1st Annual Finishing in the Future Meeting

Santa Fe, New Mexico
May 4th-5th, 2006

Los Alamos NATIONAL LABORATORY EST.1943

JGI

Contents

The 2006 "Finishing in the Future" Organizing Committee:

* *Chris Detter, Ph.D., Genomics Team Leader, LANL*
* *David Bruce, DOE-JGI Microbial Project Manager, LANL*
* *Patrick Chain, Finishing Group Leader, LLNL*
* *Cliff Han, Ph.D. Finishing Team Leader, LANL*
* *Alla Lapidus, Ph.D., Microbial Genomics Group Leader, LBNL-JGI*
* *Jeremy Schmutz, Informatics Group Leader, Stanford HGC*

Agenda

5/4/06 - Thurs.				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	La Fonda Breakfast Buffet (eggs, pancakes, bacon, etc.)	x
830 - 845	Intro	x	Welcome Intro	Paul Gilna
845 - 930	Keynote	FF048	Why finish anyway? What more do we get from finished sequence?	Julian Parkhill
930 - 1000	Speaker 1	FF028	Genomes Finishing Process at TIGR	Hoda Khouri
1000 - 1030	Speaker 2	FF036	Finishing bacterial genome sequences in the Pathogen Sequencing Unit of the Sanger Institute	David Harris
1030 - 1100	Break	x	Beverages & snacks provided	x
1100 - 1130	Speaker 3	FF046	Finishing Pipeline Developments at BCM-HGSC	Donna Muzny
1130 - 1200	Speaker 4	FF058	The Joint Genome Institute and Challenges in Microbial Genomics	Patrick Chain
1200 - 100pm	Lunch	x	Pecos Lunch Buffet (Grilled Chikcen Breast w/ Chipotle Barbecue Vinaigrette or Pan-Fried Fillet of Rainbow Trout w/ Cilantro Butter Sauce, etc.)	x
100 - 200	Lunch & Posters	x	Finish up lunch & enjoy poster session	x
200 - 230	Speaker 5	FF049	Finishing & Improvement of Whole Genome Shotgun Sequenced Eukaryotes at SHGC/JGI	Jeremy Schmutz
230 - 300	Speaker 6	FF010	Automated Finishing at the Wellcome Trust Sanger Institute	Stuart McLaren
300 - 330	Speaker 7	FF034	Quick Draft Assembly Improvement for Improved Finishing	Alex Copeland
330 - 400	Break	x	Beverages & snacks provided	x
400 - 530	Round Table #1	x	Technology Development - what's on the horizon?	Chaired by TBD
400 - 530	Round Table #2	x	Computational Finishing Methods	Chaired by Johar Ali
400 - 530pm	Round Table #3	x	Laboratory Finishing Methods	Chaired by Alla Lapidus
530 - bedtime	Dinner & entertainment on your own	x	x	x

5/5/06 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	Santa Fe Breakfast Buffet (eggs, chiliquiles, tortillas, bacon, etc.)	x
830 - 845	Intro	x	Welcome Back Intro	Jim Bristow
845 - 930	Keynote	FF027	Assisted Assembly	Sante Gnerre
930-1000	Speaker 1	FF001	Dog Genome Improvement	Michael FitzGerald
1000-1030	Speaker 2	FF002	Challenges Encountered in Finishing BAC Sequences from >60 Vertebrate Species	Bob Blakesley
1030-1100	Break	x	Beverages & snacks provided	x
1100-1130	Speaker 3	FF044	Simulating human finishing decisions with Autofinish	Cliff Han
1130-1200	Speaker 4	FF012	High Performance Customizable Software for Finishing and Assembly Analysis	Andrew Zimmer
1200 - 100pm	Lunch	x	La Fonda Lunch Buffet (Chicken La Fonda or Roasted Top Round, Sliced & Served with Au Jus, etc.)	x
100-130	Speaker 5	FF035	Recent Advances and Future Directions of Genome Finishing at TIGR	Luke Tallon
130-200	Speaker 6	FF055	Assisted genome closure and comparative genomic analysis with Optical Mapping	Colin Dykes
200-230	Speaker 7	FF061	Integrating new technologies into the JGI microbial program	Paul Richardson
230-300	Break	x	Beverages & snacks provided	x
300-415	Round Table #1	x	Technology Development - what's on the horizon?	Chaired by Paul Richardson
300-415	Round Table #2	x	Computational Finishing Methods	Chaired by Sante Gnerre
300-415	Round Table #3	x	Laboratory Finishing Methods	Chaired by Chris Detter
415-430	Closing remarks	x	x	David Bruce
430pm	End of Meeting	x	End of 2006 Finishing in the Future Meeting	x

Speaker Presentations (May 4th)

Abstracts are in order of presentation according to Agenda (page 2)

FF048 – Keynote

Why finish anyway? What more do we get from finished sequence?

Julian Parkhill

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

Given that unfinished shotgun data provides us with >99.9 % of the sequence of an organism, what more do we get from the hard work of finishing? I will try to answer this question using examples from bacterial genome projects. I will attempt to show that interesting data can reside in the gaps in shotgun sequence that variations uncovered by finishing can have biological relevance, and that even very rare single base pair errors can have biological consequences.

FF028

Genomes Finishing Process at TIGR

Hoda Khouri

The Institute of Genomic Research, Rockville, MD 20850

The finishing team at TIGR has released over 110 completely finished genomes. Some are whole genome shotgun projects (WGS) others are BAC-based projects. These genomes are very diverse (viral, bacteria, fungal and plant genomes) and present different levels of difficulties caused by high or low GC content, unclonable regions, large number of plasmids, DNA secondary structures (hard stops) and repetitive areas.

The finishing process starts after the reads are assembled with the Celera assembler. Most linked gaps are closed by sequencing linking clones using the efficient automated AutoCloser pipeline developed at TIGR. The unlinked gaps are amplified from the genomic DNA by conventional or high throughput multiplex PCR. All the repetitive areas are reviewed and verified. The remaining gaps require manual finishing and are usually caused by hard stops, homopolymer stretches, tandem repeats, mis-assemblies or large repeats.

As more finished genomes are released into the public databases, we are able to find syntheny between unfinished genomes and related complete genomes. Whole genomes alignments provide information for ordering and orientating contigs and facilitate the quality control of the final contig. Every genome (chromosome, plasmid, BAC, phage, etc…) has to satisfy the TIGR finishing criteria: each finished molecule has to be in a single contig, every base has to be covered by at least two sequence reads and spanned by at least two clones or PCR product.

Methods employed at TIGR for resolving difficult genome areas and for performing quality control will be discussed in the presentation.

FF036

Finishing bacterial genome sequences in the Pathigen Sequencing Unit of the Sanger Institute.

David Harris and the Pathogen Sequencing Unit

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

The Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute has sequenced the genomes of 50 bacteria to finished quality and is currently working on an additional 26 genomes. While each genome project has presented its own unique problems, they are now worked on according to a largely standardised pipeline. We monitor the progress of the shotgun for read and assembly quality and after its completion, we evaluate the assembly to estimate difficulty and likely timescale for finishing. Projects are submitted for automated pre-finishing and finished by the manual selection of samples for further reactions and the resolution of repeats. Finishing the more straightforward projects (those containing good quality reads and presenting few sequence problems) typically requires 2 months/Mb/finisher and an additional 5% of the number of shotgun reads are produced. More difficult projects take longer (up to 4 months/Mb/finisher) and need more finishing reads. A comparison of our recently finished bacterial genome projects to show the range of projects and the strategies used will be reported. Until recently high G+C projects have been difficult because of the presence of many small inverted sequence repeats (presumably forming stable hairpin structures in single stranded templates). These inverted repeats cause stops (sometimes called "hard stops") in standard BigDye terminator reactions and produce sequence gaps in assemblies of the reads. I will describe the use of an Amersham template amplification kit to resolve these problems.

FF046

Finishing Pipeline Developments at BCM-HGSC

Donna M. Muzny, Shannon P. Dugan, Yan Ding, Christian J. Buhay, Mike E. Holder, Aniko Sabo, Judith Hernandez, Huyen H. Dinh, Peter R. Blyth, Sandra L. Lee, Xiang Qin, Christie L. Kovar-Smith, George M. Weinstock, and Richard A. Gibbs

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

The BCM-HGSC has made significant improvements to our finishing pipeline through the application of informatics tools and finishing protocols. Central to our finishing process is the BAC-fisher tool that is part of the HGSC ATLAS genome assembler and has been adapted for the BAC based finishing. The BAC-Fisher tool has been used in the Mouse finishing project as well as all Rat finishing in QTL, ENCODE and gene dense BACs to pull in WGS reads based upon single read comparisons. The process has advantages to other methods utilizing WGS reads in that it does not rely upon the WGS assembly contigs for the identified reads. Other informatics developments include modifications to the Autofinish pipeline, Tiling Path tools and Submission reports.
Finally, current efforts are focused on utilizing the 454 technology for finishing. Combined assemblies using 454 contigs/Sanger reads are now being evaluated for performance in assemblies and finishing strategies.

Over the past year, new protocols have been developed at the HGSC to resolve difficult finishing regions. The main impact to the finishing pipeline has come from the implementation of the GE Sequence Finishing Kit. Implementation of this TempliPhi (Phi29) based template generation protocol had significant labor and cost savings and has decreased the number of requested special libraries (transponons and small insert libraries) by 70%. Another protocol development having a significant role in completing the Mouse chromosomes was direct BAC walking on conventionally prepared DNA or BAC DNA generated by the GE Finishing Kit. Current developments in this area include a low cost ($0.08-$0.10), high through-put production template prep for BACs and Fosmids that can be used for direct BAC based finishing. These protocol developments have a direct application to BAC finishing projects and WGS finishing where BAC templates and primer walks could be used as the primary finishing method. The above informatics tools and finishing protocols will be discussed as well as their role in the BCM-HGSC finishing pipeline.

