi-BAC contigs is central to the aforementioned improvement process. BACs and genome chunks (sections of WGA pulled for direct finishing work) targeting large gaps, ENCODE regions, and uncertified regions (areas of questionable integrity) are finished. Overlapping BACs and chunks are then identified as work regions, which are in turn integrated into multi-BAC assemblies. This important step minimizes the draft finished sequence interfaces during finished data integration into the genome assembly. These assemblies undergo a quality control step and are then patched into the genome assembly. Finally, the chromosome that is patched undergoes another QC analysis. Through this process, a near-finished quality genome can be produced without the expense of generating a shotgun BAC library for the entire genome. We will describe our process and results from generating these contigs.

FF0030

Primer Walking in the Finishing Lab at the Broad Institute

Xiaohong Liu, Mostafa Benamara, Tashi Lokyitsang, Charles Matthews, Tamrat Negash, Thu Nguyen, Nicole Allen, Daniel Bessette, Michael FitzGerald

Broad Institute of Harvard & MIT, Cambridge, MA 02141

Primer walking at the broad is an effective strategy to sequence gaps or resolve small regions of difficult sequence. The DNA of interest may be a plasmid (4kb) or a fosmid (40kb) insert. The initial sequencing is performed by Production Sequencing from each end using universal primers. In order to completely sequence the region of interest, Computer Finishers analyze the initial data, identify the directions of sequence, and design custom oligonucleotide primers near the end of the known sequence. The first sequencing reaction is carried out and, based on the reads found, a second primer can be designed off of the new sequence to extend farther into the gap. Multiple rounds of this can be repeated until the gap is filled or questionable sequence can be resolved. In the Finishing Lab, plasmid or fosmid 384-well glycerol plates are pulled out and specific clones are inoculated from their initial positions into a single 96-well growth plate (medium with chloramphenicol). Cloned DNA is isolated from the cells using a modified alkaline lysis procedure which has been specially adapted for high-throughput DNA purification. A PicoGreen fluorescence DNA quantitation assay is done to determine DNA concentration and the template is normalized. Custom primers are ordered automatically in a 96-well plate format and three chemistries (TB, GB, GA) can be chosen for sequencing. Primer Walk reads in the Finishing Lab have, on average, an NHGRI pass rate of ~75% and average Q20 around 500.

FF0034

Fosmid Transposition Solves Large Sequence Gaps

Chelsea Dunbar, Nicole Allen, Daniel Bessette, Mike FitzGerald, Tashi Lokyitsang, Charles Matthews, Tamrat Negash, Thu Nguyen

Finishing Laboratory Group, Broad Institute, Cambridge, MA, USA

Fosmid transposition is a workflow utilized by the Finishing Laboratory group to successfully capture large sequence gaps and rapidly provide a complete sequence for tiling path clones. This process is also useful when fosmids act as the only capturing clone. Since transposons insert themselves randomly throughout a target clone, thorough coverage can be achieved rapidly, often within two weeks. In our process, a single bacterial colony is isolated from a sample of the respective fosmid library glycerol. We electroporate the transposed fosmids in order to recover the 672 colonies required for sequencing. The transposed fosmid clones are sequenced bidirectionally from primer binding sites on the transposons. To improve the accuracy of our process we have implemented the end-sequencing of several fosmid library colonies to confirm clone identity prior to transposition. In the future, TempliPhi amplification may replace our alkaline lysis DNA prep to improve time and cost efficiency of amplification and sequencing. This process currently yields 1344 reads of approximately 600bp each that have been used by the Computer Finishing group to successfully close many gaps in the Human, Mouse, and Magnaporthe grisea genomes.

FF0040

Finishing Platform for 454 Based Assemblies

Shannon P. Dugan-Rocha, Donna Muzny, Yan Ding, Christian J. Buhay, Aniko Sabo, Mike E. Holder, Xiang Qin, Huyen Dinh, Peter Blyth, Christie Kovar, Sandra Lee, Lynne Nazereth, Erica Sodergren, George M. Weinstock and Richard Gibbs

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

A robust and versatile 454 sequencing pipeline has been established at the BCM-HGSC focused on the application, optimization and throughput of new 454 sequencing technologies. To date, 41 microbial genomes have been sequenced utilizing the 454 sequence technology with various levels of Sanger read coverage. Of these genomes, *Staphylococcus aureus USA-300* has been completed and five additional genomes, *Pantoea stewartii, Bacillus pumilus, Treponema paraluiscuniculi*, *Moraxella bovis* and *Enterococcus faecalis OG1RF* are now in the finishing pipeline. Implementation of the 454 paired-end protocols have further enhanced the finishing process. The microbial assemblies using WGS and paired end sequences have been excellent with the number of scaffolds ranging from 8-29 for 2-3Mb genomes. With 454 assemblies at 25x coverage now comparable to Sanger assemblies at 8x coverage, the challenge is now to develop a finishing pipeline to deal with the unique issues related to finishing these 454 based assemblies. Creating a new pipeline for finishing these assemblies would thus require additional bench tools as well as modifications to our existing informatics pipeline.

Informatics tools were created to display the 454 read pair data within CONSED assembly view and Autofinish parameters were adjusted to create and place longer primers(~25mers) that could be used for PCR as well as direct genome walking. Further optimization of the primer picking process included the implementation of a BLAST database to identify various types of repeats as well as areas prone to misassembly, such as ribosomal RNA regions.

Sequencing and closure techniques were also modified due to the lack of template availability. These regions now had to be completed using various PCR methods and direct primer walking from genomic DNA. Generating PCR for many gaps would require optimization of long-range PCR using the GeneAmp XL PCR kit. Primers designed on contigs without order or orientation information could also be added to a multiplex PCR reaction and then used to generate sequence from any resulting amplicons. Direct primer walking from genomic DNA was also used to extend sequence where other methods had failed. Implementation of these techniques dramatically reduced the number of scaffolds in the 454/Sanger mixed assembly for *P. stewartii*. Here Autofinish and primer walking were used to close over 100 gaps. Multiplex PCR reactions were then used to close an additional 20 gaps. In assemblies with no Sanger coverage, such as *T. paraluiscaniculi*, Autofinish was used to create primers that generated sequence to close 63 of the initial 76 gaps. Final assemblies for these genomes will be compared to related strains and validated using restriction enzymes.

