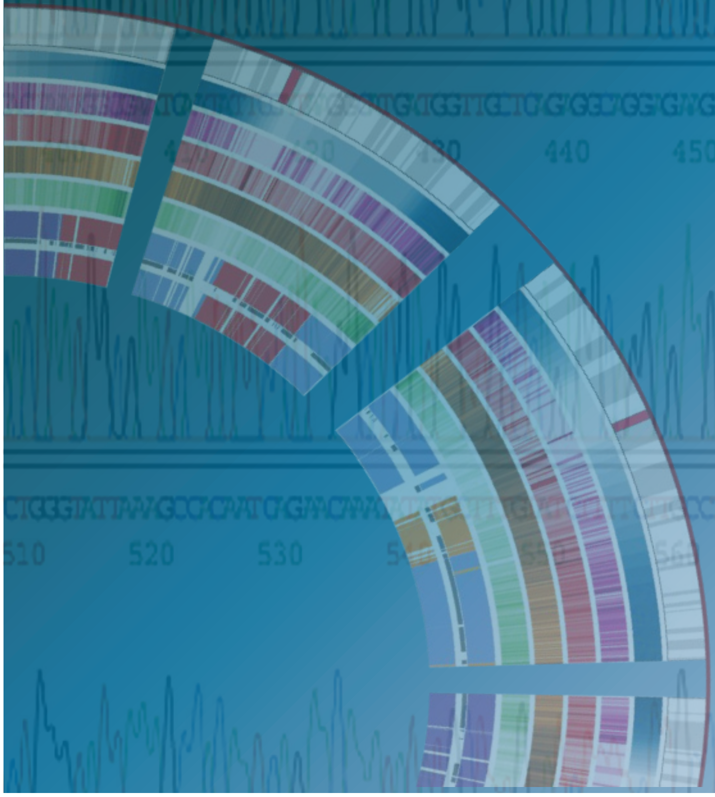
7th Annual Sequencing, Finishing, Analysis in the Future Meeting



Santa Fe, New Mexico
June 5-7, 2012

Office of Science
U.S. DEPARTMENT OF ENERGY

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Los Alamos
NATIONAL LABORATORY
EST. 1943

Contents

The 2012 "Sequencing, Finishing and Analysis in the Future" Organizing Committee:

** Chris Detter, Ph.D., JGI- LANL, Genomics Center Director, LANL*
** Johar Ali, Ph.D., Cancer Genomics Team Leader, OICR*
** Patrick Chain, Metagenomics Team Leader, LANL*
** Michael FitzGerald, Finishing Manager, Broad Institute*
** Bob Fulton, M.S., Sequence Improvement Group Leader, WashU*
** Tina Graves, Team Leader, WashU*
** Darren Grafham, Team Leader Illumina Bespoke Team, Sanger Institute*
** Alla Lapidus, Ph.D., Director Bioinformatics, IPM, FCCC*
** Donna Muzny, M.S., Director of Operations, BCM*
** Yu-Hui Rogers, VP Core Technology Development, JCVI*
** Nadia Fedorova, Team Leader Genome Finishing & Analysis, JCVI*

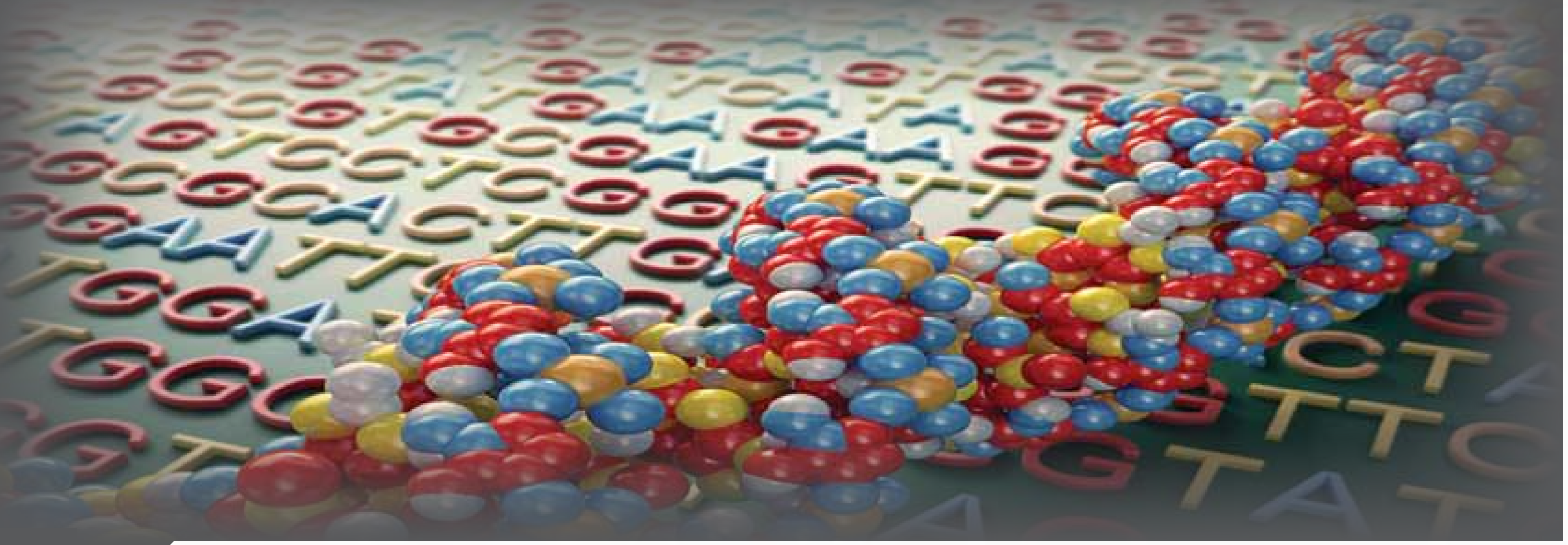
06/05/2012 - Tuesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**Breakfast**	**x**	**La Fonda Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	**x**	Welcome Intro from Los Alamos National Laboratory	**TBD**
x	Session Chair	**x**	Session Chairs	Chair - Johar Ali Chair - Donna Muzny
8:45 - 9:30	**Keynote**	**FF0032**	**Towards the Perfect Genome Sequence**	**Dr. George Weinstock**
9:30 – 9:50	Speaker 1	**FF0119**	Building the DOE Systems Biology Knowledgebase	**Tom Brettin**
9:50 - 10:10	Speaker 2	**FF0106**	Genome Sequencing of a Mapping Population Reveals Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen Phytophthora capsici	**Joann Mudge**
10:10 – 10:30	**Break**	**x**	**Beverages and Snacks Provided**	**Sponsored by OpGen**
10:30 – 10:50	Speaker 3	**FF0159**	Assembling with Longer Reads and Higher Depths	**Jim Knight - Roche**
10:50 – 11:10	Speaker 4	**FF0122**	Next Generation Sequencing Improvements	**Haley Fiske - illumina**
11:10 – 11:30	Speaker 5	**FF0239**	Using the Ion Torrent PGM for *de novo* Sequencing	**Tim Harkins - LifeTech**
11:30 – 11:50	Speaker 6	**FF0038**	Exploiting Single-Molecule Real-Time DNA Sequencing for Improved Genome Assembly and Methylome Analysis	**Steve Turner - PacBio**
11:50 - 12:40	**Panel Discussion**	**x**	**Next Generation Sequencing Technology Panel Discussion**	**Chair - Bob Fulton** **Chair - Patrick Chain**
12:40 – 2:00pm	**Lunch**	**x**	**Coronado Lunch Buffet**	**Sponsored by illumina**
x	Session Chair	**x**	Session Chairs	Chair - Alla Lapidus Chair - Bob Fulton
2:00 – 2:20	Speaker 7	**FF0060**	Ion Torrent Semiconductor Sequencing Allows Rapid, Low Cost Sequencing of the Human Exome	**David Jenkins**
2:20 – 2:40	Speaker 8	**FF0209**	En Route to the Clinic: Diagnostic Sequencing Applications Using the Ion Torrent	**Donna Muzny**
2:40 – 2:55	Speaker 9	**FF0047**	Next Generation Sequencing: Possible Application for Forensic DNA Analysis. What does the Person of Interest Look Like?	**Tom Callaghan**
2:55 – 3:10	Speaker 10	**FF0136**	Forensic DNA Standards for Next Generation Sequencing Platforms	**Pete Vallone**
3:10 – 3:30	**Break**	**x**	**Beverages and Snacks Provided**	**Sponsored by OpGen**
3:30 - 5:50pm	**Tech Time Talks** **(15 min each)**	**FF0149**	Challenges in Genomic Cloud Computing	**Daniel Bozinov**
		FF0126	NGS for the Masses: Empowering Biologists to Improve Bioinformatic Productivity	**Kashef Qaadri**
		FF0070	The PerkinElmer Omics Laboratory	**Todd Smith**
		FF0120	The Best Finish First: Sequence Finishing with Whole Genome Mapping	**Deacon Sweeney**
		FF0299	High Throughput Plasmid Sequencing with Illumina and CLC bio	**Ajay Athavale**
		FF0213a	Engineered Polymerases Provide Improved NGS Library Amplification and Enable Novel Sequencing Applications	**Maryke Appel**
		FF0144	Beyond Basic Target Enrichment: New Tools to Fuel Your NGS Research	**Jennifer Carter**
		FF0019	Better Computing for Better Bioinformatics	**George Vacek**
		FF0296	RAPID: Ultra High Throughput Sequencing Data Analysis for Quick Microbial Identification	**Robert Yamamoto**
6:00 – 7:30pm	**Posters - even #s Meet & Greet Party**	**EVEN #s**	**Poster Session with Meet & Greet Party (Sponsored by Roche)** **Food & Drinks**	**Sponsored by Roche** **6:00pm- 9:00pm**
7:30 - 9:00pm	**Posters - Odd #s Meet & Greet Party**	**ODD #s**	**Poster Session with Meet & Greet Party (Sponsored by Roche)** **Food & Drinks**	**Sponsored by Roche** **6:00pm- 9:00pm**
9:00 - bedtime	**on your own**	**x**	**Night on Your Own - Enjoy!!!**	**x**

06/06/2012 - Wednesday

Time	Type	Abstract #	Title	Speaker
7:30 - 8:30am	**Breakfast**	**x**	**Santa Fe Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	X	Welcome Back	**TBD**
x	Session Chair	x	Session Chairs	Chair - Mike Fitzgerald Chair - Tina Graves
8:45 - 9:30	**Keynote**	**FF0043**	**Plague: A Highly Fit Clonal Pathogen Emerges and Shapes Human History**	**Dr. Paul Keim**
9:30 - 9:50	Speaker 1	**FF0101**	Finishing and Special Motifs: Lessons Learned From CRISPR Analysis Using Next Generation Draft Sequences	**Catherine Campbell**
9:50 - 10:10	Speaker 2	**FF0160**	An Analysis of the Genomic Architecture at Risk Loci for SLE	**Ward Wakeland**
10:10 - 10:30	Speaker 3	**FF0279**	Resolve the Cancer Heterogeneity by Single Cell Sequencing	**Xun Xu**
10:30 - 11:00	**Break**	**x**	**Beverages and Snacks Provided**	**x**
11:00 - 11:20	Speaker 4	**FF0075**	Consed and BamView for Next-gen Sequencing	**David Gordon**
11:20 - 11:40	Speaker 5	**FF0065**	Integrating Data from Multiple Human Genome Sequencing Platforms and Bioinformatic Methods to Analyze their Error Profiles and Form Consensus Variant Calls	**Justin Zook**
11:40 - 12:00	Speaker 6	**FF0088**	One Chromosome, One Contig: Hybrid Error Correction and *de novo* Assembly of Single-Molecule Sequencing Reads	**Sergey Koren**
12:00 - 1:20pm	**Lunch**	**x**	**New Mexican Lunch Buffet**	**Sponsored by Beckman Coulter**
x	Session Chair	x	Session Chairs	Chair - Donna Muzny Chair - Johar Ali
1:20 - 1:40	Speaker 7	**FF0108**	Recent Advances in High-Throughput, Low-Latency Interfacing for Fast Scanning and Metrology in Genomics Applications	**Scott Jordan**
1:40- 2:00	Speaker 8	**FF0186**	Pilon Assembly Improvement Software	**Bruce Walker**
2:00 - 2:20	Speaker 9	**FF0188**	Putting the Pieces Together: From Assembly to Analysis	**Sean Sykes**
2:20 - 2:40	Speaker 10	**FF0045a**	Finding the Perfect Recipe for *de novo* Plant Genome Assembly: A Platform Bake-off	**Dan Ader**
2:40 - 3:00	Speaker 11	**FF0170**	Finished Prokaryotic Genome Assemblies From a Low-Cost Combination of Short and Long Reads	**Shuangye Yin**
3:00 - 3:20	Speaker 12	**FF0211**	Mercury: A Next Generation Sequencing Data Analysis and Annotation Pipeline	**David Sexton**
3:20 - 3:35	Speaker 13	**FF0004**	NCGR Informatics	**John Chow**
3:35 - 3:50	Speaker 14	**FF0174**	DTRA Algorithm Prize	**Christian Whitchurch**
3:50 - 5:15pm	**Break** **&** **Round Table** **Discussion** **(Topics TBD)**		**Beverages and Snacks Provided for the Round Table** Topics TBD: attendees to select from a few choices the week before the meeting	**x**
5:45 - 8:00pm	**Happy Hour**	**x**	**Happy Hour at Cowgirls Cafe - Sponsored by LifeTech - Map Will be Provided**	**Sponsored by LifeTech**
8:00 - bedtime	**on your own**	**x**	**Dinner and Night on Your Own - Enjoy!!!**	**x**

06/07/2011 - Thursday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**Breakfast**	**x**	**Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	**x**	Welcome Back	**Chris Detter**
x	Session Chair	**x**	Session Chairs	Chair - Patrick Chain Chair - Nadia Fedorova
8:45 - 9:30	**Keynote**	**FF0042**	**Environmental Reservoirs of Human Pathogens: The Vibrio cholerae Paradigm**	**Dr. Rita Colwell**
9:30 – 9:50	Speaker 1	**FF0185**	A Rapid Whole Genome Sequencing and Analysis System Supporting Genomic Epidemiology	**Mike FitzGerald**
9:50 – 10:10	Speaker 2	**FF0173**	Endosymbiont Hunting in the Metagenome of Asian Citrus Psyllid (*Diaphorina citri*)	**Surya Saha**
10:10 – 10:30	Speaker 3	**FF0221**	SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing	**Glenn Tesler**
10:30 – 10:50	**Break**	**x**	**Beverages and Snacks Provided**	**x**
10:50 – 11:10	Speaker 4	**FF0263**	Assembly of Large Metagenome Data Sets Using a Convey HC-1 Hybrid-Core Computer	**Alex Copeland**
11:10 – 11:30	Speaker 5	**FF0034**	Metagenomic Assembly: Challenges, Successes, and Validation	**Matt Scholz**
11:30 – 11:50	Speaker 6	**FF0208**	Metagenomics for Etiological Agent Discovery	**Matthew Ross**
11:50 – 12:10	Speaker 7	**FF0207**	Nearly Finished Genomes Produced Using Gel Microdroplet Culturing Reveals Substantial Intraspecies Diversity within the Human Microbiome	**Michael Fitzsimons**
12:10 - 1:30pm	**Lunch**	**x**	**La Fiesta Plaza Lunch**	**Sponsored by Agilent**
x	Session Chair	**x**	Session Chairs	Chair - Mike Fitzgerald Chair - Alla Lapidus
1:30 - 1:50	Speaker 8	**FF0006**	Rapid Phylogenetic and Functional Classification of Short Genomic Fragments with Signature Peptides	**Ben McMahon**
1:50 - 2:10	Speaker 9	**FF0229**	PanFunPro: Pan-Genome Analysis Based on the Functional Profiles	**Oksana Lukjancenko**
2:10 - 2:30	Speaker 10	**FF0114**	Preparation of Nucleic Acid Libraries for Personalized Sequencing Systems Using an Integrated Microfluidic Hub Technology	**Kamlesh Patel**
2:30 - 2:50	Speaker 11	**FF0142**	Capturing Native Long-Range Contiguity by *in situ* Library Construction and Optical Sequencing	**Jerrod Schwartz**
2:50 - 3:10	Speaker 12	**FF0256**	Fosmid Cre-LoxP Inverse PCR Paired-End (Fosmid CLIP-PE), A Novel Method for Generating Fosmid Pair-End Library	**Ze Peng**
3:10 - 3:30	Speaker 13	**FF0282**	Automated Sequencing Library Preparation and Suppression for Rapid Pathogen Characterization	**Todd Lane**
3:30 - 3:50	Speaker 14	**FF0109**	Evaluation of Multiplexed 16S rRNA Microbial Population Surveys Using llumina MiSeq Platform	**Julien Tremblay**
3:50 - 4:00pm	**Closing Discussions**	**x**	**Closing Discussions for General Meeting - Discuss Next Year's Meeting**	**Chair - Chris Detter**
		x	**Reminder for those interested there is a special Forensic's Session Friday from 8:00am - 12:30pm**	**x**

06/08/2011 - Friday			**Forensic Friday**	
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**x**	**x**	**Breakfast on your own**	**x**
8:30 - 8:35	Intro	**x**	Welcome Intro - Session Chair (LANL)	**Cathy Cleland**
8:35 - 8:45	Intro	x	Welcome Intro - Session Chair (US Army)	**Ken Kroupa** **Jeff Salyards**
8:45 – 9:05	Speaker 1	**FF0047**	Next Generation Sequencing; Possible Application for Forensic DNA Analysis. What does the Person of Interest Look Like?	**Tom Callaghan**
9:05 – 9:25	Speaker 2	**FF0136**	Forensic DNA Standards for Next Generation Sequencing Platforms	**Pete Vallone**
9:25 – 9:45	Speaker 3	**FF0216**	Short Tandem Repeat (STR) Analysis from Short Read Sequencing Data	**Daniel Bornman**
9:45 – 10:05	Speaker 4	**FF0191**	High Sensitivity Detection and Typing of Mixed Contributor DNA Samples Using Massively-Parallel Deep Amplicon Pyrosequencing	**Jared Latiolais**
10:05 – 10:20	**Break**	**x**	**Break**	**x**
10:20 – 10:40	Speaker 5	**FF0114**	Preparation of Nucleic Acid Libraries for Personalized Sequencing Systems Using an Integrated Microfluidic Hub Technology	**Ken Patel**
10:40 – 11:00	Speaker 6	**FF0223**	Forensic Genomics using Next Generation Sequencing by Synthesis (SBS)	**Cydne Holt**
11:00 – 11:20	Speaker 7	**FF0153**	A Highly Configurable SNP Caller for the Ion Torrent Personal Genome Machine	**Christian Buhay**
11:20 - 11:40	Speaker 9	**FF0248**	Short Tandem Repeat Sequencing on the 454 Platform	**Melissa Scheible**
11:40 - 12:00	Speaker 10	**FF0280**	STR Profiling From Personal Genomes: Happy Surprises	**Yaniv Erlich**
12:00 - 12:20	**Closing Panel Discussions**	**x**	**Panel Discussion on Forensic Applications of Next Generation Sequencing (US Army and LANL)**	**Jeff Salyards** **Chris Detter**
12:20 - 12:30	x	**x**	**Thank you**	**Cathy Cleland**



06/05/2012 - Tuesday				
Time	Type	Abstract #	Title	Speaker
7:30 - 8:30am	Breakfast	x	La Fonda Breakfast Buffet	Sponsored by NEB
8:30 - 8:45	Intro	x	Welcome Intro from Los Alamos National Laboratory	TBD
x	Session Chair	x	Session Chairs	Chair - Johar Ali Chair - Donna Muzny
8:45 - 9:30	Keynote	FF0032	Towards the Perfect Genome Sequence	Dr. George Weinstock
9:30 – 9:50	Speaker 1	FF0119	Building the DOE Systems Biology Knowledgebase	Tom Brettin
9:50 - 10:10	Speaker 2	FF0106	Genome Sequencing of a Mapping Population Reveals Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen Phytophthora capsici	Joann Mudge
10:10 – 10:30	Break	x	Beverages and Snacks Provided	Sponsored by OpGen
10:30 – 10:50	Speaker 3	FF0159	Assembling with Longer Reads and Higher Depths	Jim Knight - Roche
10:50 – 11:10	Speaker 4	FF0122	Next Generation Sequencing Improvements	Haley Fiske - illumina
11:10 – 11:30	Speaker 5	FF0239	Using the Ion Torrent PGM for *de novo* Sequencing	Tim Harkins - LifeTech
11:30 – 11:50	Speaker 6	FF0038	Exploiting Single-Molecule Real-Time DNA Sequencing for Improved Genome Assembly and Methylome Analysis	Steve Turner - PacBio
11:50 - 12:40	Panel Discussion	x	Next Generation Sequencing Technology Panel Discussion	Chair - Bob Fulton Chair - Patrick Chain
12:40 – 2:00pm	Lunch	x	Coronado Lunch Buffet	Sponsored by illumina
x	Session Chair	x	Session Chairs	Chair - Alla Lapidus Chair - Bob Fulton
2:00 – 2:20	Speaker 7	FF0060	Ion Torrent Semiconductor Sequencing Allows Rapid, Low Cost Sequencing of the Human Exome	David Jenkins
2:20 – 2:40	Speaker 8	FF0209	En Route to the Clinic: Diagnostic Sequencing Applications Using the Ion Torrent	Donna Muzny
2:40 – 2:55	Speaker 9	FF0047	Next Generation Sequencing: Possible Application for Forensic DNA Analysis. What does the Person of Interest Look Like?	Tom Callaghan
2:55 – 3:10	Speaker 10	FF0136	Forensic DNA Standards for Next Generation Sequencing Platforms	Pete Vallone
3:10 – 3:30	Break	x	Beverages and Snacks Provided	Sponsored by OpGen
3:30 - 5:50pm	Tech Time Talks (15 min each)	FF0149	Challenges in Genomic Cloud Computing	Daniel Bozinov
		FF0126	NGS for the Masses: Empowering Biologists to Improve Bioinformatic Productivity	Kashef Qaadri
		FF0070	The PerkinElmer Omics Laboratory	Todd Smith
		FF0120	The Best Finish First: Sequence Finishing with Whole Genome Mapping	Deacon Sweeney
		FF0299	High Throughput Plasmid Sequencing with Illumina and CLC bio	Ajay Athavale
		FF0213a	Engineered Polymerases Provide Improved NGS Library Amplification and Enable Novel Sequencing Applications	Maryke Appel
		FF0144	Beyond Basic Target Enrichment: New Tools to Fuel Your NGS Research	Jennifer Carter
		FF0019	Better Computing for Better Bioinformatics	George Vacek
		FF0296	RAPID: Ultra High Throughput Sequencing Data Analysis for Quick Microbial Identification	Robert Yamamoto
6:00 – 7:30pm	Posters - even #s Meet & Greet Party	EVEN #s	Poster Session with Meet & Greet Party (Sponsored by Roche) Food & Drinks	Sponsored by Roche 6:00pm- 9:00pm
7:30 - 9:00pm	Posters - Odd #s Meet & Greet Party	ODD #s	Poster Session with Meet & Greet Party (Sponsored by Roche) Food & Drinks	Sponsored by Roche 6:00pm- 9:00pm
9:00 - bedtime	on your own	x	Night on Your Own - Enjoy!!!	x

NOTES

Speaker Presentations (June 5th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0032

Towards the Perfect Genome Sequence

George Weinstock

The Genome Institute at Washington University

In the 17 years since the first bacterial genome sequence appeared, there have been enormous advances in DNA sequencing, informatics, and related technologies. Yet a typical microbial genome assembly does not look that much different from the first genomes that were done, containing sequence errors, gaps, misassemblies. Assembly software still puts repeats and other ambiguous sequences aside and leaves gaps in the sequence. What will it take to routinely produce "closed" genomes or to remove the base sequence errors, misassemblies, and other defects without the large investment in manual finishing? What strategies should be used to deal with repeats and other special situations?

NOTES

FF0119

Building the DOE Systems Biology Knowledgebase

Tom Brettin[1], Rick Stevens[2]

[1]Oak Ridge National Laboratory
[2]Argonne National Laboratory, University of Chicago

This presentation will step the audience through the development of the DOE Systems Biology Knowledgebase (KBase), a large-scale development project led by Argonne, Berkeley, Brookhaven and Oak Ridge National Laboratories, and includes participation by Cold Spring Harbor Laboratory and multiple university partners. Started in 2011, the KBase project is building the first multi domain systems biology knowledge base aimed at advancing predictive biology in microbes, microbial communities and plants. The KBase project is integrating data from many existing sources, building tools and services that will support complex workflows enabling modeling of microbes, reconciling experimental data with computational predictions, and providing a large-number of computational services that go beyond existing integrated biological databases. KBase will be deployed on a purpose-built infrastructure spanning four laboratories that collectively house multiple petabytes of data, and that will support scalable computing resources on both cloud and cluster environments. End users will be able to access many thousands of public genomes and related datasets for microbes. They will also gain access to tens of thousands of metagenomic samples and dozens of plant genomes and phenotype datasets. In addition to providing web and programmatic interfaces to these data, the KBase will enable users to upload their own private data and virtually integrate it with the public datasets for comparative analysis and development of models. The KBase is aiming to enable collaborative workflows and multiple ways of sharing. The KBase development team is integrating resources such as MicrobeOnline, The SEED, RAST, Model SEED, MG-RAST and other systems into a coherent user-oriented computing environment with a unified API. The first public release of the system is targeted for February 2013.

FF0106

Genome Sequencing of a Mapping Population Reveals Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen *Phytophthora capsici*

Joann Mudge[1]†, Kurt H. Lamour[2]†, Daniel Gobena[2], Oscar P. Hurtado-Gonzales[3], Jeremy Schmutz[4,5], Alan Kuo[4], Neil A. Miller[6], Brandon J. Rice[1], Sylvain Raffaele[7], Liliana Cano[7], Arvind K. Bharti[1], Ryan S. Donahoo[8], Sabra Finley[2], Edgar Huitema[9,10], Jon Hulvey[11], Darren Platt[4], Asaf Salamov[4], Alon Savidor[12], Rahul Sharma[13-15], Remco Stam[9,10], Dylan Storey[2], Marco Thines[13-15], Joe Win[7], Brian J. Haas[16], Darrell L. Dinwiddie[6,17], Jerry Jenkins[4,5], James R. Knight[18], Jason P. Affourtit[18], Cliff S. Han[19], Olga Chertkov[19], Erika A. Lindquist[4], Chris Detter[19], Igor V. Grigoriev[4], Sophien Kamoun[7], Stephen F. Kingsmore[6,17]

[1]National Center for Genome Resources, Santa Fe, NM 87505, USA.
[2]University of Tennessee, Department of Entomology and Plant Pathology, Knoxville, TN 37996, USA.
[3]Pioneer Hi-Bred International, Johnston, IA 50131, USA.
[4]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA.
[5]Hudson Alpha Institute of Biotechnology, Huntsville, AL 35806, USA.
[6]Children's Mercy Hospital, Kansas City, MO 64108, USA.
[7]The Sainsbury Laboratory, Norwich NR4 7UH, UK.
[8]University of Florida, IFAS-SWFREC, Immokalee, FL 34142, USA.
[9]Division of Plant Science, University of Dundee, Invergowrie, Dundee DD2 5DA, UK.
[10]Plant Pathology Program, James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK.
[11]University of Massachusetts-Amherst, Dept. of Plant, Soil, and Insect Sciences, Amherst, MA 01003, USA.
[12]Department of Molecular Biology and Ecology of Plants, Tel Aviv University, Tel Aviv 69978, Israel.
[13]Biodiversity and Climate Research Centre, D-60325 Frankfurt am Main, Germany.
[14]Senckenberg Gesellschaft für Naturforschung, D-60325 Frankfurt am Main, Germany.
[15]Goethe University, Department of Biological Sciences, Institute of Ecology, Evolution and Diversity, D-60323, Frankfurt am Main, Germany.
[16]Broad Institute, Cambridge, MA 02142, USA.
[17]School of Medicine, University of Missouri-Kansas City, Kansas City, MO 64108, USA.
[18]Roche Applied Science, Branford, Connecticut 06405, USA.
[19]Los Alamos National Laboratory, Department of Energy's Joint Genome Institute, Los Alamos, NM 87545, USA.

The oomycete vegetable pathogen *Phytophthora capsici* has shown remarkable adaptation to fungicides and new hosts. A member of the stramenopile kingdom, *P. capsici* inflicts several billion dollars in losses worldwide each year. *P. capsici* often features annual meiosis between two mating types, producing thick-walled oospores that can remain dormant for years. Epidemics reflect massive mitosis by deciduous sporangia, releasing swimming zoospores. Genome sequencing, development of a high-density genetic map, and integrative genomic/genetic characterization of *P. capsici* field isolates and intercross progeny revealed significant mitotic loss of heterozygosity (LOH) in diverse individuals. LOH was detected in clonally propagated field isolates and sexual progeny, cumulatively affecting >30% of the genome. LOH altered genotypes for more than 11,000 single nucleotide variant sites and showed a strong association with changes in mating type and pathogenicity. Overall, it appears that LOH may provide a rapid mechanism for fixing alleles and may be an important component of adaptability for *P. capsici*.

FF0159

Assembling with Longer Reads and Higher Depths

Jim Knight

Roche

This talk will present some of the issues and improvements involved in dealing with the higher depths of transcriptome assemblies, and in handling mixes of long and short reads in hybrid 454-Illumina large genome assemblies.

FF0122

Next Generation Sequencing Improvements

Haley Fiske

Illumina, Inc., 5200 Research Place, San Diego, CA 92122 USA

Illumina continues to drive improvements to both the MiSeq and HiSeq sequencing systems. Throughput increases, reduced run times and improved sample prep methods will be discussed.

FF0239

Using the Ion Torrent PGM for *de novo* Sequencing

Timothy Harkins

Life Technologies

The next generation sequencing market continues to evolve and change, and the newest member to the community is the Ion Torrent PGM. Unlike other platforms, that use an optical system to detect base incorporation events, the PGM utilizes semiconductor technology analogous to what is found in your home computer. The semiconductor chip is fabricated with multiple wells with its own discrete sensor. In each well, a single bead with clonally amplified DNA is deposited. Sequence detection then occurs as each nucleotide flows across the chip in series, with base incorporate events detected as change in voltage due to the release of hydrogen ions.

In this talk, technical advancements with the PGM will be presented; including 400 bp read chemistry and the use of long mate pair sequencing. The application of these advancements for *de novo* assembly of both simple microbial genomes and complex genomes, will be highlighted.

FF0038

Exploiting Single-Molecule Real-Time DNA Sequencing for Improved Genome Assembly and Methylome Analysis

Steve Turner

Pacific Biosciences

In the past 9 months, the readlength of Pacific Biosciences' single-molecule real-time (SMRT®) DNA sequencing technology has doubled, the input DNA sample requirements have halved, and consensus error rate has decreased 10-fold. These factors have played an important role in the use of SMRT sequencing in de novo genome assembly, but even more important has been the emergence of new bioinformatics approaches to working with the long reads and rich kinetic information that SMRT sequencing produces. We start with DNA from a previously unfinished bacterium and walk through the sample preparation, sequencing and informatics steps required to both *de-novo* assemble the genome and map its methylome to identify previously unknown methyltransferase specificities from the same data. We conclude with a discussion of the ongoing improvements at PacBio as well as a perspective on opportunities for further ways to exploit these long read data.

NOTES

NOTES

Lunch

12:40 – 2:00pm

Sponsored by

illumina®

NOTES

FF0060

Ion Torrent Semiconductor Sequencing Allows Rapid, Low Cost Sequencing of the Human Exome

David Jenkins, Justin Johnson, and Joy Adigun

EdgeBio, Gaithersburg, MD

High throughput sequencing utilizing dye terminator technologies can produce a high quantity of sequencing data, but these machines are often too expensive for a single laboratory and too slow to provide data on a clinically useful timeline. With the advent of semiconductor sequencing it is now possible to produce up to 1 GB of sequencing data in hours utilizing the Ion Torrent PGM. The sequencing data can then be used to identify point mutations and small indels across the entire exome, leading to quick identification of variants.

Utilizing Ion Torrent 318 chips with 200bp read chemistry EdgeBio has sequenced one of the first publically available exomes on the Ion Torrent PGM sequencer. The data was analyzed using the Torrent Suite software package, aligned with tmap, variants were detected with the Ion Torrent unified variant caller, and variants were annotated with the snpEff tool. The speed of the Ion Torrent PGM shows real promise for providing DNA sequencing on a clinically relevant timeline. If a sample can be processed through a sequencing machine, aligned, and analyzed in hours rather than days, physicians can quickly get a wealth of information about their patients in a short amount of time. With the cost of sequencing dropping significantly Ion Torrent shows real promise for becoming the first sequencing machine to be regularly used in the clinic. By sequencing an exome on the Ion Torrent PGM we aim to underline the ability of the technology to provide exomic information quickly, accurately, and efficiently.

FF0209

En route to the Clinic: Diagnostic Sequencing Applications Using the Ion Torrent

Donna Muzny[1], Xia Wang[1], Yuanqing Wu[1], Christian Buhay[1], Mark Wang[1], Huyen Dinh[1], Jeff Reid[1], David Wheeler[1], Luca Lotta[2], Eric Boerwinkle[3], Rui Chen[1] and Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030
[2]Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, Milan, Italy
[3]University of Texas Health Science Center at Houston, School of Public Health, Houston, TX 77030

Since its introduction to the Human Genome Sequencing Center (HGSC) in January 2011, the Ion Torrent PGM platform has demonstrated a tremendous capacity growth, as well as great flexibility. The platform has been ideal for amplicon experiments. These have been largely focused on validation of putative variants discovered on other platforms. The amplicon sequencing to characterize more than 4,500 variant sites from different cancer and genetic discovery projects has been straightforward and has only been limited by the capacity to produce the multiple PCR products. To extend the utility of the device we have now performed targeted gene panel capture sequencing on the PGM platform. This required modifications to the library construction process to improve complexity, incorporate multiplexing and allow for the use of sub-microgram DNA amounts in library and hybridization.

Capture variant discovery has been demonstrated using a custom design (~1Mb), targeting exons from all 167 known retinal disease genes. The procedure was calibrated using a well characterized (hapmap) sample, yielding 99.59% (1181/1185) overall accuracy, with a total of just four erroneous SNP calls identified. PGM target sequencing was then applied to diagnosis of two retinitis pigmentosa (RP) patient families, MOGL43 and MOGL510. Disease causing mutations in *Ush2A* have been identified and validated in both RP families representing the first reported human patient cases diagnosed on the PGM. In addition, seven regional capture designs as well as exome sequencing have now been evaluated using the PGM pipeline ranging in size from 0.1 to 43Mb in target regions. Most recently both Thrombosis gene panel capture sequencing and whole exome sequencing have resulted in the identification of causative variants. These experiments have been supported by advances in a highly configurable variant caller specifically for Ion Torrent validation applications. The platform is now demonstrated to be suitable for a clinical setting where accuracy and rapid cycle time are essential.

Angelo Bianchi Bonomi Hemophilia and Thrombosis Center, U.O.S. Dipartimentale per la Diagnosi e la Terapia delle Coagulopatie, Fondazione IRCCS Cà Granda - Ospedale Maggiore Policlinico, Università degli Studi di Milano and Luigi Villa Foundation, via Pace 9, Milan, Italy, Zip code 20122.