FF058

The Joint Genome Institute and Challenges in Microbial Genomics

Patrick Chain

Lawrence Livermore National Lab and DOE Joint Genome Institute, Livermore, CA

The Department of Energy's Joint Genome Institute somewhat naively transitioned from their role in the human genome sequencing project to tackling microbial genomes of relevance to the DOE's core missions, namely carbon sequestration and cycling, bioremediation and bioenergy. After a period where little to no finishing was performed on microbial genome projects, the JGI has revisited its stance and has established a dedicated pipeline for microbial genomics which includes microbial genome finishing performed at three institutions. Despite its tentative entrée into the microbial sequencing arena and the recent implementation of finishing these genomes, the JGI now boasts ~100 completed microbial genomes along with another ~100 projects underway.

A general outline of the JGI microbial genomics partners and the finishing process will be discussed, along with some of the current (and future) challenges encountered while finishing. Also to be touched upon will be what is certain to be recurring themes throughout this meeting: the importance of completed genome sequences (contribution to understanding genome structure, allowing thorough genome comparisons like estimating rates of evolution, identifying polymorphisms of biological importance) and the intricacies in obtaining completed genomes (such as observed cloning bias – regions recalcitrant to cloning resulting in uncaptured/unspanned gaps, hard stop gaps, tandem and large repetitive elements).

FF049

Finishing and Improvement of Whole Genome Shotgun Sequenced Eukaryotes at SHGC/JGI

Jeremy Schmutz, Jane Grimwood, SHGC Finishing Group, SHGC Sequencing Group, SHGC Informatics Group and Richard M. Myers.

Stanford Human Genome Center / Joint Genome Institute, Palo Alto, CA 94304.

After we completed the finished human genome territory allocated to the Department of Energy, we began to explore scaling clone based finishing techniques to whole genome shotgun assemblies (WGSA). We conducted several large-scale experiments where we attempted to "treat" a whole genome shotgun assembly as a large clone and feed them through a modified form of our clone-based finishing pipeline. We also explored "target" based strategies where we only attempted to fix identified potential problem areas in the WGSA and treated the rest of the consensus sequence as accurate. We have now worked on enough genomes to identify four confounding factors that we find predictive of the difficulty of finishing or substantially improving WGSAs:

1. Polymorphism rate
2. Size and frequency of identical repetitive structures
3. High GC / difficult sequence content
4. Regions that do not clone in *E.coli*

While all genome sequences contain some amount of these confounding factors, the amount of effort necessary to finish a WGSA increases substantially with these factors.

We will present a series of vignettes about whole genome finishing and improvement covering a few of our major projects and explain how our finishing strategies have been modified to account for these confounding factors, sometimes with unexpected results. We will discuss our whole genome improvement/finishing pipeline for haploid and diploid genomes and how we have adapted clone-based experimental and computational finishing techniques to WGSAs. We will also introduce the new version of our assembly viewer ORCHID that has been updated to be scaffold friendly.

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FC02-99ER62873.

FF010

Automated Finishing at the Wellcome Trust Sanger Institute

Stuart McLaren

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

The Sanger Institute has a unique commitment to finish the Zebrafish Genome (Danio rerio), http://www.sanger.ac.uk/Projects/D_rerio/ , as well as involvement in other genomes and projects. As part of the high throughput approach all clone based projects undergo an automated first parse round of reactions after the initial shotgun phase. We support both robotic systems as well as established manual sequencing. The techniques are customised according to the genome strategy. The suite of software used combines Staden and in-house programmes and packages and all stages are tracked via the on site oracle database.

The pipeline is as follows;

- Project selected from production complete buffer
- DNA plates acquired and bar-code scanned into system
- Any relevant Whole Genome Shotgun data is screened and incorporated
- Software used to query problems and gaps in Gap4 (Staden) databases
- Derivative of primer3 used to chose oligo-nucleotide primers to sequence through problems and extend into gaps
- Oligo-nucleotides arranged into 96well plate format and automatically ordered through ftp site for external synthesis
- Plates returned bar-coded by synthesis company with primers dried in wells
- Individual map of DNA to primer is printed from bar-code
- Sequencing reaction performed according to genome strategy
- When all reactions for specific projects are returned it is then automatically re-assembled using Phrap (P. Green)
- These reactions are then QA'd and project is released to relevant groups for further analysis

This system has also been adapted for large Whole Genome Shotgun pathogen projects and for automated high throughput finishing of projects such as cDNA's and Open Reading Frames for which iterative steps are undergone until the entire insert is complete.

This automated finishing approach is performed in a dedicated team, http://www.sanger.ac.uk/Teams/Team58/ .

FF034

Quick Draft Assembly Improvement for Improved Finishing

Alex Copeland

Production Genomics Facility, Joint Genome Institute, Walnut Creek, CA

While prevention is a more cost effective means toward quality improvement than correction, in production sequencing, working with contaminated DNA samples is inevitable. Although we cannot always control the quality of the DNA we agree to sequence, we can control, to some extent, contamination in assemblies we pass along to finishing. We have noted that relatively modest improvements to assembler input sometimes makes large improvements in the resulting assemblies which has direct benefits for all users of the data, especially finishers.

As part of an effort to improve the quality of sequence data released by the Production Genomics Facility we created a process, and supporting software we call 'QD', to quickly identify and exclude low quality and contaminant reads from assemblies. We have applied this process to all historical microbial projects, and it is used on all new projects before they are sent for finishing.

The initial development of this process was driven by recognition that the JGI had draft sequenced a large number of microbes, whose data was not being utilized to its full potential by the research community, at least in part, due to the low quality of the assemblies. During development, we focused on the twin goals of keeping the process sufficiently simple as to make it possible to apply to all of our legacy projects, and doing no harm. The end result has made a significant difference in the quality of our assemblies and has simplified finishing these genomes.

Misc. Notes

Misc. Notes

Round Table Discussion Notes

Round Table Discussion Notes

Speaker Presentations (May 5th)

Abstracts are in order of presentation according to Agenda (page 3)

FF027 – Keynote

Assisted Assembly

Sante Gnerre

Broad Institute of MIT and Harvard, Cambridge, MA

The quality of an assembly of a given genome is generally correlated with the depth of coverage of the WGS sequence. However when high quality sequence from related organisms exist, such information can be used to substantially enhance an initial low coverage assembly.

We developed an "assisted assembly" methodology, which exploits conserved synteny to related organisms, by aligning the WGS reads of the genome we want to assemble onto the "reference" genome, and using the alignments to improve an initial de novo assembly. The information derived from such a process must be used cautiously: for example, undetected syntenic breaks would introduce misassemblies.

Here we discuss how assisted assemblies have been used as part of an NHGRI initiative to annotate the human genome: many low coverage mammalian genomes will be aligned to the human genome and used to identify conserved features such as genes and regulatory elements. To date we have sequenced and assembled nine mammals at 2X coverage, all females, using a mix of 4 Kb plasmids and 40 Kb Fosmid libraries. On average, assisting the initial intial de novo assembly resulted in an improvement of the total contig length of 20%, and in a four-fold increase of the N50 scaffold length.

The assisted assembly algorithms have also been useful for assembling low coverage Drosophilas, and the high coverage but AT rich genome Plasmodium Falciparum, strain HB3.

FF001

Dog Genome Improvement

FitzGerald, Michael, Abouelleil, A, Aftuck, L, Arachchi, H, Ayotte, L, Berlin, A, Brown, A, Cook, A, Gearin, G, Lindblad-Toh, K, Lui, A, Macdonald, P, Pirun, M, Priest, M, Russell, L, Shea, T, Sykes, S

The Broad Institute, Cambridge, MA, USA

Completely finished genomes have allowed us to answer a multitude of scientific questions. However, our ability to efficiently generate draft assemblies has exceeded our capacity to completely finish genomes. The community needs to improve upon the efficiency of the traditional tiling path based approach. We have an NHGRI approved plan to produce a near finished version of the dog genome with high structural integrity and near complete gene content. Considerable effort was expended to find targets most likely to improve the genome's utility, though CanFan2.0 is already at high quality representing ~99% of the euchromatic sequence. Analysis of coding content indicates that approximately 4000 genes have small sequence errors and that ~1000 exons are located within the 27,000 sequence gaps. We have targeted smaller gaps with an automated primer walk strategy using fosmid templates. Approximately 99.5% of the assembly is considered Arachne "certified", based on haplotype and linked read analysis. We will target up to 80Mb of traditional finishing on shotgun sequenced BAC and fosmid clones to resolve uncertified areas, including large gaps, potential haplotype differences and the ENCODE regions. This approach of targeted improvement will serve as a model for future genomes. The resulting assembly will allow dog to stand beside human and mouse as the third mammalian reference genome.

FF002

Challenges Encountered in Finishing BAC Sequences from >60 Vertebrate Species

R. W. Blakesley[1,2], N. F. Hansen[1], NISC Comparative Sequencing Program[1,2], G. G. Bouffard[1,2], and E. D. Green[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

In the process of finishing BAC sequences from >60 different vertebrates, a number of unexpected problems have been encountered. For example, BACs from owl monkey, colombus monkey, and shrew all showed higher than average strand bias in many stretches of 2 kb or more, causing up to 20% of the BAC sequences to be represented by reads from only one strand. Hedgehog BACs have been found to be very unstable when grown in large-volume cultures prior to DNA purification, deleting as much as 50-100 kb of their inserts. Dog BAC sequences often have a gap composed of very high GC content (up to 90%), requiring extraordinary sequencing effort to complete. Platypus BACs have been the most universally difficult, averaging three times as many sequence gaps per full-shotgun assembly and three times the number of unspanned gaps as BACs from other species; these gaps also average twice the median size as gaps from all other species examined. Finally, baboon BACs frequently contain repeated sequences that cause misassembly of the shotgun reads. Interestingly, some of these species-specific characteristics are problematic for comparative-grade sequence finishing (Blakesley *et al. Genome Res* 14, 2235-2244, 2004), while others affect the path toward human-grade finishing. Our experience confirms the notion that the nuances of sequence finishing vary from species to species because of the characteristic genomic architecture encountered in each case.