FF0044

Finishing Methods for 454/Sanger Mixed Assembly

Yan Ding, Donna M. Muzny, Shannon P. Dugan-Rocha, Guan Chen, Alicia A. Hawes, Lesette M. Perez, Yih-Shin Liu, Zhangwan Li, Suzhen Wang, Judith Hernandez, Geroge M. Weinstock and Richard A. Gibbs

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

Several microbial genomes including *Staphylococcus aureus USA-300, Bacillus pumilus, Pantoea stewartii*, *Treponema paraluiscuniculi, Moraxella bovis* and *Enterococcus faecalis OG1RF* have recently been sequenced at BCM-HGSC using a combined Sanger and 454 platform. To date, the *S.aureus* genome has been completed and the remaining genomes are in our finishing pipeline with varied levels of Sanger coverage. Although the implementation of 454 paired-end protocols significantly reduces the number of contigs without order and orientation, this combined assembly method carries some additional challenges and leaves traditional closure strategies limited by template availability and uncertainty for some low quality regions represented only by 454 reads. In addition, ribosomal RNA regions add more difficulty for finishing. Therefore we have implemented closure methods including direct genomic sequencing by using GenomiPhi, multiplex PCR sequencing and long range PCR sequencing to deal with these new challenges.

Direct genomic sequencing has been designed for closing or extending gap regions where template is not available. This method has contributed to the completion of *S. aureus* and initial results with the *B. pumilus* genome have shown success rate of 60% with an average Phred20 above 580 bp. Current efforts center on optimization of direct genomic sequencing protocols for microbial genomes of different size and composition. Multiplex PCR sequencing has also been used for bridging contigs without order and orientation in complicated microbes such as *P.stewartii.* The initial assembly of *P.stewartii* contains about 181 unscaffolded contigs. Here autofinishing and primer walking were used to close over 100 gaps leaving 70 contigs without any order and orientation. A series set of 43 multiplex PCR reactions were then used to close additional 20 gaps in the assembly. Finally, long range PCR is also applied for sequencing some highly repetitive regions such as ribosomal RNA which have been identified with our repeat tagging tool. The results and protocols of these strategies will be presented.

FF0048

The Rearray Manager

Alicia Clum, Eugene Goltsman, Steve Lowry, Hui Sun, Brian Foster, Stephan Trong, Paul Richardson, Alla Lapidus

DOE Joint Genome Institute, Walnut Creek, CA

The Rearray Manager is a web-based software application designed at the Joint Genome Institute for microbial finishing. The Rearray Manager allows us to track primers and reaction plates by name, date, library, or reaction type. One of the most critical functions of the Rearray Manager is its interaction with the Tecan robot. Our Tecan machines have been programmed to cherry pick from multiple source plates into a single destination plate. This is very important to our group because as finishers often we only need a handful of clones from a 384 well plate. The Rearray Manager tells the Tecan which wells to pick from and dispense into and dictates the deck layout.

The Rearray Manager is intergraded with consed so primers and reaction lists are automatically generated. Alternatively, such lists can be entered manually. The Rearray Manager organizes clones first by library, then by clone number unless you indicate otherwise. This flexibility in sorting is extremely useful when there are multiple libraries combined in one plate.

From the Rearray Manager one can also look up plates (primer plates or reaction plates), check on the status of plates, edit or delete plates, and make sample sheets so plates can be loaded onto an ABI 3730. Finishing plates and their source plates can be visualized in 384 or 96 well format. Rearray Manager also generates a list of source plates which is essential to ordering necessary source plates from production storage. Use of the Rearray Manager allows us to effectively perform lab experiments for a large amount of projects.

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 Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396

FF0050

CTM – a Web Based Genome Closure Task Management System

Xinyue (Jerry) Liu, Hean Koo, Luke J. Tallon

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville MD U.S.A.

CTM, an acronym for Closure Task Manager, is a generic Web based project management system used by the JCVI Genome Finishing group to manage tasks within and between genome finishing (Closure) projects, to track project progress, to manage project teams, and to provide project summary / statistics to team leaders and collaborators. CTM also contains many project specific add-ons to facilitate automated generation of task description, and provides convenient links to other JCVI software / resources to facilitate high throughput genome finishing. CTM are being actively used in all JCVI genome finishing projects including Eukaryotic BAC based projects, Sample based projects such as Influenza genome sequencing, and Whole Genome Shotgun (WGS) based projects.

CTM has been essential in the success of many large scale genome sequencing and finishing projects such as the Influenza Genome Sequencing project and other BAC projects, which typically need to manage thousands of tasks in their project life cycle. Currently there are about 30 JCVI genome finishing projects that are using CTM for daily task management.

CTM adopts a design of 3-tiered architecture that is composed of the front-end web interface, the PERL/CGI middle tier, and the Sybase database backend. By design, CTM can be easily modified to meet general task management needs for non-genome finishing projects as well.

FF0052

Minerva, Arcturus, Artemis: The development and use of new finishing tools at the Wellcome Trust Sanger Institute Pathogen Sequencing Unit

Danielle Walker and the Pathogen Sequencing Unit

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

The techniques and software used in Finishing have made considerable changes and progression over the last few years. The Sanger Institute's Pathogen Sequencing Unit has committed to sequencing more comparative projects and significantly larger genomes. The changes associated with these, for a finisher, are the ability to use comparative genomes when finishing a project and the introduction of the Minerva/Arcturus interface and assembly data management system.

Completed comparative genome sequences have been of particular use in reducing the amount of time it takes to finish a project and in highlighting differences between related organisms. Artemis: a DNA sequence viewer and annotation tool, and ACT (Artemis Comparison Tool): a DNA sequence comparison viewer, have been used in annotation for some time and are now being used with great success as finishing tools. An example of this has been the continuing work at the Sanger Institute finishing several different Streptococcal genomes. Co-assembly of *Streptococcus suis* SC84 using Sanger capillary shotgun data and 454 Shotgun data is also utilising these comparative tools.

There has been good progress made over the last year on the Minerva interface, which has been used successfully on several large projects using the Arcturus assembly data management system which was developed in house. Minerva is a Java application that will enable visualisation of a current genome assembly and scaffold. Of note recently, work on *Plasmodium knowlesi, Candida parapsilosis* and various *Haemonchus contortus* clones has successfully utilised this system. We envisage that it will be useful for prospective Helminth projects which are significantly larger than the genomes we currently finish.

FF0056

Advanced Closure Techniques for Whole Genome Alignment

Heather Forberger, Xinyue Liu, Bradley Toms, Luke Tallon, Hoda Khouri, Nadia Fedorova

J. Craig Venter Institute, Rockville, MD

The emphasis on sequencing multiple strains of the same organism have been on the rise, as learning about conserved regions has proven to be key for many genomic applications such as vaccine development. Following the completion of shotgun sequencing, a genome is assembled via the Celera assembler and scaffolded. The resulting assembly has many physical gaps. By taking advantage of the similarities between strains and using previously finished strains as a reference, we can expedite the finishing of a genome, by applying various new bioinformatics and automated techniques.

Physical/unlinked gaps typically consume a large portion of the closure process. As in theory, each end would have to be compared via PCR from whole genomic DNA to every other physical end within the genome to discover its proper orientation within the complete DNA strand. Although in recent years, the implementation of multiplex PCR has increased the efficiency of pairing ends, it is still an involved procedure. Genomes that have many unclonable regions will result in higher number of physical gaps, increasing the complexity and time spent orienting scaffolds.