FF0047

Next Generation Sequencing; Possible Application for Forensic DNA Analysis. What does the Person of Interest Look Like?

Thomas Callaghan and James Robertson

FBI Laboratory, Quantico, Virginia, USA

The forensic DNA community is interested in applying Next Generation Sequencing (NGS) technology to samples recovered from unsolved violent crimes. An overview of the FBI's Combined DNA Index System (CODIS) will be presented along with possible applications of NGS technology to produce investigative leads for criminal investigations.

Forensic DNA techniques have been used for over 25 years to aid criminal investigations. Initially, Restriction Fragment Length Polymorphism (RFLP) techniques were used to include or exclude and person of interest. Next, PCR technology provided increased sensitivity and automation to forensic DNA analysis, allowing the efficient use of DNA databasing. Today, NGS seems poised to aide investigations when the perpetrator is not included in the database. NGS promises to help answer the question "What does the person of interest look like?"

In order to describe the person of interest and serve as the next generation technique for forensic DNA analysis, NGS will need to address mixture resolution, federal quality assurance standards and the possibility of sub-microgram levels of DNA for analysis. This presentation is intended to introduce the sequencing community to the operation and requirements of forensic DNA analysis.

FF0136

Forensic DNA Standards for Next Generation Sequencing Platforms

Peter M. Vallone, Carolyn R. Hill, Erica L.R. Butts, David L. Duewer, John M. Butler, and Margaret C. Kline

Applied Genetics Group, Biochemical Science Division, National Institute of Standards and Technology, 100 Bureau Drive Gaithersburg, MD 20899-8314

Over the past 22 years the Applied Genetics group at NIST has been providing DNA-based Standard Reference Materials (SRMs) for the human identity community. These forensic SRMs are required by the FBI DNA Advisory Board (Standard 9.5) to calibrate DNA typing procedures performed in forensic laboratories. The SRMs typically consist of genomic DNAs (≈100 ng in 50 μL) that have been highly characterized for forensically relevant markers: core autosomal and Y-chromosome short tandem repeat (STRs) in additional to mitochondrial genome sequence. To date the characterization of the forensic markers of interest is a combination of Sanger sequencing and fragment size analysis.

With the current interest in applying Next Generation Sequencing (NGS) technologies to forensic problems we are beginning to explore the next generation of forensic DNA-based SRMs. The considerations for candidate new materials include: source of genomic DNAs, amount of DNA required, genetic markers to be characterized, inter-laboratory testing, and specific needs of the forensic community. This talk will review the past SRMs and identify requirements for future forensic reference materials.

FF0149

Challenges in Genomic Cloud Computing

Dan Bozinov, Beata Wlodarczyk

Genimbi.com, Seattle, WA, USA

The imminent emergence of genomic medicine creates an unprecedented challenge to the computational processing of genomic data. An exponential decline in cost of DNA sequencing itself has greatly outpaced Moore's law for several years now. However, in the near future our sense of magnitude for the current volume of accumulating data may very well be dwarfed by the rapidly approaching data deluge from clinical applications. In a truly post-genomic era, the genetic basis and granularity for disease will undoubtedly be redefined. Oncologists for example will shift their focus from investigating individual genomes or genes to routinely studying spatiotemporal patterns of genome populations in order to interpret complex evolutionary processes that drive cancerous growth. In addition to wide spread personalized treatment regimens, new understanding of structural variations will be obtained from massive population genomics studies. The computational needs will drastically increase in speed and complexity. Cloud computing presents an obvious framework to address current or future demands of storing and processing big data. However, genomic cloud computing does exhibit its own unique challenges, which need to be addressed accordingly. Today the remote transmission of raw sequencing data is still virtually impractical. To achieve true “web scale”, a highly distributed storage array specifically customized towards genomic applications is required. Customized data compression methods need to be incorporated as well. These must take full advantage of the unique properties of redundant short sequence reads in the context of the underlying target genomes. Lastly, a highly intuitive interface and non-frustrating user experience (UX) is pivotal to facilitate complex bioinformatics tools to clinicians as well as non-expert researchers. Herein we introduce a cloud framework that extends genome processing beyond a specialized Platform as a Service (PaaS) system to deliver a comprehensive workspace for genomic health professionals, who can remain agnostic to the underlying technology.

FF0126

NGS for the Masses: Empowering Biologists to Improve Bioinformatic Productivity

Kashef Qaadri

Biomatters, Inc.

Numerous enterprise-scale organizations spanning academic institutions, government facilities and commercial companies rely on a mixture of custom, open-source and proprietary commercial software for their NGS bioinformatics workflows. Dedicated bioinformaticians have been crucial to developing the tools, infrastructure and automated workflows available today, but are increasingly swamped by having to perform fairly routine bioinformatic tasks that could be done by biologists. To successfully engage biologists in mainstream NGS analysis requires a simplification of the software landscape with a data delivery platform and an interface that empowers biologists to access powerful hardware and software. This talk will focus on use cases that demonstrate how to build a flexible infrastructure for sequence-based research that combines commercial, open source and custom tools to improve ROI and accelerate discovery.

FF0070

The PerkinElmer Omics Laboratory

Todd Smith

PerkinElmer

In today’s biology, systems are studied over discrete biochemical reactions and pathways. Global datasets that combine genetic data with a diverse array of “omics” data (transcriptional, epigenetic, proteomic, metabolomic) are collected using high throughput data generation platforms that include high content screening, imaging, flow cytometry, mass spectrometry, nucleic acid sequencing. Of these, the next generation DNA sequencing platforms predominate because they provide an inexpensive and scalable way to interrogate genetic differences, gene expression, and a myriad of factors that control gene expression.

The challenge ahead is automating sample preparation and informatics so that biologists can collect data from large numbers of samples prepared with standardized methods and interrogate the resulting datasets in ways that increase biological signal while decreasing experimental and computational noise. Over the past year PerkinElmer has brought laboratory and high-through data analysis software systems from Geospiza and sample preparation and analysis instrumentation from Caliper together into a global leading company having an extensive portfolio of imaging technologies, assays, reagents, and analytical systems to create a single source for omics research.

FF0120

The Best Finish First: Sequence Finishing with Whole Genome Mapping

Deacon Sweeney, Emily Zentz, Nianqing Xiao, Vadim Sapiro

OpGen, Inc., Gaithersburg, Maryland, USA

Even the best sequencing efforts can result in undetected sequence gaps and assembly errors. These challenges compromise the accuracy of microbial identification and confound genomic comparisons. Additionally, sequence finishing with libraries is an expensive and time-consuming interruption to a high-throughput pipeline. OpGen's unique Whole Genome Mapping platform offers a global view of genomic structure, by which contigs can be ordered and mis-assemblies detected. This approach results in faster, more accurate, and often less expensive finishing of the DNA sequence from microbial genomes.

For a Sequence Placement project, a *de novo* Whole Genome Map (WGM) is first assembled using OpGen's Argus® platform. Utilizing the WGM and set of sequence contigs as input, the Sequence Placement application leverages a sophisticated scoring system to confidently identify their locations in the genomic map. Placements, gap sizes, and mis-assembled contigs are automatically identified and reported. The software typically runs in less than a minute and was designed for easy integration into sequencing pipelines.

The algorithm operates by a best-first approach, where aligned contigs are sorted by confidence score and placed sequentially, with highest confidence first. Once a region of the map is occupied by a contig, the region is no longer available for additional placements. This approach results in a step-wise elimination of opportunities for spurious placements due to reduction and segmentation of the search space. The best-first approach confidently places contigs with as few as three internal restriction fragments, and typically places between 70 and 90% of the sequence contigs (depending on quality and length).

This poster will demonstrate both the algorithm and its implementation for sequence placement. We will illustrate its operation on a model microbial genome, and provide summary statistics collected from training and testing sets.

FF0299

High Throughput Plasmid Sequencing with Illumina and CLC bio.

Jing Lu, Stacie Norton, Ajay Athavale, Susan Johnson, Karen Martin, Dan Ader

Monsanto Company, St. Louis, MO

The Genomic Analysis Center at Monsanto plays an integral role in product development by supplying sequencing and finishing resources to the companies R&D pipeline. A key component of these efforts are the sequencing and analysis of plasmids destined for plant transformation, ensuring the accuracy of inputs to the transgenic plant pipeline. This workflow has been supported by Sanger sequencing for >10 years; however the advent and productionization of new sequencing technologies has presented an opportunity to increase throughput and reduce costs for this workflow. Illumina was identified as the optimal sequencing platform for these efforts based on cost, throughput, accuracy and “finishability” after a comparative study of the Illumina, SOLiD and 454 technologies. Adoption of the Illumina platform also required a new assembly and analysis platform which was found in CLC Bio. Monsanto’s finishing team have worked collaboratively with CLC to refine the software suite for improved finishing and analysis in both the Sanger and Illumina workflows. The new workflow represents a significant savings on both reagents and FTE, while reducing the overall turnaround time for plasmid sequencing.

FF0213a

Engineered Polymerases Provide Improved NGS Library Amplification and Enable Novel Sequencing Applications

Maryke Appel, Eric Van Der Walt, Ross I.M. Wadsworth, Gavin J. Rush, John F. Foskett III, and Paul J. McEwan

Kapa Biosystems, Woburn, MA

PCR has been in use for nearly 25 years and over this time the technology has matured and evolved, spawning a bewildering array of related techniques and enabling modern molecular biology as we know it. Despite the importance and commercial value of PCR, the vast majority of protocols utilize a handful of wildtype DNA polymerases; improvements to existing applications and the development of novel PCR applications have primarily been realized via innovative changes to all of the reaction components except the polymerases themselves. More recently, enzyme engineering has given rise to a new generation of DNA polymerases displaying improved performance in PCR.

We previously engineered a number of DNA polymerases including a high-fidelity, proof-reading polymerase called “KAPA HiFi” to enable robust amplification of diverse targets spanning a wide range of sizes and GCcontent. We have demonstrated that KAPA HiFi produces high yields with remarkably little bias when used for NGS library amplification in a range of applications. These characteristics have enabled dramatically improved sequence coverage uniformity in standard NGS workflows. With improved Methyl-seq performance in mind, we further engineered KAPA HiFi to allow the efficient amplification of bisulfite-converted DNA libraries containing high proportions of uracil. Using a number of case studies, we illustrate that this new generation of engineered DNA polymerases are providing better sequence coverage and are enabling the development of novel and challenging NGS applications.

FF0144

Beyond Basic Target Enrichment: New Tools to fuel your NGS Research!

Jennifer Carter

Agilent Technologies

As next generation sequencing technology rapidly evolves, so must strategies to optimize sequencing efficiency through enrichment. In an effort to expand both the applications as well as the effectiveness of targeted enrichment, Agilent has developed new tools for researchers to examine the genome. To address epigenetic questions while reducing cost and data complexity, Agilent has developed a Methylseq platform with the most comprehensive CpG island coverage of any reductive method. Going beyond RNAseq, Agilent also offers RNA enrichment, which provides the sequencing depth required to detect low abundance novel splice variants and gene fusions that are frequently missed in transcriptome sequencing while maintaining relative gene expression level assessment. Finally, for rapid resequencing of target genes without the disadvantages of a multiplex PCR, Agilent has further developed HaloPlex technology to provide the benefits of hybridization enrichment with the ease and speed of PCR strategies. In these three Agilent tools, researches are able to bridge the gap between genetic and epigenetic modifications and the subsequent transcriptional consequence, fueling integrated biology research in the next generation sequencing era.

FF0019

Better Computing for Better Bioinformatics

George Vacek[1,2], Kirby Collins[2]

[1] Corresponding author: gvacek@conveycomputer.com
[2] Convey Computer Corporation, Richardson, TX, USA.

Advances in sequencing technology have significantly increased data generation, requiring similar computational advances for bioinformatics analysis. Advanced architectures based on reconfigurable computing can reduce application run times from hours to minutes and address problems unapproachable with commodity servers. The increased capability also improves research quality by allowing more accurate, previously impractical, approaches. This work describes the use of Convey's Hybrid-Core (HC) computing architecture, which combines a traditional x86 environment with a reconfigurable coprocessor, to deliver high-performance for genomic data analysis workflows including reference mapping, de novo assembly, functional annotation, and variant analysis.

Bioinformatics applications consist of large numbers of relatively simple operations on large randomly accessed data structures. Conventional architectures lack sufficient parallelism in the core processing elements and the memory subsystem to efficiently execute these algorithms. The Convey hybrid-core (HC) architecture incorporates both a highly parallel processing architecture and a highly parallel, randomly accessible memory subsystem. Algorithms also benefit from massively parallel implementations of application-appropriate-data-type operations, which use logic gates more efficiently than commodity servers.

A translated search to screen short-read DNA sequences against a small protein database (e.g. a known toxin database or patented protein database) is useful to quickly identify those reads coming from a gene associated with the proteins of interest. This can be particularly useful for gene sequences which have multiple copies of highly conserved regions in a genome that would otherwise be difficult to assemble with whole-genome assembly, or when the goal is simply to identify one microbe out of a metagenome for further study. SWSearch, a search and alignment program using the Smith-Waterman algorithm, dramatically reduces the time to perform large numbers of local alignments. SWSearch on an HC-2^{ex} server is 14.5x faster than the fastest software implementation on a commodity x86 system. When searching Illumina reads against a database of protein sequences SWSearch on an HC-2^{ex} server is more than 7 times faster than NCBI BLASTx. Furthermore, BLAST uses a heuristic filter and matches only about 1/3 as many as found by the full Smith-Waterman approach. Since matches indicate a read that is part of a gene of interest, any miss could be significant.

Convey has also developed a personality that improves the performance of the BWA processing pipeline that allow HC systems to dramatically reduce time to solution and increase throughput for read mapping, for instance, showing more than 18x improvement for a full BWA paired-end mapping of Genome of The Netherlands data. Integrated BAM file generation leads to additional workflow optimization. Graph Constructor reduces not only run time for Velvet, but also required memory, making it capable of larger assemblies. Additional performance and workflow optimizations will be discussed, including a fast kmer counting tool that allows quick identification of optimal kmer length and coverage cutoffs for de novo assembly.

FF0296

RAPID: Ultra High Throughput Sequencing Data Analysis for Quick Microbial Identification

Robert T. Yamamoto, Matthew Kidd, Daniel Sphar, Milena Martinez

FLIR La Jolla, California, USA

FLIR is developing automated sample prep, nucleic acid clutter mitigation modules and data analyses pipelines to facilitate the use of Ultra High Throughput Sequencing (UHTS; e.g. HiSeq, SOLiD, PacBio) for comprehensive species identification. For fast data analysis we have developed RAPID-*MG* (Robust Analysis Pipeline and Identification Database for Metagenomics) for automated querying of UHTS data of unknown samples for comprehensive taxonomical identification – current capability >80,000 NCBI annotated bacterial, viral and fungal species. RAPID-*MG* (along with its parent RAPID-*BT* (BioThreats)) is an analysis pipeline and relational database that can be operational wholly contained on a laptop or a large multi-processor computer network. With a simple click on a data file, RAPID-*MG* mines the data in a logical, tiered approach and outputs reports containing successively increasing analysis resolution. Within minutes an initial report outputs all identified microbial taxonomies (order, family, genus, species) detected; followed by a second report includes all threat markers found (e.g. antibiotic resistance mutations/genes etc.); finally, as an option triggered at the start of analysis, the nearest sequenced strain for each species found is determined by comprehensively mapping sequence reads to all complete genome sequences available for the species found. Traditional methods (i.e. BLAST to Genbank) can take at least a week of super-computing time. RAPID-*MG* identification takes place at several taxonomical levels, so novel threat species which have common genus/family sequences (e.g. Coronavirus SARS) can be identified. Clinical and environmental sample reports will be presented along with updates of our recent progress on sample prep and clutter mitigation.

Tech Time Notes

Meet and Greet Party

600pm – 900pm, June 5th

Sponsored by Roche Diagnostics

Enjoy!!!

Roche

Poster Presentations (June 5th)
Even #'s 6:00-7:30pm, Odd #'s 7:30-9:00pm

FF0015

Influenza Surveillance Sequencing Using the Ion Torrent Platform

Dhwani Batra [1], Catherine Smith [2], John Barnes [2], Jeffery Sabina [3], Michelle Milham [3], Elizabeth Neuhaus [3], Michael Shaw [3].

[1] SRA International, Affiliation (Influenza Sequence Activity, Influenza Division, Centers for Disease Control and Prevention), Atlanta, GA, USA; [2]Influenza Sequence Activity, Influenza Division, Centers for Disease Control and Prevention, Atlanta, GA, USA; [3]Life Technologies, Guilford, CT, USA

NextGen sequencing (NGS) technologies have enabled the possibility of rapid and inexpensive generation of influenza whole genome sequence data. Cost-effective, whole genome sequencing (WGS) strategies may enhance surveillance programs by enabling a triage-based, targeted approach to virus characterization. WGS analysis enables rapid identification and quantification of circulating variants while sequence comparison between new and previously characterized strains with known biochemical profiles allows researchers to strategically select specimens for further isolation and characterization. Enhancement of the virologic characterization pipeline via use of WGS will also provide additional benefit such as establishing native levels of re-assortment and co-infection among influenza viruses.

The Ion Torrent Personal Genome Machine was investigated for use as a proof-of-concept WGS platform because of its affordability, small laboratory footprint, and ease-of-use. Initial efforts to sequence negative-stranded RNA genomes have been challenging. In order to successfully sequence the RNA genome of the influenza virus, multiple modifications of standard WGS experimental protocols were required, mainly to balance the large number of reads per segment in order to ensure equal representation of each gene segment. Once the bias among the segments was balanced, it was possible to establish a level of multiplexing that would confidently provide high quality consensus sequence regardless of original sample type, i.e. clinical specimen or an egg- or cell-isolate. In addition, successful identification of the presence of Influenza A co-infections was possible for some specimen samples.

The consensus sequences were obtained by scaffolding against a panel of representative whole genome sequences of different influenza virus subtypes and lineages. Only the consensus sequences that met coverage thresholds were considered. Initial runs included 1 sample per ION 314^{TM} chip and analysis indicated 100% coverage of all 8 segments, with varied depth of coverage from 300-240,000X. Even at minimum 15X coverage, randomly down-sampled reads (1/50) gave high confidence consensus sequence when reassembled. Final runs successfully multiplexed 12 samples on a single chip with 25-50X coverage to identify low-level variants of influenza viruses. Samples known to have co-infections produce higher coverage consensus to more than one reference, due to the high degree of conservation found among the internal genes of influenza. The two external genes, hemagglutinin (HA) and neuraminidase (NA), were found to have two distinct scaffolds. In the future, it will be necessary to establish an optimized threshold for detecting mixed infections of influenza A viruses.

FF0026

SMRT Sequencing Provides Insight into the Diversity of the Bovine Immunoglobulin Heavy Chain Repertoire

Peter A. Larsen and Timothy P. Smith

USDA, ARS, US Meat Animal Research Center, PO Box 166, Clay Center, NE 68933, USA

The vertebrate immune system produces a diverse antibody repertoire capable of responding to a vast array of antigens. This diversity is generated through a multifaceted process of gene segment recombination and somatic hypermutation or gene conversion. Recent advances in high-throughput sequencing technology permit the sequencing of antibody repertoires at previously unattainable depths of coverage and therefore allow researchers to better explore antibody diversity and selection within individuals. Moreover, next-generation sequencing methods provide unique approaches to a number of immuno-based research areas including antibody discovery and engineering, disease surveillance, and host immune response to vaccines. It is within this framework that we approached the bovine antibody repertoire. Here, we present single-molecule real-time (SMRT) sequencing data of the expressed bovine immunoglobulin G (IgG) heavy-chain repertoire. We generated high quality circular consensus reads of the entire VDJ region of IgG cDNA libraries. Our results indicate that the number of functional germline V segments hypothesized within *Bos taurus* is likely underestimated. Moreover, we provide data which reinforce previous hypotheses regarding the preferential usage of a single germline J segment. Our experimental design provides the foundation for future studies important to livestock research including host immune response to infections and vaccines.

The USDA is an equal opportunity employer and provider.

FF0039

Novel Purification Reagents for the Study of the Human Microbiome

Fiona Stewart, George R Feehery, Eileen T Dimalanta, Brad Langhorst, Lynne Apone, Pingfang Liu, Daniela Munafo, Christine Sumner, Sriharsa Pradhan, and Theodore Davis

New England Biolabs, 240 County Road, Ipswich, MA, 01938

Nucleic-acid based techniques such as hybridization, PCR, qPCR and Next-Gen sequencing offer a rapid and highly sensitive option for detecting pathogenic bacteria directly from specimens when compared with culture-based techniques. Aside from the inherent limitations of amplification and identification of biological samples, the human genome itself may interfere with the detection and diagnosis of pathogens partly due to the higher percentage of human genomic DNA relative to the target microbiome. As a consequence, analyses of a metagenome or microbiome from clinical samples by Next-Gen sequencing or PCR are inefficient, difficult, and time consuming. To address this problem, we have developed a unique method for the separation of large pieces (~20 kb) of human DNA from similar sizes (~20 Kb) of bacterial DNA using a Eukaryotic Binding Protein (EPB). This protein binds to hydrophobic Protein A magnetic beads that havebeen engineered to exhibit minimal non-specific DNA binding. As a demonstration of the efficacy of this methodology, DNA extracted from human saliva was added to the EPB beads andincubated for a period of time. A magnetic field was then used to separate the human DNA bound to beads from the supernatant containing bacterial DNA. The enriched bacterial DNA from the supernatant was used for sequencing on different NGS platforms and was found to have a dramatic increase in the number of reads matching bacterial genomes. Likewise, a substantial drop was seen in the number of reads matching the human genome. This simple methodology can be used to analyze entire microbiomes in a cost-effective manner utilizing established Next-Gen sequencing platforms, as well as newer single molecule sequencing technologies.

FF0045b

Minding the Gaps in the Arabidopsis Genome

Dan Ader[1], Mitch Sudkamp[1] (presenter), Xuefeng Zhou[1], TJ Corrigan[1], Randy Kerstetter[1], Zijin Du[1], Rosa Ye[2], Megan Wagner[2], Amanda Kerowicz[2], David Mead[2], Chengcang Wu[2], Todd Michael[1]

[1]Monsanto Company, St. Louis, MO
[2]Lucigen Corporation, Middleton, WI

The *Arabidopsis thaliana* genome, first sequenced in 2000, is considered a gold standard plant reference genome. Despite this standing, the reference contains numerous gaps due to repetitive sequences and inherent biases in the techniques utilized to sequence the genome. The marriage of novel sequencing library preparation methods with different biases and new sequencing technologies may allow us to better sequence and assemble these challenging genomic regions. To this end, Lucigen Corporation created a randomly sheared *A. thaliana* BAC library and screened it for the flanking sequences near the known gaps in the *A. thaliana* reference genome. These BACs were then sequenced using both Illumina short reads and Pacific Biosciences long reads. This approach has enabled us to close or extend into many of the known gaps and provide resolution and annotation in these previously uncharacterized regions of the genome.

FF0051

Using the Argus® Whole Genome Mapping System to Improve DNA Sequenced Genomes and Allow for Structural Variant Analysis

[1]Trevor Wagner, [2]Matthew Dunn, [2]Karen Brooks, [1]Nianqing Xiao, [1]Bin Zhu, [1]Deacon Sweeney, [1]Yunhu Wan, [1]Erin Newburn, [1]Rich Moore

[1]OpGen, Inc., Gaithersburg, Maryland, USA.
[2]Welcome Trust Sanger Institute, Hinxton, UK.

Achieving completely assembled chromosomes from genomes > 100 Mb using next generation sequencing technologies is difficult due to length of sequencing reads, genome structure, and assembly algorithms. Whole genome sequencing projects result in tens to tens of thousands of contigs/scaffolds thereby making the analysis of relevant genetics complex. Whole Genome Mapping (WGM) is a rapid sequence-independent technology that produces map read lengths of 150 kb to ~3 Mb for finishing of complete chromosomes of 100+ Mb genomes.

Whole Genome Maps of Helminth genomes > 100 Mb were generated by the Welcome Trust Sanger Institute using the Argus System. WGM resulted in complete Helminth chromosome maps allowing for genome size identification, whole genome DNA sequence assembly, and sequence-independent validation. WGM of *Echinococcus multilocularis* resulted in 31 map scaffolds with a total size of 111 Mb and over 85% agreement with the DNA sequence. Mapping identified the order and orientation of DNA scaffolds, distance between scaffolds, potential misassemblies, and locations where scaffolds should be joined. The project has progressed to 9 chromosomes represented by 11 major scaffolds after utilizing WGM data. This represents the first *de novo* Whole Genome Mapping of Helminths on the Argus® System.

WGM data was also generated at OpGen for plant, animal, and human genomes of > 100 Mb. Genome-Builder was used to combine mapping and sequence data of the organisms in a hybrid approach to produce long-range scaffolding of genomes. The resulting assemblies contained higher N50 and N90 statistics and fewer scaffolds compared to performing sequencing alone. In a comparative genomics study, we demonstrated that WGM, at the level of human chromosome arms, can be constructed and detect structural variations.

WGM can be used on genomes greater >100 Mb to help finish, validate, and analyze complex genomes that would be difficult with current DNA sequencing technologies alone.

FF0063

New NEBNext® Reagents and Adaptors for NGS Library Preparation

Landon Merrill, Pingfang Liu, Erbay Yigit, Eric Cantor, Bradley W Langhorst, Lynne M Apone, Daniela B Munafo, Christine Sumner, Fiona J Stewart, Thomas C Evans Jr., Eileen T Dimalanta, and Theodore B Davis

New England Biolabs

Next generation sequencing (NGS) technologies are continually evolving to increase the throughput, enabling deeper and faster sequencing of large genomes such as human. Alternatively, higher sequencing throughput allows for the simultaneous sequencing of multiple small genomes, targeted loci on large genomes, or hundreds to thousands of amplicons. Sequencing of multiple samples has the added advantage of significantly reducing per sample cost. This broad range of applications enabled by NGS requires flexible library construction workflows that have minimal bias and generate high library yields. To address this need, we have developed a series of new reagents and adaptors. These new NEBNext reagents simplify library preparation workflows and increase library yield. In addition, novel NEBNext adaptors, compatible with sequencing on the Illumina platform, offer significant flexibilities in singleplex or multiplex library preparation.

FF0078

Rapid Conservative DNA Shearing for Long Paired End Library Construction

Stacey Broomall

US Army, BioSensors Branch (RDCB-DRB-S), Edgewood Chemical Biological Center

Sample preparation for paired-end library construction in genome finishing, a process necessary for microbial forensic applications, relies on the use of commercial hydrodynamic shearing platforms. This method can be cumbersome as the available instruments have been subject to frequent clogging, air bubble formation, and substantial loss of DNA (up to 50%). Sonication and nebulization of DNA for PE purposes is not an option as they are only amenable for fragments up to 5kb and 1.5kb, respectively. Recently, ECBC functioned as a beta test bed for the Covaris g-TUBE and provided feedback for the PE library preparation process and sequencing data on the Roche 454 FLX Titanium sequencing platform.

Results indicate that this disposable tube is capable of shearing DNA from 6-20kbp using only a table top centrifuge in under 5 minutes with a 95-99% recovery. Compared with the 15-20μg of DNA required for some of the commercial shearing systems, this process required only 7.5μg to produce adequate input for the 8kb insert for the sequencing process. Post-sequencing, 99.09% of the reads mapped to our finished reference sequence with average insert size of 8124.6 bp ± 2031.2 bp. This protocol change eliminates the need for additional shearing equipment, conserves the precious DNA sample from large losses, and saves time and energy spent during the fragmentation process.

Further evaluations presented here compared the reproducibility of the process, especially using differing organisms with varying GC content. This work will be important in the microbial forensics arena or any community that has a need to finish genomes to identify and compare genomic signatures and their differences.

FF0079

Algal Genomes Drafting and Computational Finishing with 2nd Generation Technologies. How Much Data is Enough?

O. Chertkov, C. Han, S. Starkenburg

Los Alamos National Laboratory

New generation technologies allow drafting algal genomes in a very short time with a very good coverage at low cost. These big amounts of data have to be processed with appropriate assembler on a cluster with enough memory to generate good assembly. Different types of data require different coverage; they may have some cloning bias. Resulting assemblies may need to be merged together to generate better assembly.

We have explored 454, Illumina and pacbio data as drafting tools and newbler, velvet and Celera assemblers to produce final genomes. We tried to optimize cost, coverage, type of data and assembler to produce better genomes.

Our Improved High Quality drafts (IHQD) for possible diploid algal genomes have about 1000 contigs, about 100 scaffolds for ~20 Mb genomes.

FF0080

Bacterial Biodiversity and Function in a Cold Desert Ecosystem

Cristina Takacs-Vesbach[1], David Van Horn[1], Michael Gooseff[2], and John Barrett[3]

[1]Department of Biology University of New Mexico Albuquerque, NM 87131
[2]Department of Civil & Environmental Engineering Pennsylvania State University University Park, PA 16802
[3]Department of Biological Sciences Virginia Tech Blacksburg, VA 24061

For many decades the soils of the McMurdo Dry Valleys, Antarctica were thought to be essentially aseptic. We now know that this is an ecosystem that is dominated by microorganisms, however, early cultivation efforts failed to detect the apparently high diversity of the region's poorly weathered, low organic-matter soils. Initial surveys of microbial diversity using 16S rRNA gene sequencing revealed a surprising bacterial richness, including representatives from at least ten different phyla, and a high proportion of unique and rare sequences. Yet, initial surveys of microbial diversity were not exhaustive and little information was gained about the function of the detected microorganisms. Furthermore, given the low rates of microbial activity and decomposition rates, the question of whether this richness represents functioning vs dormant members of the community has been raised. We have conducted an exhaustive survey of the microbial richness, function, and activity of soil bacteria across gradients of moisture and salinity using pyrosequencing of 16S rRNA bacterial tag-encoded FLX amplicons (bTEFAP) and environmental DNA (metagenomics) combined with extracellular enzyme assays. Our metagenomic analysis included approximately 1 Gb of DNA sequences from four samples and represents a first step in linking community diversity and function, an essential step in this model ecosystem as well as soil ecosystems worldwide. Comparisons of the microbial communities detected by both methods reveal a soil biodiversity that is dominated by Actinobacteria, Proteobacteris, Firmicutes, and Acidobacteria. However, even our metagenomic analysis pointed to a moderate level of diversity for these samples, including many singleton species and ranging from 550 to 780 OTUs per sample. A majority of the metagenomic sequence was assignable to a putative function, including a large proportion of metabolic genes. The potential microbial function of dry valley soil will be discussed with the ultimate goal of understanding the role of bacteria in cold arid soils.

FF0083

JCVI Viral Finishing Pipeline: A Winning Combination of Advanced Sequencing Technologies, Software Development and Automated Data Processing

Nadia Fedorova, Danny Katzel, Tim Stockwell, Peter Edworthy, Rebecca Halpin, and David E. Wentworth

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

JCVI viral projects are supported by the NIAID Genomic Sequencing Center for Infectious Disease (GSCID). Viral sequencing and finishing pipeline at JCVI combines next generation sequencing technologies with automated data processing. This allowed us to complete over 1800 viral genomes in the last 12 months, and almost 8800 genomes since 2005.

Our NextGen pipeline, which utilizes SISPA-generated libraries with Roche/454 and Illumina sequencing, enables us to complete a wide variety of viral genomes including challenging samples. Automated assembly pipeline employs CLCbio command-line tools and JCVI cas2consed, a cas to ace assembly format conversion tool. Our complimentary Sanger pipeline software is currently being integrated with the NextGen pipeline. This will improve our data processing and will allow us to use validation software (autoTasker) more efficiently.

During the past year we have found that novel viruses, repetitive genomes, and mixed infection samples could not be easily integrated with our high-throughput assembly pipeline. We have developed an assembly and finishing process that utilizes components of the high-throughput pipeline and combines them with manual reference selection and editing. Using this approach we completed novel adenovirus genomes and mixed-infection avian influenza genomes, and improved assemblies of previously unknown arbovirus genomes. We are currently working on optimizing and automating this new pipeline.

Repetitive genomes have long been known to present great challenges during assembly and finishing. We are presenting a new approach to assembly and finishing of repetitive varicella genome that is based on separating it into overlapping PCR amplicons followed by merging sequenced amplicons during assembly.

To streamline our viral pipelines, we have fully integrated them with JCVI's LIMS and JIRA Workflow Management to create a semi-automated tracking interface that follows the progress of viral samples from acquisition through to NCBI submission. This allows us to process a large volume of samples with limited manual interaction and, at the same time, gives us flexibility to work on challenging and novel genomes.

FF0090

Differential Expression Analysis of RNA-Seq Data for Fuel-Producing Cyanobacteria

Anne M. Ruffing

Sandia National Laboratories, Bioenergy & Defense Technologies, PO Box 5800, MS 1413, Albuquerque, NM 87185, USA

Biologically-derived fuels, such as those produced by photosynthetic microalgae, are leading candidates to replace conventional petroleum-based transportation fuels. While microalgal fuels represent a renewable energy source with a potential for reduced carbon emissions, microalgal biofuel production has yet to be demonstrated at scale and is often deemed to be unsustainable and uneconomical based on life cycle and technoeconomic analyses. Genetic engineering techniques offer the potential to overcome many of the challenges facing microalgal fuel production by improving the rate of lipid production and introducing other desirable traits into industrial microalgal strains. However, metabolic engineering strategies are limited by our insufficient knowledge of microalgal metabolic networks, regulatory mechanisms, and cellular physiology. With the affordable cost of next-gen sequencing, RNA-seq is quickly becoming a common laboratory tool for deciphering systems-level responses of fuel-producing microalgae. In this study, RNA-seq is applied to assess the genetic responses of a model cyanobacterium, *Synechococcus elongatus* PCC 7942, which has been engineered to produce and excrete free fatty acids (FFA) as biodiesel precursors.

Using 17 RNA-seq samples from fuel-excreting cyanobacteria, this study analyzes the influence of data analysis strategies on the resulting differential gene expression values. The assessed strategies include data preprocessing techniques, assignment of reads aligning to multiple genes, read count normalization, and the importance of biological replicates for statistical analysis. The samples were extracted from 3 strains with varying levels of FFA production: the wild type with no FFA excretion, an engineered strain (SE01) with low FFA excretion during the exponential growth phase and high FFA excretion during the stationary phase, and an engineered strain (SE02) with high FFA excretion during both the exponential and stationary growth phases. Differential gene expression analyses will investigate changes in gene expression between these 3 strains as well as between different growth phases.