FF044

Simulating human finishing decisions with Autofinish

Cliff S. Han[1,2], Roxanne Tapia[1,2], Tom Brettin[1,2], Patrick Chain[1,3], Alla Lapidus[1]

[1]Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, California 94598, USA. [2]University of California, Los Alamos National Laboratory, Biosciences Division, M888, Los Alamos, NM 87545. [3]Lawrence Livermore National Laboratory, Biosciences Directorate, 7000 East Ave. L-441, Livermore, CA 94550

Shotgun sequencing has become the standard strategy in most genome sequencing centers, whether they involve small bacterial genomes or bigger eukaryotic ones. Compare to the rapid pace in shotgun sequencing, the finishing process is relatively slow. Autofinish program in Phred/Phrap/Consed package has been a great help in speed up the finishing task. However, the reactions picked by Autofinish are some time very different compare to those picked by human finishers. The cost of finishing process at our genome center, which involves Autofinish, is higher than those at the other genome center where no Autofinish are used. In order to reduce cost while keep the finishing speed of Autofinish, we did this simulation study in attempt to find the best parameters for Autofinish that can mimic human finishing decision. We acquired six bacterial genome projects finished at Lawrence Livermore National Laboratory and at Joint Genome Institutes at Walnut Creek. Start with shotgun reads only, finishing reactions were picked with Autofinish. Faked reads were generated for those reactions with quality scores from real sequencing data. The simulation will be stopped when a genome is finished or reaching maximal ten cycles. The research is ongoing and detailed results will be reported during the meeting.

FF012

High Performance Customizable Software for Finishing and Assembly Analysis

Andrew R. Zimmer, Sampath Settipalli, Jerome Naylor, and Toby Bloom

Broad Institute of Harvard and MIT

A particularly vexing issue posed by any high throughput finishing operation is data consistency. The parallelized "divide and conquer" nature of many finishing projects leads to data mutation, making coherent progress tracking extremely difficult.

Bluefin, the Broad Institute's next generation finishing and assembly software infrastructure, solves this problem by providing a single data source and flexible java APIs that allow disparate applications to access the same up-to-date data. Applications such as gap4 and primer3 operate harmoniously with other assembly analysis and reporting tools such as Cognos without interfering with the day-to-day operations of computer finishers and the finishing lab. In addition, because Bluefin can be linked in to various LIMS systems, it provides customizable algorithms for high throughput reagent selection and is capable of automatically integrating many types of finishing reads.

Bluefin is used for traditional BAC tiling path finishing, whole genome scale finishing, and "pick and choose"-style genome improvement projects. Bluefin keeps basepair-resolution coordinates that enable accurate, automatic merging of parallelized finishing operations.

Bluefin maintains structured assembly data in a query-able relational database, simplifying the development of assembly analysis tools, visualization applications, and automatic reagent selection.

FF035

Recent Advances and Future Directions of Genome Finishing at TIGR

Luke Tallon

The Institute of Genomic Research, Rockville, MD 20850

Traditionally a very manual phase of complete genome sequencing, finishing is an expensive and time-consuming process. TIGR continues to build tools and strategies to improve finishing efficiency. Each year, improved tools and techniques allow TIGR to finish genomes more quickly and at less cost than the previous year. Through the use of an in-house project management tool, finishing projects are divided into discrete tasks with managed progress and strategy that drives each successive round of tasks through project completion.

Using a growing number of finished genomes as alignment references, the efficiency of finishing related genomes can be improved by eliminating much of the effort to order and orient scaffolds and close inter-scaffold (physical) gaps. We use the MUMmer suite of tools for rapid alignment and prediction of scaffold order and orientation.

The most marked process improvement has been the automation of simple finishing tasks, such as resolving intra-scaffold (sequence) gaps, low coverage features, and simple repeat units. TIGR's autoCloser pipeline automates feature recognition, classification, primer design, clone selection, and reaction work-order generation. Coupled with laboratory robotics, we automate targeted resequencing of select clones with custom primers to resolve genome features with little hands-on effort.

Improved genome viewers have also assisted in the endeavor to better identify features and resolve them. We can project any feature (gap, coverage histogram, repeat units, ORFs, primer sites, etc.) onto contigs along with constituent reads and clones to allow detailed examination of the relationship of features to the contig and to each other. This assists the finisher in determining the best strategy for feature resolution.

As new non-clone based sequencing technologies become available, we will continue to improve finishing strategy and tools. TIGR is currently analyzing genomes sequenced using both traditional Sanger sequencing methods and 454 Life Sciences sequencing technology to determine the most efficient combination of sequencing and finishing that will produce high quality complete genomes faster and at a lower cost.

Improvements resulting from these advances and future directions for genome finishing at TIGR will be discussed.

FF055

Assisted genome closure and comparative genomic analysis with Optical Mapping

Colin W. Dykes[1]

[1]OpGen, Inc.

Whole-genome sequencing strategies typically involve bidirectional shotgun sequencing of variable-sized insert libraries followed by contig ordering and gap closure. The latter stages remain labor intensive and costly and are complicated by repetitive sequences and “unclonable” genomic segments. Optical Mapping – a technique for rapidly generating whole-genome, ordered restriction maps from single DNA molecules – addresses these issues by providing high-resolution chromosome scaffolds to which contigs are anchored and sorted for rapid genome finishing. By ordering contigs, sequence gaps are readily identified. Importantly, comparison of sequence contigs to Optical Maps finds misassemblies and other structural anomalies which obfuscate sequence finishing. Optical Maps have been constructed for scores of organisms, including mixed samples and microorganisms which are difficult to culture. The detail provided by Optical Mapping enables genome-wide comparisons between closely related organisms, identifying insertions, amplifications and other structural differences without any requirement for prior sequence information.

FF061

Integrating new technologies into the JGI microbial program

Paul Richardson

DOE Joint Genome Institute, Walnut Creek, CA

Misc. Notes

Misc. Notes

Round Table Discussion Notes

Round Table Discussion Notes

Poster Presentations

FF003

"Cloneview": An assembly viewer for microbial genome Finishing

S.Trong[1], E. Goltsman[1], S. Malfatti[2], P. Chain[2], and A. L. Lapidus[1]

[1] Joint Genome Institute Production Genomics Facility, Walnut Creek, CA
[2] Lawrence Livermore National Laboratory, Livermore, CA

The process of improving and completing (Finishing) draft genome assemblies using whole genome shotgun approach involves many complex iterative steps. The effectiveness of this process depends on many factors, such as the complexity of the genome, the quality of the DNA sequence and the accuracy of the assembled data. Although many efforts have been attempted to automate Finishing, it is still a relatively manual process. One such effort involves the identification and correction of regions that are incorrectly pieced together by the assembly program. To aid the finishers in visualizing and identifying these problematic areas, we have developed a visualization tool called Cloneview for displaying mate pair information in an assembly.

Our goal was to provide a graphical tool for displaying assembly information in a way that could be used to easily identify misassembled regions, repeat regions and possible biases in library creation.

Cloneview's ability to highlight read pairs that are inconsistently placed in the assembly allows the finishers to easily detect regions that are misassembled. By extracting the names of the reads to a file, further processing to correct the misjoined areas can be achieved with minimal effort. Other features of Cloneview include displaying read and clone depth, read pairs that span gaps, library-specific reads and repeat regions. In addition to the viewer, we have developed an accompanying software program to detect and report misassembled regions by looking for areas where violations in mate pairing predominately outweigh valid ones. This tool along with the viewer further enhances the finishers' ability to resolve misassemblies more efficiently.

The Microbial Finishing groups at the JGI /PGF and JGI/LLNL have incorporated these tools into their pipeline as part of an effort to rapidly finish microbial genomes.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.

FF006

The Genome Finishing Laboratory at the Broad Institute

Daniel Bessette, Nicole Allen, Mostafa Benamara, Chelsea Foley, Rakela Lubonja, Xiaohong Liu, Tashi Lokyitsang, Charles Matthews, Glen Munson, Tamrat Negash,Thu Nguyen, Keith O'Neill

The Broad Institute, Cambridge, MA, USA

The Broad Institute (rhymes with 'code'), a joint venture of MIT, Harvard, and the Whitehead Institute, has long been intimately associated with genome sequencing. The Sequencing Platform has many facets and functions but consists mainly of three core groups: Production Sequencing, Computer Finishing, and Laboratory Finishing. The Laboratory Finishing group will be the focus here.

Using a combination of high-throughput automation and small-scale custom techniques, the Laboratory Finishing group improves rough draft genomes by providing new sequencing data to fill in gaps and resolve conflicts.

Working in collaboration with the Computer Finishing and Production Sequencing teams, we fulfill work orders that include Fosmid and Plasmid Primer Walking, Fosmid and Plasmid Transposition, custom PCR, and Shatter Library creation. Built into these general workflows are special chemistries to address difficult-to-sequence regions.

This group of 12 people produces ~150,000 reads per month with an average NHGRI pass rate of 75%, which reflects the difficulty of the regions that we are working on.

There are a number of genomes currently in the lab including Tuberculosis, Neurospora, and Magnaporthe. Each genome presents unique challenges and we will present some of these issues.

FF007

QD Extension for Microbial Finishing

Benjamin Horowitz[1], Stephan Trong[1], Eugene Goltsman[1], Alex Copeland[1], Vasanth Singan[1], Steve Lowry[1], Stephanie Malfatti[2], Patrick Chain[2], Pat Kale[1], Alla Lapidus[1]

[1] Joint Genome Institute Production Genomics Facility, Walnut Creek, CA
[2] Lawrence Livermore National Laboratory, Livermore, CA

The Production Quality Control (QC) produces a series of assemblies, screening shotgun reads for contamination by vector sequence, for low-quality reads, and, in the case of microbial projects, for contamination by eukaryotic sequence. In coordination with the Microbial Finishing group lead at the PGF, an assembly that is "good enough" is named the Quality-controlled Draft (QD) assembly. Once the QD assembly is produced, the project passes from Production QC to one of the finishing groups at the JGI. During the finishing process, misassemblies are resolved, gaps between contigs are closed, and confirming reads are gathered for thinly covered areas.