We are going to explore several advanced techniques for aligning and closing these physical ends in a more direct manner using a reference genome. We have developed a bioinformatics pipeline using Perl, shell scripts, and Primer3 to achieve high-throughput design of PCR primer pairs at physical gap ends, screening out valid pairs based on their MUMMER alignment with the reference genome. Furthermore, 454 sequencing and subsequent hybrid assembly, as well as optical mapping show to be useful methods in which the strain of interest can be used as its own reference for Sanger sequence only contig alignment.

Although these techniques will prove to significantly reduce the time and labor associated with aligning physical ends, it may never work at 100%, as variations in sequence are typically present between strains. Some of the common differences include, but are not limited to insertions, deletions and recombination of DNA sequences. We will discuss the results of various trials conducted with each technique as well as the effectiveness in saving time and money in genome closure. This poster will reflect the advancements for each of these techniques, as well as their appropriate future applications

FF0058

Project Management at the Joint Genome Institute

Lynne Goodwin[1], David Bruce[1], Kerrie Barry[2], Tijana Glavina del Rio[2], Susannah Tringe[2],

[1]Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM/USA
[2] JGI Production Genomics Facility, Walnut Creek, CA/USA

As high throughput sequencing centers move from managing a small number of large projects to managing many simultaneous small projects, the ability to govern schedule, cost, quality, and project specification becomes more difficult. The JGI has implemented a formal project management system that simplifies controlling multiple small projects.

The Department of Energy Joint Genome Institute (JGI) high throughput sequencing and computational analysis group consists of teams at Oak Ridge National Laboratory, DOE Production Genomics Facility at Walnut Creek, Ca, Los Alamos National Laboratory, Lawrence Livermore National Laboratories and the Stanford University, SHGC. Historically the JGI had a small number of large projects. Today, the JGI sequencing capacity is dedicated to many small projects (< 10 MB) such as microbial, both small eubacteria and eukaryotic genomic, environmental metagenomic, and large eukaryotic projects. The JGI Project Management Office is the primary point of contact for JGI sequencing project information for both the internal and external stakeholders.

FF0068

Direct Sequencing of Large Insert Size Clones Using Templates Generated by Rolling Circle Amplification

Damon Tighe, Nancy Hammon, Susan Lucas, and Jan-Fang Cheng

US Department of Energy Joint Genome Institute, Walnut Creek, CA

Rolling circle amplification (RCA) has been widely used in production sequencing facilities for preparing high quality sequencing templates from small insert size (3 and 8 Kb) clones. This approach, however, has not been successful in preparing sequencing templates from large insert size clones (fosmids and BACs) due to the inconsistency of generating high quality reads. In the attempt to optimize this process, we have tested several conditions including heat lysis of cells, lysis buffer, addition of DMSO, premix to cell lysate volume ratio, and cycle sequencing. We will describe in detail how the various conditions influence the quality of the reads. The results show that a 10 second heat lysis at 95C on a 1 to 1 ratio of induced glycerol stock to TE containing MgCl2, 6.66% DMSO, 2 to 1 ratio of Templiphi version2 premix to cell lysate, and 38 cycles of sequencing reaction give the best sequencing quality. We also found that the induction of the fosmid copy number may actually lead to a decrease of sequence quality in particular fosmid libraries. The differences in sequencing quality resulted from different libraries are being investigated. We have begun to test conditions that are suitable for BAC end sequencing using the RCA templates. The preliminary data shows that it is possible to sequence up to 700 bp directly from BACs using the RCA products.

FF0082

The Prefinishing Pipeline at Washington University's Genome Sequencing Center

Amy Reily, Robert S. Fulton, and Richard K. Wilson

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

The Prefinishing group at the Genome Sequencing Center (GSC) offers a convenient, efficient and cost-effective method for improving the quality of shotgun sequence data. The GSC uses Autofinish (Gordon, D., University of Washington) to choose directed reactions to improve low quality regions, close gaps, or both, in a variety of assemblies. The Prefinishing process is used on entire BAC clones or whole genome assemblies of varying sizes, ranging from megabase size bacterial projects, to gigibase size eukaryotic genomes. The process can also be utilized on regions of either BACs or whole genome assemblies such as gene regions or non-repetitive regions.

The prefinishing pipeline is highly successful in closing gaps and improving low quality regions in all projects. For the clone-based Maize project, the number of contigs is reduced by 50% during the prefinishing process (average starting contigs=28, average ending contigs=15). For NHGRI BACs, the results are even better. The average contig number at the completion of the shotgun assembly is 6 while the average contig number after prefinishing is 2.5. In addition, 40% of these BACs are contiguous after the prefinishing process. This prefinishing process can stand alone as an improvement to BACs or whole genome assemblies, or can significantly reduce finishing time and costs for projects scheduled for additional improvement efforts.

FF0092

Progress in the Finishing the Communally-sequenced Organism, *Accumulibacter phosphatis*, from Waste-treatment Sludge

Stephen Lowry, Hector Garcia Martin, Phil Hugenholtz, Alicia Clum, Paul Richardson, Alla Lapidus

DOE Joint Genome Institute, Walnut Creek, CA

Accumulibacter Phosphatis species are the principal actors in the sequestration of inorganic phosphate as intracellular polyphosphate in waste-water treatment facilities. They are thus very important bioremediators.

In the present case, since bioreactor cultures yielded an 80% enrichment for *A. phosphates* species*,* it was hoped that a substantially complete genome might be sought without requiring a pure clonal culture. Communal DNA was shotgun sequenced from subclone libraries of 3, 8 and 40 kb. The sequence was further enriched for the dominant organism by considering read-depth, GC content and clade-characteristic features. The reduced set of reads was assembled in parallel with both JAZZ and Phrap4 and later, ARACHNE assembly tools, yielding approximately 8-fold coverage of an apparent 5.6±0.2 Mb of genomic DNA. The chromosomal sequence was seen as a single scaffold*.

A smaller plasmid sequence constitutes about 168 kb, and the chromosomal scaffold 5.23 Mb. Gaps between major contigs, numbering about 125, were virtually all bridged by subclone read-pairs.

Efforts to complete the *Accumulibacter phosphates* genome will be presented.