FF0091

Novel Metabolic Attributes of *Cyanothece*, a Group of Unicellular Nitrogen Fixing Cyanobacteria

Louis A. Sherman[1], Anindita Bandyopadhyay[2], Thanura Elvitigala[2], Eric Welsh[3], Jana Stöckel[2], Michelle Liberton[2], Hongtao Min[1], Himadri B. Pakrasi[2]

[1]Department of Biological Sciences, Purdue University, W. Lafayette, IN 47907
[2]Department of Biology, Washington University, St. Louis, MO 63130; [3]Biomedical Informatics Core, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612

The genus *Cyanothece* comprises unicellular cyanobacteria that are morphologically diverse and ecologically versatile. Studies over the last decade have established members of this genus as important components of the marine ecosystem, contributing significantly to the nitrogen and carbon cycle. Systems level studies of *Cyanothece* 51142, a prototypic member of this group, revealed many interesting metabolic attributes. This includes the fact that these strains exhibit high rates of N_2 fixation and H_2 production under aerobic conditions To identify the metabolic traits that define this class of cyanobacteria, six additional *Cyanothece* strains were sequenced to completion. The presence of a large, contiguous nitrogenase gene cluster and the ability to carry out aerobic nitrogen fixation distinguish *Cyanothece* as a genus of unicellular, aerobic nitrogen fixing cyanobacteria. *Cyanothece* cells can create an anoxic intracellular environment at night, allowing oxygen-sensitive processes to take place in these oxygenic organisms. Large carbohydrate reserves accumulate in the cells during the day, ensuring sufficient energy for the processes that require the anoxic phase of the cells. Our study indicates that this genus maintains a plastic genome, incorporating new metabolic capabilities while simultaneously retaining archaic metabolic traits, a unique combination which provides the flexibility to adapt to various ecological and environmental conditions. Each strain contains a large, contiguous cluster of nitrogenase genes that are regulated in a circadian fashion. However, rearrangement of the nitrogenase cluster in *Cyanothece* 7425 and the concomitant loss of its aerobic nitrogen fixing ability suggest that some cyanobacterial strains may have eventually lost their nitrogen fixing ability. The genomes of some *Cyanothece* strains are quite unique in that there are linear elements in addition to a large circular chromosome. *Cyanothece* 51142 has one linear chromosome, whereas *Cyanothece* 7822 has two large linear chromosomes and a linear plasmid. This work was supported by funding from DOE-BER (DE-FC02-07ER64694).

The work conducted by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We would like to acknowledge the excellent work of the following Joint Genome Institute personnel. From the Los Alamos National Lab: Olga Chertkov, David Sims, David Bruce, Chris Detter, Elizabeth Saunders (deceased), Roxanne Tapia and Shunsheng Han; from JGI-Walnut Creek: James Han, Tanja Woyke, Sam Pitluck, Len Pennacchio and Matt Nolan; and annotation authors Natalia Ivanova, Miriam Land (Oak Ridge National Lab), Natalia Mikhailova, Galina Ovchinikova and Amrita Pati.

FF0095

Finishing Complex Microbial Genomes Via Hybrid Assembly & Optical Mapping

Stacie Norton, Dan Ader, Randy Kerstetter, Xuefeng Zhou, Shawn Stricklin, Barry Goldman, Todd P. Michael

Monsanto Company, St. Louis, MO

Second and third generation sequencing technologies have facilitated the finishing of microbial genomes for a fraction of the previous cost. Advances in both the sequencing technologies and assembly algorithms allow closure or near closure of microbial chromosomes even at the extremes of G/C content. This has predominantly been driven by advanced sequencing library construction of Illumina paired end and long mate pair libraries or hybrid assembly with long PacBio single molecule reads. However, there are still challenges in finishing microbial genomes with numerous and diverse plasmids of different sizes and copy numbers. These microbial strains are sometimes more similar to metagenomic assembly problems and require additional attention in terms of sequencing library types and algorithm development. In addition, tools such as optical maps will provide power to develop a better understanding of these complex microbial genomes. We will discuss our work finishing complex microbial 'metagenomes' using Illumina, Pacbio and optical maps.

FF0098

Improving the Human Reference Genome

Tina Graves and the Genome Reference Consortium

Washington University School of Medicine

The current human reference sequence provides a foundation for genome-wide studies of human variation, genome structure, evolutionary biology, and human disease. Many of these studies have revealed however, that there are regions of the human reference genome that are not represented correctly. When the reference genome was labeled “finished”, it was understood that there were some regions recalcitrant to closure with the existing technology, and resources. At that time, however, the degree to which structural variation affected the ability to produce a truly representative genome sequence at some of these loci was not clear. Many of these loci have now been identified as regions of focus for the Genome Reference Consortium (GRC). These recalcitrant trouble spots tend to be associated with repetitive sequences therefore, distinguishing repeat copies from allelic differences is very difficult. One resource that is currently being utilized for these problematic regions is the Hydatidiform Mole BAC library (CHORI-17), which is a single haplotype library. Selecting a clone path from this library allows us to discriminate between different haplotypes and repeat copies. To date, we have resolved several of these loci by using this resource and have many other regions under investigation. In particular, one such problematic region in the current reference assembly (GRCh37), 1q21, has been improved significantly by these efforts. In this presentation we compare the original multi-allele derived reference sequence of the 1q21 region, which is very fragmented, poorly assembled, and incomplete, to a newly developed representation, based on the single haplotype sequence of the same area. This presentation also highlights other similar efforts currently underway by the GRC to improve the reference and enhance the understanding of the human genome.

FF0099

Variant Validation, Extension, and Interpretation Methods at the Genome Institute at Washington University

Bob Fulton

The Genome Institute at Washington University School of Medicine, 4444 Forest Park Blvd., St. Louis MO 63108

With the ever-increasing throughput of next generation sequencing, variant validation is increasingly critical to understanding the mutational spectrum of the sequenced genomes. Validation provides confirmation of putative variant calls, thus helping to improve variant calling algorithms. In addition to confirmation of putative calls, the validation process provides a deeper understanding of variant frequency, and helps with the interpretation of the impact of the variation. For somatic mutations, variant frequencies provide clues to tumor purity, and clonality, and help identify likely driver events, or variants critical to the progression or metastasis of this disease.

These methods not only provide validation, but also can be used to extend putative variants across other samples, to identify commonly mutated genes across sample panels. This presentation will outline validation/extension methods and decision processes utilized for large and small-scale variant confirmation.

FF0102

High-Throughput Mitochondrial Genome Sequencing for Loggerhead Sea Turtle Conservation Genetics

Andrew D. Farmer[1], Connor T. Cameron[1], Ernest F. Retzel[1], James Huntley[2], Steven Gao[2], Mark G. Dodd[3], Brian M. Shamblin[4], Campbell J. Nairn[4]

[1] National Center for Genome Research, Santa Fe, New Mexico
[2] University of Colorado BioFrontiers Institute, Boulder, Colorado
[3] Georgia Department of Natural Resources, Wildlife Resources Division, Brunswick, Georgia
[4] University of Georgia, Warnell School of Forestry and Natural Resources, Athens, Georgia

The loggerhead sea turtle (*Caretta caretta*) is an endangered species that is distributed globally in warm temperate and tropical waters. The southeastern United States (US) hosts one of two large nesting aggregates globally and is therefore critical to conservation and recovery efforts for the species. To date, population structure has been defined using either a 380 or 817 bp region of the 16 kb mitochondrial genome. Sharing of the short haplotype sequences among distinct subpopulations confounds resolution of stock structure and population estimates. To better understand the phylogeography of loggerhead genetic distribution, nesting behavior, and composition of mixed foraging aggregations, an effort is underway to analyze the mitochondrial genome and to develop full mitogenome haplotype markers that can be applied to studies of the loggerhead population at global, regional, and local scales. Until now, mitochondrial genomes have been sequenced one at a time. In this project, samples were collected from egg shells of 134 nesting females representing the loggerhead turtle nesting distribution in the southeastern US. These samples were amplified using a protocol developed at the University of Georgia (Nairn). An 817 bp region used for current haplotype assignments was sequenced using Sanger methods. From these, 21 pools of individual samples representing distinct nesting beaches were assembled. Sequencing libraries were prepared using a modified Illumina library preparation protocol (Huntley) to produce 21 indexed libraries from samples with as little as 4 nanograms of DNA. The libraries were pooled (12 and 9 libraries in two pools respectively) and sequenced at the University of Colorado BioFrontiers Institute. Sequencing of the pools was accomplished on an Illumina HiSeq 2000 (1X50 Single Read) yielding an average of 140M reads per pool. Sequencing accuracy was a paramount consideration because some haplotypes differ by as little as 1 bp over the 817 bp control region. The resulting sequences ranged from 4,000-88,000x coverage. The sequence data was analyzed at NCGR using bioinformatics protocols modified to accommodate the deep coverage. Candidate synonymous and non-synonymous SNPs and indels were called using methods developed at NCGR, and extended haplotypes encompassing the entire genome are presently being refined. Informative variation identified by comparative analyses of the full mitochondrial genome sequences will provide increased resolution among subpopulations and mixed foraging aggregations.

FF0109

Evaluation of Multiplexed 16S rRNA Microbial Population Surveys Using llumina MiSeq Platform

Julien Tremblay, Edward S Kirton, Kanwar Singh, Feng Chen and Susannah G Tringe

DOE Joint Genome Institute, Walnut Creek, CA, 94598, USA

In recent years, microbial community surveys extensively relied on 454 pyrosequencing technology (pyrotags). Recently, the Illumina sequencing platform HiSeq2000 has largely surpassed 454 in terms of read quantity and quality with typical yields of up to 600 Gb of paired-end 150 bases reads in one 18 day run. Yet many labs still rely on pyrotags for community profiling because the HiSeq throughput exceeds their needs, the run time is long, and accumulating sufficient samples to effectively utilize a full run introduces significant delay. Illumina recently introduced the new mid-range MiSeq sequencing platform which gives an output of 1 Gb of paired-end 150 base reads in a single day run. With its moderately-high throughput and support for massive multiplexing (barcoding), this platform represents a promising alternative to 454 technology to perform 16S rRNA-based microbial population surveys.

A workflow was therefore developed to confirm that Illumina MiSeq is a suitable platform to accurately characterize microbial communities. We surveyed microbial populations coming from various environments by targeting the 16S rRNA hypervariable region V4 which generated amplicons size of about 290 bases. These amplicons were sequenced with the MiSeq platform from both 5' and 3' ends followed by *in silico* assembling using their shared overlapping part. Downstream analyses through our Itags pipeline are also described, including a novel clustering strategy generating fast and accurate distribution of bacterial operational taxonomic units (OTUs).

Our results suggest that the MiSeq sequencing platform successfully recaptures known biological results and should provide a useful tool for 16S rRNA characterization of microbial communities.

FF0110

Finishing and Sequence Improvement Pipeline at The Genome Institute at Washington University

Aye Wollam, Tina Lindsay, and Bob Fulton

The Genome Institute at Washington University, St. Louis, MO

With the establishment of the massively parallel next generation sequencing platforms, finishing and sequence improvement process at The Genome Institute at Washington University in St Louis has undergone major changes. Both the clone-based pipeline and the microbial (whole genome) pipelines have now been fully transitioned to utilize 454 and Illumina data, in addition to the Sanger data in select projects. In the clone-based pipeline, all active clones are pooled and sequenced on the Illumina platform as a strategy to cheaply and efficiently enhance the quality of draft assemblies comprised of Sanger or 454 sequence data. In this presentation, we will discuss the focus of our finishing efforts and outline the steps involved in our clone-based and the whole genome finishing pipelines, where both manual and automated efforts are used to close gaps, resolve misassemblies and correct consensus errors.

FF0111

Genome Assembly and Finishing Using CLC bio Tools

Marta Matvienko, Cecilie Boysen, Joe Salvatore, Rob Mervis, David Michaels, and Jannick Bendtsen

CLC bio

De novo genome assembly and genome finishing is becoming increasingly important with the massive amounts of data being generated by next generation sequencing. The enormous amounts of sequencing data by new sequencing methods still leave room for Sanger sequencing data in finishing projects for validation and closure. The hybrid data sources can complement each other to generate high quality assemblies.

We present the Finishing Module, an integrated software package, which adds functionality to the CLC Genomics Workbench and CLC Genomics Server to aid the process of genome finishing. The Finishing Module seamlessly integrates into CLC Genomics Workbench and provides additional functionality for aligning contigs to reference, contig analysis, modifications, and coverage annotations. Moreover, the module includes tools for automated primer design, gap closure, contig extension, manual and automated joining of contigs.

Here we demonstrate how the contigs from *de novo* assembled bacterial genomes (Pseudomonas and Salmonella) were processed in the CLC bio Finishing Module. The contigs were assembled using our new *de novo* assembler that allows for optimization of assemblies using different combinations of word and bubble sizes. The resulting contigs were analyzed and aligned to the closest reference genomes in the Finishing Module. The primers for gap closure were designed using the automated primer design tool. This semi-automated workflow helps replace many manual tasks in genome finishing and closure.

FF0118

Using the MicroScope Web-based Platform to Analyze RNA-Seq Data

Marion Weiman, David Vallenet, Gregory Salvignol, Clauding Medigue, and Stephane Cruveiller

Laboratorie D'Analyses Bioinformatiques pour la Genomique et le Metabolisme, Umr 8030 CEA/Genoscope – CNRS – Universite d'Evry, 2 rue Gaston Cremieux, 91057, Evry, Cedex, France

MicroScope is a well-established web-based platform primarily dedicated to microbial genome (re)annotation and comparative genomics [1]. It integrates several databases and software tools allowing advanced automated genome annotation and provides user-friendly web interfaces to query and curate gene annotations. MicroScope also provides several layers of analytical tools focused on (i) comparative genomics, (ii) the reconstruction and analysis of metabolic networks, (iii) the integration of functional genomics data (e.g. mutant phenotypes [2]), and, more recently, (iv) the analysis of bacterial polymorphism evolution from high-throughput sequencing (HTS) data [3].

Among the applications of HTS, RNA sequencing (RNA-Seq) offers significant improvements over microarrays [4]. RNA-Seq methods provide direct access to transcript structure, are not limited to a predefined list of transcripts, and cover a larger dynamic range of expression levels. As RNA-Seq data could provide value added to primary annotations, at both syntactic and functional levels, we recently integrated an RNA-Seq analysis and visualization module in the platform. This module is made of two main components: a HTS data analysis pipeline coupled with a database storing results, and a web-based visualization interface.

RNA-Seq data analysis is a rapidly evolving field. Experimental protocols are still under optimization and data processing methods are not fully settled yet. For instance, quantification and normalization of expression levels are still debated [5] and some biases introduced by experimental protocols are not properly handled [6]. Part of improvements we will implement in MicroScope RNA-Seq module will therefore follow the state of the art in this field. In addition, we plan to extend the module to include analyses of the structure of transcripts. We currently design an analysis pipeline aiming at locating Transcription Starting Sites (TSS) from TSS specific RNA-Seq experiments. Results from this pipeline will then be combined with transcript coverage data to build global transcription maps.

[1] D. Vallenet, S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Medigue, MicroScope: a platform for microbial genome annotation and comparative genomics, *Database*, 2009:bap021, 2009.

[2] E. Giraud, L. Moulin, D. Vallenet, V. Barbe, E. Cytryn, J. Avarre, M. Jaubert, D. Simon, F. Cartieaux, Y. Prin, G. Bena, L. Hannibal, J. Fardoux, M. Kojadinovic, L. Vuillet, A. Lajus, S. Cruveiller, Z. Rouy, S. Mangenot, B. Segurens, C. Dossat, W.L. Franck, W. Chang, E. Saunders, D. Bruce, P. Richardson, P. Normand, B. Dreyfus, D. Pignol, G. Stacey, D. Emerich, A. Verméglio, C. Médigue, and M. Sadowsky, Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia, *Science*, 316:1307-1312, 2007.

[3] S. Wielgoss, J.E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R.E. Lenski and D. Schneider. Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3*, 1(3):183-186, 2011.

[4] F. Ozsolak and P.M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nature Reviews. Genetics*, 12:87-98, 2011.

[5] J.H. Bullard, E. Purdom, K.D. Hansen and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[6] J. Li, H. Jiang, and W.H. Wong, Modeling non-uniformity in short-read rates in RNA-Seq data, *Genome Biology*, 11:R50, 2010.

FF0120

The Best Finish First: Sequence Finishing with Whole Genome Mapping

Deacon Sweeney, Emily Zentz, Nianqing Xiao, Vadim Sapiro

OpGen, Inc., Gaithersburg, Maryland, USA

Even the best sequencing efforts can result in undetected sequence gaps and assembly errors. These challenges compromise the accuracy of microbial identification and confound genomic comparisons. Additionally, sequence finishing with libraries is an expensive and time-consuming interruption to a high-throughput pipeline. OpGen's unique Whole Genome Mapping platform offers a global view of genomic structure, by which contigs can be ordered and mis-assemblies detected. This approach results in faster, more accurate, and often less expensive finishing of the DNA sequence from microbial genomes.

For a Sequence Placement project, a *de novo* Whole Genome Map (WGM) is first assembled using OpGen's Argus® platform. Utilizing the WGM and set of sequence contigs as input, the Sequence Placement application leverages a sophisticated scoring system to confidently identify their locations in the genomic map. Placements, gap sizes, and mis-assembled contigs are automatically identified and reported. The software typically runs in less than a minute and was designed for easy integration into sequencing pipelines.

The algorithm operates by a best-first approach, where aligned contigs are sorted by confidence score and placed sequentially, with highest confidence first. Once a region of the map is occupied by a contig, the region is no longer available for additional placements. This approach results in a step-wise elimination of opportunities for spurious placements due to reduction and segmentation of the search space. The best-first approach confidently places contigs with as few as three internal restriction fragments, and typically places between 70 and 90% of the sequence contigs (depending on quality and length).

This poster will demonstrate both the algorithm and its implementation for sequence placement. We will illustrate its operation on a model microbial genome, and provide summary statistics collected from training and testing sets.

FF0127

The Genome Reference Consortium

Kate Auger and Jonathan Wood, on behalf of the Genome Reference Consortium

Wellcome Trust Sanger Institute, Cambridge, UK

The Genome Reference Consortium (GRC1) was formed by the Wellcome Trust Sanger Institute (WTSI), the European Bioinformatics Institute (EBI), The Genome Institute at Washington University (TGI) and the National Center for Biotechnology Information (NCBI) to ensure that the human, mouse and zebrafish reference assemblies are biologically relevant by producing alternative assemblies of structurally variant loci where necessary, alongside the closure of gaps and correction of misrepresented regions. The efforts of the GRC culminate in the release of assemblies that better represent regions with complex allelic diversity, ultimately providing more robust substrates for genome analysis within a chromosomal context (Church DM, et al2). Whilst WTSI and TGI hold equal responsibility for the human and mouse genomes, the zebrafish genome is solely attributable to WTSI.

The reference assemblies are clone-based, with the vast majority of clones sequenced, assembled and finished using the established Sanger capillary pipeline. In the era of next generation sequencing, WTSI are utilising the strengths of the Illumina sequencing platform to create and release draft assemblies of clones to supplement the genome reference assemblies (see poster by J. Wood). Protocols for finishing using Illumina assemblies have also been developed.

The GRC has created operation procedures to improve the quality of an assembly. These include reporting, tracking, assessing and correcting errors discoverable with the genome analysis tools provided by the partners, plus separate customised analyses to aid the evaluation of an assembly against a certain data type. In addition, we are receiving user information on existing issues and subsequently integrate them into our workflow. These procedures are made transparent by the available issue listings on the GRC home page and the highlighting in Mapviewer3 and Ensembl4 genome browsers. We also provide public access to PGPviewer5, the genome evaluation browser.
To ensure modernised regions are accessible by the research community, the GRC has generated a system to update current reference sequences via the release of genome patches in a minor assembly update occurring between full major assembly releases. At the time of release, patch scaffolds are not integrated into the reference chromosomes and therefore do not disrupt the coordinate system of the reference assembly. The system has initially been employed for the human reference assembly and patches can be viewed in popular genome browsers, such as Ensembl, Mapviewer and UCSC6. There are two types of genome patches: FIX and NOVEL, and examples of both will be presented.

1genomereference.org, 2Church DM, et al. (2011) Modernizing Reference Genome Assemblies. PLoS Biol 9(7): e1001091.doi:10.1371/journal.pbio.1001091, 3ncbi.nlm.nih.gov/mapview, 4ensembl.org, 5pgpviewer.ensembl.org, 6genome.ucsc.edu

FF0128

The Illumina Clone Assembly Pipeline at the Wellcome Trust Sanger Institute

Jonathan Wood and Kate Auger

Wellcome Trust Sanger Institute, Cambridge, UK

At the Wellcome Trust Sanger Institute (WTSI) the capillary sequencing pipeline has successfully generated clone sequence data for many years. Using an established shotgun read assembler, Phrap1, and a proven approach to manually finishing the resultant clone assemblies; there had not been a reason to change the processes until recently, with the rapid development of next generation technologies. The Illumina platform provides a method in which to generate a large amount of data, relatively cheaply, to generate draft clone assemblies as a substrate for further improvement.
The Illumina Clone Assembly (ICA) (Andrew Whitwham, WTSI) pipeline was designed to assemble exclusively Illumina sequence data from indexed clones, across all species, in the absence of whole genome shotgun (WGS) capillary read assemblies. The purpose of the pipeline is to maximise the quality of automatic assemblies for submission as draft sequence prior to the finishing process. It utilises two assemblers, SOAPdenovo2, to maximise the assembled contig size independent of WGS scaffold information, and ABySS3 to address potential errors in the SOAP assembly caused by aggressive read clipping.

Although PCR-free Illumina libraries are preferable, the process is not currently automated, consequently precluding its use for highly indexed pools. Therefore, limited-PCR libraries, using KapaHiFi DNA polymerase, were created from individually prepped BAC and fosmid clone DNA. These indexed libraries, with an insert size of 450-550bp, were then pooled prior to cluster formation and sequencing of 100bp paired end reads on the Illumina HiSeq2000 platform.

The generated Illumina data is manually fed in to the ICA pipeline where the process is run individually for each clone. The data is screened for vector and E.Coli contamination, quality filtered by Phusion4 and then passed to SOAPdenovo and ABySS for assembling; these two assemblies are then merged together. Once the merged assembly is created, SMALT5 indexes and maps reads to the contigs and generates SAM files for viewing in the Gap5 editor6, Each assembly takes about 15 minutes and a 96 clone multiplexed lane run on a cluster takes a maximum of two hours. Gap5 is a genome assembly editor created specifically to handle the large volume of short read data generated, enabling it to display the complete assembly and allowing the traditional finishing process to proceed.

Although plans for future developments are not definite, ICA could theoretically assemble whole genomes as well as data sets from a combination of platforms, rendering it invaluable within the clone sequencing pipeline at WTSI.
1phrap.org
2De novo assembly of human genomes with massively parallel short read sequencing (Li R et al 2009) Genome Res
3ABySS: A parallel assembler for short read sequence data (Simpson JT et al 2009) Genome Res
4The phusion assembler (Mullikin JC, Ning Z 2003) Genome Res
5SMALT - http://www.sanger.ac.uk/resources/software/smalt/ (Ponstingl H, Ning Z 2010)
6Gap5—editing the billion fragment sequence assembly (Bonfield JK 2010) Bioinformatics

FF0140

Development of Forensic-Quality Mitochondrial DNA Data on the Illumina Platform

Rebecca S. Just[1,2,] Mark Whitten[3], Mingkun Li[3], Elizabeth A. Lyons[1,2], James P. Ross[1,2], Odile M. Loreille[1,2], Jodi A. Irwin[1,2]

[1]American Registry of Pathology, Rockville, MD, USA
[2]Armed Forces DNA Identification Laboratory, Dover AFB, DE, USA
[3] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The development of forensic-quality mitochondrial DNA (mtDNA) reference profiles is essential to the use of mtDNA in casework scenarios. However, the generation of such reference data by Sanger sequencing is laborious and costly, requiring thoughtful strategy design, automated laboratory processing, fully electronic data handling and extensive data review to avoid errors. Next-generation sequencing (NGS) promises rapid, cost-effective data production when samples are sequenced in parallel, and has potential utility for the development of forensic reference profiles. But NGS presents substantial challenges to consensus sequence determination at the level of surety required in forensics, given the massive amount of data generated by NGS platforms, sequencing errors in general and technology-specific drawbacks (e.g. 454 homopolymer-length errors), and considerations related to heteroplasmy detection and reporting.

As a first step toward developing NGS data handling strategies for forensic mtDNA applications, entire mitochondrial genome (mtGenome) consensus sequences were developed from eighty-five anonymous population samples using parallel tagged sequencing in a single Illumina GA-IIx run, and the resulting profiles were compared to Sanger sequence data for portions of the mtGenome. The results indicated that even with high depth of coverage, standard secondary analysis of the Illumina reads will produce some ambiguous positions, as well as discrepancies with the Sanger consensus sequences in regions prone to length heteroplasmy. However, sequence ambiguities due to point heteroplasmy could be accurately distinguished from an excess of sequencing error in the short-read data, and in most instances of length heteroplasmy the majority molecule could be correctly inferred by applying region-specific alignment parameters. The opportunity to process tens or hundreds of samples in parallel using next generation sequencing technologies should eventually accelerate the production of mtGenome reference data for forensic purposes, provided effective alignment and data evaluation strategies are developed to ensure forensic data quality standards are maintained.

FF0150

Next Generation Pathogen Characterization Workbench for Global Biosurveillance

Helen Cui, Tracy Erkkila, and Patrick Chain,

Los Alamos National Laboratory

Responding to the rapid growth of biological data generated by high throughput biotechnologies, in particular, nucleotide sequencing, microarray expression and pathway analysis, and large-scale modeling and simulation, innumerable databases and analytic tools have been developed and continue to be developed. These databases and tools are widely distributed providing extensive utility to a spectrum of audiences including individual, organization, and the public.

The data, information, and knowledge accumulated by these databases and analytic tools are tremendous assets to the public health and biodefense community to support pathogen characterization, infectious disease diagnosis and spread prediction, forensics and attribution analysis, and integrated global biosurveillance. However, a community information technology architecture is lacking for analysts, researchers and decision makers to access the ever growing body of data and tools through a shared platform or workbench to make the most of such assets.

Los Alamos National Laboratory has initiated the development of preliminary information technology platform and components to provide access to existing information and analytic tools, and to generate actionable inference and recommendations. The fundamental scientific knowledge of host-pathogen-environment interaction guides the design of the platform and its implementation. Numerous analytic algorithms and tools are being developed throughout the community at large for genomic and metagenomic sample analysis, epidemiology modeling, and other applications. Samples of developed analytic tools are being adapted by and applied to the workbench by flexible and scalable approaches, with user oriented interface and security. Multi-step data processing can be accomplished with a workflow manager such as Galaxy that provides a flexible way to organize and execute a computational analysis pipeline. We are exploring its production pipeline utility and developed a server hosting more than 800 bioinformatics tools that can be rapidly accessed and linked to construct computational pipelines.

Establishing a next generation pathogen characterization pipeline and workbench is a highly desired core interagency capability that enables rapid analysis and action upon biothreat and infectious disease data, information and knowledge.

This presentation combines the work supported by the US Department of State and Defense Threat Reduction Agency.

FF0153

A Highly Configurable SNP Caller for the Ion Torrent Personal Genome Machine

Christian Buhay[1], Qiaoyan Wang[1], Huyen Dinh[1], Imad Khalil[1], Michael Holder[1], Yuan-Qing Wu[1], Mark Wang[1], Lora Lewis[1], Christie Kovar[1], David Wheeler[1], Donna Muzny[1], Eric Boerwinkle[1,2] and Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030
[2]University of Texas Health Science Center at Houston, School of Public Health, Houston, TX 77030

The BCM-HGSC has been evaluating Life Technologies' Ion Torrent Personal Genome Machine (PGM) sequencing platform throughout 2011. Orthogonal validation is one of the principal applications for the PGM. We have developed a highly configurable SNP variant caller purpose-built for Ion Torrent sequence. Preliminary data suggests validation rates of 90% or better using purely automated means.

The PGM's performance as a viable validation tool was evaluated using five amplicon pools totaling roughly 3000 sites. Utilizing the Ion Torrent mapper (TMAP), more than 94% of all reads produced aligned to the targeted pool. Average coverage across the pools was 850X, with 98% of all target bases covered at 40X or better. These projects were originally sequenced on the SOLiD platform then orthogonally validated with Roche/454. The intersection of automated PGM validation results with 454 was initially 70% - 85%. Many of the initial variant calls were incorrect due to insertions or deletions at or around the validation site. We developed a modified variant caller that builds on Samtools pileup, parses and filters the pileup for various user-defined parameters such as overall site coverage, reference and minor allele site coverage, and allele frequency. Two TCGA projects were re-validated using the modified variant caller. In an interrogation spanning 2000 sites, automated Ion Torrent concordance with 454 rose to 94%. Additional analysis on whole exome capture and Ion's Ampliseq panels on the PGM are underway.

FF0156

Marine Microbial Eukaryote Transcriptome Sequencing Project

Arvind K. Bharti, Robin Kramer, Connor Cameron, Ken A. Seal, Alex G. Rice, Kathy Myers, Greg D. May, John A. Crow and Callum J. Bell

National Center for Genome Resources (NCGR), Santa Fe, NM 87505 USA

Members of the kingdom protista are believed to be earth's first eukaryotes. The concept of their evolution and relationship have been changing as new protists are continuously being discovered. Marine microbial eukaryotes comprise a vast array of single-celled, nucleated microbes, including diatoms, dinoflagellates, amoeba, ciliates and water molds. These organisms fill numerous ecological roles ranging from photosynthetic primary producers (base of aquatic food webs) to heterotrophic consumers of pre-formed organic compounds. Phytoplankton is responsible for generation of up to half of the world's oxygen. Despite their great abundance and importance, the gene content of oceanic microbial eukaryotes has not been studied extensively. Therefore one of the major goals of this G.B. Moore Foundation funded project is to generate a catalogue of expressed genes from 750 such microbes. So far 742 samples representing 358 unique species have been approved for sequencing, assembly and annotation. The RNA libraries will be sequenced from both ends on the Illumina Hi-Seq 2000 platform, assembled and annotated based on searches against the PFam, SUPERFAMILY and TIGRFAMs HMM libraries. All paired-end sequence reads will be deposited in the ENA (European Nucleotide Archive) while the assemblies, annotations and metadata will be made available through CAMERA (camera.calit2.net). Primary focus of this project is to increase the research community's scientific knowledgebase and also improve metagenomic analyses of complex marine communities. For more details, please visit the project web site MarineMicroEukaryotes.org.

FF0157

Repeat Reduction in Illumina Libraries Using Duplex-Specific Nuclease Prior to Sequencing

Alexander Kozik[1], Lutz Froenicke[1], Marta Matvienko[1,2], Dean Lavelle[1], Belinda Martineau[1], and Richard W. Michelmore[1]

[1]The Genome Center, University of California, Davis, CA, 95616, USA
[2]CLC bio [http://www.clcbio.com]

Current DNA sequencing technologies provide opportunities to generate massive amounts of sequence data. Analyses of large plant and animal genomes have been complicated by the presence of repetitive sequences of varying degrees of complexity and sequence divergence. Several uses of sequence data, such as gene and SNP discovery as well as genotyping, would benefit from libraries with reduced abundance of repeated sequences. We refined a method for reducing the high-copy components in libraries prior to sequencing on Illumina GA and HiSeq platforms. DNA libraries are denatured to single strands and then allowed to partially reanneal. Treatment with a thermostable duplex-specific nuclease (DSN) after an appropriate reannealment period results in the selective destruction of the more rapidly reannealing high-copy sequences, leaving the low-copy component to be amplified and sequenced. As a part of the Compositae Genome Project http://compgenomics.ucdavis.edu, the lettuce transcriptome and gene space have been sequenced using this repeat-reduction approach, assembled and analyzed using CLC Genomics Workbench. Experiments were designed to investigate the consequences of variables in the DSN protocol. These demonstrate that 2- to 3-fold enrichment of gene space can be achieved for large plant genomes such as lettuce (2.7 Gb) that are comprised of ~70% repeated sequences.

FF0160

An Analysis of the Genomic Architecture at Risk Loci for SLE

Ekta Rai[1], Benjamin Wakeland[1], Chaoying Liang[1], Prithvi R Sharma[1], Kasthuribai Viswanathan[1], David Karp[1], Nancy Olsen[1], Igor Dozmorov, Laurie Davis, Pratik Doshi, Graham Wiley, Ken Kaufman[2], John Harley[2], Patrick Gaffney[2], Edward K. Wakeland[1]

[1]Department of Immunology, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, USA
[2]Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA.

Susceptibility to SLE is impacted by both genetic and environmental factors. More that 30 SLE susceptible loci have been identified, however, the causal variants responsible for these associations are largely unknown. We performed targeted resequencing of 4.3 Mbases in 30 SLE associated LD blocks in 192 Caucasians (107 SLE cases and 85 controls). In these targeted regions, an average of ~99% bases was captured by at least one non-redundant read and ~93.4% bases by at least 15 non-redundant reads, yielding average fold coverage of ~107X. The high quality sequencing calls were confirmed by having >99% concordance with the Immunochip array data. A total of ~21,000 variations (SNPs and Indels) were identified, of which ~38% were novel. Of the total variations, ~32% potentially impacted function, categorized as ~5% non-synonymous; ~4% synonymous; ~13% UTR; ~2% deleterious; ~1% splice; and 37% cis-eQTL, Most of the non-synonymous and deleterious variations were rare, suggesting that either they are newly evolved or have been subjected to purifying selection. Interestingly, a high accumulation of rare deleterious variations restricted to cases were observed in the loci reported to have major effect on SLE susceptibility in humans or animal models (C1Q; TREX1; C2-CFB; MSH5; PTPN22 etc). To explore the allelic architecture of functional variations in tight LD with SLE tagging SNPs, phylogenic networks were drawn using the Median joining network system. This analysis identified a specific CLADE of alleles containing multiple SLE associated SNPs in several regions (BLK, TNIP1, JAZF1, TNFAIP3, LYN etc.) that have strong cis-eQTL impact. These results indicate that many risk loci for this autoimmune disease contain a diverse array of allelic haplotypes in tight LD with disease “tagging” SNPs with variable contributions to disease susceptibility.

FF0163

Microbial Identification Through 16S Sequencing on the Ion Torrent PGM

Ginger A. Metcalf, Embriette R. Alicki, Michael E. Holder, Huyen Dinh, Mark Wang, Joseph F. Petrosino, Donna M. Muzny and Richard A. Gibbs

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

In early 2011, the Human Genome Sequencing Center at Baylor College of Medicine began evaluating Life Technologies' Ion Torrent Personal Genome Machine (PGM) sequencing platform. The PGM has proven to be a platform capable of producing sequence data for a variety of applications with a relatively short turn-around time.

Recent experiments at the HGSC have shown the platform to be ideal for amplicon sequencing. To date these amplicon experiments have been largely focused on validation of known variants discovered on other platforms. Here we evaluate possible metagenomic applications of the Ion Torrent PGM by sequencing amplicons generated from the V4 region of 16S ribosomal RNA. The V4 region is one of 9 variable regions found in prokaryotic rRNA that are used to identify the numerous microbial species found in metagenomic samples. In this study, amplicons were generated from metagenomic DNA obtained from mouse stool samples and sequenced in a multiplexed Ion Torrent PGM run. The test samples were used in a previous study which exhibited a significant difference in the microbial communities in the intestines of mice with IgA deficiency compared to those isolated from wild type mice. This comparison was performed through the analysis of 16S sequence data generated through pyrosequencing on a 454 GS instrument.

The sequence data generated from the Ion Torrent PGM was analyzed using the CloVR 16S pipeline. The output of the CloVR pipeline was used to evaluate the differences in the relative abundance of the various microbial taxa present across each sample. This information is to be compared to the results that were generated using the sequencing data generated on the 454 instrument for the same group of samples. This comparison will help to assess the utility of the Ion Torrent PGM in successfully sequencing metagenomic samples for taxonomic identification.