Judging whether a project is ready for finishing is a complex task, and requires answering numerous questions, such as: Is sufficient read and clone coverage present? Will existing contig linking information (scaffolding information) allow the closing of gaps between contigs? Has high or low GC content contributed to difficult-to-sequence or difficult-to-clone areas? Do the libraries have unusual features (e.g., being absent from certain areas of the genome, or having widely-spread insert size distributions)? Are contig GC content distributions consistent with project GC content distribution? How many repeats are present in the genome? Have repeats contributed to misassemblies?

In order to facilitate the answering of these questions, we undertook the QD Extension project. QD Extension enhances the existing microbial finishing pipeline with results from the Phrap, PGA and Arachne assemblers, with scaffold information, and with coherent, informative web reports. By providing such additional information in a browsable, web-based format, we help the Microbial Finishing group lead at PGF more accurately and quickly determine whether a project is ready for finishing. This project facilitated the identification of a set of metrics that help to shed light on the finishing task that lies ahead of the biologist assigned to finish a prokaryotic genome.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.

FF009

Zebrafish Finishing at the Wellcome Trust Sanger Institute

Darren Grafham

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

In Spring 2001 the Sanger Institute started sequencing the 1.6GB genome of the zebrafish (*Danio rerio*) following two strategies: clone mapping and sequencing from BAC and PAC libraries and whole genome shotgun sequencing with subsequent assembly. http://www.sanger.ac.uk/Projects/D_rerio/ . Zebrafish is an important biological organism for embryology due to rapid development as most major organs are present at 24 hours, Imaging since translucency allows accessibility to studying developmental processes and genetically through forward and reverse genetics.

The genome is due for completion late 2007 and currently has 1.0Gb of finished sequence. The project has raised a number of questions regarding genome sequencing from multiple individuals and presented unique challenges for finishing in particular. The challenges within clones has included very large tandem arrays up to 100kb, almost impossible to sequence repeat elements such as the 407bp dr284 element which skips 350bp and sub-clone variation. At the chromosome level distinguishing haplotypes from duplications has required software development. Current progress and the issues faced will be discussed along with new tools developed to assist genome finishing.

FF011

Genomic Sequence Refinement for Comparative Analyses

J. Gupta[1], S. Y. Brooks[1], N. F. Hansen[1], 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

Several years ago, we investigated approaches for refining sequence assemblies of BAC-derived shotgun reads to an intermediate grade appropriate for multispecies sequence comparisons (Blakesley *et al. Genome Res* 14, 2235-2244, 2004). Since then, 'comparative-grade' sequence finishing has matured at NISC into a stable process. All BAC-derived shotgun sequences are routinely processed through several steps to refine the assemblies to produce comparative-grade finished sequence, characterized by: (1) typically eight-fold average redundancy in Q20 bases; (2) removal of low-quality sequence data; and 3) order and orientation of all contigs >2 kb in size, which is verified by independent data. Since most of the process employs computational tools (and rarely requires the generation of additional sequence reads), this refinement process costs significantly less (~10-fold) than the near-perfect sequence used for the human genome. Our analyses indicate that 99% of the sequence is present with <1 error per 10^4 bases, and half of the missing bases reside within simple repeats. To date, we have produced comparative-grade finished sequence for >5000 BACs. These have proven to be of recognizably improved value over full-shotgun (Phase 1) sequence assemblies for numerous studies, including the ENCODE Project.

FF015

An Overview of the Microbial Genome Finishing Procedures Utilized at the Joint Genome Institute - Los Alamos National Laboratory (JGI-LANL)

D. R. Sims[1], O. Chertkov[1], A.C. Munk[1], H. Kiss[1], E. Saunders[1], L.S. Thompson[1], P. Gilna[1], and C. S. Han[1]

Los Alamos National Laboratory, Los Alamos, NM

The mission of the JGI is to provide integrated high-throughput sequencing and computational analysis to enable genomic-scale/systems-based scientific approaches to DOE-relevant challenges in energy and the environment. One of the steps in accomplishing this mission is to produce a finished microbial genome sequence from drafted sequences. The most commonly used protocols employed by genome finishers at JGI-LANL are herein described. Methods: Upon receipt of a drafted genome, two cycles of lanlAutoFinish are run in which attempts are made to solve duplications and to close captured gaps. dupFinisher uses paired reads to resolve repetitive regions. When the repeats are solved, dupFinisher creates fake reads spanning the repeats. Additionally, dupFinisher selects primers designed to bridge unresolved repeats. Primer walks are used to resolve gap and coverage issues within scaffolds using a variety of chemistries. When a project has been resolved to no more than six scaffolds, PCR primers near the end of each scaffold are paired in all possible combinations to close uncaptured gaps. Hard GC stops, duplications and misassemblies are the most common challenges to the finishing process, each requiring a different approach. For hard stops in the sequence, finishers take a graduated approach from least to most expensive: DMSO, dGTP, and a commercial kit [Amersham Biosciences, PN# 25-6401-01]. If still unsuccessful, DNA and primers are sent to an outside entity for sequencing. Finally, a shatter library is requested. A powerful tool for solving difficult duplications is the transposon bomb. The Consed program is the primary computer aid in identifying and resolving misassemblies. Results: To date, over 30 complete microbial genomes have been finished using these methods, representing approximately 150 Mb of microbial DNA sequence with an average error rate nearing 0.0 per 10,000 bases. The median number of repeats solved per genome was 61. An average of 475 primer walk reactions were required per megabase of genome. Conclusion: JGI-LANL has developed an effective and efficient method for finishing bacterial genomes.

FF016

From Highly Repetitive Sequences to Unclonable Regions: Challenges to Finishing Extremely Difficult Microbial Genomes

O. Chertkov, H. Kiss, A.C. Munk, E. Saunders, D.R. Sims, L.S. Thompson, P. Gilna, C. S. Han

Los Alamos National Laboratory, Los Alamos, NM; Joint Genome Institute, Walnut Creek, CA

Sequencing and gene annotation are important tools for understanding how pathogens work in biomedical and environmental sciences and other research areas. High-quality draft sequences are useful, but finished genomes are required for detailed analysis of genome functions. Finishing efforts are sometimes complicated by highly repetitive and unclonable regions that resist standard techniques such as primer walks. We tried to resolve these problematic regions and came up with few new techniques. Methods: The random shotgun method was used in sequencing genomes (up to 8-10x coverage). Repetitive areas were resolved with DupFinisher (Cliff Han, unpublished) or with transposone bombs of bridging clones. Custom primer walks or PCR amplification closed gaps between contigs. Adaptor-PCR was used to close scaffold gaps not covered by drafted reads. Results: JGI-LANL has finished approximately 25 microbial genomes; all finished sequences have been uploaded into public databases. Some of these genomes have more than 100 repeats (e.g., *Trichodesmium erythraeum*, 254 repeats; *Methanosarcina barkeri*, 150 repeats), most of which were resolved by DupFinisher. *Alkaliphillus metalliredigenes* had 80 uncuptured gaps after draft stage, *Trichoderma reesei*-about 150. Unclonable regions were covered by reads generated from adaptor-PCR, whose longest PCR product measured 6 kb (*Alkaliphillus metalliredigenes*). Conclusion: Most bacterial genomes can be finished with our current technology. However, very high numbers of repeats (>500 repeats) and poor shotgun library coverage (from low-quality DNA) continue to interfere with efficient finishing.

FF017

Finishing genomes of closely related strains of *Shewanella*

H. Kiss, O. Chertkov, A. C. Munk, D. R. Sims, E. Saunders, G. C. Detter, L. S. Thompson, R. Tapia, T. S. Brettin, P. Gilna, C. S. Han

Los Alamos National Laboratory, Los Alamos, NM, Joint Genome Institute, Walnut Creek, CA,

Members of the metal-reducing genus *Shewanella* can grow almost anywhere and do not cause disease in any organism. Several species effectively reduce polyvalent metals and radionuclides and consequently are important for confining and cleaning up contaminated DOE sites. The Joint Genome Institute (JGI) has generated draft sequences for 10 *Shewanella* species; seven of these are being finished at JGI-LANL. Methods: In addition to standard finishing methods, such as closing captured gaps with primer walks and resolving duplications, we used the MUMer program package to help close uncaptured gaps. First, we generated a phylogenetic tree using the 16S rRNA from the seven *Shewanella* species. Then we used the nucmer program from the MUMmer program package to align the closest neighbors. The output of the nucmer program was used to generate two different kinds of graphical views (Mummerplot and MapView) of the alignment. Results: According to the generated phylogenetic tree, *Shewanella sp.* MR-4-0 and *Shewanella sp.* MR-7 are very closely related. Using the draft sequence of *Shewanella sp.* MR-4-0, which consisted of 10 contigs in one scaffold, we were able to order and orient four scaffolds in *Shewanella sp.* MR-7. This enabled us to use PCR to close the predicted scaffold gaps with a minimum time and effort. We were not able to draw a clear conclusion when we compared more distantly related species, because of too many genomic rearrangements between them. The genome size of the seven *Shewanella* species varies from 4.4 Mb to 5.2 Mb. The GC content ranges from 42% to 54%. Five of the seven *Shewanella* species being finished at JGI-LANL have been fully sequenced. Conclusion: The use of sequencing data from closely related organisms to determine the orientation and the location of contigs in a query sequence can significantly speed up the finishing of drafted genomes.

FF018

Whole Genome Shotgun Library Approach For Microbial Sequencing Projects at the JGI

Eileen M. Dalin, Doug Smith, Hope Tice, Kerrie Barry, David Bruce and Paul M. Richardson.

US Department of Energy Joint Genome Institute, Walnut Creek, California 94598 USA.