*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.*

06/20/2007 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
745 - 845am	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
845 - 900	Intro	x	Welcome Back Intro - New Technologies	Paul Richardson
900 - 930	Speaker 1	FF0033d	New Amplification and Cloning Tools for Finishing Genomes	David Mead
930 - 950	Speaker 2	FF0102	TaxSorter: A Solution to Metagenomic Projects	Li Liu
950 -1020	Break	x	Beverages and snacks provided	x
1020 - 1040	Speaker 3	FF0047a	Metagenomic Assembly QC	Alla Lapidus
1040 -1100	Speaker 4	FF0065	Evaluation of New methods and Approaches for Comparative Metagenomic Studies	Emmanuel Mongodin
1100 -1200	Panel Discussion	x	**Panel Discussion** - New Technologies	Chair - Alla Lapidus
1200 - 130pm	Lunch & Close of meeting	x	**La Fiesta Plaza Lunch Buffet - sponsored by illumina** (Cheese enchiladas served with “Christmas” (red and green) chile, 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 (June 20th)

Abstracts are in order of presentation according to Agenda

FF0033d

New Amplification and Cloning Tools for Finishing Genomes

David Mead

Lucigen Corp, Middleton WI 53562

Accurate cloning and sequence assembly are hindered by several major hurdles, including: 1) restriction-site bias in BAC cloning, 2) clone bias against structure rich sequences, repeats or AT-rich DNA, 3) chimeric clones in shotgun libraries, 4) cloning of trace amounts of template, and 5) non-specific products in whole genome amplification. Our group has recently developed technologies to address each of these problems. Centromeric and other highly repetitive genomic regions are absent or vastly under-represented in typical partial-digest BAC libraries. We have developed a method to randomly shear and clone DNA fragments of >100 kb. “Random Shear” BAC libraries show uniform coverage over regions that are over- or under-represented in conventional BAC libraries. For cloning fragments of up to 30 kb, the novel “pJAZZ” linear cloning vector provides unprecedented ability to maintain regions that are unclonable in circular plasmids. Examples include large, highly AT-rich fragments of 20-30 kb and regions of di-, tri-, and tetra-nucleotide repeats. The pJAZZ linear vector has been used to sequence and assembly an AT rich genome and the results demonstrate bias free closure without the use of BAC or fosmid libraries, saving considerable time and expense.

We have also developed a method of “GC Cloning” to minimize chimeric inserts in shotgun libraries. A simple tailing reaction appends a 3’-G residue to target fragments, which are ligated to a vector with a 3’-C tail. The frequency of chimeras in the resulting libraries is less than 1%, simplifying the process of sequence assembly. Another bottleneck in genomic analysis is cloning low amounts of DNA. The high efficiency of GC Cloning allows direct cloning of nanogram amounts of DNA without template amplification. For library construction from even smaller samples, e.g., isolated cells, rare microbes, or metagenomic communities, we demonstrate a method to anonymously amplify and clone picogram amounts of DNA. Inserting the amplified DNA into transcription-free pSMART vectors reduced cloning bias against toxic sequences.

Finally, a new thermostable phage DNA polymerase allows isothermal whole genome amplification at elevated temperatures. Based on strand displacement from nicked DNA rather than from random primers, this amplification method eliminates the background associated with de novo synthesis from exogenous primers. These methods and vectors provide improved cloning of fragments from all sizes and from diverse sources.

FF0102

TaxaSorter: A Solution to Metagenomic Projects

Li Liu, Yijun Sun, Fahong Yu, William G. Farmerie

Interdisciplinary Center for Biotechnology Research, University of Florida, Florida, U.S.A.

A metagenomic project generates a large collection of heterogeneous sequences. Assigning the correct taxonomic origin for each and every sequence is the key to answer all questions metagenomic projects try to address. We have developed a program, TaxaSorter that classifies sequences into a hierarchical taxonomy tree based on sequence similarities. The algorithm is based on BLAST search. However, instead of simply taking the taxonomic origin of the first hit, probability values of all taxonomic origins from the top 100 hits are calculated for every query sequence. If only one taxonomic origin is assumed for each sequence, the one with the highest probability is selected. Otherwise, all possible origins are retained with their corresponding probability values. Then, all sequences are sorted through a hierarchical taxonomy tree where each taxon node is associated with the total count of linked and sub-linked sequences. TaxaSorter also includes functions to compare two taxonomy trees on all taxon nodes to identify nodes that are significantly over-represented or under-represented. Simulations were conducted to estimate the sensitivity and specificity. TaxaSorter was applied on several metagenomic data sets, including both traditional Sanger sequences and 454 sequences. The sample result from a soil metagenomic project was presented. TaxaSorter is implemented as a module in BlastQuest, available at UF/ICBR genomics server.

FF0047

Metagenomic Assembly QC

Alla Lapidus, Alex Copeland, Natalia Ivanova

DOE Joint Genome Institute, Walnut Creek, CA 94598

The whole-genome shotgun sequencing approach was successfully used for a number of microbial community projects, however useful quality control and assembly of these data require reassessing methods developed to handle relatively uniform sequences derived from isolate microbes. In addition to the typically very large size of metagenomic sequencing projects, potential problems observed in metagenomic assemblies include chimeric contigs produced by co-assembly of sequencing reads originating from different species, and non-uniform sequence coverage resulting in significant under- and over-representation of certain community members. Depending on the assembly algorithm and sequencing read depth, some fragments are resolved into strain-specific contigs corresponding to different haplotypes, while others are co-assembled into a composite scaffold with strain-specific variations appearing as single-nucleotide polymorphisms. Large-scale genome rearrangements and the presence of mobile genetic elements (phages, transposons) in the abundant community members result in assembly break points in the areas of synteny breakdown. Many parameters of metagenomic assemblies remain unknown, including the nature and extent of chimeric contigs , influence of the level of polymorphism on co-assembly of reads into composite contigs, overall quality of assembies and binning, etc. Assembly correctness is essential for metagenomic projects since it influences much more the subsequent analysis and interpretation of metagenomic data than in case of individual microbes due to coverage related increases in consensus error rate. In addition, there are no tools specifically designed for metagenomic annotation.

We believe that to improve the quality of metagenomic assemblies, reads should be stringently quality and vector trimmed prior to assembly. We report on results of using different trimming tools and different combinations of such tools and assemblers on a number of metagenomic projects sequenced at JGI. Resulting assemblies were visually inspected and analyzed using IMG/M system developed at JGI for microbial genome analysis.

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 Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396

FF0065

Evaluation of New Methods and Approaches for Comparative Metagenomic Studies

Emmanuel Mongodin

J. Craig Venter Institute, Rockville, MD 20850, USA

Metagenomics is a rapidly emerging field of research for studying microbial communities. The term "metagenomics" was first coined by Jo Handelsman (University of Wisconsin) and is defined as "the application of modern genomics techniques to the study of complex communities of microbial organisms directly in their natural environments". In complex environments where most species are not cultivable, key questions such as "who's there?" and "what are they doing?" can be answered through metagenomics approaches without the need for isolation and lab cultivation of individual species. Metagenomics aims to capture the full measure of microbial diversity by recovering communities of microbial genes directly from the environment where these organisms live, in order to shed light on the biological processes present in the environment studied. Most of the current tools and methods used to analyze metagenomic data sets are adapted from the ones developed for the analysis of single genomes. But the genetic and taxonomic complexity (genomes of multiple strains and species present in highly variable abundance) of metagenomic data sets imposes new challenges on existing analysis tools. Therefore, in order to fully and comprehensively analyze the growing number of metagenomic projects, development of new analysis tools and methods specifically dedicated to metagenomics sets is needed.