FF0164

Reproductive Genetics and Development in the Fungus *Myceliophthora heterothallica,* a Thermophilic Model for the Chaetomiaceae

Miriam I. Hutchinson[1], Amy J. Powell[2], Kylea J. Parchert[2], Joanna L. Redfern[1], Andrea Martinez[1], Martha Perez-Arriaga[1], Randy M. Berka[3], Adrian Tsang[4], Eric Ackerman[2], Blake Simmons[5], Igor V. Grigoriev[6], Stephen R. Decker[7], Michael E. Himmel[7] and Donald O. Natvig[1]

[1]Department of Biology, University of New Mexico, Albuquerque NM 87131; [2]Sandia National Laboratories, Albuquerque, NM 87123; [3]Novozymes, Inc., Davis, CA 95618; [4]Concordia University, Montreal, Quebec, Canada; [5]Sandia National Laboratories, Livermore, CA 94551; [6]DOE Joint Genome Institute, Walnut Creek, CA 94598; and [7]National Renewable Energy Laboratory Biosciences Center, Golden, CO 80401

Members of the Chaetomiaceae are among the most reported fungi in studies of biomass degradation. They are of interest for their ability to produce thermostable carbohydrate-active enzymes. This has led to the sequencing of genomes from two thermophilic species, *Myceliophthora thermophila* and *Thielavia terrestris*. Until now, there has been no genetically tractable model either for this family, or more generally, for thermophilic fungi. We have characterized reproduction in the thermophile *Myceliophthora heterothallica* toward the goals of establishing this organism as a model for the group and developing it as an expression platform. *M. heterothallica* was reported to be heterothallic based on the fact that matings between two strains resulted in the production of fruiting bodies and ascospores. Prior to our work, however, heterothallism had not been confirmed with independent assortment of mating loci and autosomal genes. We speculate that this lack of confirmation of true heterothallism resulted from a failure to obtain ascospore germination. We found that ascospores are resistant to germination at temperatures below 47-50°C. This discovery allowed us to confirm heterothallism and analyze the segregation of markers in crosses. Sequences from opposite mating types show that mating regions are conserved relative to other Sordariales. Interestingly, different stages of development have different temperature optima: ascospore germination occurs at 47°C and above, ascocarp formation is optimal at 30°C, and growth is optimal at 45°C. We have successfully crossed *M. heterothallica* strains from Indiana, New Mexico and Germany, and we are expanding the number of known strains by surveying across latitudinal and elevation gradients. In addition, we are developing methods for transformation and gene replacement. Our goal is to develop *M. heterothallica* as a model organism to study fundamental aspects of thermophily and the biology of Chaetomiaceae.

FF0165

Improvement of Microbial Genomes with Evolving Sequencing Technologies

A.C.Munk, K. Davenport, O. Chertkov, H. Daligault, W.Gu, H. Teshima, X. Zhang, D., Bruce, C. Detter, Y. Xu, R. Tapia, T. Yilk, B. Quintana, K. Reitenga, Y. Kunde, L. Green, T. Erkkila, C.Han, and P. Chain

Joint Genome Center-Los Alamos National Laboratory

As DNA sequencing technology evolves and improves, JGI-LANL's microbial genome sequencing pipeline has been (is being) adapted to use Illumina data as a draft. To take advantage of the strengths of different assembly programs, Illumina short -insert (270 bp) libraries and long-insert (4-12kbp) libraries are assembled with Allpaths and also with Velvet. The resulting assemblies are combined in a Phrap assembly. Consed is used to design PCR reactions to span remaining gaps. PCR products are pooled and sequenced with PacBio. Filtered subreads are collected for each PCR product and assembled separately, and appropriate contigs are added to the Phrap assembly to close gaps. Finally, reads from the Illumina draft data are mapped back to the assembly to correct for SNPs and indels.

Alternatively, we are exploring whole-genome PacBio sequencing as an additional draft data set. Draft assemblies will be generated using both Illumina and PacBio data. Gaps may be closed using contig-end sequences to enrich for PacBio reads which extend across the gaps.

Future plans include analyzing the gaps remaining in Illumina-only draft data, exploring the possibility of closing gaps by mapping reads back to the draft assembly, and using longer PacBio reads to replace long-insert mate-pair libraries by linking Illumina contigs to make scaffolds.

FF0172

The 3BSEQ Project : Getting the Most of Newbler® (Assembler/Assemblies)

Alberti Adriana[1], Couloux Arnaud[1], Choulet Frederic[2], Theil Sebastien[2], Mangenot Sophie[1], Vacherie Benoît[1], Magdelenat Ghislaine[1], Barbe Valerie[1], Feuillet Catherine[2], Wincker Patrick[1]

[1]CEA/DSV/IG/Genoscope, LF, Evry, France
[2]INRA, UMR1095 Génétique Diversité et Ecophysiologie des Céréales, F-63100 Clermont-Ferrand, France

The 3BSEQ project aims at obtaining an annotated sequence of chromosome 3B, the largest bread wheat chromosome (1Gb) , and at exploiting this knowledge to develop tilling arrays of the 3B gene space for further functional and structural characterizations .The project takes advantage of the potential offered by the next generation sequencing to develop an original strategy and deliver a high quality draft sequence of the chromosome.

We sequenced the ~10'000 BAC clones of the minimal tiling path established during the construction of the physical map. As sequencing each BAC individually would have been too costly and labour-intensive, we used a pooling strategy to combine BACs from the same physical contigs in one or more pools of all, DNA was extracted from each BAC individually using a standard method and then, pools corresponding to 10 equimolar amounts of BAC DNA were mixed. Each pool was used to create a mate pair library of 454-Roche Titanium fragments through the re-circularization of 8kb fragments. In addition, we produced pair end illumina reads of sorted chromosome 3B at 44x coverage. The first 454 sequence assembly resulted in about 16 000 scaffolds (293 000 contigs) with a N50 of 270 kb and an average of 5 scaffolds per physical contigs. To reduce the number of contigs and improve the scaffolding, we developed a strategy to improve the continuity based on Newbler work files. The method uses the Illumina reads to fill gaps and correct scaffold consensus errors. Details of the method will be presented.

FF0173

Endosymbiont Hunting in the Metagenome of Asian Citrus psyllid (*Diaphorina citri)*

Surya Saha[1], Wayne B. Hunter[2] and Magdalen Lindeberg[1]

[1] Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, NY, USA
[2] USDA-ARS, US Horticultural Research Lab, Fort Peirce, FL, USA

The Asian citrus psyllid (D. citri Kuwayama or ACP) is host to 7+ bacterial endosymbionts and is the insect vector of Ca. liberibacter asiaticus (Las), causal agent of citrus greening. To gain a better understanding of endosymbiont and pathogen ecology and develop improved detection strategies for Las, DNA from D. citri was sequenced to 108X coverage. Initial analyses have focused on Wolbachia, an alpha-proteobacterial primary endosymbiont typically found in the reproductive tissues of ACP and other arthropods. The metagenomic sequences were mined for wACP reads using BLAST and 4 sequenced Wolbachia genomes as bait. Putative wACP reads were then assembled using Velvet and MIRA3 assemblers over a range of parameter settings. The resulting wACP contigs were annotated using the RAST pipeline and compared to Wolbachia endosymbiont of Culex quinquefasciatus (wPip). MIRA3 was able to reconstruct a majority of the wPip CDS regions and was selected for scaffolding with Minimus2, SSPACE and SOPRA using large insert mate-pair libraries. The wACP scaffolds were compared to wPip using Abacas and Mauve contig mover to orient and order the contigs. The functional annotation of scaffolds was evaluated by comparing it to wPip genome using RAST. The draft assembly was verified using an OrthoMCL based comparison to the 4 sequenced Wolbachia genomes. We expanded the scope of endosymbiont characterization beyond wACP using 16S rDNA and partial 23S rDNA analysis as a guide. Results will be presented regarding endosymbionts, their potential interactions and their impact on the disease of citrus greening.

FF0174

DTRA Algorithm Prize

Christian Whitchurch

Defense Threat Reduction Agency

As *n*th generation DNA sequencing technology moves out of the research lab and closer to the diagnostician's desktop, the process bottleneck will quickly become information processing. The Defense Threat Reduction Agency (DTRA) and the Department of Defense are interested in averting this logjam by fostering the development of new diagnostic algorithms capable of processing sequence data rapidly in a realistic, moderate-to-low resource setting. With this goal in mind, DTRA is sponsoring an algorithm development prize.

The Challenge:

Given raw sequence read data from a complex diagnostic sample, what algorithm can most rapidly and accurately characterize the sample, with the least computational overhead?

The Stakes:

$1,000,000

Prize details and sequencing datasets will be made available this Fall. Monitor http://www.dtra.mil/Business.aspx for updates on this program.

FF0176

Phenotypic to Genotypic Characterization of *Bacillus anthracis* BACI293 Identifies Bacteriophage Insertions

Carson Baldwin[1], Susan Coyne[1], Michelle Shipley[1], Christine Munk[2], Shannon Johnson[2], Mark Wolcott[1], Jeff Koehler[1] and Tim Minogue[1]

[1]Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases, 1425 Porter Street, Fort Detrick, MD, 21702
[2]Los Alamos National Laboratory, Bikini Atoll Rd., SM 30, Los Alamos, NM 87545

Pathogen identification and characterization is a critical aspect to the diagnostic process. Multiple technologies can be applied towards the goal of linking phenotypic observation with genotypic correlates. Whole genome optical mapping is one such technology and can provide a highly detailed restriction map of a genome. During this process, high molecular weight DNA molecules are purified and digested with a specific restriction enzyme. These fragments are resolved based on length, and the fragments are assembled into a highly detailed, whole genome map. This technology has been applied in multiple aspects of the diagnostic process including pathogen identification, sequence assembly, and comparative genomics. In the specific context of comparative genomics, we characterized a panel of *Bacillus anthraces* strains by numerous phenotypic and genotypic methodologies. During this characterization, one isolate that displayed unique morphologies on selective media, BACI293, contained two inserts identified by optical mapping not observed in the other *B. anthracis* clade C strains. We conducted draft sequencing and preliminary assembly using Illumina and 454 sequencing data, resulting in 41 scaffolds and 461 contigs. These contigs were analyzed *in silico* using the OpGen optical mapper and aligned against BACI293 and BACI008, a strain that does not contain the two inserts. Through this process, we identified a contig that contained both of the inserts. The sequence for the two inserts was identified, and a BLAST analysis identified 90-100% query coverage for several *B. anthracis* bacteriophages. Further analyses of the two inserts could identify gene expression cassettes contained within the bacteriophage inserts or disruption of *B. anthracis* genes resulting from the bacteriophage insertion that resulted in the altered morphology observed on selective media.

FF0177

Use of Selective Sequencing Probes for Pathogen Identification by Next-Generation Sequencing

Adrienne Hall[1], Jeff Koehler[1], Alex Rolfe[2], Tim Minogue[1]

[1]Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases, 1425 Porter Street, Fort Detrick, MD, 21702
[2]Pathogenica, 27 Drydock Avenue, Boston, MA, 02210.

Next-generation sequencing (NGS) is an important tool for pathogen identification from clinical and environmental samples. One problem often encountered is trying to identify pathogen sequence from the background of host or metagenomic sequences. One potential solution is the use of selective sequencing to target specific genomic signatures, and Pathogenica's DxSeq probes use this approach to capture targeted genomic regions for NGS. Specifically, DxSeq probes are designed with complementary ends that hybridize to the flanking regions of the desired target DNA. The gap between the 5' and 3' ends of the probe is filled by a polymerase, and the probe is circularized through ligation. PCR amplification across the captured sequence from the now circularized probe amplifies the target signature and adds sample barcodes and sequencing adapters to the PCR product. This approach is expandable, capable of including thousands of probes against potentially hundreds of pathogens. Using this technology, a panel of sequence capture probes was designed bioinformatically to discriminate multiple filoviruses including Ebola strains Sudan, Zaire, Reston, Cote d'Ivoire, and Bundibugyo as well as Marburg strains Angola, Musoke, and Ci67. Each probe was screened for a PCR amplification product using cDNA from each virus. From the initial 90 probes designed, 69 probes generated a detectable PCR product. Amplification products were combined and sequenced using Roche's GS Junior, and the resulting reads were aligned against the reference sequences for each virus. From this sequencing analysis, 42 probes resulted in at least 10 reads aligning to the targeted capture sequence. Overall, this process resulted in probe coverage for almost all of the filoviruses. This sequence capture technology coupled to sequencing on small-throughput, clinical-targeted sequencer represents a viable option to screen for, in a single reaction, multiple pathogens.

FF0184

Approaches for *de novo* Assembly of Difficult Genomes

Sarah Young, Carsten Russ, Iain MacCallum, Filipe J. Ribeiro, Aaron M. Berlin, Sean Sykes, Terrence P. Shea, Sakina Saif , Ted Sharpe, Shuangye Yin, Sante Gnerre, Dariusz Przybylski, The Sequencing Platform, Bruce J Walker, David B. Jaffe, Chad Nusbaum

Broad Institute of MIT and Harvard, Cambridge, MA, USA

Advances in sequencing and assembly methods have enabled routine generation of high quality *de novo* assemblies of genomes from short reads. However, some genome projects still pose significant challenges due to limitations in sample quantity or quality, aspects of genome composition or limitations of sequencing technology:

- Poor sample quality can have a deleterious effect on library construction
- Unculturable genomes require whole genome amplification, which introduces bias
- Genomic repeats larger than the library insert length are unresolvable
- Low complexity sequences longer than a read length are difficult to resolve
- Samples can contain mixed populations of organisms due to the nature of the sample or due to contaminants

To address some of the challenges described here, we have developed and applied a number of laboratory and analytical techniques:

- Complementary sequencing methods can overcome biases and increase resolution of genomic structure
- Jumping libraries with wide insert size ranges help resolve repeats
- Data normalization ameliorates effects of extreme sequence bias
- Removing polymorphism pre-assembly and introducing it back in post-assembly can provide longer more contiguous sequences

We applied these methods and assessed the assemblies using an analysis package developed to help identify and interpret key assembly qualities (see abstract by Sykes, *et al*). Even on complex data sets, our automated processes now can generate high quality genome assemblies, useful as a backbone for genome improvement, or directly for genomic research.

FF0187

Genome Mis-assemblies – A Challenge to Characterize

Sakina Saif, Sean Sykes, Bruce Walker, Sarah Young, Aaron Berlin, Terrance Shea, Sante Gnerre, Amr Abouelleil, Alma Imamovic, Harindra Arachchi, Jerome Naylor, Carsten Russ , Chad Nusbaum

The Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, MA, USA

Mis-assemblies result from incorrect reconstruction of the genome. They are often identified when regions of assembled contigs fail to agree with same region in a closely related reference genome. However in *de-novo* genome assemblies, mis-assemblies are much more challenging to identify.
Various characteristics of genomes contribute to mis-assemblies and result in assembly algorithms placing sequencing reads incorrectly:

- **a)** Duplicated sequences
- **b)** Nonidentical copies of repeats that are hard to differentiate from the sequencing errors
- **c)** Regions of the genome that are undercovered due to sequencing bias

Two types of structural mis-assemblies commonly occur:

- **a)** Local mis-joins at erroneous insertion or deletion sites
- **b)** Extensive mis-joins at erroneous rearrangements and inversions

Through this work, we attempted to identify signatures of genome mis-assemblies as described above and understand their characteristics with the help of computational tools (see abstract by Walker, *et al*.). For our analysis we chose the following genomes for their diverse nature and availability of finished references (*): *Escherichia coli* MG1655*, *Mycobacterium tuberculosis* KZN 1435*, *Rhodobacter Sphaeroides* 2.4.1*, *Staphylococcus aureus* USA300* and *Bacillus cereus* BMG1.7.
For each of these genomes, we generated ALLPATHS-LG assemblies using Illumina 101b paired-end reads from 180bp fragment and 3kb jump libraries (see abstract by Young, *et al*.) and analyzed them in detail (see abstract by Sykes, *et al*.). With the goal of identifying signatures in the data that would indicate the presence of a misassembly, we then examined genomic datasets and annotations such as Nucmer repeats, kmer coverage, read coverage, percent GC and ambiguity across the genome using the Integrated Genomics Viewer (IGV) application.

We intend to generate a set of mis-assembly metrics for each genome that include mis-assembly rate, mis-assembled contig and base numbers, and annotations that indicate strong evidence of mis-joins. This will give us a better understanding of why these mis-assemblies happen and suggest ways for automated fixes.

FF0192

High Sensitivity Detection and Typing of Mixed Contributor DNA Samples Using Massively-Parallel Deep Amplicon Pyrosequencing

Jared Latiolais[1], Andrew Feldman[2], Jeffrey Lin[2], Plamen Demirev[2], Ishwar Sivakumar[2], Thomas Mehoke[2], and Robert Bever[1]

[1]The Bode Technology Group, Lorton, VA
[2]The Johns Hopkins Applied Physics Laboratory, Laurel, MD

We demonstrate that Roche 454 Titanium chemistry based deep sequencing of individual PCR amplicons yields greater levels of sensitivity for detecting minor contributor's in a mixed forensic sample than conventional sequencing methods.

We present data from a series of mixture de-convolution experiments using both mitochondrial DNA and nuclear STR amplicons. The data sets show highly sensitive detection of low copy contributor STRs well below the 1-10% level, which is the approximate limit for gel electrophoresis. Dilutions of human DNA in mixtures from 1-5 different contributors at different relative concentrations were prepared for PCR amplification and subsequent searching by the 454 system. In addition, “touch” samples were also collected and sequenced from objects touched by multiple individuals. We fabricated fusion primers for 8 CODIS STR loci routinely sued in forensic casework. Following PCR amplification, emulsion PCR is performed and clonally amplified in a massively-parallel fashion. Flow-based pyrosequencing is then performed to determine ~500 bp of the sequence of DNA coupled to each bead. Each bead sequence correlates to one single molecule of DNA which represents one individual contributor. The ratios of the major to minor contributor sequence reads at each locus enables linkage of detected alleles across different loci to form the respective genotypes. A typical 454 sequencing run yields up to 100 thousand reads, thus offering minor contributors detection thresholds at less than one part per thousand.

We also observed that while pyrosequencing is not error-free, the impact of these errors on tandem repeat analysis is minimized through modest bioinformatics analysis of sequences against all known alleles. The system also offers the potential to both detect and sequence new alleles in the course of routine forensic casework and thus can be an important feature tool in the boarded forensics community.

FF0199

Efficient and Accurate Metagenomics Search Using a Desktop Computer and a Large Scalable Persistent Memory Device

Jonathan Allen, Sasha Ames, David Hysom, Kevin Mcloughlin, Shea Gardner, Scott Llloyd, Maya Gokhale, Tom Slezak

Livermore National Laboratory

The increased use of bench top sequencing and increase in sequencer output rates is generating a growing demand for the high performance computing resources required to annotate and analyze large metagenomic data sets. While large compute clusters provide a critical resource for realizing highly parallelized database sequence searches, new methods for efficient annotation of metagenomic data is needed to keep up with growing use of sequencing. Here we give preliminary results showing the use of custom built k-mer genome databases annotated with detailed taxonomic information to support efficient and accurate taxonomy labeling of fungal, microbial and viral genomes down to the strain level. The approach is limited by the amount of available single address space memory but this limitation is circumvented through the use of lower cost more scalable persistent memory (flash drive) devices. Results show that a large metagenomic dataset can be searched against a k-mer database that exceeds 512GB of memory to efficiently label the metagenomic contents on a desktop computer with a flash drive.

FF0201

A Transposon-Based Bacterial Pathogen Gap Closure Method Using Ion Torrent

Anna Montmayeur, Harindra M. Arachchi, Caryn McCowan, Gary Gearin, Amr Abouelleil, J. Pendexter Macdonald, Niall Lennon, Rachel Erlich, Dana Robbins, Bruce W. Birren, Jennifer Wortman, and Michael G. FitzGerald

The Broad Institute, 320 Charles Street, Cambridge, MA

Finished genome sequence is an important component of our bacterial pathogen research program. Complete sequence facilitates identification of drug resistance and virulence mechanisms, epidemiological work, and provides a complete reference for mapping related strains. Since the advent of the next-generation sequencing revolution we continue to work toward development of laboratory finishing approaches for our bacterial pathogen genomes that take advantage of the high coverage resulting from next generation sequencing platforms.

Both labor-intensive sample library construction processes and abundance of data inherent to next-generation sequencing combine to make it inapplicable for finishing the small numbers of gaps currently observed in near-finished Illumina or hybrid assemblies. Instead, our laboratory finishing protocol is now exclusively based on the sequencing of PCR amplicons generated around gap regions using Sanger. In cases of scaffold gaps or large captured gaps, iterative walking along PCR amplicons is often necessary when no known reference exists, and thus becomes an unacceptably long process that we are continuously working to advance.

The Ion Torrent sequencing platform has given us the means to cost-effectively process a moderate quantity of large amplicons in a very time efficient manner, using limited amounts of starting material. Coupling this technology to an existing transposon-based library construction kit (Nextera) has also greatly reduced the time consuming steps of sample barcoding and adapter addition. Our pilot sample pool consisted of 11 amplicons, ranging in size from 3.9 to 10 kb, generated around gaps from assemblies, including that of the antibiotic resistant *Enterococcus casseliflavus* strain EC20 (899205). Here we will describe the laboratory procedure developed for generating a multiplexed sample pool using the Nextera transposon-based library construction kit in conjunction with the Ion Torrent sequencer.

We also describe a novel assembly strategy developed specifically to make use of the ultra-deep coverage obtained for each of our samples. In depth analyses of gap closure success rates using complete Ion Torrent data versus random samplings of the data is detailed. All amplicons were analyzed against reference data available from our production (Sanger) finishing process. Preliminary results demonstrate our ability to obtain accurate sequence data at very high coverage for all of our test amplicons.

This project has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900018C.

FF0210

Mind the Gap: Upgrading Reference Genomes with the Pacific Biosciences RS Long-Read Sequencing Technology.

Adam English, Stephen Richards, Yi Han, Jiaxin Qu, Xiang Qin, Vanesa Vee, Mark Wang, Kim Worley, Eric Boerwinkle, Donna Muzny, Jeffrey Reid, and Richard Gibbs

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

We describe a method to upgrade draft genome references and new genome assemblies using long-sequence reads. Many organisms have had their genomes “sequenced” to high-quality draft status using whole genome assembly techniques. High-quality draft genome sequences contain gaps and other imprecision due to several factors: imperfect whole genome assembly techniques, sequencing biases, repetitive genomic features, and polymorphism. Traditionally, draft genomes were improved using Sanger based manual finishing processes. For more facile assembly and automated finishing of draft genomes, we used long Pacific Biosciences RS reads (PacBio), where 7kb reads are not uncommon. Using PacBio reads, we upgrade the draft genome sequences of a simulated Drosophila melanogaster, the version 2 draft Drosophila pseudoobscura (D. pseudo) genome sequence, and a submission of the Assemblathon 2.0 parrot data. Using 5.6x coverage of PacBio long reads, we saw reads address 62.8% of gaps in our parrot assembly, which closed 26% and improved 67% of gaps. With 20x mapped coverage of PacBio long reads, we addressed 99% of gaps and were able to close 69% and improve 16% of gaps in D. pseudo. Results were validated using Sanger sequencing on 96 randomly selected gaps in D. pseudo. Software; to both automate the finishing process using long read data and provide “lift-over” co-ordinate tables to easily port existing annotations to the upgraded assembly, is publically available in an open source repository.

FF0213b

A High-Throughput qPCR Based Method for the Quantification and Quality Control of Human Genomic DNA Samples

Maryke Appel, Michelle Nderu, Eric Van Der Walt, Liesl Joubert, Gavin J. Rush, John F. Foskett Iii, and Paul J. Mcewan

Kapa Biosystems, Woburn, MA

A number of innovations have reduced the amount of input DNA required for next-generation sequencing (NGS) library construction. Large sequencing facilities have scaled protocols down to cut reagent costs and to reduce sample input requirements to 100 ng DNA or less. Similarly, transposase-mediated Nextera™ library construction typically requires 50 ng DNA or less. Some sample types – such as formalin-fixed paraffin-embedded (FFPE) tissue or forensic samples – provide DNA at low concentrations and of variable quality.

Current standard methods for assessing DNA at low concentrations for NGS such as PicoGreen® and NanoDrop, have significant limitations, namely poor sensitivity, contaminant interference, and poor power for predicting successful library construction. Without reliable methods for determining the concentration and quality of input DNA, library construction remains inconsistent and inefficient. We have developed and validated a qPCR-based method for the quantification and qualification of human genomic DNA samples prior to NGS library construction. This method utilizes a set of quality-controlled, prediluted DNA quantification standards and a number of separate PCR primer pairs defining amplicons of various sizes within a highly conserved human target sequence. The assay can be adapted for forensic applications such as sex typing and mitochondrial DNA typing.

FF0216

Short Tandem Repeat (STR) Analysis from Short Read Sequencing Data

Daniel Bornman, M.S., Seth Faith, Ph.D., Brian Young, Ph.D.

Battelle Memorial Institute

The cost and speed of nucleic acid sequencing is rapidly decreasing due to next-generation sequencing technologies, enabling novel opportunities for collecting numerous forensically relevant data from single data sets. In this study, we demonstrate interrogation of human saliva samples with massively parallel sequencing technologies (Illumina) for evaluating the allele status of the 13 core CODIS short tandem repeat (STR) loci. To enable the analysis of STR from next generation sequencing (NGS) data we modified the approach for standard reference genome alignment yet preserved the ability to leverage open-source short read alignment software. Following an enrichment PCR step to amplify a ~2 kb region surrounding each CODIS STR site, we performed multiplexed sequencing on an Illumina GAIIx of five individuals, plus an equal parts mixture on two of the five samples. This sequencing run generated read lengths of 150 nucleotides that were aligned to a modified reference genome specifically constructed to enable analysis of diploid STR polymorphic patterns. Reference alignment was performed using the Bowtie short read alignment program that generated standard sequencing file formats for downstream analysis and interpretation. Analysis of the final sequence alignment data enabled us to correctly estimate (with complete concordance to CE) STR allele status for the entire set of core CODIS loci from each sample including the single 1:1 mixture sample. Some exceptions included two instances where the repeat pattern for the D21S11 locus had repeat lengths greater than could be covered by the current maximum 150 bp read length. Data are also presented that demonstrates the use of a novel filtering method that preprocesses the raw sequencing data for enrichment of short reads containing STR patterns. This filtering method provides up to an 8-fold increase in speed for completion of the STR calling method from NGS data and enables the data to be processed on computing resources comparable to a desktop computer.

FF0219

The Loblolly Pine Genome Project

Daniela Puiu

Johns Hopkins University

The Loblolly Pine Genome Project (LPGP) is part of the larger Pinerefseq project and has as goal the sequencing, assembly and annotation of the Loblolly pine (Pinus taeda), the most ecologically and economically important conifer in SE United States. Development of a high-quality reference genome will serve as a model for assembling two other conifers: Douglas fir (Pseudotsuga menziesii) and Sugar pine (Pinus lambertiana).

With 12 chromosomes, a genome size of ~24Gbp, and high repeat content, this genome presents a major challenge for genome assembly. We are sequencing the genome using the Illumina technology from a combination of whole-genome shotgun and fosmid pool libraries.

Johns Hopkins School of Medicine along with Univ. of Maryland College Park are currently working on sequence library qc, read correction, whole genome & fosmid pooled assembly and chloroplast & mitochondrion finishing. We are evaluating existing read correction and assembly software as well as developing new methods targeted towards this particular genome. We will present our data processing, assembly strategy and preliminary results.

FF0220

Sequencing and Informatics Core at NCGR

Peter Ngam, Jennifer Jacobi, Pooja Umale, Anitha Sundarajan, Juanita Martinez, Dominique Aló, Ingrid E. Lindquist, Connor Cameron, Thiru Ramaraj, Patricia Mena, Nico Devitt, Robin Kramer, Alex Rice, John Crow, Faye Schilkey, and Callum Bell

National Center for Genome Resources, Santa Fe, NM

The National Center for Genome Resources (NCGR) is a non-profit research institute whose mission is to improve human health and nutrition through genome sequencing and analysis. The NCGR Sequencing and Informatics Core offers next generation sequencing, genotyping, and bioinformatics services in the state, regionally and nationally. The Illumina CS-Pro certified lab houses the Illumina HiSeq and GAIIx, and the PacBio RS for sequencing and Illumina's BeadExpress and iScan for genotyping. Sequencing services include whole genome and transcriptome shotgun sequencing, ChIP, small RNA, long-insert paired end, ultra-low sample input (1ng) DNA, and targeted exome sequencing. NCGR provides follow-up informatics analysis to these experiments including *de novo* assembly (transcriptome and genome), read count-based expression and SNP reports, and differential expression analysis. The core also develops cutting edge bioinformatics to solve challenging analysis problems.

NCGR has a longstanding reputation for developing effective bioinformatics tools and performing cutting-edge scientific research, for example*: Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing,* Sci Transl Med. 2011 and *Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis,* Nature 2010 Apr 29. NCGR received the 2009 Bio-IT World Best Practices award for basic research in Schizophrenia and was the Computerworld 2009 Laureate in recognition of its Alpheus® system, a pipeline for variant and expression detection of next generation sequencing data.

Providing additional leverage to NCGR's own research, sequencing and analysis experience, this core provides an unprecedented knowledgebase and facility to researchers locally, regionally, nationally and around the globe. Please contact Patricia Mena at seq@ncgr.org to effectively integrate these technologies and analyses into your own work.

Grant Support: The National Institutes of Health's National Center for Research Resources (5P20RR016480-12) and the National Institute of General Medical Sciences (8P20GM103451-12).

FF0235

Hybrid Illumina PacBio De Novo Microbial and Fungal Assemblies

Alicia Clum, Matt Blow, Igor Grigoriev, Jeffery Martin, Len Pennacchio, Hui Sun, Tanja Woyke, Alex Copeland

DOE Joint Genome Institute, Walnut Creek, California 94598, United States of America

At JGI we sequence genomes to support DOE missions. JGI continues to test new protocols, technologies, and assemblers to facilitate efficient and effective de novo assembly. This poster focuses on hybrid Illumina and PacBio assemblies of microbial and fungal genomes. Currently, microbial and fungal projects are assembled using ALLPATHS-LG (Gnerre, S. et al.,), a whole-genome shotgun assembler which utilizes short read data like that generated by the Illumina platform, and more recently, long read data from the PacBio platform. We have demonstrated a reduction in the number of contigs and an increase in the contig N50 of assemblies with the inclusion of PacBio data for microbial and fungal sized projects. The number of contigs is reduced by 28% in fungal genomes and 60% in microbial genomes. As technology improves, goals change as well. With the emergence of PacBio data and the ability of ALLPATHS-LG to hybrid assemble it along with short read data, the goal is to achieve a closed microbial genome and the best possible assembly for fungal genomes.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

FF0250

Whole Mitochondrial Genome Sequencing Utilizing Probe Capture and Next Generation Sequencing

Valerie McClain[1], Cassandra Calloway[2], George Sensabaugh[3]

1-University of California, Davis, Davis, California, 2-Children's Hospital Oakland Research Institute, Oakland, California, 3-University of California, Berkeley, Berkeley, California

Mitochondrial DNA (mtDNA) testing is most useful in forensic cases when samples are degraded and nuclear STR testing cannot produce a complete discriminating profile. Current approaches include sequencing the hypervariable regions (HVI/HVII) with Sanger sequencing. The maternal inheritance pattern of mtDNA and sequence information from only the hypervariable regions provides limited discrimination power, particularly in the Caucasian population. Sanger sequencing is also limiting in that it often fails to detect heteroplasmy as well as mixtures which are common in forensic samples.

We propose to apply Next Generation Sequencing coupled with the sequence capture system to overcome these limitations. The probe capture technique would allow the entire mitochondrial sequence to be analyzed and has the potential to be applied to limited and degraded samples since it is not dependent on specific primer sequences. Next generation sequencing has proven useful to detect mixtures including heteroplasmy at lower levels than current Sanger sequencing methods.

Typical next generation sequencing methods require starting DNA amounts higher than a typical forensic sample; however, we have successfully produced libraries for capture with significantly less starting material (100pg) and are working to optimize for even more limited DNA amounts. Here we show the development of a method for whole mitochondrial genome sequencing using liquid phase probe capture followed by 454 next generation sequencing for the application to forensic science.

FF0257

JGI Resequencing: Recognizing False Positives, False Negatives and Structural Variation in User Data

Anna Lipzen, Wendy Schackwitz, Joel Martin, and Len Pennacchio.

Lawrence Berkeley Laboratory United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598

The technological advances responsible for continuous increases in sequencing yield and reduction in cost-per-base of Next Generation Sequencers present new demands for data analysis. Evaluation of high sequence data volume should be systematic, rapid and (ideally) fully-automated; however it should also permit easy customization and flexibility to accommodate various project types. The JGI Resequencing team performs detection of single nucleotide polymorphisms, short insertions and deletions and structural variations in a wide-range of bioenergy-relevant organisms. The team assists with upfront project design and aids the collaborator with interpretation of sequencing results. The resequencing pipeline consists of an automated component that allows for fast, high-throughput project turnaround and a custom component where analysis parameters are optimized for collaborator's specific organism and where results are manually perused. While the automatic tools (maq, samtools, BreakDancer) are great for variant discovery, false positives and false negatives still pose a problem. Common sources of false positive calls include: collapsed repeats, ambiguously mapped reads, edges of structural variation and Illumina sequence specific errors. Sources of false negative calls include: library bias and sequence divergence. Additionally, the available tools often misidentify the exact breakpoints of structural variants. For this reason, when feasible, we provide users with custom analysis that flags probable false calls and successfully resolves breakpoints of structural variations. We will present some of the techniques used in our custom analysis here.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

FF0259

Production Sequencing Platforms at the DOE Joint Genome Institute

Chris Daum[1], Matt Zane[1], Alison Fern[1], Diane Bauer[1], Laura Sandor[2], Megan Kennedy[1], Mojgan Amirebrahimi[2], Nicole Shapiro[2], JGI Sequencing Platforms Team, & Chia-Lin Wei[2]

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

The U.S. Department of Energy (DOE) Joint Genome Institute's (JGI) Sequencing Technologies group is committed to the generation of high-quality genomic DNA sequence to support and accelerate DOE user's science in the mission areas of renewable energy generation, global carbon management, and environmental characterization and clean-up.

Within the JGI's Sequencing Technologies group a robust Production Sequencing Platforms group has been established, which utilizes state-of-the-art DNA sequencing technologies to provide low cost and high throughput sequencing capacity for JGI users. The Sequencing Platforms group operates an automated sample library prep process on the Caliper Sciclone G3 platform, which then supplies samples for sequencing on the Group's Ilumina HiSeq & Miseq and Pacific Biosciences RS sequence analyzer platforms. Optimization of these platforms has been ongoing with the aim of continual process improvement of the laboratory workflow, reducing operational costs and project cycle times to increase sample throughput, and improving the overall quality of the sequence generated. A sequence QC analysis pipeline has been implemented to automatically generate read and assembly level quality metrics.

An overview of these sequencing platform technologies and the foremost of these optimization projects, along with sequencing and operational strategies, throughput numbers, and sequencing quality results will be presented.