The US Department of Energy's Joint Genome Institute is a high-throughput sequencing center and user facility that has sequenced a large number of microbial genomes. The strategy for most projects calls for construction of whole genome shotgun libraries from high-molecular weight DNA isolated from an axenic culture. In general, the JGI produces 3 insert size-selected libraries for all whole genome shotgun projects. We generate a 3kb high-copy pUC18 library, an 8kb low-copy pMCL200 library, and a 40kb pCC1FOS fosmid library. The DNA is randomly sheared, fragments are end-repaired for blunt-end cloning, and then size selected on an agarose gel, extracted and purified. 3 & 8kb inserts are cloned into the appropriate vector and transformed into *E. coli.* 40kb inserts are cloned, packaged and infected by phage into *E. coli.* PCR using primers flanking the inserts are used to determine the percentage of clones with inserts for both the 3 and 8kb libraries, before proceeding to production sequencing. Clones (10-384-well plates) from each of the 3 & 8kb libraries are initially sequenced and library quality is assessed at this stage before full sequencing is completed. Both 3 & 8kb libraries are sequenced to 4x sequencing coverage and the 40kb library is sequenced to 30x clone coverage. The 3 library approach generally results in more complete genome coverage at the draft stage, and pairing information allows for contig order and orientation and repeat resolution in the sequence. Finishing using standard methods is also facilitated by this approach.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, and by the University of California.

FF020

Arcturus: an assembly data management system

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

The Wellcome Trust Sanger Institute Hinxton Cambridge UK

In recent years, the Sanger Institute's Pathogen Sequencing Unit (PSU) has expanded its sequencing programme to include protozoan pathogens and model organisms. This movement towards sequencing and finishing larger genomes (>20Mb) using a whole genome shotgun approach, has highlighted the need to manage large amounts of shotgun and assembly data in a flexible manner. In order to achieve this, Arcturus, an assembly data management system is being developed.

Arcturus is based upon the popular open-source MySQL database system. The core software is written in Perl and consists of a set of scripts which allows Arcturus to be integrated into the existing assembly pipeline and to interact with tools such as Gap4.

A novel feature of the core system is that each contig formed by joining existing contigs is automatically linked to its “parent” contigs in the database. This allows tracking of every contig's “family history” and it enables gene annotations and tagged features to be re-mapped to each new generation of contigs.

A graphical user interface is currently under development. This is a Java application which will enable finishers and project managers to visualise the current state of an assembly. It will include an interactive scaffolding tool which will utilise the information from short range inserts (e.g. pUC and pMAQ read-pairs) and longer-range information (e.g. from BAC or fosmid end sequences), which it will then display graphically.

Future plans for Arcturus include integration with PSU annotation tools and with other Sanger Institute databases such as FingerPrint Contig (FPC). Arcturus will give our finishers the flexibility to work on large genomes, our biologists the ability to track and update gene annotations on unfinished genomes, and our project managers the ability to track the cost and progress of sequencing projects.

FF021

A High Throughput Bovine Full Length cDNA Sequencing Pipeline

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

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

Full length cDNA (fl-cDNA) sequence finishing can be time consuming and expensive due to the manual work involved by highly skilled finishers. To reduce cost and accelerate fl-cDNA finishing, we have implemented a high throughput automated fl-cDNA sequence finishing pipeline. The sequence finishing pipeline is designed to import any number of reads to the MySQL data base. The sequenced reads obtained using vector primers are assembled using PHRAP and checked for errors in an automated fashion. The automated checks are performed for contiguity, high quality base discrepancies, low quality bases, chimerism, single strand coverage, 5' and 3' cloning tag absence as well as the integrity of the tag, mis-rearray (contamination, clone picking error), problems in the DNA purification process, and absence or presence of open reading frame (ORF). Clones which fail these checks are flagged and processed appropriately. Depending on the nature of the problem(s), fl-cDNA clones are either eliminated from the finishing pipeline, manually checked, or subjected to a further round of finishing by primer walking, using Primer3[1] in an automated fashion to design primers. The regions requiring manual intervention are also detected by our automated pipeline and conveniently navigated for manual editing or further processing. Finished fl-cDNA clones are aligned to the 6.2X bovine genome in an automated fashion where, besides serving as a source of annotation, they may serve to help improve the genome assembly or may be identified as chimeras. By using our automated system we were able to finish and annotate 4124 distinct fl-cDNA transcripts and completed over 300,000 EST's since May 2005.

References:

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

FF023

Microbial Finishing at DOE JGI/PGF: Sequencing Difficult DNA Templates

Michele Martinez, Paul Richardson, and Alla Lapidus

US Department of Energy Joint Genome Institute, Walnut Creek, California 94598 USA.

The US DOE Joint Genome Institute's (JGI) mission is to provide the scientific community with high-quality finished genomes. Approximately 300 microbial genomes are currently in the JGI pipeline and to date, 79 have been completed. The objective of the Microbial Finishing laboratory is to process sequencing reactions in order to increase the quality of reads, to close physical gaps and/or sequence gaps. Most genomes contain complex regions which are difficult to sequence with standard protocols, so the laboratory must use a multitude of techniques specialized for each project. Problematic regions include, for example, GC-rich areas, hairpin loops, homopolymer stretches, and tandem repeats of variable length. Gap closer in such regions is expensive as well as time-consuming, since it requires extensive troubleshooting strategies. Approaches include, optimizing reaction conditions, applying various sequencing chemistries, and additional manual editing. In an effort to greatly reduce the amount of areas requiring special analysis, genomes with >65% GC content are processed with a four step-approach: 5% DMSO, Sequence Finishing Kit (SFK), PCR, and shatter libraries. As a result of this strategy, JGI/PGF's Microbial Genome Finishing Group has been able to complete a number of complex microbial projects, such as, Frankia (~75% GC-rich) and Thermobifida fusca (~68% GC rich).

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. W-7405-ENG-36.

FF024

Microbial Finishing at PGF/JGI

E. Goltsman, M. Martinez, V. Singan, S. Lowry, S. Trong, B. Horowitz, P. Kale, A. Copeland, P. Richardson, A. L. Lapidus.

DOE Joint Genome Institute, Walnut Creek, CA.

DOE JGI is involved in number of Microbial programs: U.S. DOE Microbial Genome Programs, GTL, Community Sequencing Program. Each program includes sequencing, finishing and a detailed analysis of genomes of the different representatives of the microbial world. The completion of the draft sequence is followed by the finishing effort, which includes identification and resolution of misassemblies and repetitive regions, gap closure, and improvement of low-quality regions. JGI's goal is to produce finished sequences for all the prokaryotic projects in the JGI pipeline.

The finishing strategy at the JGI's PGF and LLNL facilities start with automated repeat resolution using software developed in-house to correct misassembled regions followed by automated primer design using Consed's Autofinish to extend contigs and resolve low quality areas. Next, we perform an additional round of Autofinishing and manually tackle repeat resolution in difficult regions. Once the genome has been correctly assembled, we design experiments to close the remaining gaps. Our last phase, polishing, ensures that the final error rate is <1 per 50 Kb with a minimum of 2x depth coverage throughout the assembly. We have developed a finishing software system to automate many of the finishing tasks and are continually improving our pipeline to meet our demanding goals.

During the last two years, the combined efforts of the three groups (PGF, LANL, LLNL) allowed us to finish 80 microbial projects (28 project are done by PGF, Walnut Creek). Our goal is to fulfill the needs of all of the projects undertaken by JGI and to make the finishing step less time consuming by developing and implementing new strategies and approaches.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.

FF025

Assembler Complementation Tool

Vasanth Singan, Stephan Trong, Eugene Goltsman, Alla Lapidus

Joint Genome Institute – Production Genomics Facility,
2800 Mitchell Dr, Walnut Creek, CA-94598

A plethora of genome assemblers exist, that use various techniques to generate the consensus. In most cases, no one specific assembler can give the ideal assembly. Recent comparison on assemblers showed that assemblers that perform well for certain genomes do not necessarily perform in the same manner with all the projects. In order to facilitate finishers with choosing the optimal assembly, we propose a tool to help identify an initial draft assembly from multiple assemblers so that the finishers can choose the best assembly to start with.

The Assembler Complementation tool is a graphical visualization program that will not only serve to decide the best assembly but also help to collectively use the best consensus for comparative finishing. This tool will aid in identifying misassemblies, gaps, and rearrangements in assemblies relative to one another. Gaps can be visually identified by comparing a fragmented assembly against a better assembled one and conversely; Misassemblies can be identified by comparing a conserved assembly against a less stringent one.

The contigs of the two assemblies to be compared are blasted against one another and based on the reference genome selected, for each of the contig in the reference genome, hits from contigs of the query genome are shown in a tiling view format with a different color for each contig. Information regarding the region of misassemblies, hits, etc can also be obtained interactively.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.

FF026

Testing fidelity of assembly, binning and gene calling using synthetic metagenomic datasets

Mavromatis K[1], Goltsman E[1], Barry K[1], Shapiro H[1], Korzeniewski F[1], Rigoutsos I, Salamov A[1], Hugenholtz P[1], Kyrpides N[1]

Joint Genome Institute. 2800 Mitchell Drive Walnut Creek, CA 94598
IBM TJ Watson Research Center, Yorktown Heights, New York 10598

It is well known that draft isolate genomes contain misassemblies and incorrect gene calls that are identified and corrected during finishing and annotation QC respectively. Draft metagenomes also contain misassemblies and bad gene calls that are compounded to an unknown degree by the presence of multiple species and strains. Since it is currently unfeasible to progress metagenomes beyond draft assemblies (with the possible exception of dominant populations) quantifying the extent and character of misassemblies and miscalls is a useful exercise.

We constructed synthetic metagenomic datasets to mimic a number of real metagenomic datasets by combining reads from a selection of 113 isolate genome sequencing projects available through the Joint Genome Institute.

Isolate genomes were selected to represent populations in metagenomic datasets based on similar patterns of genome size, GC content and phylogenetic position. Reads were randomly sampled from the selected genomes to match the read depth of their corresponding populations in the metagenomic assemblies. Sampled reads were then assembled and annotated using availabe programs and same parameters used to assemble and annotate real metagenomic data. The extent of chimeric assembly (assembly of multiple genomes into the same contig) could then be quantified since the source of each read in a given contig was known.