Using the Gut Microbiome data set (Gill *et al*, Science 2006) as a test-case, we evaluated new comparative metagenomics tools and methods. The random shotgun reads from the distal gut microbiome of subjects 7 and 8 were assembled separately using the Celera Assembler. Accuracy of ORF prediction was tested using an in-house BlastX-based pipeline, as well as the recently developed ORF finder, Metagene. Phylogenetic assignments were done using Blast and the MEGAN software, at the level of sequencing read, contigs and proteins. Comparative analyses showed that only 4% of the proteins seems to be conserved between the metaproteomes of subject 7 and subject 8. However, comparative protein domain analysis showed that 88% of the protein domains are shared between the two metaproteomes. Therefore, despite significant differences in the microbial genetic diversity, most likely at the level of the strains and species, it seems that there is a high conservation of function between the gut metagenomes of these two healthy individuals. The results presented here highlight the need for novel approaches for metagenomic analyses, with accurate gene and protein predictions, but also function and pathway predictions together with phylogenetic assignments. This study should lay the foundation for the development of comparative metagenomic data annotation and analysis pipelines.

Panel Discussion Notes

Panel Discussion Notes

Close of Meeting Lunch

12pm - 130pm, June 20th

Sponsored by illumina

Enjoy!!!

Misc. Notes

Misc. Notes

Misc. Notes

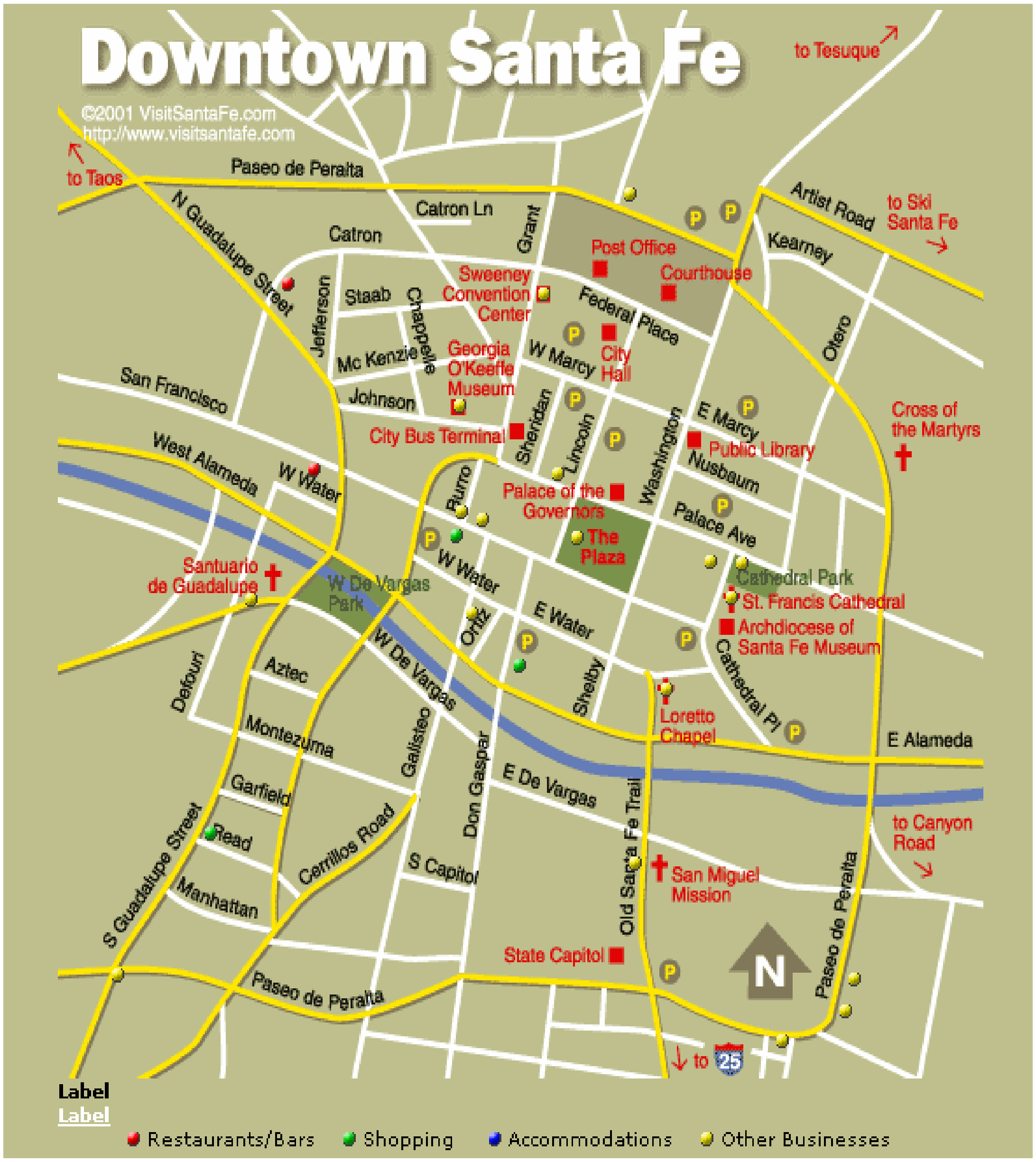
Attendees

FF #	Name	Affiliation	email	Talk / Poster / Neither
1	**Chris Detter**	Los Alamos National Laboratory - JGI	cdetter@lanl.gov	N
2	**Beth Nelson**	Novozymes Inc., Davis, CA	BANE@novozymes.com	N
3	**Lou Sherman**	Purdue University, West Lafayette, IN	lsherman@purdue.edu	P
4	**Haibao Tang**	Plant Genome Mapping Laboratory, University of Georgia	bao@uga.edu	P
5	**Stephan Trong**	DOE Joint Genome Institute (LLNL), Walnut Creek, CA	trong1@llnl.gov	T
6	**Chris Munk**	Los Alamos National Laboratory - JGI	cmunk@lanl.gov	P
7	**Liz Saunders**	Los Alamos National Laboratory - JGI	ehs@lanl.gov	N
8	**David Harris**	The Wellcome Trust Sanger Institute, Hinxton, Cambridge	deh@sanger.ac.uk	P
9	**Olga Chertkov**	Los Alamos National Laboratory - JGI	ochrtkv@lanl.gov	P
10	**David Sims**	Los Alamos National Laboratory - JGI	dsims@lanl.gov	P
11	**Mihai Pop**	University of Maryland, Center for Bioinformatics & Computational Bio.	mpop@umiacs.umd.edu	KEYNOTE
12	**Alice Young**	NIH Intramural Sequencing Center (NISC), NHGRI, Rockville, MD	alicey@nhgri.nih.gov	N
13	**Jennifer Vogt**	NIH Intramural Sequencing Center (NISC), NHGRI, Bethesda, MD	jvogt@mail.nih.gov	N
14	**Jyoti Gupta**	NIH Intramural Sequencing Center (NISC), NHGRI, Bethesda, MD	jyotig@mail.nih.gov	P
15	**Beatrice Barnabas**	NIH Intramural Sequencing Center (NISC), NHGRI, Bethesda, MD	bbenjam@nhgri.nih.gov	N
16	**Pat Kale**	DOE Joint Genome Institute (LLNL), Walnut Creek, CA	kale1@llnl.gov	N
17	**Shelise Brooks**	NIH Intramural Sequencing Center (NISC), NHGRI, Bethesda, MD	sbrooks@mail.nih.gov	N