FF0260

Multiple Long Mate Pair Approaches to Facilitate Short Read Based *de novo* Genome Assembly

Julianna Chow[1], Jaya Rajamani[1], James Han[2], Alicia Clum[1], Alex Copeland[1], Shweta Deshpande[1], and Chia-lin Wei[1]

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

Short reads based *de-novo* assembly is challenging for complex genomes with variable homologous regions and considerable repetitive elements. Long-Mate Paired (LMP) libraries of different jumping sizes offer potential solutions to bring short reads of different gap sizes into contigs and improve the intactness of the draft genome assembly. Furthermore, LMP sequences can facilitate the ordering of contigs into scaffolds and detect structural variations like indel and translocations. Here, we report the developments of two complementary LMP library construction methods: Cre-Lox Inverse PCRIllumina Paired-End (CLIP-PE) and Ligation Free Paired-End (LFPE). In CLIP-PE, libraries are created by ligation of biotin-LoxP adaptors to the ends of fixed sizes gDNA and circularized by Cre recombinase mediated intra-molecule recombination. The circularized product is fragmented by a selection of 4-base pair cutting enzyme (*NlaIII, MseI or HpyCH4IV*). The ends of fragmented DNA are self-ligated, biotin-streptavidin selected and enriched by inverse PCR. Mate-Pairs generated by CLIPPE libraries are identified by the 4-base pair enzyme cutting site. LFPE libraries are created by ligation of internal adaptors lacking 5' phosphate to the ends of gDNA fragments and circularization by hybridization. The circularized product is nick translated, digested by T7 Exonuclease/ S1 Nuclease into short paired tags of fix span sizes, biotin-streptavidin selected and finally, enriched by PCR. Mate-Pairs generated from LFPE libraries are identified by the internal linker junction site. Two organisms *Mycosphaerella fijiensis CIRAD86* and *Phycomyces blakesleeanus NRRL1555* were selected to test the effectiveness of these two LMP library creation methods. Sequencing data generated from the two methods were analyzed to evaluate the resulted assembled genomes for their qualities, coverage, redundancy, bias and accuracy. We conclude that both methods have their unique advantages and disadvantages for various genome assembly applications.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

FF0262

Fungal Genome Improvement

Hui Sun, Alicia Clum, Kurt LaButti, Igor Grigoriev, Alex Copeland

DOE Joint Genome Institute, Walnut Creek, California 94598, United States of America

This poster focuses on fungal genome improvement methods at the JGI. We compare results generated from Illumina only ALLPATHS-LG assemblies, hybrid Illumina and PacBio ALLPATHS-LG assemblies, hybrid Roche (454) and Illumina Newbler assemblies, and hybrid Roche and Illumina Newbler assemblies improved with Fosmid primer walk data. Adding Sanger fosmid primer walk data is a traditional method used in genome improvement and has shown to reduce the number of contigs and increase contig N50. However, this process can be labor intensive and time consuming. Our goal is to develop genome improvement methods which do not rely on Sanger data. Long read data from the PacBio platform and the ability of ALLPATHS-LG to hybrid assemble it with short read data or post assemble it has shown to have effectively and efficiently improved assembly stats. The number of contigs has reduced and the contig N50 has increased after PacBio data was incorporated for the fungal genomes we tested and this resulting data can be used for comparative studies of fungal biology.

FF0270

Evolution of Young Gene Duplicates in the Human Genome

Lijing Bu, Jieying Yang and Vaishali Katju

Department of Biology, University of New Mexico

The duplication of existing DNA is frequent and ubiquitous in all three domains of life. Gene duplication is widely accepted as a driving force in the evolution of large, complex genomes from smaller, simpler ones. While one gene copy suffices to perform the ancestral function, the extra copy creates opportunities for the generation of evolutionary novelties. Evolutionary analyses focusing on the early life stages of gene duplicates would advance our understanding of how a new functional gene is generated following a duplication event. Studies of two model organisms, the nematode *C. elegans* and the yeast *S. cerevisiae* have revealed divergent patterns of young duplicate formation and retention. In humans, duplicated genes account for more than one third of the total gene number, and play a role not only in shaping phenotypic novelty, but in increased incidence of some genetic diseases. This ongoing study aims to investigate the early evolutionary history of recently originated duplicates in the human genome. First, the GenomeHistory program was implemented to screen for and identify human gene duplicates. Putative young gene duplicates generated in the human genome post-dating the human-chimpanzee split were identified using synonymous divergence cutoff of < 0.1. Data on several features of these human paralogs are being collected (the chromosome location, transcriptional orientation, degree of synonymous and nonsynonymous divergence, exon-intron structure, duplication span and extent of upstream and downstream flanking region homology) to elucidate the structural features, genomic characteristics, and early evolution of these new entrants into the human genome.

FF0298

Investigating Host-Pathogen Interactions via Transcriptome Sequencing — A case study using a *Yersinia pestis*-THP-1 cell infection model

Bin Hu[1], Christopher J. Stubben[2], and Patrick S. Chain[3]

[1, 3] Genome Science Group (B-6), Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, 87545
[2] Biosecurity & Public Health, Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, 87545

A global understanding of the expression of virulence genes during pathogen-host interactions will provide critical insights into the transcriptional regulatory networks of infectious diseases. We analyzed RNA-seq data from experiments that mimic the infection of Yersinia pestis. RNAs from Y. pestis CO92 strain cultured at 26°C and 37°C were enriched, and rRNAs were depleted by using the Epicentre RiboZero kit prior to sequencing with a lane of Illumina. We also analyzed the data of infecting human THP-1 cells with Y. pestis, to better understand the pathogen's response to extracellular contact as well as intracellular survival. All together these experiments generated about 24 billion reads, from which we identified many up- and down-regulated Y. pestis genes. The data also allow the identification of both previously predicted and new putative small RNA candidates. Furthermore, during the THP-1 infection model allows us to investigate human cell transcriptional response to Y. pestis infection. Our results suggest RNA-seq provides an unprecedented resolution to study the transcriptome regulation in both pathogen and host cells during infection and allows the identification of novel candidate small RNAs whose role in pathogenicity has yet to be determined.

FF0299

High Throughput Plasmid Sequencing with Illumina and CLC bio.

Jing Lu, Stacie Norton, Ajay Athavale, Susan Johnson, Karen Martin, Dan Ader

Monsanto Company, St. Louis, MO

The Genomic Analysis Center at Monsanto plays an integral role in product development by supplying sequencing and finishing resources to the companies R&D pipeline. A key component of these efforts are the sequencing and analysis of plasmids destined for plant transformation, ensuring the accuracy of inputs to the transgenic plant pipeline. This workflow has been supported by Sanger sequencing for >10 years; however the advent and productionization of new sequencing technologies has presented an opportunity to increase throughput and reduce costs for this workflow. Illumina was identified as the optimal sequencing platform for these efforts based on cost, throughput, accuracy and “finishability” after a comparative study of the Illumina, SOLiD and 454 technologies. Adoption of the Illumina platform also required a new assembly and analysis platform which was found in CLC Bio. Monsanto’s finishing team have worked collaboratively with CLC to refine the software suite for improved finishing and analysis in both the Sanger and Illumina workflows. The new workflow represents a significant savings on both reagents and FTE, while reducing the overall turnaround time for plasmid sequencing.

FF0300

Burkholderia pseudomallei Genome Sequencing, Comparisons and Data Management

Shannon Johnson, Todd Yilk, Tracy Erkkila, Matt Scholz, Sanaa Ahmed, Patrick Chain, & Chris Detter
LANL B6 Genome Science Group, Los Alamos National Laboratory, USA

Special acknowledgement to our project collaborators at BYU, LLNL, NAU, & NBACC in the USA and DSTO, Griffith, JCU, Menzies, & PathWest in Australia as well as our primary sponsor Traci Pals at DHS Science & Technology, USA.

Burkholderia pseudomallei is primarily a soil bacterium, in humans it causes melioidosis (also called Whitmore’s disease) a disease with a mortality rate of 20-50%. The bacterium is considered endemic to Southeast Asia and Northern Australia. Previous studies have found B. pseudomallei to exhibit a great deal of phylogenetic diversity between environmental andclinical isolates. In order to gain a better reference database of genomic sequences for clinical isolates from across Northern Australia we set out to sequence 25 clinical strains to Improved High Quality Draft (IHQD) status.

To date 22 strains have been sequenced and assembled to the level of IHQD. A cross comparison with a fully finished reference B. pseudomallei strain from GenBank,found 125,584 SNPs of which 51K are unique to a single organism and very few are shared between the remaining strains sequenced here. This indicates that the high level of sequence diversity found in preliminary studies is reflected in the whole genome sequencing data. Many of the rearrangements and SNPs located in this comparison of IHQD sequenced genomes may be validated or refuted if the sequencing level is brought up to fully finished. There are between 200 and 700 genomic rearrangements per strain, between the assemblies from this project and the selected reference.

A website to provide collaborative access to the data from this project has been developed. The website provides an organized approach to hosting the data and analysis results, provides some analysis capabilities, and allows for data and results to be visualized. Users can access the website to browse data and results by project and by sample. Data files can be uploaded and downloaded, BLAST queries against data on the website can be performed. In the future, posted results will be visualized by Jbrowse, and a discussion forum will provide a collaborative space to track conversations on projects and results.

FF0304

Inducing Artificial Polyploidy in *Bacillus subtilis* Improves Genomic Recovery: A Novel Method to Advance Single Cell Genomics

Armand E K Dichosa[b], Michael S Fitzsimons[b], Chien-Chi Lo[a, b], Lea L Weston[b], Lara G Preteska[b], Jeremy P Snook[b], Xiaojing Zhang[a, b], Wei Gu[a, b], Kim McMurry[a, b], Lance D Green[a, b], Patrick S Chain[a, b], J Chris Detter[a, b], and Cliff S Han[a, b]

[a]The DOE Joint Genome Institute, [b]Bioscience Division, Los Alamos National Laboratory, MS M888, Los Alamos, New Mexico 87545, USA.

Genomic amplification through multiple displacement amplification (MDA) has allowed researchers to yield micrograms of genomic amplicons from a single chromosome copy. As such, MDA with single cell isolation and much sequencing efforts have made it possible to determine bacterial phylotypes and metagenomic profiles from an environmental sample *in situ* without the need for cultivation. However, low genomic coverage and gaps observed in the assembled genome are inherent issues surrounding MDA when using a single chromosomal template, making it difficult to obtain an accurate genomic profile. To complete the genomic assembly, additional genomic template and labor are needed.

Our prior investigations have shown that increasing the number of chromosomal template greatly reduces gaps in the assembled genome, thereby producing a more complete genomic profile. However, isolating an adequate number of the exact species/strain from an environmental sample without cultivation is in itself difficult. Consequently, we hypothesized that simultaneously inhibiting bacterial cells from completing cytokinesis while maintaining viability will phenotypically result in larger-than-normal cells possessing at least two complete chromosomes. To test our hypothesis, we used PC190723, an antimicrobial inhibitor of the bacterial and euryarchaeal cell division protein FtsZ, to prevent *Bacillus subtilis* ATCC 6633 from completing cytokinesis. Cytographic data via flow cytometry have shown that PC190723-treated *B. subtilis* increased to near double cell size, while qPCR analyses quantified more genomic content and four times less amplification bias from the FtsZ-inhibited cells compared to the untreated controls. Ultimately, *de novo* genomic assemblies suggest that a single polyploid *B. subtilis* cell contributes 17% more genomic coverage than an untreated cell. Our study implies that similar FtsZ-inhibitors, whether singly or in combination, can be used to induce artificial polyploidy in a microbial community and, when utilized with high-throughput, FACS-based analyses, a more complete genomic and/or metagenomic profile can be achieved with greater efficiency.

Poster Session Notes

Poster Session Notes

06/06/2012 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	Breakfast	x	Santa Fe Breakfast Buffet	Sponsored by NEB
8:30 - 8:45	Intro	X	Welcome Back	TBD
x	Session Chair	x	Session Chairs	Chair - Mike Fitzgerald Chair - Tina Graves
8:45 - 9:30	Keynote	FF0043	Plague: A Highly Fit Clonal Pathogen Emerges and Shapes Human History	Dr. Paul Keim
9:30 - 9:50	Speaker 1	FF0101	Finishing and Special Motifs: Lessons Learned From CRISPR Analysis Using Next Generation Draft Sequences	Catherine Campbell
9:50 - 10:10	Speaker 2	FF0160	An Analysis of the Genomic Architecture at Risk Loci for SLE	Ward Wakeland
10:10 - 10:30	Speaker 3	FF0279	Resolve the Cancer Heterogeneity by Single Cell Sequencing	Xun Xu
10:30 - 11:00	Break	x	Beverages and Snacks Provided	x
11:00 - 11:20	Speaker 4	FF0075	Consed and BamView for Next-gen Sequencing	David Gordon
11:20 - 11:40	Speaker 5	FF0065	Integrating Data from Multiple Human Genome Sequencing Platforms and Bioinformatic Methods to Analyze their Error Profiles and Form Consensus Variant Calls	Justin Zook
11:40 - 12:00	Speaker 6	FF0088	One Chromosome, One Contig: Hybrid Error Correction and *de novo* Assembly of Single-Molecule Sequencing Reads	Sergey Koren
12:00 - 1:20pm	Lunch	x	New Mexican Lunch Buffet	Sponsored by Beckman Coulter
x	Session Chair	x	Session Chairs	Chair - Donna Muzny Chair - Johar Ali
1:20 - 1:40	Speaker 7	FF0108	Recent Advances in High-Throughput, Low-Latency Interfacing for Fast Scanning and Metrology in Genomics Applications	Scott Jordan
1:40- 2:00	Speaker 8	FF0186	Pilon Assembly Improvement Software	Bruce Walker
2:00 - 2:20	Speaker 9	FF0188	Putting the Pieces Together: From Assembly to Analysis	Sean Sykes
2:20 - 2:40	Speaker 10	FF0045a	Finding the Perfect Recipe for *de novo* Plant Genome Assembly: A Platform Bake-off	Dan Ader
2:40 - 3:00	Speaker 11	FF0170	Finished Prokaryotic Genome Assemblies From a Low-Cost Combination of Short and Long Reads	Shuangye Yin
3:00 - 3:20	Speaker 12	FF0211	Mercury: A Next Generation Sequencing Data Analysis and Annotation Pipeline	David Sexton
3:20 - 3:35	Speaker 13	FF0004	NCGR Informatics	John Chow
3:35 - 3:50	Speaker 14	FF0174	DTRA Algorithm Prize	Christian Whitchurch
3:50 - 5:15pm	Break & Round Table Discussion (Topics TBD)		Beverages and Snacks Provided for the Round Table Topics TBD: attendees to select from a few choices the week before the meeting	x
5:45 - 8:00pm	Happy Hour	x	Happy Hour at Cowgirls Cafe - Sponsored by LifeTech - Map Will be Provided	Sponsored by LifeTech
8:00 - bedtime	on your own	x	Dinner and Night on Your Own - Enjoy!!!	x

NOTES

Speaker Presentations (June 6th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0043

Plague: A Highly Fit Clonal Pathogen Emerges and Shapes Human History

Paul Keim

Northern Arizona University

Yersinia pestis a highly fit clonal pathogen that has emerged from natural reservoirs to plague humans for thousands of years. Genomic analysis has been a highly productive approach towards understanding it past and current epidemiology. Whole genome sequences, multiple locus genotyping and large databases have allowed the reconstruction of historical and current plague epidemics, as well as, allowing the forensic analysis of a laboratory acquired case. The historical data are consistent with multiple epidemic waves with the epicenter in Asia, particular China. Genomic analysis of trace material from the Middle Ages has allowed a greater understanding of the Black Death. Plague represents a model for tackling other clonal pathogens.

NOTES

FF0101

Finishing and Special Motifs: Lessons Learned From CRISPR Analysis Using Next Generation Draft Sequences

Catherine Campbell, Mitchell Holland, Katharine Jennings, and Matthew McCoy

Noblis, 3150 Fairview Park Drive South, Falls Church VA

In the context of finishing and analyzing genomic sequences, special attention should be paid to regions in the genome that contain repetitive elements that can be used for strain characterization (eg., MLVA, MLST, and CRIPSRs). These regions can be prone to misassembly, especially in draft sequences, and phylogenetic context can be affected if the assemblies in these regions are inaccurate. Clustered regularly interspaced short palindromic repeats (CRISPRs) are a family of repetitive elements in bacteria that consist of sets of repeated DNA sequences (direct repeats or DR) separated by short, unique DNA sequences (spacers). These CRISPR systems are thought to act as an adaptive immune system against invading phages and plasmids. The sequences of the spacers are usually highly polymorphic among strains of the same species and are hypothesized to consist of pieces of phage or plasmid DNA that the host bacterium incorporates into the chromosome to help target incoming foreign DNA for destruction. The DR sequences are often highly conserved within a species and are used to define CRISPR type. CRISPRs therefore have the potential to provide information regarding the environmental history for a bacterial sample, and may be able to provide some context for phylogenetic relationships among strains. We have examined CRISPR sequences in over 60 *Yersinia pestis* strains, some of which were newly sequenced for this project. Using Roche 454 shotgun WGS reads, we encountered some difficulties with the *de novo* assembly of spacers and DR in the CRISPR region, which resulted in the apparent deletion of these spacers in the assembled genome. When used to create phylogenetic trees, these misassembled spacer sequences may cause inaccurate comparisons between strains, and may indicate divergence where none truly exists. Therefore in these experiments we implemented specialized methods to identify misassembled known spacers. Although useful for performing quality checks on assemblies, this technique cannot be used to confirm whether all novel spacers have been identified.

FF0160

An Analysis of the Genomic Architecture at Risk Loci for SLE

Ekta Rai[1], Benjamin Wakeland[1], Chaoying Liang[1], Prithvi R Sharma[1,] Kasthuribai Viswanathan[1], David Karp[1], Nancy Olsen[1], Igor Dozmorov, Laurie Davis, Pratik Doshi, Graham Wiley, Ken Kaufman[2], John Harley[2], Patrick Gaffney[2], Edward K. Wakeland[1]

[1]Department of Immunology, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, USA
[2]Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA.

Susceptibility to SLE is impacted by both genetic and environmental factors. More that 30 SLE susceptible loci have been identified, however, the causal variants responsible for these associations are largely unknown. We performed targeted resequencing of 4.3 Mbases in 30 SLE associated LD blocks in 192 Caucasians (107 SLE cases and 85 controls). In these targeted regions, an average of ~99% bases was captured by at least one non-redundant read and ~93.4% bases by at least 15 non-redundant reads, yielding average fold coverage of ~107X. The high quality sequencing calls were confirmed by having >99% concordance with the Immunochip array data. A total of ~21,000 variations (SNPs and Indels) were identified, of which ~38% were novel. Of the total variations, ~32% potentially impacted function, categorized as ~5% non-synonymous; ~4% synonymous; ~13% UTR; ~2% deleterious; ~1% splice; and 37% cis-eQTL, Most of the non-synonymous and deleterious variations were rare, suggesting that either they are newly evolved or have been subjected to purifying selection. Interestingly, a high accumulation of rare deleterious variations restricted to cases were observed in the loci reported to have major effect on SLE susceptibility in humans or animal models (C1Q; TREX1; C2-CFB; MSH5; PTPN22 etc). To explore the allelic architecture of functional variations in tight LD with SLE tagging SNPs, phylogenic networks were drawn using the Median joining network system. This analysis identified a specific CLADE of alleles containing multiple SLE associated SNPs in several regions (BLK, TNIP1, JAZF1, TNFAIP3, LYN etc.) that have strong cis-eQTL impact. These results indicate that many risk loci for this autoimmune disease contain a diverse array of allelic haplotypes in tight LD with disease “tagging” SNPs with variable contributions to disease susceptibility.

FF0279

Resolve the Cancer Heterogeneity by Single Cell Sequencing

Xun Xu

Beijing Genome Institute

Tumor heterogeneity presents a challenge for inferring clonal evolution and driver gene identification. So far, deep sequencing of tumor tissues or of cancer cell lines has provided an incomplete understanding of the genetic information and mechanisms involved in tumor progression. To provide a more comprehensive picture of the genetic changes that occur in tumors, we have carried out genome sequencing at single cell level of common cancers and assessed the genetic changes that had occurred within these cells. We carried out whole-exome single-cell sequencing on a JAK2- negative myeloproliferative neoplasm patient, indicating a monoclonal evolution in this case. We identified essential thrombocythemia (ET)- related candidate mutations such as SESN2 and NTRK1, which may be involved in neoplasm progression. We carried out single-cell exome sequencing on a ccRCC tumor and its adjacent kidney tissue. Our data showed no significant clonal subpopulations diverse mutation spectrums, and suggested that ccRCC may be more genetically complex than previously thought. We further apply the single cell analysis on different other three cancers, including gastric cancer, liver cancer, bladder cancer, all these provides initial characterization of the disease-related genetic architecture at the single-cell nucleotide level. The single-cell sequencing method that opens new ways to investigate individual tumors, with the aim of developing more effective cellular targeted therapies.

FF0075

Consed and BamView for Next-gen Sequencing

David Gordon and Phil Green

University of Washington, Seattle, Washington, USA

BamView is a new program that displays an overview of reads in a BAM file using little memory, and can bring up Consed to view and edit targeted regions. It plots read depth (with 0-read-depth regions labelled), depth of reads having inconsistently mapped mates, and discrepancy rates (including indels), allowing problem areas to be pinpointed.

Consed can now create an ace file from a BAM file, either for a targeted region or the entire reference sequence. When creating the ace file, Consed's "shallowerDepth" feature allows selection of a representative, more manageable subset of reads in regions of extremely high coverage depth. Cigar strings are now shown.

Consed now has a probability-based method of finding SNPS. This can be run in batch or allowing manual review of each location.

Assembly View now incorporates improved filtering of suspect links.

A read or reads can be fixed at the top of the Aligned Reads window to facilitate scrutiny of particular reads while scrolling.

Consensus bases can be recalculated in batch, with an option to allow manual review of all changes. Consensus bases are recalculated when reads are removed.

Miniassembly is now faster and doesn't duplicate tags.

These and other features will be discussed.

FF0065

Integrating Data from Multiple Human Genome Sequencing Platforms and Bioinformatic Methods to Analyze their Error Profiles and Form Consensus Variant Calls

Justin Zook, Daniel Samarov, Marc Salit

National Institute of Standards and Technology, Gaithersburg, MD

Systematic errors and biases become very important as it is becoming possible to sequence even large whole genomes deeply. Integrating data from multiple platforms and algorithms for the same sample improves understanding of errors and biases and greatly increases the likelihood of accurate variant calls. To generate well-characterized human genome reference materials that will help enable clinical validation of sequencing, we are comparing and integrating the large diversity of data from multiple versions of Illumina, Complete Genomics, SOLiD, and 454 on the human CEU cell line NA12878. Hundreds of thousands of differences exist between whole genome variant calls on this sample from the variety of sequencing platforms, mapping algorithms, and variant calling pipelines. To help understand and resolve these differences, we are comparing the data on the CEU sample and investigating characteristics of the concordant and discordant calls. We have found that many discordant calls have characteristics such as relatively high strand bias, coverage, read position bias, mapping quality, and/or allele balance that distinguish them from most concordant variants. We have developed a Bayesian statistical model with heuristics that can be used to compare and optimize integration of data from multiple platforms, algorithms, and runs, using the information gained about systematic errors. We expect this model to be useful for sequencing applications requiring both high specificity and high sensitivity, as well as for samples with variants at small allele fractions, such as in tumor samples, mitochondrial heteroplasmy, RNA editing, and bacterial strain mixtures. By integrating the large amount of current and future data for NA12878, we expect to form a highly sensitive and specific set of variant calls for this sample that can be used as a reference by clinical and research laboratories to validate and improve sample preparation, sequencing, and bioinformatic pipelines.

FF0088

One Chromosome, One Contig : Hybrid Error Correction and *de novo* Assembly of Single-molecule Sequencing Reads

Adam M. Phillippy[1], Sergey Koren[1,2], Michael C. Schatz[3], Brian P. Walenz[4], Jeffrey Martin[5], Jason Howard[6], Ganeshkumar Ganapathy[6], Zhong Wang[5], David A. Rasko[7], W. Richard McCombie[3], and Erich D. Jarvis[6]

[1]National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702, USA
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
[3]Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724, USA
[4]The J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, MD 20850, USA
[5]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
[6]Howard Hughes Medical Institute, Duke University Medical Center, Department of Neurobiology, Durham, NC 27710, USA
[7]Institute for Genome Sciences, Department of Microbiology & Immunology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Emerging single-molecule sequencing instruments can generate multi-kilobase sequences with the potential to dramatically improve genome and transcriptome assembly. However, the high error rate of single-molecule reads is challenging, and has limited their use to resequencing bacteria. To address this limitation, we introduce a novel correction algorithm and assembly strategy that utilizes shorter, high-identity sequences to correct the error in single-molecule sequences. We demonstrate the utility of this approach on Pacbio RS reads of phage, prokaryotic, and eukaryotic whole genomes, including the novel genome of the parrot *Melopsittacus undulatus*, as well as for RNA-seq reads of the corn (*Zea mays*) transcriptome. Our approach achieves over 99.9% read correction accuracy and produces substantially better assemblies than current sequencing strategies: in the best examples, producing automatically closed bacterial chromosomes without the use of paired ends.

NOTES

NOTES

Lunch

12:00 – 1:20pm

Sponsored by

BECKMAN
COULTER®

Notes

FF0108

Recent Advances in High-Throughput, Low-Latency Interfacing for Fast Scanning and Metrology in Genomics Applications

Scott Jordan

Physik Instrumente

We have entered an era in which new interfacing techniques are enablers, in their own right, for novel scanning, imaging and metrology techniques.
For example, clever leveraging of new interfacing technologies has yielded nanoscale stabilization and resolution enhancement as well as enabling high-throughput processing of scanned specimens.

To assist in choosing and implementing interfacing approaches which maximize performance and enable new capabilities, we review available interfaces such as USB2, GPIB and Ethernet against the specific needs of positioning for the scanned-imaging community. We spotlight new developments such as LabVIEW FPGA, which allows non-specialists to quickly devise custom logic and interfaces of unprecedentedly high performance and parallelism. Notable applications from the leading edge of nanoscience are reviewed, including a clever amalgamation of AFM and optical tweezers, and a picometer-scale-accuracy interferometer devised for ultrafine positioning validation. We note the Serial Peripheral Interface (SPI), emerging as a high-speed/low-latency instrumentation interface. The utility of instrument-specific parallel (PIO) and TTL sync/trigger (DIO) interfaces is also discussed. Requirements of tracking and autofocus are reviewed against the time-critical needs of typical applications (to avoid, for example, photobleaching). A novel planarization approach is reviewed, providing a nanoscale-accurate datum plane over large scan areas (up to 800x800um) without scan-line flattening. Finally, not to be overlooked is the original real-time interface: analog I/O, with novel capabilities introduced in recent months. Here additional advancements are discussed, including a resolution-enhancing technique for analog voltage generation and a useful combination of high-speed block-mode and single-point data acquisitions.

FF0186

Pilon Assembly Improvement Software

Bruce J. Walker, Sarah Young, Sean Sykes, Sakina Saif, Aaron Berlin, Terrance Shea, Sante Gnerre, Alma Imamovic, Amr Abouelleil, Chad Nusbaum

Broad Institute, 320 Charles St., Cambridge MA

We introduce a software program called Pilon, which uses read alignment analysis to diagnose, report, and potentially fix several classes of problems with *de novo* genome assemblies, Pilon is part of a larger suite of tools we have developed for assembly analysis and improvement (see abstract by Sykes, *et al*.). We will show examples of Pilon applied to several recent bacterial assemblies and demonstrate the types of issues it is capable of identifying and resolving.

Prior to running Pilon, raw sequencing reads are aligned to the input assembly with an alignment tool such as Burrows-Wheeler Aligner (BWA) or Bowtie 2 to create aligned Binary sequence Alignment Map (BAM) files. Pilon uses the BAM files along with the assembly and builds pileups and summary statistics at every base position in the assembly. Statistics computed for every locus include counts of “good” versus “bad” read pair coverage, physical coverage, mean insert size of pairs covering the base, mapping quality and base call quality summaries, etc.

Pilon then uses the pileup statistics to confirm good relations of the genome and identify potential problem areas. The first issues identified are base mismatches and small indels identifiable within alignments. Pilon then attempts to identify larger structural problems, such as potential larger insertions or deletions, mis-joins at the contig or scaffold levels, and collapsed repeats. If jumping libraries (mate pairs) are available, Pilon will attempt to report on circularity of scaffold structures (e.g., bacterial chromosomes or plasmids). Pilon summarizes these statistics and identified features, saving a set of files containing tracks of suitable for viewing with tools such as the Integrative Genomics Viewer (IGV). These tracks provide and experienced assembly analyst or finisher with indicators of potential problems areas in the assembly for manual inspection of improvement (see abstract by Saif, *et al.*)

Having identified and reported potential issues, Pilon optionally attempts to fix as many of them as possible. Resolving base mismatches and small indels is usually straightforward. For larger insertions or deletions, Pilon will attempt to find a solution consistent with the input data or to create contig breaks at the suspected locations. Pilon will also break contigs and scaffolds where it identifies mis-assemblies, and will attempt to resolve mixed-haplotypes of repeats by using jumping libraries. Finally, Pilon will attempt to fill in gaps, or at least extend contig ends, by using well-anchored reads near the ends of the contigs. After Pilon applies these fixes and produces a new assembly, the original reads can be re-aligned to the new assembly, and Pilon can be run again. This process can be repeated, potentially confirming more sequence and fixing fewer errors with each subsequent iteration.

FF0188

Putting the Pieces Together: From Assembly to Analysis

Sean Sykes, Sarah Young, Bruce Walker, Terrance Shea, Harindra Arachchi, Sante Gnerre, Jerome Naylor, Aaron Berlin, Sakina Saif, Amr Abouelleil, Alma Imamovic, Chad Nusbaum.

The Broad Institute, 320 Charles Street, Cambridge, MA

Assembling genomes using next-gen sequencing has now become standard practice, and there are roughly a dozen tools available to do it. After a genome has been assembled, there are many standalone analysis tools that one can apply. However, each assembler has its own set of outputs, and each analysis tool has its own input requirements, so that moving data from one to the next becomes a special handling exercise for the user. To simplify and standardize this process, we have built an assembly analysis package that takes output from any assembler, applies a wide-ranging suite of analytical tools (converting data formats where necessary), and produces a standardized set of output metrics, statistics, and graphics.

The analysis suite includes a broad set of generalized analytical modules that fall into several broad categories, including analysis modules that:

- Generate and collect basic assembly statistics and metrics, such as measures of contig and scaffold sizes, gap metrics, ambiguities, repeat content, *etc.*;
- Measure the accuracy of the assembly against a reference genome, when available, at various scales from base-level accuracy to scaffold accuracy
- Align input sequence read data back to the assembly in order to generate coverage metrics and provide assessment and visualizations of integrity and data quality for *de novo* assemblies.

Key elements of this analysis package are standardized outputs, standardized inputs and simplified and automated handling of data. Easily parsable standardized data formats make it more straightforward to build additional modules that can be integrated into the analysis pipeline. With this analysis package, we can easily compare results from multiple assemblers, multiple data types, or multiple levels of sequence coverage to determine the best approach to delivering quality assemblies for a given genome.

The final outputs of these analysis modules are integrated into a comprehensive assembly analysis report, which serves as a powerful tool for assessing the quality and veracity of assemblies, for visualizing potential problems in assemblies (see abstract by Saif*, et al*.), and for driving downstream analyses.

FF0045a

Finding the Perfect Recipe for *de novo* Plant Genome Assembly: A Platform Bake-off

Dan Ader, Ashton Bell, Mitch Sudkamp, Shiaw-Pyng Yang, Randy Kerstetter, Todd P. Michael

Monsanto Company, St. Louis, MO

Sedum album (White stonecrop) is a facultative Crassulacean Acid Metabolism (CAM) plant with the ability to switch between C_3 and CAM photosynthesis to fix carbon dioxide (CO_2). Under normal conditions of water and temperature, *S. album* fixes CO_2 during the daytime (light) through C_3 photosynthesis, while under conditions of limiting water and high temperatures *S. album* switches to CAM photosynthesis and restricts CO_2 fixation to the nighttime (dark) to reduce water loss. *S. album* is an ideal system to understand the mechanisms controlling stress induced C_3-CAM switching, but there are currently no genomic resources available in this species. *S. album* has a small genome similar in size to that of *Arabidopsis thaliana* (142 Mb), and therefore we used it as a test case for how well second and third generation sequencing technologies would perform in *de novo* assembly of a small plant genome. We generated sequence on second generation sequencing platforms Illumina HiSeq2000, LIFE 5500xl, and Roche 454; bench top sequencers Illumina MiSeq and LIFE ION Torrent PGM; and third generation PacBio RS. We compared the error models and sequence bias of each platform and how well each contributed to a *S. album* de novo assembly.

FF0170

Finished Prokaryotic Genome Assemblies From a Low-cost Combination of Short and Long Reads

Shuangye Yin[1], Iain MacCallum[1], Filipe J Ribeiro[1], Dariusz Przybylski[1], Sante Gnerre[1], Ted Sharpe[1], Terrance P Shea[1], Sarah Young[1], Bruce J Walker[1], Nick Patterson[1], Carsten Russ[1], Chad Nusbaum[1], David B Jaffe[1]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, 02142

Draft genome assemblies for prokaryotic organisms are currently generated at low cost using whole genome shotgun sequencing. However, these assemblies are incomplete, typically having many gaps – potentially resulting in incorrect or incomplete biological analyses. It is possible to close all these gaps through 'finishing', an iterative manual process of data analysis and directed sequencing. Such manual finishing remains the gold standard, but the process is slow and, more importantly, very expensive, typically costing 25-100 fold more than the shotgun sequencing data. We have developed an effective formula for generation of finished quality prokaryotic genomes at low cost, and without manual analysis or additional labwork. This formula employs a mixture of data from two complementary technologies: 'short' 100 base reads generated on the Illumina platform with an error rate of ~1%, and 'long' ~1000 base reads generated on the Pacific Biosciences platform with an error rate of ~15%. The long reads are used both to patch amplification-induced gaps in the short read coverage, and to resolve repeats in the range 100-1000 bases. We applied this method to sixteen bacterial genomes, yielding essentially perfect assemblies. For example, the two chromosomes of the GC-rich bacterium *Rhodobacter sphaeroides* are present as circular contigs in the assembly, as are three plasmids. Two additional plasmids are present as a single composite circle, joined along essentially perfect 15 kb repeats that cannot be resolved using the data. These assemblies have less than one error per million bases. We also assess the accuracy of the previously finished references for three of our assemblies, showing that in fact our assemblies are of higher quality. This essentially automated method for sequencing and assembly produces prokaryotic genomes that are finished, at a fraction of the cost of traditional approaches.

FF0211

Mercury: A Next Generation Sequencing Data Analysis and Annotation Pipeline

David P. Sexton, Eric Boerwinkle, Peter Pham, Matthew Bainbridge, Danny Challis, Fuli Yu, Jeffery Reid, Richard Gibbs

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

The use of whole exome capture and whole genome sequencing employing next generation sequencing technology has created immense quantities of data, which must be analyzed and annotated for biological significance. Increasingly, these techniques are being used in the clinical setting in order to diagnose and manage disease cases with a significant genetic component. The resulting terabyte-scale data generated from next generation sequencing precludes a labor-intensive approach to data analysis and annotation. For these reasons, we have developed Mercury, a semi-automated pipeline for clinical next-generation sequence data analysis and annotation.

The Mercury pipeline is designed to function in three distinct settings: clinical diagnostics, cancer analysis, and research projects focusing on a specific disease. Achieving this range of functions requires a flexible platform with the ability to use software tools, which are common across all three areas and tools specific to each category. This assortment of software tools are controlled by a master script, which places these tools into a logical analysis pipeline and controls their execution. There are five modules within the script, which are responsible for each phase of analysis and annotation. Base calling and read qualities are handled by Illumina's CASAVA software package, which creates fastq files. The mapping phase is handled by BWA, which creates a BAM file as a final product. Variant calling uses two software packages, ATLAS2SNP and ATLAS2Indel. These packages create a VCF file containing all variant calls. Finally, annotation occurs through the internally developed Cassandra annotation suite. A filtered and ranked VCF is the final product of the pipeline. Using this framework it is easy to incorporate new analysis and annotation tools as required.