The effect of multiple genomes vs single genomes on ab initio gene calling could also be assessed. An additional benefit of having synthetic metagenomic datasets for which the identity of all contigs is known is to provide benchmarks for binning methods. The results of this analysis will be presented and discussed in relation to the use of the data, e.g. metabolic reconstruction vs population structure.

FF029

454 Assemblies of Microbial and Mammalian Genomes

Christian J. Buhay, Donna M. Muzny, Alicia C. Hawes, Xiang Qin, Peter R. Blyth, Huyen H. Dinh, Sandra L. Lee, Lynne V. Nazareth, Christie L. Kovar-Smith, Joseph, Huaiyang Jiang, Erica Sodergren, Michael E. Holder, 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. 454 Life Sciences has introduced an alternate sequencing method using emulsion-based methods of isolation and amplification, and an instrument that performs pyro-sequencing in pico liter sized wells. At the BCM-HGSC, we are evaluating sequencing and assembly strategies to integrate the 454 sequencing technology into our Sanger sequencing and finishing pipelines.

Microbial 454 sequencing and assemblies have resulted in assemblies comparable toor better than that of Sanger sequencing with N-50 contig sizes of 25 Kb. To evaluate how well the 454 sequencing technology would perform on mammalian genomes, we sequenced a pool of 10 rat BAC clones and 10 Human BAC clones using 454 methods. 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 assembly coverage to be about 90%, however the N-50 contig length was low at 3-5Kb. Further investigations of utilizing the 454 and Sanger reads in mixed assemblies are in progress. Initial results have shown that in mixed assemblies using 454 contigs and Sanger reads, the 454 contigs have been shown to close gaps. Optimization curves have been developed to determine the optimal number of Sanger clones necessary to close all gaps and finish in a combined 454/Sanger mixed assembly. Conventional finishing tools such as Autofinish can then be used to close gaps. These 454/Sanger finishing techniques have direct applications in microbial finishing as well as upgrading mammalian genome regions.

FF037

Resolving Repetitive Regions with DupFinisher

E. Saunders, O. Chertkov, H. Kiss, A. C. Munk, D. Sims. L. S. Thompson, P. Gilna, S.C. Han

JGI-Los Alamos, Los Alamos National Laboratory, Los Alamos, NM, USA

Background: Genome finishers at LANL receive bacterial projects containing only shotgun reads. To finish these genomes, it is necessary to close gaps and verify repeats, both potentially time- and labor-intensive processes. The DupFinisher (C. Han, unpublished) software tool has been developed to streamline some of these finishing tasks. DupFinisher can automatically detect repetitive regions, assemble each repeat individually using paired draft reads and primer walk reads, check the quality of these subassemblies, create artificial joins for finished and properly assembled repeats, and run automated gap closure scripts on unfinished subassemblies. Methods: DupFinisher consists of two components: a parameter file and a PERL program. DupFinisher uses BLAST (Altschul et al, 1997) to automatically detect repeats, and then takes subclone and primer walk reads from unique areas on one side of a repeat and finds the pair for each unique read chosen. These reads are collected and stored. If at least 3 common reads are found in another set of reads, they are combined into a project called Dupxxx. The program continues in this manner through all the contigs. The output from DupFinisher includes individual assemblies for each repeat and for unpaired repetitive ends. These subassemblies are collected in a single directory called dfRun#. This directory also contains three text files that summarize the results, a directory containing consensus sequences of the repeats, and a directory containing files of finishing reactions for unfinished repeats.

Results: We have found that DupFinisher correctly resolves 70-90% of the duplications in a bacterial genome. Two automated rounds of DupFinisher can decrease the number of contigs by 22-24%.

Conclusions: DupFinisher is a powerful tool for genome finishing which can substantially decrease the time and effort needed to finish genomes which contain many repeats.

FF039

Ensuring Transparent and Consistent High Quality in cDNA Finishing

Diana Palmquist, Peiming Huang, Brian Wynhoven, Elizabeth Chun, Robert Kirkpatrick, Johar Ali, Asim Sidiqqui, Robert Holt, Marco Marra, Steven Jones.

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

Creating high quality reference sequences is a critical backbone of the genomics world. The genome sequencing efforts have established specific rules by which a "gold standard" of finished sequence is achieved [1]. This standard was logically extended to guide finished sequence generation for expressed genes [3,4]. The differing nature of cDNA finishing however, results in some ambiguity in the application of this standard. Often there exist additional rules which are applied to cDNA sequence finishing that are not reported in the final analysis, but speak to the increased complexities of cDNA finishing. Together with the limitation of the genomic driven tool development, these factors result in multiple interpretations of high quality.

To address the concern in the variability of quality definitions, and to better protect our sequence sets from error, we have examined multiple problems which can plague cDNA projects. Typical problems include biological errors such as clone contamination, procedural errors such as clone misrearray and sequence errors such as base miscalls. Through the development of standards, customized tools and consistent application practices we have minimized these risks, enabling us to produce full-length cDNA sequence sets of transparent and uniform high quality.

Here we report on errors and sources of quality variability which may occur, suggest strategies to mitigate these, and illustrate an accurate and efficient method of obtaining high quality cDNA sequence. In a field where there is little opportunity for direct peer review of data [2] and production driven pressure to rapidly produce and release this data, such detailed attention to quality consistency and error avoidance is essential. With vigilant attention to these issues, both the amount and the reliability of the data derivable from these sequence sets will be greatly improved.

References

http://www.genome.gov/10000923; Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project - September 7, 2001

Felsenfeld, A., Peterson, J., Schloss, J., Guyer, M. 1999. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* 9:1-4.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S. et al. 2002. A Drosophila full-length cDNA resource. *Genome Biology. 3(12):* research0084.1–84.20.

Strausberg, R.L., Feingold, E.A., Klausner, R.D., Collins, F.S. 1999. The Mammalian Gene Collection *Science.* 286(5439): 455-457

FF040

Useful Tools for Finishing Difficult Microbial Genomes

A. C. Munk, O. Chertkov, H. Kiss, E. H. Saunders, D. R. Sims, L. S. Thompson, L. J. Meincke, Y. C. Rogers, P. Gilna, C. S. Han

Los Alamos National Laboratory, Los Alamos, NM, USA

Two challenges faced in finishing microbial genomic sequences are strings of repeats and large numbers of uncloned regions (scaffold gaps).

Strings of similar repeats are often resolved incorrectly by assembly software. MergeTBpairs.pl software was written to help resolve these repeats in transposon-bombed clones. Transposon sequences are inserted randomly into clone DNA using Epicentre's EZ::TN^{TM} <KAN-2> Insertion Kit. Transposons are inserted such that a 9 bp direct repeat from the subclone flanks the transposon on either side. Bombed subclones are sequenced with universal primers that extend in both directions from the transposon; both reads begin with a 9 bp overlap. The mergeTBpairs.pl program detects this overlap and converts the two reads into a single read, thereby doubling the read length from ~600 to ~1200 bp. Longer reads generated by mergeTBpairs.pl help the assembly program resolve repeats in *Shewanella denitrificans.*

Connecting scaffolds across gaps lacking clone links can require many expensive PCR reactions. LANL's adaptor-PCR method can reduce the number of PCR reactions required. Microbial genomic DNA is digested with a restriction enzyme that leaves a 5' overhang and results in the bulk of fragments ranging in size from 5 to 10 kb. A chimeric adaptor, with 5' sequence complementary to the overhang and 3' sequence complementary to the adaptor primer, is ligated to the genomic restriction fragments. PCR is done using the adaptor primer and a custom primer from the scaffold end. PCR products are end sequenced; if >1500 bp, they are subcloned and sequenced. Adaptor-PCR reduces the number of scaffold gaps from 36 to 26 or fewer in *Alkaliphillus metalliredigenes.* Scaffold gaps that remain after adaptor-PCR can be tackled with a reduced number of pairwise PCR reactions or with another round of adaptor-PCR using a different set of enzymes.

FF041

Genome Sequence Analysis of *Cyanothece* sp. ATCC 51142, a Unicellular Nitrogen-fixing Cyanobacterium.

Eric A. Welsh[1], Louis A. Sherman[2], Himadri B. Pakrasi[1]

[1]Department of Biology, Washington University, St. Louis, MO 63130
[2]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907

The unicellular, diazotrophic cyanobacterium, *Cyanothece* sp. ATCC 51142 is an interesting organism, in that it has the ability to perform both oxygenic photosynthesis and nitrogen fixation within the same cell. Oxygenic photosynthesis is detrimental to nitrogen fixation due to the high oxygen sensitivity of nitrogenases, the enzymes involved in nitrogen fixation. Several filamentous cyanobacterial strains solve this problem by separating the two processes in space, by forming specialized heterocyst cells at intervals along the filaments. *Cyanothece*, a unicellular organism, solves this problem by separating the processes in time. *Cyanothece* exhibits a robust circadian rhythm, with cyclic behavior evident at all levels, from cellular morphology to RNA and protein expression levels. *Cyanothece* 51142 is the target organism for a system biology level Membrane Biology Grand Challenge project funded by EMSL/PNNL-DOE.

The genome of *Cyanothece* is currently being sequenced at the Washington University Genome Sequencing Center. Preliminary analysis predicts ~5600 protein coding genes, ~43% of which have been observed via high throughput proteomics. An annotation pipeline has been developed to auto-annotate the genes and provide as much useful information as possible for the final manual annotation and curation process. We will discuss this annotation pipeline and the website developed to make this data easily accessible and interpretable.

This project was funded by the Danforth Foundation Funds at Washington University, and the membrane Biology 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.

FF042

Novel Bioinformatics Methods for Troubleshooting of Genomic Shotgun Data.