18	**Granger Sutton**	J Craig Venter Institute / TIGR, Rockville, MD	GSutton@venterinstitute.org	T
19	**Hoda Khouri**	J Craig Venter Institute / TIGR, Rockville, MD	HKhouri@jcvi.org	P
20	**Mandy Sanders**	The Wellcome Trust Sanger Institute, Hinxton, Cambridge	mjs@sanger.ac.uk	P
21	**Angie Lackey**	Roche Diagnostics, Indianapolis, IN	angie.lackey@roche.com	N
22	**Omayma Al-Awar**	Edge BioSystems, Inc.	oalawar@edgebio.com	N
23	**Brad Toms**	J Craig Venter Institute / TIGR, Rockville, MD	btoms@tigr.ORG	P
24	**Margaret Priest**	The Broad Institute - MIT, Cambridge, MA	mpriest@broad.mit.edu	P
25	**Linda Meincke**	Los Alamos National Laboratory - JGI	meincke@lanl.gov	N
26	**Harindra Arachchi**	The Broad Institute - MIT, Cambridge, MA	harindra@broad.mit.edu	T
27	**Adam Brown**	The Broad Institute - MIT, Cambridge, MA	abrown@broad.mit.edu	P
28	**Amr Abouelleil**	The Broad Institute - MIT, Cambridge, MA	amr@broad.mit.edu	P
29	**Hilary Morrison**	Marine Biological Laboratory, Woods Hole, MA	morrison@mbl.edu	N
30	**Xiaohong Liu**	The Broad Institute - MIT, Cambridge, MA	xliu@broad.mit.edu	P
31	**Daniel Bessette**	The Broad Institute - MIT, Cambridge, MA	danielb@broad.mit.edu	P
32	**Sean Sykes**	The Broad Institute - MIT, Cambridge, MA	ssykes@broad.mit.edu	T
33 a	**David Mead**	Lucigen Corp., Middleton, WI	dmead@lucigen.com	P
33 b	**David Mead**	Lucigen Corp., Middleton, WI	dmead@lucigen.com	P
33 c	**David Mead**	Lucigen Corp., Middleton, WI	dmead@lucigen.com	P
33 d	**David Mead**	Lucigen Corp., Middleton, WI	dmead@lucigen.com	T
34	**Chelsea Dunbar**	The Broad Institute - MIT, Cambridge, MA	cfoley@broad.mit.edu	P
35	**Christian Buhay**	Baylor College of Medicine - HGSC, Houston, TX	cbuhay@bcm.tmc.edu	P
36	**Mike FitzGerald**	The Broad Institute - MIT, Cambridge, MA	fitz@broad.mit.edu	N
37	**Pat Minx**	Washington University Genome Sequencing Center, St. Louis, MO	pminx@watson.wustl.edu	N
38	**Lynn Carmichael**	Washington University Genome Sequencing Center, St. Louis, MO	lcarmich@watson.wustl.edu	N
39	**Joann Mudge**	National Center for Genome Resources, Santa Fe, NM	jm@ncgr.org	T
40	**Shannon Dugan-Rocha**	Baylor College of Medicine - HGSC, Houston, TX	sdugan@bcm.tmc.edu	P
41	**Stephen Kingsmore**	National Center for Genome Resources, Santa Fe, NM	sfk@ncgr.org	N
42	**Jeremy Schmutz**	Stanford Human Genome Center (JGI), Palo Alto, CA	jeremy@paxil.stanford.edu	T
43	**x**	**x**	x	x
44	**Yan Ding**	Baylor College of Medicine - HGSC, Houston, TX	yding@bcm.tmc.edu	P
45	**George Weinstock**	Baylor College of Medicine - HGSC, Houston, TX	gwstock@bcm.tmc.edu	KEYNOTE
46	**Hajni Kiss**	Los Alamos National Laboratory - JGI	hajkis@lanl.gov	N
47a	**Alla Lapidus**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	alapidus@lbl.gov	T
47b	**Alla Lapidus**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	alapidus@lbl.gov	P
48	**Alicia Clum**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	AClum@lbl.gov	P
49	**Michael Holder**	Baylor College of Medicine - HGSC, Houston, TX	mholder@bcm.tmc.edu	N
50	**Jerry Liu**	J Craig Venter Institute / TIGR, Rockville, MD	xliu@tigr.ORG	P