FF0004

NCGR Informatics

Andrew D. Farmer, Thiruvarangan Ramaraj, Robin S. Kramer, Connor T. Cameron, Ken A. Seal, Alex J. Rice, Charles F. Black, Kathy Myers, John A. Crow

National Center for Genome Resources

Santa Fe, New Mexico USA

NCGR Informatics supports the data analysis and interpretation needs of NCGR scientists, develops research support information systems and data processing pipelines, and provides bioinformatics partnerships with industry. The group comprises teams specializing in bioinformatics, software development, and the administration of high-performance computing resources.

The Bioinformatics team has developed strengths in experimental design, methodology development, and the analysis of NGS genomic and RNA-Seq data, including: assembly and annotation of prokaryotic and eukaryotic genomes, SNP and small indel prediction, expression quantitation and differential analysis, and the assembly and annotation of transcriptome references. Applications include the development of transcriptome assemblies for 750 marine microbial eukaryotes under funding from the Gordon and Betty Moore Foundation.

Our Software Development team is particularly strong in the design and implementation of research support systems, for example: online databases, laboratory information management systems, and high volume processing of data from NCGR's Sequencing Center. The team has been responsible for developing the Legume Information System (comparative-legumes.org) in cooperation with partners at USDA-ARS (Ames). It is improving the internal pipelines used for processing and quality assessment of Illumina sequencing runs.

NCGR's continued scientific progress is made possible by our Information Technology team, which works closely with our scientists, analysts, and developers. The team manages the substantial storage resources required for working with NGS technologies, the computational resources (large memory compute servers, compute clusters, database servers, web servers) for data processing, and the external and internal networks for large scale, high speed data transfer.

FF0174

DTRA Algorithm Prize

Christian Whitchurch

Defense Threat Reduction Agency

As *n*th generation DNA sequencing technology moves out of the research lab and closer to the diagnostician's desktop, the process bottleneck will quickly become information processing. The Defense Threat Reduction Agency (DTRA) and the Department of Defense are interested in averting this logjam by fostering the development of new diagnostic algorithms capable of processing sequence data rapidly in a realistic, moderate-to-low resource setting. With this goal in mind, DTRA is sponsoring an algorithm development prize.

The Challenge:

Given raw sequence read data from a complex diagnostic sample, what algorithm can most rapidly and accurately characterize the sample, with the least computational overhead?

The Stakes:

$1,000,000

Prize details and sequencing datasets will be made available this Fall. Monitor http://www.dtra.mil/Business.aspx for updates on this program.

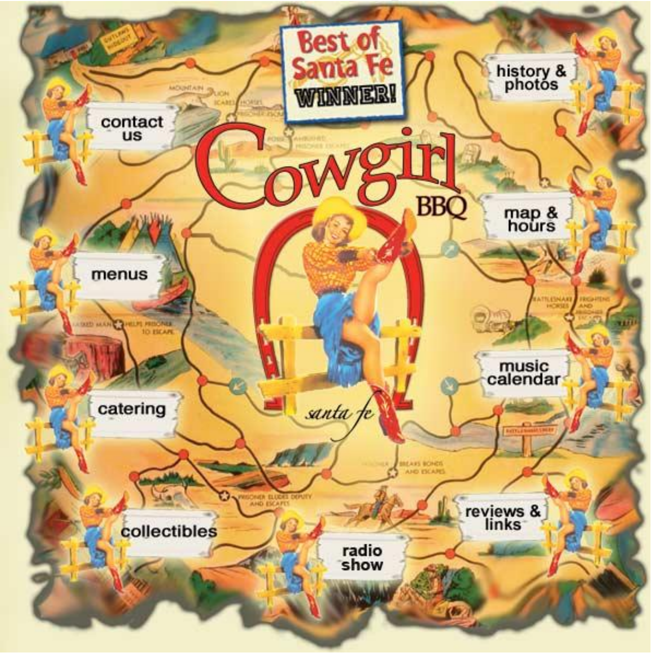
Round Table Discussion Notes

Round Table Discussion Notes

Happy Hour(s)

Cowgirl BBQ

505.982.2565   319 S. Guadalupe St   Santa Fe, NM



See map on next page!

5:45pm – 8:00pm, June 6th

Drink tickets (margaritas, beer, sodas) will be provided

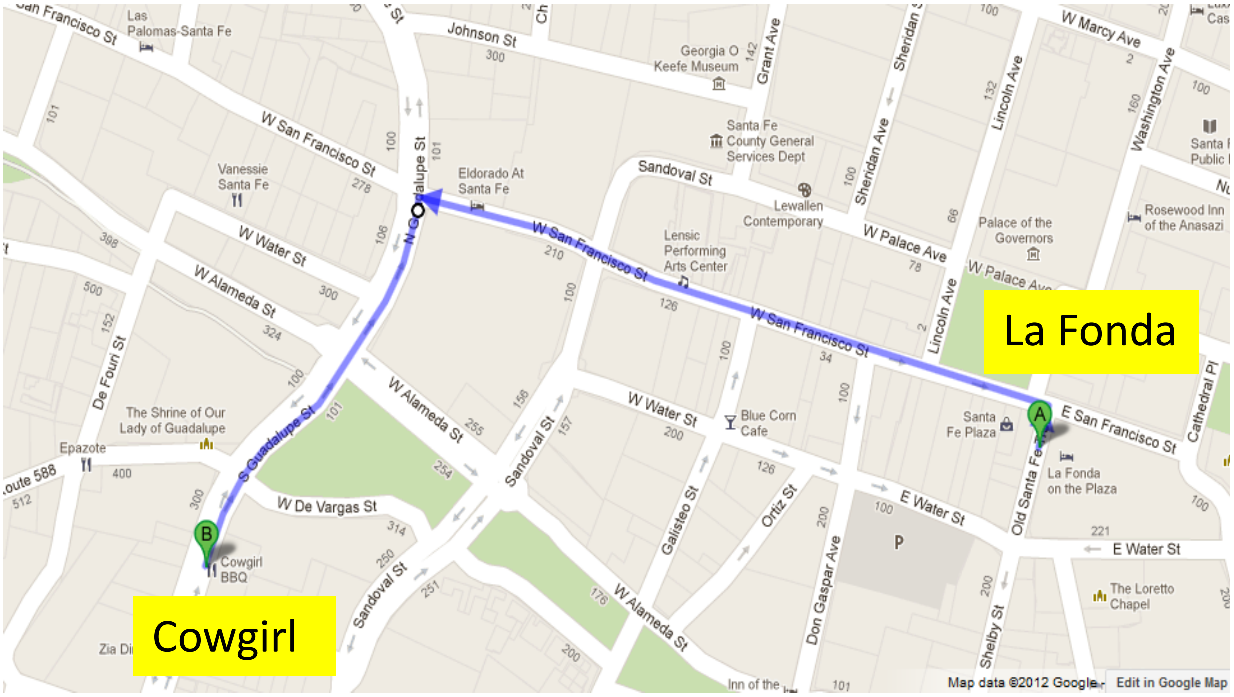
Sponsored by Life Technologies

Enjoy!!!

life technologies™

Map to Cowgirls BBQ

505.982.2565   319 S. Guadalupe St   Santa Fe, NM



Total Walking Distance

0.5 miles, 10 minutes

The Legend...

Many years ago, when the cattle roamed free and Cowpokes and Cowgirls rode the range, a sassy young Cowgirl figured out that she could have as much fun smokin’ meats and baking fine confections as she could bustin’ broncs and rounding up outlaws. So she pulled into the fine bustling city of Santa Fe and noticed that nobody in town was making Barbeque the way she learned out on the range. She built herself a Texas-style barbecue pit and soon enough the sweet and pungent scent of mesquite smoke was wafting down Guadalupe street and within no time at all folks from far and near were lining up for heaping portions of tender mesquite-smoked brisket, ribs and chicken. Never one to sit on her laurels, our intrepid Cowgirl figured out that all those folks chowing down on her now-famous BBQ need something to wash it all down with. Remembering a long-forgotten recipe from the fabled beaches of Mexico, she began making the now-legendary Frozen Margarita and the rest, as we say, is History. Before you could say “Tequila!” the musicians were out playing on the Cowgirl Patio and the party was in full swing.

06/07/2011 - Thursday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**Breakfast**	**x**	**Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	**x**	Welcome Back	**Chris Detter**
x	Session Chair	**x**	Session Chairs	Chair - Patrick Chain Chair - Nadia Fedorova
8:45 - 9:30	**Keynote**	**FF0042**	**Environmental Reservoirs of Human Pathogens: The Vibrio cholerae Paradigm**	**Dr. Rita Colwell**
9:30 – 9:50	Speaker 1	**FF0185**	A Rapid Whole Genome Sequencing and Analysis System Supporting Genomic Epidemiology	**Mike FitzGerald**
9:50 – 10:10	Speaker 2	**FF0173**	Endosymbiont Hunting in the Metagenome of Asian Citrus Psyllid (*Diaphorina citri*)	**Surya Saha**
10:10 – 10:30	Speaker 3	**FF0221**	SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing	**Glenn Tesler**
10:30 – 10:50	**Break**	**x**	**Beverages and Snacks Provided**	**x**
10:50 – 11:10	Speaker 4	**FF0263**	Assembly of Large Metagenome Data Sets Using a Convey HC-1 Hybrid-Core Computer	**Alex Copeland**
11:10 – 11:30	Speaker 5	**FF0034**	Metagenomic Assembly: Challenges, Successes, and Validation	**Matt Scholz**
11:30 – 11:50	Speaker 6	**FF0208**	Metagenomics for Etiological Agent Discovery	**Matthew Ross**
11:50 – 12:10	Speaker 7	**FF0207**	Nearly Finished Genomes Produced Using Gel Microdroplet Culturing Reveals Substantial Intraspecies Diversity within the Human Microbiome	**Michael Fitzsimons**
12:10 - 1:30pm	**Lunch**	**x**	**La Fiesta Plaza Lunch**	**Sponsored by Agilent**
x	Session Chair	**x**	Session Chairs	Chair - Mike Fitzgerald Chair - Alla Lapidus
1:30 - 1:50	Speaker 8	**FF0006**	Rapid Phylogenetic and Functional Classification of Short Genomic Fragments with Signature Peptides	**Ben McMahon**
1:50 - 2:10	Speaker 9	**FF0229**	PanFunPro: Pan-Genome Analysis Based on the Functional Profiles	**Oksana Lukjancenko**
2:10 - 2:30	Speaker 10	**FF0114**	Preparation of Nucleic Acid Libraries for Personalized Sequencing Systems Using an Integrated Microfluidic Hub Technology	**Kamlesh Patel**
2:30 - 2:50	Speaker 11	**FF0142**	Capturing Native Long-Range Contiguity by *in situ* Library Construction and Optical Sequencing	**Jerrod Schwartz**
2:50 - 3:10	Speaker 12	**FF0256**	Fosmid Cre-LoxP Inverse PCR Paired-End (Fosmid CLIP-PE), A Novel Method for Generating Fosmid Pair-End Library	**Ze Peng**
3:10 - 3:30	Speaker 13	**FF0282**	Automated Sequencing Library Preparation and Suppression for Rapid Pathogen Characterization	**Todd Lane**
3:30 - 3:50	Speaker 14	**FF0109**	Evaluation of Multiplexed 16S rRNA Microbial Population Surveys Using llumina MiSeq Platform	**Julien Tremblay**
3:50 - 4:00pm	**Closing Discussions**	**x**	**Closing Discussions for General Meeting - Discuss Next Year's Meeting**	**Chair - Chris Detter**
		x	**Reminder for those interested there is a special Forensic's Session Friday from 8:00am - 12:30pm**	**x**

NOTES

Speaker Presentations (June 7th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0042

Environmental Reservoirs of Human Pathogens: The *Vibrio cholerae* Paradigm

Dr. Rita R. Colwell

University of Maryland, College Park, Maryland

Since the mid-1980's when utilization of satellite sensors to monitor land and oceans for purposes of understanding climate, weather, and vegetation distribution and seasonal variations became possible, refinement of the inter-relationships of the environment and infectious diseases was accomplished, both qualitatively and quantitatively. Seasonality of diseases like malaria and cholera has been documented by epidemiologists, but new research is providing knowledge of a very close interaction of the environment and infectious disease. With satellite sensors, these relationships can be quantified and comparatively analyzed. Recent studies of cholera provide models, both retrospective and prospective, for understanding and predicting disease epidemics, notably those that are vectorborne. Cholera outbreaks can be predicted by monitoring selected environmental parameters. Zooplankton carry cholera bacteria as a component of their natural flora. Genomics of the cholera bacterium provide a tracking tool and with remote sensing, an early warning system for public health and for measuring effects of climate change on public health.

NOTES

FF0185

A Rapid Whole Genome Sequencing and Analysis System Supporting Genomic Epidemiology

Michael G. FitzGerald, Grad Y, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, DeSmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Moller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Hanage WP, Hung DT, and Birren BW.

Broad Institute of MIT and Harvard, Cambridge, MA, USA

During May and June of 2011, two bloody diarrhea and hemolytic uremic syndrome (HUS) outbreaks occurred in Europe. 3816 cases were reported in Germany, including 845 with HUS leading to 50 deaths. The French outbreak reported 15 bloody diarrhea cases with 9 demonstrating HUS. Both outbreaks were tied to a Shiga-toxin producing *Escherichia coli* O104:H4 strain carrying enteroaggregative and beta-lactam antibiotic resistance plasmids. The bacterial isolates are indistinguishable using standard tests. We undertook a rapid, whole genome sequencing epidemiology approach using isolates from both outbreaks. Modification of our standard procedures resulted in sequence, assembly, annotation and analysis for 17 strains within 10 weeks. Data were generated using the Illumina, Pacific Biosciences and Roche-454 platforms. Analysis of the Illumina data allowed us to identify 21 SNPs and all were validated by targeted Sanger sequencing.

We will detail subsequent developments to our rapid sequencing pipeline, including the role of sequencing platforms like the Illumina MiSeq which can sequence a single bacterial genome per day and internal developments like automation of hybrid Illumina-PacBio assembly (10 genomes/day capacity) and SNP calling pipelines (25 genomes/day capacity). This work clearly demonstrates the power of whole genome sequence to delineate closely related organisms and the importance of high quality primary data.

This project has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900018C.

FF0173

Endosymbiont Hunting in the Metagenome of Asian Citrus psyllid (*Diaphorina citri)*

Surya Saha[1], Wayne B. Hunter[2] and Magdalen Lindeberg[1]

[1] Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, NY, USA
[2] USDA-ARS, US Horticultural Research Lab, Fort Peirce, FL, USA

The Asian citrus psyllid (D. citri Kuwayama or ACP) is host to 7+ bacterial endosymbionts and is the insect vector of Ca. liberibacter asiaticus (Las), causal agent of citrus greening. To gain a better understanding of endosymbiont and pathogen ecology and develop improved detection strategies for Las, DNA from D. citri was sequenced to 108X coverage. Initial analyses have focused on Wolbachia, an alpha-proteobacterial primary endosymbiont typically found in the reproductive tissues of ACP and other arthropods. The metagenomic sequences were mined for wACP reads using BLAST and 4 sequenced Wolbachia genomes as bait. Putative wACP reads were then assembled using Velvet and MIRA3 assemblers over a range of parameter settings. The resulting wACP contigs were annotated using the RAST pipeline and compared to Wolbachia endosymbiont of Culex quinquefasciatus (wPip). MIRA3 was able to reconstruct a majority of the wPip CDS regions and was selected for scaffolding with Minimus2, SSPACE and SOPRA using large insert mate-pair libraries. The wACP scaffolds were compared to wPip using Abacas and Mauve contig mover to orient and order the contigs. The functional annotation of scaffolds was evaluated by comparing it to wPip genome using RAST. The draft assembly was verified using an OrthoMCL based comparison to the 4 sequenced Wolbachia genomes. We expanded the scope of endosymbiont characterization beyond wACP using 16S rDNA and partial 23S rDNA analysis as a guide. Results will be presented regarding endosymbionts, their potential interactions and their impact on the disease of citrus greening.

FF0221

SPAdes: a New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing

Glenn Tesler

University of California, San Diego

The lion's share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies. A major goal of single-cell genomics is to complement gene-centric metagenomic data with whole-genome assemblies of uncultivated organisms.

Assembly of single-cell data is challenging because of highly non-uniform read coverage as well as elevated levels of sequencing errors and chimeric reads. We describe SPAdes, a new assembler for both single-cell and standard (multicell) assembly, and demonstrate that it improves on the recently released E+V-SC assembler (specialized for single-cell data) and on popular assemblers Velvet and SoapDeNovo (for multicell data). SPAdes generates single-cell assemblies, providing information about genomes of uncultivatable bacteria that vastly exceeds what may be obtained via traditional metagenomics studies.

This is a joint work with Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey Gurevich, Mikhail Dvorkin, Alexander Kulikov, Valery Lesin, Sergey Nikolenko, Son Pham, Andrey Prjibelski, Alexey Pyshkin, Alexander Sirotkin, Nikolay Vyahhi, Max Alekseyev, and Pavel Pevzner.

Software: http://bioinf.spbau.ru/spades

Article: Bankevich et al (2012), Journal of Computational Biology, 19(5):
455-477, doi:10.1089/cmb.2012.0021.

FF0263

Assembly of Large Metagenome Data Sets Using a Convey HC-1 Hybrid-core Computer

A. Copeland, K. Labutti, B. Foster, S. Tringe, J. Jansson

DOE Joint Genome Institute, Walnut Creek, California 94598, United States of America

Assembly of metagenome data provides substantial benefit to downstream analysis by reducing the size and improving the accuracy of the input data, and providing contigs which are longer substrates for annotation and other analysis. Metagenomes appear to be an excellent fit for the Illumina HiSeq 2000 platform. It is only with the very large base output of this platform that one can hope to produce sufficient depth to assemble communities containing thousands of unique genomes. As early as 2009, it was apparent that analysis at these depths would be a major challenge (Nature Methods 6, 623 (2009)). As predicted, once a sufficiently large metagenome sequence data set has been produced, it is very difficult, or even impossible, to assemble due to the very large memory requirements. In this poster, we describe using Convey's graph constructor on an HC-1 server to reduce the memory and run time requirements of Velvet's (Genome Res. 2008. 18: 821-829) assembly workflow without prior data reduction. We will present assembly results from several large data sets (>300Gb) including the Great Prairie Grand Challenge project.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

FF0034

Metagenomic Assembly: Challenges, Successes, and Validation

Matthew Scholz

Los Alamos National Laboratory

Metagenome assembly is a constantly evolving process, with data input, data types, amounts of data, available software, and available hardware changing (optimistically) on a monthly basis. As algorithms for DeBruijn graph based assemblers improve to handle more data, and assembly strategies improve to allow binning of data into smaller pieces, it has not yet become clear that any method is a widely applicable, best approach for assembly of any one type of metagenome. This has led to development of a pipeline at LANL for post-assembly merging of multiple assemblies of the same data to produce improved assemblies over any of the input assemblies. This process uses out of the box assembly tools to perform a series of filtering and similarity based assembly steps. This process has been in continuous use at LANL for more than a year, and has resulted in improved assembly of hundreds of metagenome assemblies. There are also a number of steps that must be taken in this process to validate the produced contigs, including read-mapping, and inra-assembly comparisons.

FF0208

Metagenomics for Etiological Agent Discovery

Matthew Ross[1]; Khanh Thi-Thuy Nguyen[2]; David Wheeler[3]; Joseph Petrosino[1]

[1]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX; [2]The University of Texas MD Anderson Cancer Center, Houston, TX; [3]The Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

In the past, identification of etiologic agents, both viral and bacterial, typically relied on propagation of the suspected agent in the laboratory. As technology progressed, researchers were able to characterize some newly discovered viruses and bacteria by purely molecular methods, but these still required some knowledge of the agent being sought. There are still numerous diseases whose epidemiology suggests an infectious cause, but where an etiological agent has not been identified through traditional means. We are utilizing both bacterial and viral metagenomics in an attempt to identify agents associated with diseases where an infectious agent is suspected. These include diseases such as Kawasaki Disease, where epidemiological evidence suggests a viral agent, and unknown agents of sepsis in immunocompromised subjects (e.g. those undergoing chemotherapy).

A metagenomics approach has multiple theoretical advantages over past methods. There is no requirement for propagation of the suspected agent in the lab, nor is there any need for prior knowledge of the agent at all. We¹ve also found it to be very sensitive, able to detect sub-PFU levels of viruses in spiked clinical samples. For suspected bacterial agents we are using a 16S rRNA gene survey of samples and for suspected viral agents we are generating randomly primed cDNA libraries followed by next generation sequencing. Sequences from cDNA libraries are assembled and queried against a custom viral database. Using these methods, we are demonstrating the ability to detect bacterial and viral agents where traditional diagnostic measures have failed.

FF0207

Nearly Finished Genomes Produced using Gel Microdroplet Culturing Reveals Substantial Intraspecies Diversity within the Human Microbiome

Fitzsimons MS[1], Novotny M[2], Lo CC[1], Dichosa AEK[1], Yee-Greenbaum JL[2], Snook JP[1], Gu W[1], Chertkov O[1], Davenport KW[1], McMurry K[1], Gleasner CD[1], Wills PL[1], Parson-Quintana B[1], Chain PS[1], Detter JC[1], Lasken RS[2] and Han CS[1].

[1]Genome Science, Los Alamos National Laboratory, Los Alamos, NM
[2]J. Craig Venter Institute, San Diego, CA

The majority of microbial genomic diversity remains unexplored. This is in large part due to our inability to culture most microorganisms in isolation, which is a prerequisite for traditional sequencing. Single cell sequencing has allowed researchers to circumvent this limitation by amplifying DNA directly from a single cell using the whole genome amplification technique of multiple displacement amplification (MDA). However, MDA of a single DNA molecule suffers from amplification bias and contamination, which makes assembling a genome difficult and completely finishing a genome impossible except in extraordinary circumstances. Gel microdrop cultivation allows one to culture a diverse microbial community and provide hundreds to thousands of genetically identical cells as input into an MDA reaction. We demonstrate the utility of this approach by comparing sequencing results of gel microdroplets and single cells following MDA. The central result of this study is that bias is reduced in the MDA reaction and genome sequencing and assembly is greatly improved when utilizing gel microdroplets. We also show the range of microorganisms for which this technique can be utilized and its use for studying intraspecies genomic diversity (i.e. the pan-genome). Gel microdroplets offer a powerful and high-throughput technology for assembling whole genomes from complex samples and for probing the pan-genome of naturally occurring populations.

NOTES

Lunch

12:10 – 1:30pm

Sponsored by

Agilent Technologies

Notes

FF0006

Rapid Phylogenetic and Functional Classification of Short Genomic Fragments with Signature Peptides

Ben McMahon, Joel Berendzen, Judith Cohn, Nick Hengartner

Los Alamos National Laboratory

Classification is difficult for shotgun metagenomics data from environments such as soils, where the diversity of sequences is high and where reference sequences from close relatives may not exist. Approaches based on sequence-similarity scores must deal with the confounding effects that inheritance and functional pressures exert on the relation between scores and phylogenetic distance, while approaches based on sequence alignment and tree-building are typically limited to a small fraction of gene families. We describe an approach based on finding one or more exact matches between a read and a precomputed set of peptide 10-mers.

At even the largest phylogenetic distances, thousands of 10-mer peptide exact matches can be found between pairs of bacterial genomes. Genes that share one or more peptide 10-mers typically have high reciprocal BLAST scores. Among a set of 403 representative bacterial genomes, some 20 million 10-mer peptides were found to be shared. We assign each of these peptides as a signature of a particular node in a phylogenetic reference tree based on the RNA polymerase genes. We classify the phylogeny of a genomic fragment (e.g., read) at the most specific node on the reference tree that is consistent with the phylogeny of observed signature peptides it contains. Using both synthetic data from four newly-sequenced soil-bacterium genomes and ten real soil metagenomics data sets, we demonstrate a sensitivity and specificity comparable to that of MEGAN metagenomics analysis package using BLASTX against the NR database. Phylogenetic and functional similarity metrics applied to the real metagenomics data indicates a signal-to-noise ratio of approximately 400 for distinguishing among environments. Our method assigns ~6.6 Gbp/hr on a single CPU, compared with 25 kbp/hr for methods based on BLASTX against the NR database.

Classification by exact matching against a precomputed list of signature peptides provides comparable results to existing techniques for reads longer than about 300 bp and does not degrade severely with shorter reads. Orders of magnitude faster than existing methods, the approach is suitable now for inclusion in analysis pipelines and appears to be extensible in several different directions

FF0229

PanFunPro: Pan-Genome Analysis Based on the Functional Profiles

Oksana Lukjancenko, David W. Ussery

Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitrover Building 208, Kgs. Lyngby, Denmark

The sequencing of complete bacterial strains has provided us with large volumes of data (over 1600 genomes) and using these data to analyse proteomes, find minimal microbial genome and identify antimicrobial targets are one of the science's current targets. A variety of *in silico* techniques, using both full-sequence and conserved domain comparison, were created to analyse genomic data.

A new computational approach (PanFunPro), based on proteome-grouping according to functional domain architecture, was developed to identify bacterial pan-genome, set of conserved functions; and genes, specific to a particular taxonomy level. We analyzed over 1600 microbial genomes and results are showing that the universal core is very small (<30 genes), containing only genes that encode the most basic functions for bacteria to survive; and a set of core genes for groups of genomes living in different conditions differ depending on the environment. Analysis of pan-genomes for different taxonomy-levels identified levelspecific genes, which can be useful in taxonomy prediction of unknown genomes.

Overall, our results demonstrate that this approach will provide some improvements to coregenome and pan-genome calculations, and bring a different view to taxonomy prediction.

FF0114

Preparation of Nucleic Acid Libraries for Personalized Sequencing Systems Using an Integrated Microfluidic Hub Technology

Kamlesh D. Patel, Hanyoup Kim, Numrin Thaitrong, Victoria VanderNoot, Michael Bartsch, Robert Meagher, Ronald Renzi, Steve Branda, Stanley Langevin, Zachary Bent, Joe Schoeniger, and Todd Lane

Sandia National Laboratories, Livermore CA USA

While DNA sequencing technology is advancing at an unprecedented rate, sample preparation technology often relies on manual bench-top processes, which are slow and labor-intensive. Automation of sample preparation using microfluidic techniques is well-suited to address these limitations. However, fabricating a single monolithic microfluidic device that replicates all the relevant benchtop processes can be prohibitively complicated and is not flexible to execute the protocols for processing different samples. We have designed and characterized a digital microfluidic (DMF) platform to function as a central hub for interfacing multiple lab-on-a-chip sample processing modules towards automating the preparation of clinically-derived DNA samples for Next Generation Sequencing (NGS). The automated molecular biology platform (AMB) is designed to interface directly with personalized sequencing systems to detect unknown pathogens by enriching informative nucleic acids sequences (those derived from the pathogen) and suppressing background DNA (those from the host) to maximize the sensitivity of these systems.

I will present our recent developments on the core architecture of the AMB platform, the DMF central hub, and demonstrate its flexibility in coupling droplet-based microfluidics with continuous-flow microchannel devices to prepare DNA samples for NGS. I will focus my presentation on our results for collecting fractions of nanogram amounts of host-suppressed DNA in discrete 1-□L droplets on the DMF device for processing and integrating the key sample preparations functions such as DNA suppression, fragmentation/ligation, bead-based clean-up, and PCR. Additionally, I will show our recent accomplishments in incorporating quantitative analysis of barcoded DNA libraries as part our integrated sample preparation workflow, where the resulting prepared DNA library can be directly transferred to a MiSEQ Illumina sequencer flowcell for cluster generation and fast sequencing to discover the pathogen by its genomic sequence. Beyond NGS, I will also briefly highlight two related applications for our hub platform for rapidly preparing DNA for battlefield DNA forensics and remote biosurveillance.

FF0142

Capturing native long-range contiguity by in situ library construction and optical sequencing

Jerrod J. Schwartz, Choli Lee, Joseph B. Hiatt, Andrew Adey, Jay Shendure

Department of Genome Sciences, University of Washington, Seattle, WA 98195

The relatively short read lengths associated with the most cost-effective DNA sequencing technologies have limited the quality and completeness of both de novo genome assembly and human genome sequencing. Consequently, there is a strong need for methods that capture various scales of contiguity information at a throughput commensurate with the current scale of massively parallel sequencing. We propose *in situ* library construction and optical sequencing on the flow cells of currently available next-generation sequencing platforms as an efficient means of capturing both contiguity information and primary sequence with a single technology. In this proof-of-concept study, we demonstrate basic feasibility by generating >30,000 E. coli paired-end reads separated by 1, 2, or 3 kb using in situ library construction on standard Illumina flow cells. We also show that it is possible to stretch single molecules ranging from 3 – 8 kb on the surface of a flow cell prior to in situ library construction, thereby enabling the production of clusters whose physical relationship to one another on the flow cell is related to genomic distance.

FF0256

Fosmid Cre-LoxP Inverse PCR Paired-End (Fosmid CLIP-PE), A Novel Method for Generating Fosmid Pair-End Library.

Ze Peng, Feng Chen, Jeff L. Froula, Zhiying Zhao, James Han, Alicia Clum, Alex C. Copeland and Jan-Fang Cheng

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

Fosmid end sequencing has been widely utilized in genome sequence assemblies, genome structural variation studies etc. The Illumina HiSeq sequencer has dramatically reduced the cost of sequencing a paired read to about 0.003¢, it is advantageous to have the ability to perform fosmid paired-end sequencing in this platform. We have developed a new approach to construct fosmid paired-end library that is suitable for Illumina sequencing. This approach employs a newly modified fosmid vector that contains two loxP sites and two Illumina adaptor priming sites flanking the cloning site. Fosmid DNA prepared from the library can be treated with the Cre recombinase to remove most of the vector DNA, leaving only 107bp of the vector sequence with insert DNA. Frequent cutting restriction enzymes and ligase are used to digest the fosmid DNA to small (less than a Kb) fragments and re-circularize the fosmid ends and all the internal fragments. Finally a PCR step with the Illumina primers is used to enrich the fosmid paired-ends for sequencing. The advantages of this approach are that (1) the circularization of short fragments is very efficient, therefore the success rate is higher than other approaches that attempt to join both ends of large fosmid vectors; and (2) the restriction enzyme cutting generates an identifiable junction tag for splitting the paired reads. Our results have shown that this approach has produced mostly fosmid size (30-40Kb) pairs from the targeted fungi and plant genomes and has drastically increased the scaffold sizes in the assembled genomes.

FF0282

Automated Sequencing Library Preparation and Suppression for Rapid Pathogen Characterization

Todd Lane

Sandia National Laboratories

Bioweapons and emerging infectious diseases pose serious and growing threats to our national security. Effective response to an outbreak critically depends upon rapid and accurate identification and characterization of the causative pathogen. Probe-based methods are problematic due to need for *a priori* knowledge of pathogen properties such as nucleic acid (NA) sequences. In recent years, unbiased Second-Generation Sequencing (SGS) of NA extracted from clinical samples has enabled discovery of novel pathogens. This brute-forceapproach can be powerful, but it is inefficient and frequently ineffectual, primarily because the signal-to-background (pathogen-to-host NA) ratio in clinical samples is often vanishingly small. We are developing a new automated molecular biology technology that selectively suppresses host background in NA extracts from clinical samples, and prepares the residual NA (enriched for pathogen-derived content) for NGS analysis. This microfluidics-based technology, coupled with a new bioinformatics pipeline for efficient analysis of NGS datasets, comprises a Rapid Threat Organism Recognition (RapTOR) system for focused sequencing of pathogen genome/transcriptome constituents in the context of complex host backgrounds. Application of this system will greatly accelerate identification and characterization of novel pathogens, and thereby support rational and effective response to infectious disease outbreaks.

FF0109

Evaluation of Multiplexed 16S rRNA Microbial Population Surveys Using llumina MiSeq Platform

Julien Tremblay, Edward S Kirton, Kanwar Singh, Feng Chen and Susannah G Tringe

DOE Joint Genome Institute, Walnut Creek, CA, 94598, USA

In recent years, microbial community surveys extensively relied on 454 pyrosequencing technology (pyrotags). Recently, the Illumina sequencing platform HiSeq2000 has largely surpassed 454 in terms of read quantity and quality with typical yields of up to 600 Gb of paired-end 150 bases reads in one 18 day run. Yet many labs still rely on pyrotags for community profiling because the HiSeq throughput exceeds their needs, the run time is long, and accumulating sufficient samples to effectively utilize a full run introduces significant delay. Illumina recently introduced the new mid-range MiSeq sequencing platform which gives an output of 1 Gb of paired-end 150 base reads in a single day run. With its moderately-high throughput and support for massive multiplexing (barcoding), this platform represents a promising alternative to 454 technology to perform 16S rRNA-based microbial population surveys.

A workflow was therefore developed to confirm that Illumina MiSeq is a suitable platform to accurately characterize microbial communities. We surveyed microbial populations coming from various environments by targeting the 16S rRNA hypervariable region V4 which generated amplicons size of about 290 bases. These amplicons were sequenced with the MiSeq platform from both 5' and 3' ends followed by *in silico* assembling using their shared overlapping part. Downstream analyses through our Itags pipeline are also described, including a novel clustering strategy generating fast and accurate distribution of bacterial operational taxonomic units (OTUs).

Our results suggest that the MiSeq sequencing platform successfully recaptures known biological results and should provide a useful tool for 16S rRNA characterization of microbial communities.

Discussion Notes

06/08/2011 - Friday			**Forensic Friday**	
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**x**	**x**	**Breakfast on your own**	**x**
8:30 - 8:35	Intro	**x**	Welcome Intro - Session Chair (LANL)	**Cathy Clealand**
8:35 - 8:45	Intro	x	Welcome Intro - Session Chair (US Army)	**Ken Kroupa** **Jeff Salyards**
8:45 – 9:05	Speaker 1	**FF0047**	Next Generation Sequencing; Possible Application for Forensic DNA Analysis. What does the Person of Interest Look Like?	**Tom Callaghan**
9:05 – 9:25	Speaker 2	**FF0136**	Forensic DNA Standards for Next Generation Sequencing Platforms	**Pete Vallone**
9:25 – 9:45	Speaker 3	**FF0216**	Short Tandem Repeat (STR) Analysis from Short Read Sequencing Data	**Daniel Bornman**
9:45 – 10:05	Speaker 4	**FF0191**	High Sensitivity Detection and Typing of Mixed Contributor DNA Samples Using Massively-Parallel Deep Amplicon Pyrosequencing	**Jared Latiolais**
10:05 – 10:20	**Break**	**x**	**Break**	**x**
10:20 – 10:40	Speaker 5	**FF0114**	Preparation of Nucleic Acid Libraries for Personalized Sequencing Systems Using an Integrated Microfluidic Hub Technology	**Ken Patel**
10:40 – 11:00	Speaker 6	**FF0223**	Forensic Genomics using Next Generation Sequencing by Synthesis (SBS)	**Cydne Holt**
11:00 – 11:20	Speaker 7	**FF0153**	A Highly Configurable SNP Caller for the Ion Torrent Personal Genome Machine	**Christian Buhay**
11:20 - 11:40	Speaker 9	**FF0248**	Short Tandem Repeat Sequencing on the 454 Platform	**Melissa Scheible**
11:40 - 12:00	Speaker 10	**FF0280**	STR Profiling From Personal Genomes: Happy Surprises	**Yaniv Erlich**
12:00 - 12:20	**Closing Panel Discussions**	**x**	**Panel Discussion on Forensic Applications of Next Generation Sequencing (US Army and LANL)**	**Jeff Salyards** **Chris Detter**
12:20 - 12:30	x	**x**	**Thank you**	**Cathy Clealand**

NOTES

Special Friday Forensic Session

Speaker Presentations (June 8th)

Abstracts are in order of presentation according to Agenda

Friday Session

FF0047

Next Generation Sequencing; Possible Application for Forensic DNA Analysis. What does the Person of Interest Look Like?