Eugene Goltsman[1], Randal Cox[2], Michael Mazur[3], Alla Lapidus[1], Alex Copeland[1]

[1] Joint Genome Institute, Walnut Creek, CA
[2] Department of Biochemistry and Molecular Genetics; University of Illinois at Chicago, Chicago, IL 60607.
[3] Integrated Genomics, Chicago, IL

End-sequencing of shotgun libraries of small genomic inserts is, by far, the most popular approach to Whole Genome Sequencing (WGS) today. Irregularities in WGS datasets present assembly problems that are expensive and time-consuming to solve, with cloning bias, contamination and long repeats posing the biggest challenges. Shotgun assembly data exhibit well recognizable patterns that follow certain statistical models and deviations from these models usually stem from flaws and abnormalities in the input data, which, in turn, reflect problems in the cloning protocol, chemistries, or in the DNA being sequenced. We developed several statistical and bioinformatic methods for detecting cloning bias, DNA contamination and high repeat content at early stages of the WGS project. These methods are based on analyses of a) depth of coverage distributions, b) progressive assembly dynamics and c) GC composition distribution of real and simulated shotgun datasets. We identify and describe relationships between coverage (in terms of read depth and number of gaps), and the binomial/Poisson function, and demonstrate ways to routinely identify cloning bias and contamination by relying on these relationships. Differences in GC composition between different genomes, libraries and even plates allowed us to identify cases of suspected contamination by identifying bimodal patterns in the GC distribution in the sequences of a genomic project. Routine automated application is also discussed.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.

FF045

UPGRADING THE RAT GENOME

Shannon P. Dugan-Rocha, Donna Muzny, Aniko Sabo, Yan Ding, Christian J. Buhay, Mike E. Holder, Alicia C. Hawes, Ziad M. Khan, George M. Weinstock and Richard Gibbs

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

At the BCM-HGSC we are developing and testing strategies for upgrading a number of genomes, including the Rat genome. Upgrading the Rat genome enhances overall data quality and utility by improving the assembled genome sequence in various regions of interest. One method for upgrading the genome that is currently being explored utilizes tools built to identify and tag genes within a Consed assembly.

To upgrade the Rat genome we have concentrated on gene rich BACs for finishing, with emphasis on annotation and manual curation of targeted regions. Approximately 300 gene rich Rat BACs along with 300 Rat BACs corresponding to QTLs were chosen and added to our production and finishing pipelines. The mRNAs associated with a subset of these BACs were obtained using information from the UCSC browser and their sequence placed onto our BAC sequence data using BLAT. The appropriate hits were then translated into exon tags within the individual Consed assemblies and labeled with the corresponding gene name and basepair position. Finishing of the BACs was completed to Phase3 or Phase2 “comparative” standard with the ultimate goal of completing the gene regions for annotation.

Additional features being considered include incorporating human and mouse ortholog tags and utilizing an existing BCM tool, Genboree, to graphically place and view any “unfinished” gene regions within a corresponding BAC clone.

Implemetation of these strategies ultimately facilitates high quality finishing of gene rich regions and the value of that upgraded data will be determined by comparison to manual curation standards achieved with Human Chromosomes 3 and 12.

FF047

454 Sequencing for Gap Closure in Microbial Genome Assemblies

Joseph Alessi, Hector Garcia, Falk Warnecke, Ed Kirton, Phil Hugenholtz, Paul Richardson and Feng Chen.

DOE Joint Genome Institute, 2800 Mitchell Dr. B400, Walnut Creek, CA 94598

Most microbial genome finishing projects at the Joint Genome Institute require the use of a multitude of molecular techniques to achieve a finished genome. The majority of these techniques are applied to sequencing through gaps in the assembly. Traditional shotgun sequencing is known to have difficulty in both cloning of A/T rich regions and sequencing of G/C rich regions. To help alleviate this problem we have applied the 454 sequencing platform as another tool for gap closure. Although 454

Sequencing has been shown to have difficulty with homopolymer stretches of nucleotides, it does not have the same biases as shotgun sequencing. Therefore we feel these two approaches together can be complementary. We have developed a protocol in which gap-spanning fosmids are pooled together, from one or more projects. This DNA is sequenced with 454 and assemblied using the Newbler Assembler and the resulting contigs are added into their respective projects. One 60x60 picotiter chip can yield as much as 32 Mb, allowing for the pooling of 20-28 fosmids with and average read depth of 20-28X. We chose fosmids that spanned gaps in a single microbial genome assembly as well as gap spanning fosmids within a given scaffold of a metagenomic sequencing project. We are also applying 454 technology to whole genome shotgun sequencing to assist with poorly assembled projects from Sanger sequencing alone.

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. W-7405-ENG-36.

FF052

Implementing Project Management at the Joint Genome Institute

Lynne Goodwin[1], Kerrie Barry[2], Alex Copeland[2], David Bruce[1]

[1]Los Alamos National Laboratory, Los Alamos, NM USA,
[2]Joint Genome Institute, 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. Implementation of a formal project management system 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, and Lawrence Livermore National Laboratories. Historically the JGI had a small number of large projects. Today, a large fraction of the JGI sequencing capacity is dedicated to small projects (< 10 MB) such as microbial genomic and environmental metagenomic projects. JGI projects are currently managed using a loosely coordinated set of databases, spreadsheets, text files and E-mail communications along with frequent meetings to address problems, changes and updates. To better organize these efforts, the JGI is adapting project management principles based on the industry standard Project Management Body of Knowledge (PMBOK, Project Management Institute).

The PMBOK model outlines uniform procedures for project management. The PMBOK emphasizes integration of cost, time, scope, and quality management during the project life cycle (planning, initiation, specification, execution, monitoring, change control, and close out). PMBOK principles were developed for construction projects and must be altered substantially for managing multiple simultaneous sequencing projects.

The JGI is prototyping PMBOK adaptations for a group of projects sequenced in FY2006 with full implementation planned in FY2007. Goals include improved change control processes, scheduling, stakeholder communication, and more accurate cost reporting.

FF059

Microbial Genome Finishing at Lawrence Livermore National Laboratory

Lisa Vergez and Patrick Chain

Lawrence Livermore National Lab and DOE Joint Genome Institute, Livermore, CA

As part of the US DOE's Joint Genome Institute (JGI), the Lawrence Livermore National Laboratory finishing group's goal is to complete a subset of the genomes drafted at the Production Genomics Facility (PGF). Our process begins upon receiving ≥8X coverage of high quality draft sequence for each microbial genome project. Two rounds of automated primer design experiments are processed on each genome for the purpose of gap closure, ambiguity resolution and polishing low quality sequence regions. After incorporating the data into assemblies, repeat and uncaptured gap resolution is performed. We utilize scripts that automate the step of solving spanned repeats, and rely on manual verification to solve more complex repeats. For example, in the genome *Burkholderia vietnamiensis*, one particular repeat was found approximately 55 times neighboring other repetitive regions or even interrupting other repeat elements. Additionally, a ca.25kb family of large repeats was occasionally interrupted by another 6 kb repeat. Other problems that require manual intervention include hard stop gaps (where sequencing abruptly stops), uncaptured gaps, tandem repeats, polymorphisms, homopolymeric regions, and are typically resolved using primer walking, PCR, and specialized sequencing reactions. Our approach to finishing and examples of problem regions within select genomes will be discussed.

FF065

Finishing GC-Rich Microbial Genomes

L.S.Thompson, O. Chertkov, H. Kiss, A.C. Munk, E. H. Saunders, D. R. Sims, P. Gilna, C. S. Han

Los Alamos National Laboratory, Los Alamos, NM.

The Finishing Team at Los Alamos, part of the JGI Microbial Genomics, has encountered difficulty in finishing sequencing projects involving GC-rich organimsms. These genomes with a GC content of greater than 60% contain regions of unique secondary structure which interfere with cycle sequencing reactions. When encountered, these stable secondary structures create a pause or stop in the sequencing reaction and cause a huge decrease in signal. These secondary structures are in the form of a hairpin structure which is formed when two identical yet reversely oriented sequences in the denatured single strand DNA fold back on each other. Together these inverted repeats are called a palindrome and may contain a loop consisting of unpaired bases between the identical sequences. Our graded approach to solving these regions is to try different chemistries in the following order: add DMSO, use dgtp chemistry, use the Amersham Sequence Finishing Kit, and finally send the template and primer to Sequetech for analysis. After examining these hairpin regions, we propose a sequencing strategy to follow during production sequencing of GC-rich organisms.

FF066

The *Azotobacter vinelandii* Sequencing Project

Nancy Miller[1], Phil Latreille[1], Jing Lu[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.

Misc. Notes

Misc. Notes

Misc. Notes

Misc. Notes

Misc. Notes

Attendees

Joe Alessi
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
JBAlessi@lbl.gov

Johar Ali
BC Cancer Agency
Genome Sciences Centre, Vancouver, BC
jali@bcgsc.ca

Kerrie Barry
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
KWBarry@lbl.gov

Daniel Bessette
The Broad Institute - MIT, Cambridge, MA

Bob Blakesley
NIH Intramural Sequencing Center (NISC), Bethesda, MD
rblakesl@nhgri.nih.gov

Gerry Bouffard
NIH Intramural Sequencing Center (NISC), NHGRI, Rockville, MD
bouffard@mail.nih.gov

Thomas Brettin
Los Alamos National Laboratory (JGI), Los Alamos, NM
brettin@lanl.gov

Jim Bristow
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
JBristow@lbl.gov

Shelise Brooks
NIH Intramural Sequencing Center (NISC), NHGRI, Rockville, MD
sbrooks@mail.nih.gov

David Bruce
Los Alamos National Laboratory (JGI), Los Alamos, NM
dbruce@lanl.gov

Christian Buhay
Baylor College of Medicine - HGSC, Houston, TX
cbuhay@bcm.tmc.edu

Patrick Chain
Lawrence Livermore National Laboratory (JGI), Livermore, CA
chain2@llnl.gov

Olga Chertkov
Los Alamos National Laboratory (JGI), Los Alamos, NM
ochrtkv@lanl.gov

Michael Chin
Sequetech Corporation, Mountain View, CA
mchin@sequetech.com

Alex Copeland
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
accopeland@lbl.gov