51	**Alan Tracey**	The Wellcome Trust Sanger Institute, Hinxton, Cambridge	alt@sanger.ac.uk	P
52	**Danielle Walker**	The Wellcome Trust Sanger Institute, Hinxton, Cambridge	dw2@sanger.ac.uk	P
53	**Giselle Kerry**	The Wellcome Trust Sanger Institute, Hinxton, Cambridge	gh2@sanger.ac.uk	P
54	**Darren Grafham**	The Wellcome Trust Sanger Institute, Hinxton, Cambridge	dg1@sanger.ac.uk	N
55	**Stephanie Malfatti**	Lawrence Livermore National Laboratory (JGI), Livermore, CA	malfatti3@llnl.gov	P
56	**Heather Forberger**	x	x	P
57	**Susannah Green Tringe**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	SGTringe@lbl.gov	N
58	**Lynne Goodwin**	Los Alamos National Laboratory - JGI	lynneg@lanl.gov	P
59	**Gary Resnick**	Los Alamos National Laboratory	resnick@lanl.gov	N
60	**Luke Tallon**	J Craig Venter Institute / TIGR, Rockville, MD	ljtallon@tigr.ORG	N (P #77)
61	**Paul Richardson**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	PMRichardson@lbl.gov	N
62	**Norman Doggett**	Los Alamos National Laboratory	doggett@lanl.gov	N
63	**Tom Brettin**	Los Alamos National Laboratory - JGI	brettin@lanl.gov	N
64	**Cliff Han**	Los Alamos National Laboratory - JGI	han_cliff@lanl.gov	N
65	**Emmanuel Mongodin**	J Craig Venter Institute / TIGR, Rockville, MD	mongodin@jcvi.org	T
66	**x**	x	x	x
67	**Mary Kim**	J Craig Venter Institute / TIGR, Rockville, MD	mkim@jcvi.org	P
68	**Damon Tighe**	DOE Joint Genome Institute (LLNL), Walnut Creek, CA	tighe2@llnl.gov	P
69	**Jennifer Currie**	University of Arizona, Plant Sciences Department	jcurrie@email.arizona.edu	P
70	**Yeisoo Yu**	University of Arizona, Plant Sciences Department	yeisooyu@ag.arizona.edu	N
71	**Shanmugam Rajasekar**	University of Arizona, Plant Sciences Department	shans@ag.arizona.edu	N
72	**Angelina Angelova**	University of Arizona, Plant Sciences Department	angelova@ag.arizona.edu	N
73	**Ara Ko**	University of Arizona, Plant Sciences Department	arako@ag.arizona.edu	N
74	**Teri Mueller**	University of Arizona, Plant Sciences Department	trambo@ag.arizona.edu	N
75	**Nick Sisneros**	University of Arizona, Plant Sciences Department	sisneros@ag.arizona.edu	N
76	**x**	x	x	x
77	**Jeffrey Sparenborg**	x	x	P
78	**x**	x	x	x
79	**Bob Fulton**	Washington University Genome Sequencing Center, St. Louis, MO	bfulton@watson.wustl.edu	N
80	**Aye Wollam**	Washington University Genome Sequencing Center, St. Louis, MO	awollam@watson.wustl.edu	T
81	**Laura Courtney**	Washington University Genome Sequencing Center, St. Louis, MO	lcourtne@watson.wustl.edu	P
82	**Amy Reily**	Washington University Genome Sequencing Center, St. Louis, MO	akozlowi@watson.wustl.edu	P
83	**Chad Tomlinson**	Washington University Genome Sequencing Center, St. Louis, MO	ctomlins@watson.wustl.edu	N
84	**Neha Shah**	Washington University Genome Sequencing Center, St. Louis, MO	nshah@watson.wustl.edu	N
85	**Lee Trani**	Washington University Genome Sequencing Center, St. Louis, MO	ltrani@watson.wustl.edu	N
86	**Scott Kruchowski**	Washington University Genome Sequencing Center, St. Louis, MO	skruchow@watson.wustl.edu	N
87	**Johar Ali**	BC Cancer Agency - Genome Sciences Centre, Vancouver, BC	jali@bcgsc.ca	T
88	**Jim Knight**	454 Life Sciences (Roche), Branford, CT	jknight@454.com	T
89	**Ben Faga**	Cold Spring Harbor Laboratory, NY	faga.cshl@gmail.com	P
90	**Donna Muzny**	Baylor College of Medicine - HGSC, Houston, TX	donnam@bcm.tmc.edu	T
91	**Yasmin Mohamoud**	J Craig Venter Institute / TIGR, Rockville, MD	yasminm@jcvi.org	N
92	**Steve Lowry**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	slowry@lbl.gov	P
93	**Susan Lucas**	DOE Joint Genome Institute (LLNL), Walnut Creek, CA	lucas11@llnl.gov	N
94	**Tijana Glavina del Rio**	DOE Joint Genome Institute (LLNL), Walnut Creek, CA	glavinadelrio1@llnl.gov	N
95	**Miranda Harmon-Smith**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	MLHarmon-Smith@lbl.gov	N
96	**Alex Copeland**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	accopeland@lbl.gov	N
97	**Eileen Dalin**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	e_dalin@lbl.gov	N
98	**Fang Cheng**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	jfcheng@lbl.gov	N
99	**Fiona Hyland**	Applied Biosystems, Foster City, CA	hylandfc@appliedbiosystems.com	T
100	**Clark Eason**	Applied Biosystems, Foster City, CA	Clark.Eason@appliedbiosystems.com	N
101	**Brandon Blakey**	Applied Biosystems, Foster City, CA	blakeybm@appliedbiosystems.com	N
102	**Li Liu**	University of Florida, ICBR, Gainesville, FL	liliu@biotech.ufl.edu	T
103	**Neil Miller**	National Center for Genome Resources, Santa Fe, NM	nam@ncgr.org	N
104	**Mark Lawrence**	Mississippi State University, College of Veterinary Medicine	lawrence@cvm.msstate.edu	N
105	**Ryan Weil**	Roche Applied Science, Indianapolis, IN	ryan.weil@roche.com	N
106	**Jim Bristow**	DOE Joint Genome Institute (LBNL), Walnut Creek, CA	JBristow@lbl.gov	N

107	**David Bruce**	Los Alamos National Laboratory - JGI	dbruce@lanl.gov	N
108	**John Havens**	Integrated DNA Technologies, Coralville, IA	jhavens@idtdna.com	N
109	**Ken Taylor**	Integrated DNA Technologies, Coralville, IA	ktaylor@idtdna.com	N
110	**Gary Schroth**	Illumina, Inc., Hayward, CA	GSchroth@illumina.com	T
111	**Jonathan Eisen**	UC Davis Genome Center, Davis, CA	jaeisen@ucdavis.edu	T
112	**Tonya Ehlmann**	Monsanto Company, St. Louis, MO	tonya.s.ehlmann@monsanto.com	N
113	**Jing Lu**	Monsanto Company, St. Louis, MO	jing.lu@monsanto.com	N
114	**Anjali Pradhan**	Applied Biosystems, Foster City, CA	PradhaAA@appliedbiosystems.com	N
115	**Jan Korbel**	Yale University, New Haven, CT	jan.korbel@Yale.edu	T
116	**Lori Court**	Caldera Pharmaceuticals, Inc (LANL), Los Alamos, NM	court@cpsci.com	N
117	**Jian Xu**	Washington University Genome Sequencing Center, St. Louis, MO	jxu@watson.wustl.edu	P
118	**Chad Geringer**	Illumina, Inc., Hayward, CA	cgeringer@illumina.com	N
119	**Patrick Finn**	Invitrogen Corporation, Carlsbad, CA	Patrick.Finn@invitrogen.com	N
120	**Keith Farnsworth**	Invitrogen Corporation, Carlsbad, CA	Keith.Farnsworth@invitrogen.com	N
121	**Cheryl Gleasner**	Los Alamos National Laboratory - JGI	cdgle@lanl.gov	N
122	**Mary Campbell**	Los Alamos National Laboratory - JGI	mcampbell@lanl.gov	N
123	**Beverly Parson-Quintana**	Los Alamos National Laboratory - JGI	bapq@lanl.gov	N
124	**Patti Wills**	Los Alamos National Laboratory - JGI	wills@lanl.gov	N
125	**Kim McMurry**	Los Alamos National Laboratory - JGI	kmcmurry@lanl.gov	N
126	**Roxanne Tapia**	Los Alamos National Laboratory - JGI	rox@lanl.gov	N
127	**Andy Seirp**	Los Alamos National Laboratory - JGI	aseirp@lanl.gov	N
128	**Avinash Kewalramani**	Los Alamos National Laboratory - JGI	avinash@lanl.gov	N
129	**Bill Feiereisen**	Los Alamos National Laboratory	wjf@lanl.gov	N
130	**Licen Xu**	Applied Biosystems, Foster City, CA	Licen.Xu@appliedbiosystems.com	N
131	**Jane Hutchinson**	Roche Diagnostics, Indianapolis, IN	NA	N
132	**Gary Nunn**	Illumina, Inc., Hayward, CA	gnunn@illumina.com	N
133	**Patrick Chain**	Lawrence Livermore National Laboratory (JGI), Livermore, CA	chain2@llnl.gov	N