Thomas Callaghan and James Robertson

FBI Laboratory, Quantico, Virginia, USA

The forensic DNA community is interested in applying Next Generation Sequencing (NGS) technology to samples recovered from unsolved violent crimes. An overview of the FBI's Combined DNA Index System (CODIS) will be presented along with possible applications of NGS technology to produce investigative leads for criminal investigations.

Forensic DNA techniques have been used for over 25 years to aid criminal investigations. Initially, Restriction Fragment Length Polymorphism (RFLP) techniques were used to include or exclude and person of interest. Next, PCR technology provided increased sensitivity and automation to forensic DNA analysis, allowing the efficient use of DNA databasing. Today, NGS seems poised to aide investigations when the perpetrator is not included in the database. NGS promises to help answer the question "What does the person of interest look like?"

In order to describe the person of interest and serve as the next generation technique for forensic DNA analysis, NGS will need to address mixture resolution, federal quality assurance standards and the possibility of sub-microgram levels of DNA for analysis. This presentation is intended to introduce the sequencing community to the operation and requirements of forensic DNA analysis.

FF0136

Forensic DNA Standards for Next Generation Sequencing Platforms

Peter M. Vallone, Carolyn R. Hill, Erica L.R. Butts, David L. Duewer, John M. Butler, and Margaret C. Kline

Applied Genetics Group, Biochemical Science Division, National Institute of Standards and Technology, 100 Bureau Drive Gaithersburg, MD 20899-8314

Over the past 22 years the Applied Genetics group at NIST has been providing DNA-based Standard Reference Materials (SRMs) for the human identity community. These forensic SRMs are required by the FBI DNA Advisory Board (Standard 9.5) to calibrate DNA typing procedures performed in forensic laboratories. The SRMs typically consist of genomic DNAs (≈100 ng in 50 µL) that have been highly characterized for forensically relevant markers: core autosomal and Y-chromosome short tandem repeat (STRs) in additional to mitochondrial genome sequence. To date the characterization of the forensic markers of interest is a combination of Sanger sequencing and fragment size analysis.

With the current interest in applying Next Generation Sequencing (NGS) technologies to forensic problems we are beginning to explore the next generation of forensic DNA-based SRMs. The considerations for candidate new materials include: source of genomic DNAs, amount of DNA required, genetic markers to be characterized, inter-laboratory testing, and specific needs of the forensic community. This talk will review the past SRMs and and identify requirements for future forensic reference materials.

FF0216

Short Tandem Repeat (STR) Analysis from Short Read Sequencing Data

Daniel Bornman, M.S., Seth Faith, Ph.D., Brian Young, Ph.D.

Battelle Memorial Institute

The cost and speed of nucleic acid sequencing is rapidly decreasing due to next-generation sequencing technologies, enabling novel opportunities for collecting numerous forensically relevant data from single data sets. In this study, we demonstrate interrogation of human saliva samples with massively parallel sequencing technologies (Illumina) for evaluating the allele status of the 13 core CODIS short tandem repeat (STR) loci. To enable the analysis of STR from next generation sequencing (NGS) data we modified the approach for standard reference genome alignment yet preserved the ability to leverage open-source short read alignment software. Following an enrichment PCR step to amplify a ~2 kb region surrounding each CODIS STR site, we performed multiplexed sequencing on an Illumina GAIIx of five individuals, plus an equal parts mixture on two of the five samples. This sequencing run generated read lengths of 150 nucleotides that were aligned to a modified reference genome specifically constructed to enable analysis of diploid STR polymorphic patterns. Reference alignment was performed using the Bowtie short read alignment program that generated standard sequencing file formats for downstream analysis and interpretation. Analysis of the final sequence alignment data enabled us to correctly estimate (with complete concordance to CE) STR allele status for the entire set of core CODIS loci from each sample including the single 1:1 mixture sample. Some exceptions included two instances where the repeat pattern for the D21S11 locus had repeat lengths greater than could be covered by the current maximum 150 bp read length. Data are also presented that demonstrates the use of a novel filtering method that preprocesses the raw sequencing data for enrichment of short reads containing STR patterns. This filtering method provides up to an 8-fold increase in speed for completion of the STR calling method from NGS data and enables the data to be processed on computing resources comparable to a desktop computer.

FF0191

High Sensitivity Detection and Typing of Mixed Contributor DNA Samples Using Massively-Parallel Deep Amplicon Pyrosequencing

Jared Latiolais[1], Andrew Feldman[2], Jeffrey Lin[2], Plamen Demirev[2], Ishwar Sivakumar[2], Thomas Mehoke[2], and Robert Bever[1]

[1]The Bode Technology Group, Lorton, VA
[2]The Johns Hopkins Applied Physics Laboratory, Laurel, MD

We demonstrate that Roche 454 Titanium chemistry based deep sequencing of individual PCR amplicons yields greater levels of sensitivity for detecting minor contributor's in a mixed forensic sample than conventional sequencing methods.

We present data from a series of mixture de-convolution experiments using both mitochondrial DNA and nuclear STR amplicons. The data sets show highly sensitive detection of low copy contributor STRs well below the 1-10% level, which is the approximate limit for gel electrophoresis. Dilutions of human DNA in mixtures from 1-5 different contributors at different relative concentrations were prepared for PCR amplification and subsequent searching by the 454 system. In addition, "touch" samples were also collected and sequenced from objects touched by multiple individuals. We fabricated fusion primers for 8 CODIS STR loci routinely sued in forensic casework. Following PCR amplification, emulsion PCR is performed and clonally amplified in a massively-parallel fashion. Flow-based pyrosequencing is then performed to determine ~500 bp of the sequence of DNA coupled to each bead. Each bead sequence correlates to one single molecule of DNA which represents one individual contributor. The ratios of the major to minor contributor sequence reads at each locus enables linkage of detected alleles across different loci to form the respective genotypes. A typical 454 sequencing run yields up to 100 thousand reads, thus offering minor contributors detection thresholds at less than one part per thousand.

We also observed that while pyrosequencing is not error-free, the impact of these erros on tandem repeat analysis is minimized through modest bioinformatics analysis of sequences against all known alleles. The system also offers the potential to both detect and sequence new alleles in the course of routine forensic casework and thus can be an important feature tool in the boarded forensics community.

FF0114

Preparation of Nucleic Acid Libraries for Personalized Sequencing Systems Using an Integrated Microfluidic Hub Technology

Kamlesh D. Patel, Hanyoup Kim, Numrin Thaitrong, Victoria VanderNoot, Michael Bartsch, Robert Meagher, Ronald Renzi, Steve Branda, Stanley Langevin, Zachary Bent, Joe Schoeniger, and Todd Lane

Sandia National Laboratories, Livermore CA USA

While DNA sequencing technology is advancing at an unprecedented rate, sample preparation technology often relies on manual bench-top processes, which are slow and labor-intensive. Automation of sample preparation using microfluidic techniques is well-suited to address these limitations. However, fabricating a single monolithic microfluidic device that replicates all the relevant benchtop processes can be prohibitively complicated and is not flexible to execute the protocols for processing different samples. We have designed and characterized a digital microfluidic (DMF) platform to function as a central hub for interfacing multiple lab-on-a-chip sample processing modules towards automating the preparation of clinically-derived DNA samples for Next Generation Sequencing (NGS). The automated molecular biology platform (AMB) is designed to interface directly with personalized sequencing systems to detect unknown pathogens by enriching informative nucleic acids sequences (those derived from the pathogen) and suppressing background DNA (those from the host) to maximize the sensitivity of these systems.

I will present our recent developments on the core architecture of the AMB platform, the DMF central hub, and demonstrate its flexibility in coupling droplet-based microfluidics with continuous-flow microchannel devices to prepare DNA samples for NGS. I will focus my presentation on our results for collecting fractions of nanogram amounts of host-suppressed DNA in discrete 1-□L droplets on the DMF device for processing and integrating the key sample preparations functions such as DNA suppression, fragmentation/ligation, bead-based clean-up, and PCR. Additionally, I will show our recent accomplishments in incorporating quantitative analysis of barcoded DNA libraries as part our integrated sample preparation workflow, where the resulting prepared DNA library can be directly transferred to a MiSEQ Illumina sequencer flowcell for cluster generation and fast sequencing to discover the pathogen by its genomic sequence. Beyond NGS, I will also briefly highlight two related applications for our hub platform for rapidly preparing DNA for battlefield DNA forensics and remote biosurveillance.

FF0223

Forensic Genomics using Next Generation Sequencing by Synthesis (SBS)

Cydne L. Holt and Adam Lowe

Illumina, San Diego, CA, USA

With the advent of next-generation sequencing (NGS), the spectrum of known human genomic variation has expanded at an unprecedented rate. NGS is resetting the amount and type of information available to investigative genetics. To-date, the vast majority of sequence data generated globally has been done utilizing Illumina SBS technology. In application to forensic biology, SBS has the potential to deliver a “universal" forensic DNA panel that addresses multiple disciplines simultaneously, including criminal casework, databank, parentage testing (mass disaster, missing persons), ancestry, phenotyping, death investigation and metagenomics. Practical implementation of SBS in a forensic setting is enabled by the MiSeq system, which simplifies and automates the NGS process.

Data have demonstrated the potential of NGS as a multipurpose genotyping platform. Studies of saliva samples have shown that autosomal STR genotypes, plus their internal SNPs, Y and mitochondrial haplotypes, predictive ancestry and visible traits, and metagenomic data (as investigative leads) can be done in a single sequencing run. This approach provides backward compatibility with the 13 core CODIS STR loci. A denser set of forensically relevant markers is under development.

Increased discrimination power from dense, high value forensic sequencing data allows interpretation of partially degraded and/or mixed DNA. Because NGS performs a molecule-by-molecule analysis of the original sample’s content, it is possible to view the number of observations of a given allele, and measure mixture ratios based on a count (digital) vs. a peak height (analog) result. This is expected to dramatically extend capabilities in the analysis of complex samples.

The application of these technologies to forensic genomics will be presented along with data from Illumina internal labs, and collaborations.

FF0153

A Highly Configurable SNP Caller for the Ion Torrent Personal Genome Machine

Christian Buhay[1], Qiaoyan Wang[1], Huyen Dinh[1], Imad Khalil[1], Michael Holder[1], Yuan-Qing Wu[1], Mark Wang[1], Lora Lewis[1], Christie Kovar[1], David Wheeler[1], Donna Muzny[1], Eric Boerwinkle[1,2] and Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030
[2]University of Texas Health Science Center at Houston, School of Public Health, Houston, TX 77030

The BCM-HGSC has been evaluating Life Technologies' Ion Torrent Personal Genome Machine (PGM) sequencing platform throughout 2011. Orthogonal validation is one of the principal applications for the PGM. We have developed a highly configurable SNP variant caller purpose-built for Ion Torrent sequence. Preliminary data suggests validation rates of 90% or better using purely automated means.

The PGM's performance as a viable validation tool was evaluated using five amplicon pools totaling roughly 3000 sites. Utilizing the Ion Torrent mapper (TMAP), more than 94% of all reads produced aligned to the targeted pool. Average coverage across the pools was 850X, with 98% of all target bases covered at 40X or better. These projects were originally sequenced on the SOLiD platform then orthogonally validated with Roche/454. The intersection of automated PGM validation results with 454 was initially 70% - 85%. Many of the initial variant calls were incorrect due to insertions or deletions at or around the validation site. We developed a modified variant caller that builds on Samtools pileup, parses and filters the pileup for various user-defined parameters such as overall site coverage, reference and minor allele site coverage, and allele frequency. Two TCGA projects were re-validated using the modified variant caller. In an interrogation spanning 2000 sites, automated Ion Torrent concordance with 454 rose to 94%. Additional analysis on whole exome capture and Ion's Ampliseq panels on the PGM are underway.

FF0248

Short Tandem Repeat Sequencing on the 454 Platform

Melissa K. Scheible1,2, Odile Loreille1,2, and Jodi A. Irwin1,2.

1American Registry of Pathology, Rockville, MD, USA

2Armed Forces DNA Identification Laboratory, Dover AFB, DE, USA

The development of reduced size amplicons for short tandem repeat (STR) markers has increased the success of typing samples with degraded DNA (Butler et al. 2003, Coble and Butler 2005, Grubwieser et al. 2006, Hill et al. 2008). However, when separating these alleles using capillary electrophoresis, the number of loci that can be multiplexed together is restricted by the number of available dyes. Those loci with overlapping size ranges must either be labeled with different dyes, altered with mobility modifiers or amplified in separate multiplexes (Hill et al. 2008). Nevertheless, there’s a limit to the number of markers that can be reasonably multiplexed in a single reaction. One potential solution to this limitation of standard chemistries and detection platforms is next generation sequencing technology based on clonal amplification followed by emulsion PCR and pyrosequencing (Margulies et al. 2005). With this strategy, many loci with overlapping amplicon size ranges can be multiplexed in one reaction. This not only simplifies multiplex optimization, but more importantly, allows conservation of valuable sample extract.

To investigate the feasibility of next generation sequencing technology for the multiplex detection and sequence production of short tandem repeats, PCR methods, already familiar to the forensic science community, were used to enrich common short tandem repeat markers. Samples were amplified with a variety of multiplexing strategies: a commercial kit using labeled primers, an optimized multiplex containing thirteen miniSTR markers, and a series of multiplexes containing four miniSTR markers each. Each sample multiplex was barcoded with a different sample-specific multiplex identifier (MID) for subsequent parallel tagged sequencing on the GS Junior System (454 Life Sciences, Branford, CT). Here, we present the results of our preliminary data assessments and discuss some of the challenges that have been encountered with highly multiplexed markers and fluorescently-labeled primers.

FF0280

STR Profiling From Personal Genomes: Happy Surprises

Yaniv Erlich

Whitehead Institute for Biomedical Research

Short Tandem Repeats (STRs) have a wide range of applications, including medical genetics, forensics, and genetic genealogy. High throughput sequencing (HTS) has the potential to profile hundreds of thousands of STR loci. However, mainstream bioinformatics pipelines are inadequate for the task. These pipelines treat STR mapping as gapped alignment, which results in cumbersome processing times and a biased sampling of STR alleles.

In the first part of my talk, I will present a novel algorithm, called lobSTR, to profile STR markers from High Throughput Sequencing data. We validated lobSTR's accuracy by measuring its consistency in calling STRs from whole genome sequencing of two biological replicates from the same individual, by tracing Mendelian inheritance patterns in STR alleles in whole-genome sequencing of a HapMap trio, and by comparing lobSTR results to traditional molecular techniques, including the CODIS set.

In the second part of my talk, I will show how to recover surnames from anonymous sequencing datasets using lobSTR and massive Web 2.0 genealogical databases. We demonstrate the feasibility of the technique by recovering the surname 'Venter' from Craig Venter's genome. We also find that short read datasets are amenable for surname recovery. Applying this technique to more than 1,000 US males shows a success rate of 13%-16%. This suggests that surname recovery can generate investigation leads from crime scene samples.

Discussion Notes

2012 Attendees (by FF #)

FF #	First	Last Name	Affiliation	email	Abstract
FF0001	Chris	Detter	Los Alamos National Laboratory (LANL)	cdetter@lanl.gov	No
FF0002	Roger	Barrette	U.S. Dept. of Agriculture, APHIS	Roger.W.Barrette@aphis.usda.gov	No
FF0003	Martina	Siwek	Critical Reagents Program Biosurveillance, CBMS	martina.siwek.ctr@us.army.mil	No
FF0004	John	Chow	National Center for Genome Resources (NCGR)	jac@ncgr.org	**Yes**
FF0005	Terry	Gaasterland	Scripps Genome Center UCSD	gaasterland@gmail.com	No
FF0006	Ben	McMahon	Los Alamos National Laboratory (LANL)	mcmahon@lanl.gov	**Yes**
FF0007	Michael	Rhodes	Life Technologies	Michael.Rhodes@lifetech.com	No
FF0008	Hazuki	Teshima	Los Alamos National Laboratory (LANL)	hazuki@lanl.gov	No
FF0009	Wei	Gu	Los Alamos National Laboratory (LANL)	wgu@lanl.gov	No
FF0010	Alfredo Lopez	De Leon	Novozymes, Inc.	ALLO@novozymes.com	No
FF0011	Ara	Kooser	University of New Mexico	ghashsnaga@gmail.com	No
FF0012	Jason	Aulds	National Center for Medical Intelligence	jaulds@ncmi.detrick.army.mil	No
FF0013	Robert	Baker	Texas Tech University	Robert.Baker@ttu.edu	No
FF0014	Beth	Nelson	Novozymes, Inc.	BANE@novozymes.com	No
FF0015	Dhwani	Batra	SRA International	bun3@cdc.gov	**Yes**
FF0016	Diana	Northup	University of New Mexico	dnorthup@unm.edu	No
FF0017	Lucy	Zhang	Los Alamos National Laboratory (LANL)	xlz@lanl.gov	No
FF0018	Elizabeth	Montano	University of New Mexico	emont@unm.edu	No
FF0019	George	Vacek	Convey Computer Corporation	gvacek@conveycomputer.com	**Yes**
FF0020	George	VanDegrift	Convey Computer Corporation	gvandegrift@conveycomputer.com	No
FF0021	Nur	Hasan	CosmosID	nur.hasan@cosmosid.net	No
FF0022	Jane	Hutchinson	Roche Applied Science	jane.hutchinson@roche.com	No
FF0023	Jean	Challacombe	Los Alamos National Laboratory (LANL)	jchalla@lanl.gov	No
FF0024	John	Havens	Integrated DNA Technologies	jhavens@idtdna.com	No
FF0025	Lee	Katz	Centers for Disease Control and Prevention	gzu2@cdc.gov	No
FF0026	Peter	Larsen	USDA Agricultural Research Service	Peter.Larsen@ARS.USDA.GOV	**Yes**
FF0027	Malin	Young	Sandia National Laboratories	mmyoung@sandia.gov	No
FF0028	Jon	Longmire	Los Alamos National Laboratory (LANL)	jonlongmire@lanl.gov	No
FF0029	Massie	Ballon	Joint Genome Institute (JGI)	mlballon@lbl.gov	No
FF0030	Teri	Mueller	Roche Applied Science	teri.mueller@roche.com	No
FF0031	Michael	Rey	Novozymes, Inc.	MWR@novozymes.com	No
FF0032	George	Weinstock	The Genome Institute at Washington University	gweinsto@genome.wustl.edu	**Yes**
FF0033	Caleb	Phillips	Texas Tech University	caleb.phillips@ttu.edu	No
FF0034	Matt	Scholz	Los Alamos National Laboratory (LANL)	mscholz@lanl.gov	**Yes**
FF0035	Bryce	Ricken	Sandia National Laboratories	bricken@sandia.gov	No
FF0036	Robert	Settlage	Virginia Bioinformatics Institute, Virginia Tech	rsettlage@vbi.vt.edu	No
FF0037	Scott	Sammons	Centers for Disease Control and Prevention	zno6@cdc.gov	No
FF0038	Steve	Turner	Pacific Biosciences	sturner@pacificbiosciences.com	**Yes**
FF0039	Fiona	Stewart	New England Biolabs	stewart@neb.com	**Yes**
FF0040	Cheryl	Tarr	Centers for Disease Control and Prevention	crt6@cdc.gov	No
FF0041	Ian	Watson	Defense Threat Reduction Agency	Ian.Watson@dtra.mil	No
FF0042	Rita	Colwell	University of Maryland	rcolwell@umiacs.umd.edu	**Yes**
FF0043	Paul	Keim	Northern Arizona University	Paul.Keim@nau.edu	**Yes**
FF0044	Maryann	Turnsek	Centers For Disease Control and Prevention	hud4@cdc.gov	No
FF0045a	Dan	Ader	Monsanto Company	daniel.b.ader@monsanto.com	**Yes**
FF0045b	Dan	Ader	Monsanto Company	daniel.b.ader@monsanto.com	**Yes**
FF0046	Roman	Aranda	Expeditionary Forensic Division, USACIL	RomanAranda@a-tsolutions.com	No
FF0047	Tom	Callaghan	FBI Laboratory, Quantico	Thomas.Callaghan@ic.fbi.gov	**Yes**
FF0048	Scott	Remine	JPEO	scott.remine@dtra.mil	No
FF0049	Judy	Macemon	OpGen	jmacemon@opgen.com	No
FF0050	Mike	Fitzpatrick	OpGen	mfitzpatrick@opgen.com	No
FF0051	Trevor	Wagner	OpGen	twagner@opgen.com	**Yes**
FF0052	Randy	Stratton	OpGen	rstratton@opgen.com	No
FF0053	Chad	Locklear	Integrated DNA Technologies	clocklear@idtdna.com	No
FF0054	Edward	Kirton	Joint Genome Institute (JGI)	eskirton@lbl.gov	No
FF0055	Han	Cao	BioNano Genomics, Inc	han@bionanogenomics.com	No
FF0056	Harper	VanSteenhouse	BioNano Genomics, Inc	hvansteenhouse@bionanogenomics.com	No
FF0057	Alla	Lapidus	Fox Chase Cancer Center	Alla.Lapidus@fccc.edu	No
FF0058	Michael	Requa	BioNano Genomics, Inc	MRequa@bionanogenomics.com	No
FF0059	Ron	Walters	Pacific Northwest National Laboratory (PNNL)	ron@ron-walters.com	No
FF0060	David	Jenkins	EdgeBio	DJenkins@edgebio.com	**Yes**
FF0061	David	Trees	Centers for Disease Control and Prevention	dlt1@cdc.gov	No
FF0062	Eija	Trees	Centers for Disease Control and Prevention	eih9@cdc.gov	No
FF0063	Landon	Merrill	New England Biolabs	Merrill@neb.com	**Yes**
FF0064	Eric	Ackerman	Sandia National Laboratories	eackerm@sandia.gov	No
FF0065	Justin	Zook	National Institute of Standards and Technology	justin.zook@nist.gov	**Yes**
FF0066	Xing	Yang	BioNano Genomics, Inc	xyang@bionanogenomics.com	No
FF0067	Thomas	Cebula	Johns Hopkins University	tcebula1@jhu.edu	No
FF0068	Robert	Mervis	CLC bio	rmervis@clcbio.com	No

2012 Attendees (by FF #)

FF #	First	Last Name	Affiliation	email	Abstract
FF0069	Vish	Mokashi	Naval Medical Research Center (NMRC)	Vishwesh.Mokashi@med.navy.mil	No
FF0070	Todd	Smith	Perkin Elmer	Todd.Smith@PERKINELMER.COM	Yes
FF0071	Robin	Kramer	National Center for Genome Resources (NCGR)	rsk@ncgr.org	No
FF0072	Darrell	Ricke	MIT Lincoln Laboratory	darrell.ricke@ll.mit.edu	No
FF0073	Beverly	Parson-Quintana	Los Alamos National Laboratory (LANL)	bapq@lanl.gov	No
FF0074	Mohd	Noor	Malaysia Genome Institute	emno72@gmail.com	No
FF0075	David	Gordon	University of Washington	dgordon@u.washington.edu	Yes
FF0076	Ben	Allen	Los Alamos National Laboratory (LANL)	bsa@lanl.gov	No
FF0077	Rachel	Bartholomew	Pacific Northwest National Laboratory (PNNL)	rachel.bartholomew@pnnl.gov	No
FF0078	Stacey	Broomall	Edgewood Chemical Biological Center (ECBC)	stacey.m.broomall.civ@mail.mil	Yes
FF0079	Olga	Chertkov	Los Alamos National Laboratory (LANL)	ochrtkv@lanl.gov	Yes
FF0080	Cristina	Takacs-Vesbach	University of New Mexico	cvesbach@unm.edu	Yes
FF0081	Karen	Davenport	Los Alamos National Laboratory (LANL)	kwdavenport@lanl.gov	No
FF0082	Tracy	Erkkila	Los Alamos National Laboratory (LANL)	terkkila@lanl.gov	No
FF0083	Nadia	Fedorova	J. Craig Venter Institute	NFedorov@jcvi.org	Yes
FF0084	Todd	Dickinson	BioNano Genomics, Inc	tdickinson@bionanogenomics.com	No
FF0085	Yusuf	Noor	Malaysia Genome Institute	yusufmn@gmail.com	No
FF0086	Lee	Edsall	Ludwig Institute for Cancer Research, UCSD	ledsall@ucsd.edu	No
FF0087	Suman	Pakala	J. Craig Venter Institute	SPakala@jcvi.org	No
FF0088	Sergey	Koren	NBACC	korens@nbacc.net	Yes
FF0089	James	Robertson	FBI Laboratory, Quantico	James.Robertson2@ic.fbi.gov	No
FF0090	Anne	Ruffing	Sandia National Laboratories	aruffin@sandia.gov	Yes
FF0091	Lou	Sherman	Purdue University	lsherman@purdue.edu	Yes
FF0092	Catherine	Smith	Centers for Disease Control and Prevention	cab2@cdc.gov	No
FF0093	Karen	Martin	Monsanto Company	karen.e.martin@monsanto.com	No
FF0094	Jing	Lu	Monsanto Company	jing.lu@monsanto.com	No
FF0095	Stacie	Norton	Monsanto Company	stacie.t.norton@monsanto.com	Yes
FF0096	Nicole	Rosenzweig	Edgewood Chemical Biological Center (ECBC)	carolyn.n.rosenzweig.civ@mail.mil	No
FF0097	Shanmuga	Sozhamannan	Naval Medical Research Center (NMRC)	Shanmuga.Sozhamannan@med.navy.mil	No
FF0098	Tina	Graves	Washington University School of Medicine	tgraves@genome.wustl.edu	Yes
FF0099	Bob	Fulton	Washington University School of Medicine	bfulton@genome.wustl.edu	Yes
FF0100	Anitha	Sundararajan	National Center for Genome Resources (NCGR)	asundara@ncgr.org	No
FF0101	Catherine	Campbell	Noblis	Catherine.Campbell@noblis.org	Yes
FF0102	Ernie	Retzel	National Center for Genome Resources (NCGR)	efr@ncgr.org	Yes
FF0103	Graham	Threadgill	Beckman Coulter Life Sciences	gjthreadgill@beckman.com	No
FF0104	Heather	Buelow	University of New Mexico	hnbuelow@gmail.com	No
FF0105	Ingrid	Lindquist	National Center for Genome Resources (NCGR)	iel@ncgr.org	No
FF0106	Joann	Mudge	National Center for Genome Resources (NCGR)	jm@ncgr.org	Yes
FF0107	Johar	Ali	Ontario Institute for Cancer Research	Johar.Ali@oicr.on.ca	No
FF0108	Scott	Jordan	Physik Instrumente	scottj@pi-usa.us	Yes
FF0109	Julien	Tremblay	Joint Genome Institute (JGI)	jtremblay@lbl.gov	Yes
FF0110	Aye	Wollam	Washington University School of Medicine	awollam@genome.wustl.edu	Yes
FF0111	Marta	Matvienko	CLC bio	mmatvienko@clcbio.com	Yes
FF0112	Martha	Perez-Arriaga	University of New Mexico	marperez@cs.unm.edu	No
FF0113	Lee Ann	McCue	Pacific Northwest National Laboratory (PNNL)	leeann.mccue@pnnl.gov	No
FF0114	Kamlesh	Patel	Sandia National Laboratories	kdpatel@sandia.gov	Yes
FF0115	Ahmet	Zeytun	Los Alamos National Laboratory (LANL)	azeytun@lanl.gov	No
FF0116	Amy	Powell	Sandia National Laboratories	ajpowel@sandia.gov	No
FF0117	Thiru	Ramaraj	National Center for Genome Resources (NCGR)	tr@ncgr.org	No
FF0118	Valerie	Barbe	Genoscope	vbarbe@genoscope.cns.fr	Yes
FF0119	Tom	Brettin	Oak Ridge National Laboratory (ORNL)	brettints@ornl.gov	Yes
FF0120	Deacon	Sweeney	OpGen	dsweeney@opgen.com	Yes
FF0121	Corinne	Da Silva	Genoscope	dasilva@genoscope.cns.fr	No
FF0122	Haley	Fiske	illumina, Inc.	hfiske@illumina.com	Yes
FF0123	Debra	Glidewell	U.S. Army Criminal Investigation Laboratory	debra.glidewell@us.army.mil	No
FF0124	Joe	Salvatore	CLC bio	jsalvatore@clcbio.com	No
FF0125	Elizabeth	Johnson	U.S. Army Criminal Investigation Laboratory	elizabeth.johnson4@us.army.mil	No
FF0126	Kashef	Qaadri	Biomatters, Inc.	kashef@biomatters.com	Yes
FF0127	Kate	Auger	Wellcome Trust Sanger Institute	kaa@sanger.ac.uk	Yes
FF0128	Jo	Wood	Wellcome Trust Sanger Institute	jmdw@sanger.ac.uk	Yes
FF0129	Patrick	Minx	Washington University School of Medicine	pminx@genome.wustl.edu	No
FF0130	Yuliya	Kunde	Los Alamos National Laboratory (LANL)	y.a.kunde@lanl.gov	No
FF0131	Tim	Minogue	USAMRIID	timothy.minogue@us.army.mil	No
FF0132	Nicholas	Beckloff	Case Western Reserve University	nmb64@case.edu	No
FF0133	Sterling	Thomas	Noblis	Sterling.Thomas@noblis.org	No
FF0134	Tom	Cebula	CosmosID	tom.cebula@cosmosid.net	No
FF0135	Matthew	Tobelmann	Defense Threat Reduction Agency	Matthew.Tobelmann@DTRA.Mil	No
FF0136	Pete	Vallone	National Institute of Standards and Technology	peter.vallone@nist.gov	Yes
FF0137	Edward	Wack	MIT Lincoln Laboratory	wack@ll.mit.edu	No

2012 Attendees (by FF #)

FF #	First	Last Name	Affiliation	email	Abstract
FF0138	Tim	Hunkapiller	Discovery Bio	tim@discoverybio.com	No
FF0139	Kanwar	Singh	Joint Genome Institute (JGI)	ksingh@lbl.gov	No
FF0140	Rebecca	Just	Armed Forces DNA Identification Laboratory	rebecca.s.just@us.army.mil	Yes
FF0141	Judy	Le	KAPA Biosystems	judy.le@kapabiosystems.com	No
FF0142	Jerrod	Schwartz	University of Washington	jschwar@u.washington.edu	Yes
FF0143	James	Harper	MIT Lincoln Laboratory	harper@ll.mit.edu	No
FF0144	Jennifer	Carter	Agilent Technologies	jennifer_carter@agilent.com	Yes
FF0145	Cheryl	Gleasner	Los Alamos National Laboratory (LANL)	cdgle@lanl.gov	No
FF0146	Gary	Simpson	University of New Mexico	garyl.simpson@comcast.net	No
FF0147	James	Gale	Tricore Reference Lab	James.Gale@tricore.org	No
FF0148	Don	Natvig	University of New Mexico	dnatvig@gmail.com	No
FF0149	Daniel	Bozinov	Genimbi	dbozinov@genimbi.com	Yes
FF0150	Helen	Cui	Los Alamos National Laboratory (LANL)	hhcui@lanl.gov	Yes
FF0151	Hajni	Daligault	Los Alamos National Laboratory (LANL)	hajkis@lanl.gov	No
FF0152	Christina	Chiu	RainDance Technologies	ChiuC@raindancetech.com	No
FF0153	Christian	Buhay	Baylor College of Medicine	cbuhay@bcm.edu	Yes
FF0154	Matthew	Bohn	US Army Criminal Investigation Laboratory	Matthew.Bohn@anser.org	No
FF0155	Chip	Beckwith	Caliper	Chip.Beckwith@PERKINELMER.COM	No
FF0156	Arvind	Bharti	National Center for Genome Resources (NCGR)	akb@ncgr.org	Yes
FF0157	Alexander	Kozik	UC Davis Genome Center	akozik@atgc.org	Yes
FF0158	Kristen	O'Connor	Department of Homeland Security	kristen.oconnor@associates.hq.dhs.gov	No
FF0159	Jim	Knight	Roche Applied Science	james.knight@roche.com	Yes
FF0160	Ward	Wakeland	UT Southwestern Medical Center at Dallas	Edward.Wakeland@UTSouthwestern.edu	Yes
FF0161	Kim	McMurry	Los Alamos National Laboratory (LANL)	kmcmurry@lanl.gov	No
FF0162	Maggie	Werner-Washburne	University of New Mexico	maggieww@unm.edu	No
FF0163	Ginger	Metcalf	Baylor College of Medicine	metcalf@bcm.edu	Yes
FF0164	Miriam	Hutchinson	University of New Mexico	yesterdaymail@gmail.com	Yes
FF0165	Chris	Munk	Los Alamos National Laboratory (LANL)	cmunk@lanl.gov	Yes
FF0166	Brigid	O'Brien	USACIL	brigid.obrien@us.army.mil	No
FF0167	Tootie	Tatum	Joint Genome Institute (JGI)	oltatum@lbl.gov	No
FF0168	Krista	Reitenga	Los Alamos National Laboratory (LANL)	reitenga@lanl.gov	No
FF0169	Scott	Layne	Alfred E Mann Foundation	scottl@aemf.org	No
FF0170	Shuangye	Yin	The Broad Institute	shuangye@broadinstitute.org	Yes
FF0171	Tom	Slezak	Lawrence Livermore National Laboratory (LLNL)	Slezak1@llnl.gov	No
FF0172	Sophie	Mangenot	Genoscope	mangenot@genoscope.cns.fr	Yes
FF0173	Surya	Saha	Cornell University	ss2489@cornell.edu	Yes
FF0174	Christian	Whitchurch	Defense Threat Reduction Agency	christian.whitchurch@dtra.mil	Yes
FF0175	Chris	Whitehouse	USAMRIID	chris.whitehouse@us.army.mil	No
FF0176	Jeffrey	Koehler	USAMRIID	Jeff.Koehler@amedd.army.mil	Yes
FF0177	Adrienne	Hall	USAMRIID	adrienne.t.hall@us.army.mil	Yes
FF0178	Kitty	Chase	USAMRIID	Catherine.J.Chase@us.army.mil	No
FF0179	John	Barnes	Centers for Disease Control and Prevention	fzq9@cdc.gov	No
FF0180	Holly	Barnes	Agilent Technologies	holly_barnes@agilent.com	No
FF0181	Jana	Blackett	Agilent Technologies	jana_blackett@agilent.com	No
FF0182	Jodi	Irwin	Armed Forces DNA Identification Laboratory	jodi.a.irwin@us.army.mil	No
FF0183	Odile	Loreille	Armed Forces DNA Identification Laboratory	oloreille@hotmail.com	No
FF0184	Sarah	Young	The Broad Institute	stowey@broadinstitute.org	Yes
FF0185	Mike	FitzGerald	The Broad Institute	fitz@broadinstitute.org	Yes
FF0186	Bruce	Walker	The Broad Institute	bruce@broadinstitute.org	Yes
FF0187	Sakina	Saif	The Broad Institute	ssaif@broadinstitute.org	Yes
FF0188	Sean	Sykes	The Broad Institute	ssykes@broadinstitute.org	Yes
FF0189	Dave	Michaels	CLC bio	dmichaels@clcbio.com	No
FF0190	Jon	Murray	CLC bio	jmurray@clcbio.com	No
FF0191	Jared	Latiolais	Bode Technology Group	Jared.Latiolais@bodetech.com	Yes
FF0192	Robert	Bever	Bode Technology Group	Robert.Bever@bodetech.com	Yes
FF0193	Lori	Peterson	Caldera Pharmaceuticals	peterson@cpsci.com	No
FF0194	Nicole	Touchet	Caldera Pharmaceuticals	nltouchet@yahoo.com	No
FF0195	Louise	McConnell	Life Technologies	Louise.McConnell@lifetech.com	No
FF0196	Matt	Hickenbotham	Life Technologies	Matthew.Hickenbotham@lifetech.com	No
FF0197	Kyle	O'Connor	Life Technologies	kyle.o'connor@lifetech.com	No
FF0198	Omayma	Al-Awar	illumina, Inc.	oalawar@illumina.com	No
FF0199	Jonathan	Allen	Lawrence Livermore National Laboratory (LLNL)	Allen99@llnl.gov	Yes
FF0200	Laura	Kavanaugh	Syngenta Biotechnology, Inc.	laura.kavanaugh@syngenta.com	No
FF0201	Anna	Montmayeur	The Broad Institute	annamont@broadinstitute.org	Yes
FF0202	David	Bruce	Los Alamos National Laboratory (LANL)	dbruce@lanl.gov	No
FF0203	Jennie	Hunter-Cevera	RTI International	hunterce@rti.org	No
FF0204	Toni Marie	Diegoli	Armed Forces DNA Identification Laboratory	toni.diegoli@us.army.mil	No
FF0205	Paula	Imbro	Sandia National Laboratories	pmimbro@sandia.gov	No
FF0206	James	Schupp	Translational Genomics Research Institute	jschupp@tgen.org	No