Eileen Dalin
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
e_dalin@lbl.gov

Chris Daum
DOE Joint Genome Institute (LLNL), Walnut Creek, CA

Chris Detter
Los Alamos National Laboratory (JGI), Los Alamos, NM
cdetter@lanl.gov

Norman Doggett
Los Alamos National Laboratory (JGI), Los Alamos, NM
doggett@lanl.gov

Shannon Dugan-Rocha
Baylor College of Medicine - HGSC, Houston, TX
sdugan@bcm.tmc.edu

Colin Dykes
OpGen, Inc., Madison, WI

Mike FitzGerald
The Broad Institute - MIT, Cambridge, MA
fitz@broad.mit.edu

Bob Fulton
Washington University Genome Sequencing Center, St. Louis, MO
bfulton@watson.wustl.edu

Paul Gilna
Calit2, University of California San Diego, San Diego, CA
pgil@lanl.gov

Sante Gnerre
The Broad Institute - MIT, Cambridge, MA
sante@gnerre.com

Eugene Goltsman
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
egoltsman@lbl.gov

Lynne Goodwin
Los Alamos National Laboratory (JGI), Los Alamos, NM
lynneg@lanl.gov

Darren Grafham
The Wellcome Trust Sanger Institute, Hinxton, Cambridge
dg1@sanger.ac.uk

Jyoti Gupta
NIH Intramural Sequencing Center (NISC), NHGRI, Bethesda, MD
jyotig@mail.nih.gov

Cliff Han
Los Alamos National Laboratory (JGI), Los Alamos, NM
han_cliff@lanl.gov

David Harris
The Wellcome Trust Sanger Institute, Hinxton, Cambridge
deh@sanger.ac.uk

John Henkhaus
OpGen, Inc., Madison, WI
jhenkhaus@opgen.com

Ben Horowitz
DOE Joint Genome Institute (LLNL), Walnut Creek, CA
horowitz2@llnl.gov

Pat Kale
DOE Joint Genome Institute (LLNL), Walnut Creek, CA
kale1@llnl.gov

Hoda Khouri
The Institute for Genomic Research (TIGR), Rockville, MD
hkhouri@tigr.ORG

Hajnalka Kiss
Los Alamos National Laboratory (JGI), Los Alamos, NM
hajkis@lanl.gov

Alla Lapidus
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
alapidus@lbl.gov

Heiko Liesegang
Georg-August University of Göttingen, Goettingen, Germany

Susan Lucas
DOE Joint Genome Institute (LLNL), Walnut Creek, CA
lucas11@llnl.gov

Steve Lowry
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
slowry@lbl.gov

Michele Martinez
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
MLMartinez@lbl.gov

Kostas Mavrommatis
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
KMavrommatis@lbl.gov

Stuart McLaren
The Wellcome Trust Sanger Institute, Hinxton, Cambridge
sm2@sanger.ac.uk

Nancy Miller
Monsanto Company, St. Louis, MO
nancy.m.miller@monsanto.com

Pat Minx
Washington University Genome Sequencing Center, St. Louis, MO

Karen Mungall
The Wellcome Trust Sanger Institute, Hinxton, Cambridge
klb@sanger.ac.uk

Chris Munk
Los Alamos National Laboratory (JGI), Los Alamos, NM
cmunk@lanl.gov

Donna Muzny
Baylor College of Medicine - HGSC, Houston, TX
donnam@bcm.tmc.edu

Himadri Pakrasi
Washington University, St. Louis, MO
pakrasi@biology2.wustl.edu

Diana Palmquist
BC Cancer Agency - Genome Sciences Centre, Vancouver, BC
dianap@bcgsc.ca

Julian Parkhill
The Wellcome Trust Sanger Institute, Hinxton, Cambridge
parkhill@sanger.ac.uk

Keith O'Neill
The Broad Institute - MIT, Cambridge, MA
koneill@broad.mit.edu

Gary Resnick
Los Alamos National Laboratory, Los Alamos, NM
resnick@lanl.gov

Paul Richardson
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
PMRichardson@lbl.gov

Liz Saunders
Los Alamos National Laboratory (JGI), Los Alamos, NM
ehs@lanl.gov

Jeremy Schmutz
Stanford Human Genome Center (JGI), Palo Alto, CA
jeremy@paxil.stanford.edu

Lou Sherman
Purdue University, West Lafayette, IN
lsherman@bilbo.bio.purdue.edu

Martin Shumway
The Institute for Genomic Research (TIGR), Rockville, MD
shumwaym@tigr.ORG

David Sims
Los Alamos National Laboratory (JGI), Los Alamos, NM
dsims@lanl.gov

Vasanth Singan
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
VRSingan@lbl.gov

Axel Strittmatter
Georg-August University of Göttingen, Goettingen, Germany
astritt@gwdg.de

Sean Sykes
The Broad Institute - MIT, Cambridge, MA
ssykes@broad.mit.edu

Luke Tallon
The Institute for Genomic Research (TIGR), Rockville, MD
ljtallon@tigr.ORG

Sue Thompson
Los Alamos National Laboratory (JGI), Los Alamos, NM
thompson_l_sue@lanl.gov

Stephan Trong
DOE Joint Genome Institute (LLNL), Walnut Creek, CA
trong1@llnl.gov

Kiryl Tsukerman
Integrated Genomics Inc., Chicago, IL
kiryl@integratedgenomics.com

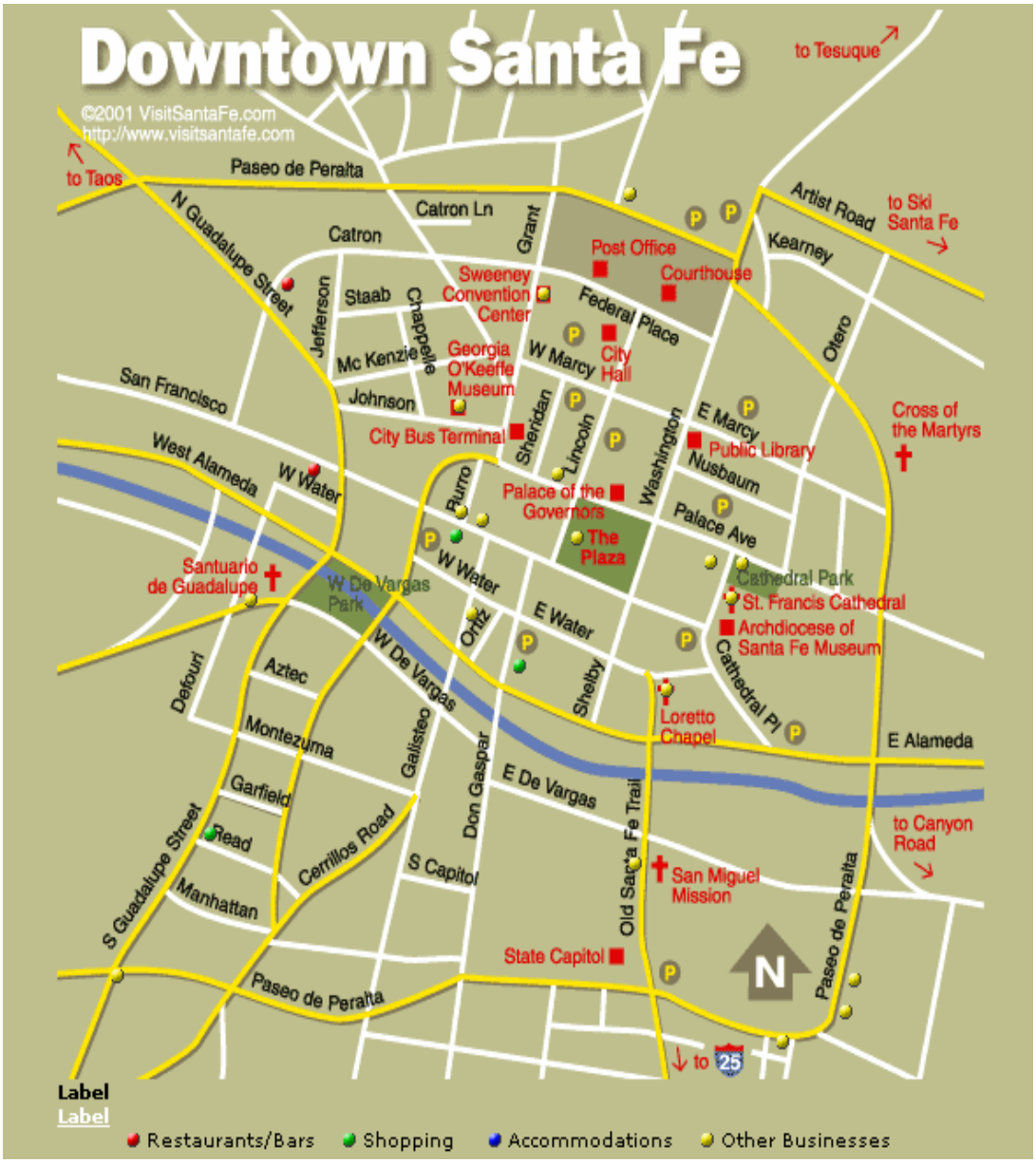
Lisa Vergez
Lawrence Livermore National Laboratory (JGI), Livermore, CA
vergez1@llnl.gov

Eric Welsh
Washington University, St. Louis, MO
ewelsh@ccb.wustl.edu

Li Weng
DOE Joint Genome Institute (LBNL), Walnut Creek, CA
lweng@lbl.gov

Andrew Zimmer
The Broad Institute - MIT, Cambridge, MA
zim@mit.edu

Map of Santa Fe, NM



History of Santa Fe, NM

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

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

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

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

Preconquest and Founding
(circa 1050 to 1607)

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

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

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

Settlement Revolt & Reconquest
(1607 to 1692)

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

Established Spanish Empire
(1692 to 1821)

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

The Mexican Period
(1821 to 1846)

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

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

Territorial Period
(1846 to 1912)

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

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

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

Statehood
(1912 to present)

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

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

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

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

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