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)

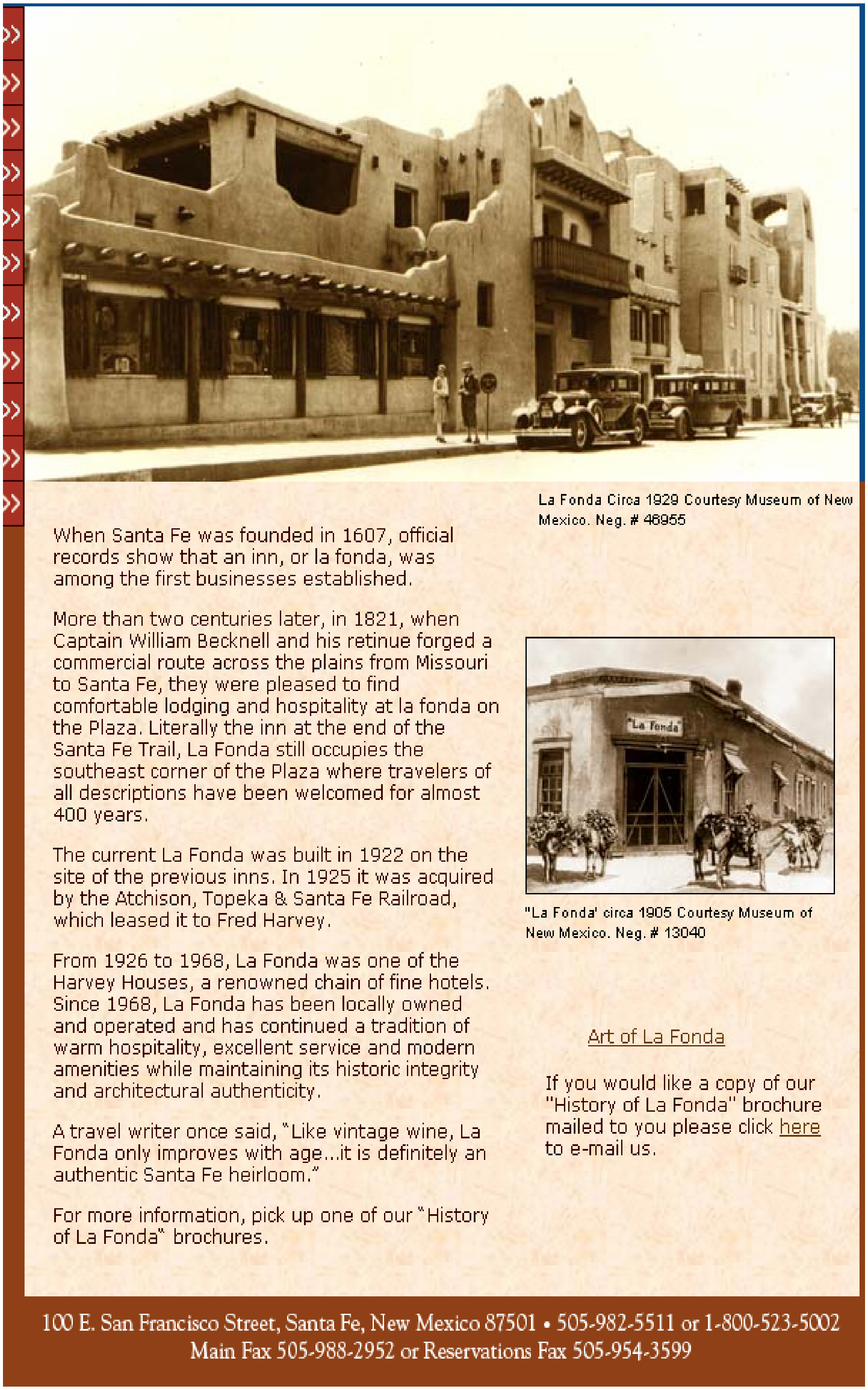
When New Mexico gained statehood in 1912, many people were drawn to Santa Fe's dry climate as a cure for tuberculosis. The Museum of New Mexico had opened in 1909, and by 1917, its Museum of Fine Arts was built. The state museum's emphasis on local history and native culture did much to reinforce Santa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

History of La Fonda



La Fonda Circa 1929 Courtesy Museum of New Mexico. Neg. # 46955

When Santa Fe was founded in 1607, official records show that an inn, or la fonda, was among the first businesses established.

More than two centuries later, in 1821, when Captain William Becknell and his retinue forged a commercial route across the plains from Missouri to Santa Fe, they were pleased to find comfortable lodging and hospitality at la fonda on the Plaza. Literally the inn at the end of the Santa Fe Trail, La Fonda still occupies the southeast corner of the Plaza where travelers of all descriptions have been welcomed for almost 400 years.

The current La Fonda was built in 1922 on the site of the previous inns. In 1925 it was acquired by the Atchison, Topeka & Santa Fe Railroad, which leased it to Fred Harvey.

From 1926 to 1968, La Fonda was one of the Harvey Houses, a renowned chain of fine hotels. Since 1968, La Fonda has been locally owned and operated and has continued a tradition of warm hospitality, excellent service and modern amenities while maintaining its historic integrity and architectural authenticity.

A travel writer once said, "Like vintage wine, La Fonda only improves with age...it is definitely an authentic Santa Fe heirloom."

For more information, pick up one of our "History of La Fonda" brochures.

"La Fonda' circa 1905 Courtesy Museum of New Mexico. Neg. # 13040

Art of La Fonda

If you would like a copy of our "History of La Fonda" brochure mailed to you please click here to e-mail us.

100 E. San Francisco Street, Santa Fe, New Mexico 87501 • 505-982-5511 or 1-800-523-5002
Main Fax 505-988-2952 or Reservations Fax 505-954-3599