2012 Attendees (by FF #)

FF #	First	Last Name	Affiliation	email	Abstract
FF0207	Michael	Fitzsimons	Los Alamos National Laboratory (LANL)	msfitzsimons@lbl.gov	Yes
FF0208	Matthew	Ross	Baylor College of Medicine	mcross@bcm.edu	Yes
FF0209	Donna	Muzny	Baylor College of Medicine	donnam@bcm.edu	Yes
FF0210	Adam	English	Baylor College of Medicine	English@bcm.edu	Yes
FF0211	David	Sexton	Baylor College of Medicine	dsexton@bcm.edu	Yes
FF0212	Michael	Holder	Baylor College of Medicine	mholder@bcm.edu	No
FF0213a	Maryke	Appel	Kapa Biosystems	maryke.appel@kapabiosystems.com	Yes
FF0213b	Maryke	Appel	Kapa Biosystems	maryke.appel@kapabiosystems.com	Yes
FF0214	Bharath	Prithiviraj	University of Colorado, Boulder	Bharath.Prithiviraj@Colorado.EDU	No
FF0215	Craig	Blackhart	Los Alamos National Laboratory (LANL)	blackhart@lanl.gov	No
FF0216	Daniel	Bornman	Battelle Memorial Institute	bornmand@battelle.org	Yes
FF0217	Brian	Foster	Lawrence Berkeley Laboratory	bfoster@lbl.gov	No
FF0218	Dan	Colman	University of New Mexico	drcolman1@gmail.com	No
FF0219	Daniela	Puiu	Johns Hopkins University	dpuiu@hotmail.com	Yes
FF0220	Faye	Schilkey	National Center for Genome Resources (NCGR)	fds@ncgr.org	Yes
FF0221	Glenn	Tesler	University of California, San Diego	gptesler@math.ucsd.edu	Yes
FF0222	Lynne	Goodwin	Los Alamos National Laboratory (LANL)	lynneg@lanl.gov	No
FF0223	Cydne	Holt	illumina, Inc.	cholt@illumina.com	Yes
FF0224	Charles	Hong	Defense Threat Reduction Agency	Charles.Hong@DTRA.MIL	No
FF0225	Jenifer	Smith	Pennsylvania State University	jas1110@psu.edu	No
FF0226	Nancy	McMillan	Battelle Memorial Institute	McMillanN@battelle.org	No
FF0227	Mary Lea	Killian	National Veterinary Services Laboratories, USDA	Mary.L.Killian@aphis.usda.gov	No
FF0228	Jonathan	Nowacki	Roche Applied Science	jonathan.nowacki@roche.com	No
FF0229	Oksana	Lukjancenko	University of Denmark	oksana@cbs.dtu.dk	Yes
FF0230	Owen	Leiser	Northern Arizona University	Owen.Leiser@nau.edu	No
FF0231	Nito	Panganiban	Tulane National Primate Research Center	apangani@tulane.edu	No
FF0232	Joshua	Santarpia	Sandia National Laboratories	jsantar@sandia.gov	No
FF0233	Brian	Young	Battelle Memorial Institute	youngb@battelle.org	No
FF0234	Harold	Lee	Pacific Biosciences	hlee@pacificbiosciences.com	No
FF0235	Alicia	Clum	Joint Genome Institute (JGI)	aclum@lbl.gov	Yes
FF0236	David	Van Horn	University of New Mexico	vanhorn@unm.edu	No
FF0237	Ellen	Paxinos	Pacific Biosciences	epaxinos@pacificbiosciences.com	No
FF0238	George	Rosenberg	University of New Mexico	ghrose@unm.edu	No
FF0239	Tim	Harkins	Life Technologies	Timothy.Harkins@lifetech.com	Yes
FF0240	Jason	Farlow	x	jasonfarlow@hotmail.com	No
FF0241	Jonathan	Bingham	Pacific Biosciences	jbingham@pacificbiosciences.com	No
FF0242	Lee	Kolakowski	x	kolakowki@mac.com	No
FF0243	Meredith	Ashby	Pacific Biosciences	mashby@pacificbiosciences.com	No
FF0244	Peter	Olsen	Novozymes, Inc.	PBO@novozymes.com	No
FF0245	Peter	Pesenti	Defense Threat Reduction Agency	Peter.Pesenti@dtra.mil	No
FF0246	Rob	Miller	University of New Mexico	rdmiller@unm.edu	No
FF0247	Sandy	Calloway	Children’s Hospital Oakland Research Institute	scalloway@chori.org	No
FF0248	Melissa	Scheible	Armed Forces DNA Identification Laboratory	melissa.scheible@us.army.mil	Yes
FF0249	Shannon	Steinfadt	Los Alamos National Laboratory (LANL)	shannon@lanl.gov	No
FF0250	Valerie	McClain	University of California, Davis	vpettebone@ucdavis.edu	Yes
FF0251	Victoria	Hansen	University of New Mexico	vhansen@unm.edu	No
FF0252	Cathy	Cleland	Los Alamos National Laboratory (LANL)	ccleland@lanl.gov	No
FF0253	Sirisha	Sunkara	Joint Genome Institute (JGI)	ssunkara@lbl.gov	No
FF0254	Nikos	Kyrpides	Joint Genome Institute (JGI)	nckyrpides@lbl.gov	No
FF0255	Stephanie	Malfatti	Joint Genome Institute (JGI)	samalfatti@lbl.gov	No
FF0256	Ze	Peng	Joint Genome Institute (JGI)	zpeng@lbl.gov	Yes
FF0257	Anna	Lipzen	Joint Genome Institute (JGI)	alipzen@lbl.gov	Yes
FF0258	Xiandong	Meng	Joint Genome Institute (JGI)	xiandongmeng@lbl.gov	No
FF0259	Chris	Daum	Joint Genome Institute (JGI)	daum1@llnl.gov	Yes
FF0260	Julianna	Chow	Joint Genome Institute (JGI)	jchow@lbl.gov	Yes
FF0261	Kurt	Labutti	Joint Genome Institute (JGI)	klabutti@lbl.gov	No
FF0262	Hui	Sun	Joint Genome Institute (JGI)	hsun@lbl.gov	Yes
FF0263	Alex	Copeland	Joint Genome Institute (JGI)	accopeland@lbl.gov	Yes
FF0264	Bry	Lingard	Defence Science and Technology Laboratory	BLINGARD@dstl.gov.uk	No
FF0265	Adam	Kotorashvili	Central Public Health Reference Laboratory	Kotorashvili@gmail.com	No
FF0266	Jeff	Froula	Joint Genome Institute (JGI)	jlfroula@lbl.gov	No
FF0267	Ken	Taylor	Integrated DNA Technologies	ktaylor@idtdna.com	No
FF0268	Keven	Stevens	Integrated DNA Technologies	kstevens@idtdna.com	No
FF0269	Joseph	Kaufman	Joseph Kaufman & Associates	joe@joekaufman.net	No
FF0270	Lijing	Bu	University of New Mexico	lijing@unm.edu	Yes
FF0271	Daniel	Mazur	Life Technologies	Daniel.Mazur@lifetech.com	No
FF0272	Timothy	McMahon	Armed Forces DNA Identification Laboratory	timothy.p.mcmahon.ctr@us.army.mil	No
FF0273	Mhlengi	Ncube	Johns Hopkins University	207512189@stu.ukzn.ac.za	No
FF0274	Minh	Nguyen	National Institute of Justice	Minh.Nguyen@usdoj.gov	No

2012 Attendees (by FF #)

FF #	First	Last Name	Affiliation	email	Abstract
FF0275	Ori	Sargsyan	Los Alamos National Laboratory (LANL)	sargsyan@lanl.gov	No
FF0276	Helena	Skar	Los Alamos National Laboratory (LANL)	skar@lanl.gov	No
FF0277	Tracey	Freitas	Los Alamos National Laboratory (LANL)	tracey.freitas@gmail.com	No
FF0278	Richard	Winegar	MRIGlobal	rwinegar@mriglobal.org	No
FF0279	Xun	Xu	Beijing Genome Institute (BGI)	xuxun@genomics.cn	**Yes**
FF0280	Yaniv	Erlich	Whitehead Institute for Biomedical Research	yaniv@wi.mit.edu	**Yes**
FF0281	David	Hirschberg	Columbia University	david.hirschberg@columbia.edu	No
FF0282	Todd	Lane	Sandia National Laboratories	twlane@sandia.gov	**Yes**
FF0283	Robert	Dietrich	Syngenta Biotechnology, Inc.	bob.dietrich@syngenta.com	No
FF0284	Callum	Bell	National Center for Genome Resources (NCGR)	cjb@ncgr.org	No
FF0285	Susan	Cropp	FBI Laboratory	Susan.Cropp@ic.fbi.gov	No
FF0286	Jeff	Clark	Integrated DNA Technologies	jclark@idtdna.com	No
FF0287	Joe	Clarke	Syngenta Biotechnology, Inc.	joseph.clarke@syngenta.com	No
FF0288	Lisa	Salow	RainDance Technologies	salowl@raindancetech.com	No
FF0289	Alex	Hutcheson	Pacific Biosciences	ahutcheson@pacificbiosciences.com	No
FF0290	Luke	Hickey	Pacific Biosciences	lhickey@pacificbiosciences.com	No
FF0291	Rich	Lussier	Convey Computer Corporation	rlussier@conveycomputer.com	No
FF0292	Holly	Ganz	University of California, Davis	hhganz@ucdavis.edu	No
FF0293	Marilee	Morgan	Lovelace Respiratory Research Institute	mmorgan@mrn.org	No
FF0294	Sarah	Buddenborg	University of New Mexico	sbuddenb@unm.edu	No
FF0295	Mike	Smith	JPEO - CRP	michael.aaron.smith1@us.army.mil	No
FF0296	Robert	Yamamoto	FLIR	robert.yamamoto@flir.com	**Yes**
FF0297	Yilin	Zhang	Elim Biopharmaceuticals, Inc.	yilin12@gmail.com	No
FF0298	Bin	Hu	Los Alamos National Laboratory (LANL)	binhu@lanl.gov	**Yes**
FF0299	Ajay	Athavale	Monsanto Company	ajay.athavale@monsanto.com	**Yes**
FF0300	Shannon	Johnson	Los Alamos National Laboratory (LANL)	shannonj@lanl.gov	**Yes**
FF0301	Martin	Simonsen	CLC bio	msimonsen@clcbio.com	No
FF0302	Tod	Stuber	U.S. Dept. of Agriculture, APHIS	Tod.P.Stuber@aphis.usda.gov	No
FF0303	Patrick	Chain	Los Alamos National Laboratory (LANL)	pchain@lanl.gov	No
FF0304	Armand	Dichosa	Los Alamos National Laboratory (LANL)	armand@lanl.gov	**Yes**
FF0305	Chien-Chi	Lo	Los Alamos National Laboratory (LANL)	chienchi@lanl.gov	No
FF0306	Jennifer	Price	Los Alamos National Laboratory (LANL)	jprice@lanl.gov	No
FF0307	Tumpa	Arefin	Los Alamos National Laboratory (LANL)	arefin_a@lanl.gov	No
FF0308	Hong	Shen	Los Alamos National Laboratory (LANL)	xshen@lanl.gov	No
FF0309	Shawn	Starkenburg	Los Alamos National Laboratory (LANL)	shawns@lanl.gov	No
FF0310	Gary	Resnick	Los Alamos National Laboratory (LANL)	gary.resnick2@gmail.com	No

NOTES

2012 Attendees (by name)

FF #	First	Last Name	Affiliation	email	Abstract
FF0064	Eric	Ackerman	Sandia National Laboratories	eackerm@sandia.gov	No
FF0045a	Dan	Ader	Monsanto Company	daniel.b.ader@monsanto.com	Yes
FF0045b	Dan	Ader	Monsanto Company	daniel.b.ader@monsanto.com	Yes
FF0198	Omayma	Al-Awar	illumina, Inc.	oalawar@illumina.com	No
FF0107	Johar	Ali	Ontario Institute for Cancer Research	Johar.Ali@oicr.on.ca	No
FF0076	Ben	Allen	Los Alamos National Laboratory (LANL)	bsa@lanl.gov	No
FF0199	Jonathan	Allen	Lawrence Livermore National Laboratory (LLNL)	Allen99@llnl.gov	Yes
FF0213a	Maryke	Appel	Kapa Biosystems	maryke.appel@kapabiosystems.com	Yes
FF0213b	Maryke	Appel	Kapa Biosystems	maryke.appel@kapabiosystems.com	Yes
FF0046	Roman	Aranda	Expeditionary Forensic Division, USACIL	RomanAranda@a-tsolutions.com	No
FF0307	Tumpa	Arefin	Los Alamos National Laboratory (LANL)	arefin_a@lanl.gov	No
FF0243	Meredith	Ashby	Pacific Biosciences	mashby@pacificbiosciences.com	No
FF0299	Ajay	Athavale	Monsanto Company	ajay.athavale@monsanto.com	Yes
FF0127	Kate	Auger	Wellcome Trust Sanger Institute	kaa@sanger.ac.uk	Yes
FF0012	Jason	Aulds	National Center for Medical Intelligence	jaulds@ncmi.detrick.army.mil	No
FF0013	Robert	Baker	Texas Tech University	Robert.Baker@ttu.edu	No
FF0029	Massie	Ballon	Joint Genome Institute (JGI)	mlballon@lbl.gov	No
FF0118	Valerie	Barbe	Genoscope	vbarbe@genoscope.cns.fr	Yes
FF0179	John	Barnes	Centers for Disease Control and Prevention	fzq9@cdc.gov	No
FF0180	Holly	Barnes	Agilent Technologies	holly_barnes@agilent.com	No
FF0002	Roger	Barrette	U.S. Dept. of Agriculture, APHIS	Roger.W.Barrette@aphis.usda.gov	No
FF0077	Rachel	Bartholomew	Pacific Northwest National Laboratory (PNNL)	rachel.bartholomew@pnnl.gov	No
FF0015	Dhwani	Batra	SRA International	bun3@cdc.gov	Yes
FF0132	Nicholas	Beckloff	Case Western Reserve University	nmb64@case.edu	No
FF0155	Chip	Beckwith	Caliper	Chip.Beckwith@PERKINELMER.COM	No
FF0284	Callum	Bell	National Center for Genome Resources (NCGR)	cjb@ncgr.org	No
FF0192	Robert	Bever	Bode Technology Group	Robert.Bever@bodetech.com	Yes
FF0156	Arvind	Bharti	National Center for Genome Resources (NCGR)	akb@ncgr.org	Yes
FF0241	Jonathan	Bingham	Pacific Biosciences	jbingham@pacificbiosciences.com	No
FF0181	Jana	Blackett	Agilent Technologies	jana_blackett@agilent.com	No
FF0215	Craig	Blackhart	Los Alamos National Laboratory (LANL)	blackhart@lanl.gov	No
FF0154	Matthew	Bohn	US Army Criminal Investigation Laboratory	Matthew.Bohn@anser.org	No
FF0216	Daniel	Bornman	Battelle Memorial Institute	bornmand@battelle.org	Yes
FF0149	Daniel	Bozinov	Genimbi	dbozinov@genimbi.com	Yes
FF0119	Tom	Brettin	Oak Ridge National Laboratory (ORNL)	brettints@ornl.gov	Yes
FF0078	Stacey	Broomall	Edgewood Chemical Biological Center (ECBC)	stacey.m.broomall.civ@mail.mil	Yes
FF0202	David	Bruce	Los Alamos National Laboratory (LANL)	dbruce@lanl.gov	No
FF0270	Lijing	Bu	University of New Mexico	lijing@unm.edu	Yes
FF0294	Sarah	Buddenborg	University of New Mexico	sbuddenb@unm.edu	No
FF0104	Heather	Buelow	University of New Mexico	hnbuelow@gmail.com	No
FF0153	Christian	Buhay	Baylor College of Medicine	cbuhay@bcm.edu	Yes
FF0047	Tom	Callaghan	FBI Laboratory, Quantico	Thomas.Callaghan@ic.fbi.gov	Yes
FF0247	Sandy	Calloway	Children’s Hospital Oakland Research Institute	scalloway@chori.org	No
FF0101	Catherine	Campbell	Noblis	Catherine.Campbell@noblis.org	Yes
FF0055	Han	Cao	BioNano Genomics, Inc	han@bionanogenomics.com	No
FF0144	Jennifer	Carter	Agilent Technologies	jennifer_carter@agilent.com	Yes
FF0067	Thomas	Cebula	Johns Hopkins University	tcebula1@jhu.edu	No
FF0134	Tom	Cebula	CosmosID	tom.cebula@cosmosid.net	No
FF0303	Patrick	Chain	Los Alamos National Laboratory (LANL)	pchain@lanl.gov	No
FF0023	Jean	Challacombe	Los Alamos National Laboratory (LANL)	jchalla@lanl.gov	No
FF0178	Kitty	Chase	USAMRIID	Catherine.J.Chase@us.army.mil	No
FF0079	Olga	Chertkov	Los Alamos National Laboratory (LANL)	ochrtkv@lanl.gov	Yes
FF0152	Christina	Chiu	RainDance Technologies	ChiuC@raindancetech.com	No
FF0004	John	Chow	National Center for Genome Resources (NCGR)	jac@ncgr.org	Yes
FF0260	Julianna	Chow	Joint Genome Institute (JGI)	jchow@lbl.gov	Yes
FF0286	Jeff	Clark	Integrated DNA Technologies	jclark@idtdna.com	No
FF0287	Joe	Clarke	Syngenta Biotechnology, Inc.	joseph.clarke@syngenta.com	No
FF0252	Cathy	Cleland	Los Alamos National Laboratory (LANL)	ccleland@lanl.gov	No
FF0235	Alicia	Clum	Joint Genome Institute (JGI)	aclum@lbl.gov	Yes
FF0218	Dan	Colman	University of New Mexico	drcolman1@gmail.com	No
FF0042	Rita	Colwell	University of Maryland	rcolwell@umiacs.umd.edu	Yes
FF0263	Alex	Copeland	Joint Genome Institute (JGI)	accopeland@lbl.gov	Yes
FF0285	Susan	Cropp	FBI Laboratory	Susan.Cropp@ic.fbi.gov	No
FF0150	Helen	Cui	Los Alamos National Laboratory (LANL)	hhcui@lanl.gov	Yes
FF0121	Corinne	Da Silva	Genoscope	dasilva@genoscope.cns.fr	No
FF0151	Hajni	Daligault	Los Alamos National Laboratory (LANL)	hajkis@lanl.gov	No
FF0259	Chris	Daum	Joint Genome Institute (JGI)	daum1@llnl.gov	Yes
FF0081	Karen	Davenport	Los Alamos National Laboratory (LANL)	kwdavenport@lanl.gov	No
FF0010	Alfredo Lopez	De Leon	Novozymes, Inc.	ALLO@novozymes.com	No

2012 Attendees (by name)

FF #	First	Last Name	Affiliation	email	Abstract
FF0001	Chris	Detter	Los Alamos National Laboratory (LANL)	cdetter@lanl.gov	No
FF0304	Armand	Dichosa	Los Alamos National Laboratory (LANL)	armand@lanl.gov	Yes
FF0084	Todd	Dickinson	BioNano Genomics, Inc	tdickinson@bionanogenomics.com	No
FF0204	Toni Marie	Diegoli	Armed Forces DNA Identification Laboratory	toni.diegoli@us.army.mil	No
FF0283	Robert	Dietrich	Syngenta Biotechnology, Inc.	bob.dietrich@syngenta.com	No
FF0086	Lee	Edsall	Ludwig Institute for Cancer Research, UCSD	ledsall@ucsd.edu	No
FF0210	Adam	English	Baylor College of Medicine	English@bcm.edu	Yes
FF0082	Tracy	Erkkila	Los Alamos National Laboratory (LANL)	terkkila@lanl.gov	No
FF0280	Yaniv	Erlich	Whitehead Institute for Biomedical Research	yaniv@wi.mit.edu	Yes
FF0240	Jason	Farlow	x	jasonfarlow@hotmail.com	No
FF0083	Nadia	Fedorova	J. Craig Venter Institute	NFedorov@jcvi.org	Yes
FF0122	Haley	Fiske	illumina, Inc.	hfiske@illumina.com	Yes
FF0185	Mike	FitzGerald	The Broad Institute	fitz@broadinstitute.org	Yes
FF0050	Mike	Fitzpatrick	OpGen	mfitzpatrick@opgen.com	No
FF0207	Michael	Fitzsimons	Los Alamos National Laboratory (LANL)	msfitzsimons@lbl.gov	Yes
FF0217	Brian	Foster	Lawrence Berkeley Laboratory	bfoster@lbl.gov	No
FF0277	Tracey	Freitas	Los Alamos National Laboratory (LANL)	tracey.freitas@gmail.com	No
FF0266	Jeff	Froula	Joint Genome Institute (JGI)	jlfroula@lbl.gov	No
FF0099	Bob	Fulton	Washington University School of Medicine	bfulton@genome.wustl.edu	Yes
FF0005	Terry	Gaasterland	Scripps Genome Center UCSD	gaasterland@gmail.com	No
FF0147	James	Gale	Tricore Reference Lab	James.Gale@tricore.org	No
FF0292	Holly	Ganz	University of California, Davis	hhganz@ucdavis.edu	No
FF0145	Cheryl	Gleasner	Los Alamos National Laboratory (LANL)	cdgle@lanl.gov	No
FF0123	Debra	Glidewell	U.S. Army Criminal Investigation Laboratory	debra.glidewell@us.army.mil	No
FF0222	Lynne	Goodwin	Los Alamos National Laboratory (LANL)	lynneg@lanl.gov	No
FF0075	David	Gordon	University of Washington	dgordon@u.washington.edu	Yes
FF0098	Tina	Graves	Washington University School of Medicine	tgraves@genome.wustl.edu	Yes
FF0009	Wei	Gu	Los Alamos National Laboratory (LANL)	wgu@lanl.gov	No
FF0177	Adrienne	Hall	USAMRIID	adrienne.t.hall@us.army.mil	Yes
FF0251	Victoria	Hansen	University of New Mexico	vhansen@unm.edu	No
FF0239	Tim	Harkins	Life Technologies	Timothy.Harkins@lifetech.com	Yes
FF0143	James	Harper	MIT Lincoln Laboratory	harper@ll.mit.edu	No
FF0021	Nur	Hasan	CosmosID	nur.hasan@cosmosid.net	No
FF0024	John	Havens	Integrated DNA Technologies	jhavens@idtdna.com	No
FF0196	Matt	Hickenbotham	Life Technologies	Matthew.Hickenbotham@lifetech.com	No
FF0290	Luke	Hickey	Pacific Biosciences	lhickey@pacificbiosciences.com	No
FF0281	David	Hirschberg	Columbia University	david.hirschberg@columbia.edu	No
FF0212	Michael	Holder	Baylor College of Medicine	mholder@bcm.edu	No
FF0223	Cydne	Holt	illumina, Inc.	cholt@illumina.com	Yes
FF0224	Charles	Hong	Defense Threat Reduction Agency	Charles.Hong@DTRA.MIL	No
FF0298	Bin	Hu	Los Alamos National Laboratory (LANL)	binhu@lanl.gov	Yes
FF0138	Tim	Hunkapiller	Discovery Bio	tim@discoverybio.com	No
FF0203	Jennie	Hunter-Cevera	RTI International	hunterce@rti.org	No
FF0289	Alex	Hutcheson	Pacific Biosciences	ahutcheson@pacificbiosciences.com	No
FF0022	Jane	Hutchinson	Roche Applied Science	jane.hutchinson@roche.com	No
FF0164	Miriam	Hutchinson	University of New Mexico	yesterdaymail@gmail.com	Yes
FF0205	Paula	Imbro	Sandia National Laboratories	pmimbro@sandia.gov	No
FF0182	Jodi	Irwin	Armed Forces DNA Identification Laboratory	jodi.a.irwin@us.army.mil	No
FF0060	David	Jenkins	EdgeBio	DJenkins@edgebio.com	Yes
FF0125	Elizabeth	Johnson	U.S. Army Criminal Investigation Laboratory	elizabeth.johnson4@us.army.mil	No
FF0300	Shannon	Johnson	Los Alamos National Laboratory (LANL)	shannonj@lanl.gov	Yes
FF0108	Scott	Jordan	Physik Instrumente	scottj@pi-usa.us	Yes
FF0140	Rebecca	Just	Armed Forces DNA Identification Laboratory	rebecca.s.just@us.army.mil	Yes
FF0025	Lee	Katz	Centers for Disease Control and Prevention	gzu2@cdc.gov	No
FF0269	Joseph	Kaufman	Joseph Kaufman & Associates	joe@joekaufman.net	No
FF0200	Laura	Kavanaugh	Syngenta Biotechnology, Inc.	laura.kavanaugh@syngenta.com	No
FF0043	Paul	Keim	Northern Arizona University	Paul.Keim@nau.edu	Yes
FF0227	Mary Lea	Killian	National Veterinary Services Laboratories, USDA	Mary.L.Killian@aphis.usda.gov	No
FF0054	Edward	Kirton	Joint Genome Institute (JGI)	eskirton@lbl.gov	No
FF0159	Jim	Knight	Roche Applied Science	james.knight@roche.com	Yes
FF0176	Jeffrey	Koehler	USAMRIID	Jeff.Koehler@amedd.army.mil	Yes
FF0242	Lee	Kolakowski	x	kolakowki@mac.com	No
FF0011	Ara	Kooser	University of New Mexico	ghashsnaga@gmail.com	No
FF0088	Sergey	Koren	NBACC	korens@nbacc.net	Yes
FF0265	Adam	Kotorashvili	Central Public Health Reference Laboratory	Kotorashvili@gmail.com	No
FF0157	Alexander	Kozik	UC Davis Genome Center	akozik@atgc.org	Yes
FF0071	Robin	Kramer	National Center for Genome Resources (NCGR)	rsk@ncgr.org	No
FF0130	Yuliya	Kunde	Los Alamos National Laboratory (LANL)	y.a.kunde@lanl.gov	No
FF0254	Nikos	Kyrpides	Joint Genome Institute (JGI)	nckyrpides@lbl.gov	No

2012 Attendees (by name)

FF #	First	Last Name	Affiliation	email	Abstract
FF0261	Kurt	Labutti	Joint Genome Institute (JGI)	klabutti@lbl.gov	No
FF0282	Todd	Lane	Sandia National Laboratories	twlane@sandia.gov	**Yes**
FF0057	Alla	Lapidus	Fox Chase Cancer Center	Alla.Lapidus@fccc.edu	No
FF0026	Peter	Larsen	USDA Agricultural Research Service	Peter.Larsen@ARS.USDA.GOV	**Yes**
FF0191	Jared	Latiolais	Bode Technology Group	Jared.Latiolais@bodetech.com	**Yes**
FF0169	Scott	Layne	Alfred E Mann Foundation	scottl@aemf.org	No
FF0141	Judy	Le	KAPA Biosystems	judy.le@kapabiosystems.com	No
FF0234	Harold	Lee	Pacific Biosciences	hlee@pacificbiosciences.com	No
FF0230	Owen	Leiser	Northern Arizona University	Owen.Leiser@nau.edu	No
FF0105	Ingrid	Lindquist	National Center for Genome Resources (NCGR)	iel@ncgr.org	No
FF0264	Bry	Lingard	Defence Science and Technology Laboratory	BLINGARD@dstl.gov.uk	No
FF0257	Anna	Lipzen	Joint Genome Institute (JGI)	alipzen@lbl.gov	**Yes**
FF0305	Chien-Chi	Lo	Los Alamos National Laboratory (LANL)	chienchi@lanl.gov	No
FF0053	Chad	Locklear	Integrated DNA Technologies	clocklear@idtdna.com	No
FF0028	Jon	Longmire	Los Alamos National Laboratory (LANL)	jonlongmire@lanl.gov	No
FF0183	Odile	Loreille	Armed Forces DNA Identification Laboratory	oloreille@hotmail.com	No
FF0094	Jing	Lu	Monsanto Company	jing.lu@monsanto.com	No
FF0229	Oksana	Lukjancenko	University of Denmark	oksana@cbs.dtu.dk	**Yes**
FF0291	Rich	Lussier	Convey Computer Corporation	rlussier@conveycomputer.com	No
FF0049	Judy	Macemon	OpGen	jmacemon@opgen.com	No
FF0255	Stephanie	Malfatti	Joint Genome Institute (JGI)	samalfatti@lbl.gov	No
FF0172	Sophie	Mangenot	Genoscope	mangenot@genoscope.cns.fr	**Yes**
FF0093	Karen	Martin	Monsanto Company	karen.e.martin@monsanto.com	No
FF0111	Marta	Matvienko	CLC bio	mmatvienko@clcbio.com	**Yes**
FF0271	Daniel	Mazur	Life Technologies	Daniel.Mazur@lifetech.com	No
FF0250	Valerie	McClain	University of California, Davis	vpettebone@ucdavis.edu	**Yes**
FF0195	Louise	McConnell	Life Technologies	Louise.McConnell@lifetech.com	No
FF0113	Lee Ann	McCue	Pacific Northwest National Laboratory (PNNL)	leeann.mccue@pnnl.gov	No
FF0006	Ben	McMahon	Los Alamos National Laboratory (LANL)	mcmahon@lanl.gov	**Yes**
FF0272	Timothy	McMahon	Armed Forces DNA Identification Laboratory	timothy.p.mcmahon.ctr@us.army.mil	No
FF0226	Nancy	McMillan	Battelle Memorial Institute	McMillanN@battelle.org	No
FF0161	Kim	McMurry	Los Alamos National Laboratory (LANL)	kmcmurry@lanl.gov	No
FF0258	Xiandong	Meng	Joint Genome Institute (JGI)	xiandongmeng@lbl.gov	No
FF0063	Landon	Merrill	New England Biolabs	Merrill@neb.com	**Yes**
FF0068	Robert	Mervis	CLC bio	rmervis@clcbio.com	No
FF0163	Ginger	Metcalf	Baylor College of Medicine	metcalf@bcm.edu	**Yes**
FF0189	Dave	Michaels	CLC bio	dmichaels@clcbio.com	No
FF0246	Rob	Miller	University of New Mexico	rdmiller@unm.edu	No
FF0131	Tim	Minogue	USAMRIID	timothy.minogue@us.army.mil	No
FF0129	Patrick	Minx	Washington University School of Medicine	pminx@genome.wustl.edu	No
FF0069	Vish	Mokashi	Naval Medical Research Center (NMRC)	Vishwesh.Mokashi@med.navy.mil	No
FF0018	Elizabeth	Montano	University of New Mexico	emont@unm.edu	No
FF0201	Anna	Montmayeur	The Broad Institute	annamont@broadinstitute.org	**Yes**
FF0293	Marilee	Morgan	Lovelace Respiratory Research Institute	mmorgan@mrn.org	No
FF0106	Joann	Mudge	National Center for Genome Resources (NCGR)	jm@ncgr.org	**Yes**
FF0030	Teri	Mueller	Roche Applied Science	teri.mueller@roche.com	No
FF0165	Chris	Munk	Los Alamos National Laboratory (LANL)	cmunk@lanl.gov	**Yes**
FF0190	Jon	Murray	CLC bio	jmurray@clcbio.com	No
FF0209	Donna	Muzny	Baylor College of Medicine	donnam@bcm.edu	**Yes**
FF0148	Don	Natvig	University of New Mexico	dnatvig@gmail.com	No
FF0273	Mhlengi	Ncube	Johns Hopkins University	207512189@stu.ukzn.ac.za	No
FF0014	Beth	Nelson	Novozymes, Inc.	BANE@novozymes.com	No
FF0274	Minh	Nguyen	National Institute of Justice	Minh.Nguyen@usdoj.gov	No
FF0074	Mohd	Noor	Malaysia Genome Institute	emno72@gmail.com	No
FF0085	Yusuf	Noor	Malaysia Genome Institute	yusufmn@gmail.com	No
FF0016	Diana	Northup	University of New Mexico	dnorthup@unm.edu	No
FF0095	Stacie	Norton	Monsanto Company	stacie.t.norton@monsanto.com	**Yes**
FF0228	Jonathan	Nowacki	Roche Applied Science	jonathan.nowacki@roche.com	No
FF0197	Kyle	O’Connor	Life Technologies	kyle.o’connor@lifetech.com	No
FF0166	Brigid	O'Brien	USACIL	brigid.obrien@us.army.mil	No
FF0158	Kristen	O'Connor	Department of Homeland Security	kristen.oconnor@associates.hq.dhs.gov	No
FF0244	Peter	Olsen	Novozymes, Inc.	PBO@novozymes.com	No
FF0087	Suman	Pakala	J. Craig Venter Institute	SPakala@jcvi.org	No
FF0231	Nito	Panganiban	Tulane National Primate Research Center	apangani@tulane.edu	No
FF0073	Beverly	Parson-Quintana	Los Alamos National Laboratory (LANL)	bapq@lanl.gov	No
FF0114	Kamlesh	Patel	Sandia National Laboratories	kdpatel@sandia.gov	**Yes**
FF0237	Ellen	Paxinos	Pacific Biosciences	epaxinos@pacificbiosciences.com	No
FF0256	Ze	Peng	Joint Genome Institute (JGI)	zpeng@lbl.gov	**Yes**
FF0112	Martha	Perez-Arriaga	University of New Mexico	marperez@cs.unm.edu	No

2012 Attendees (by name)

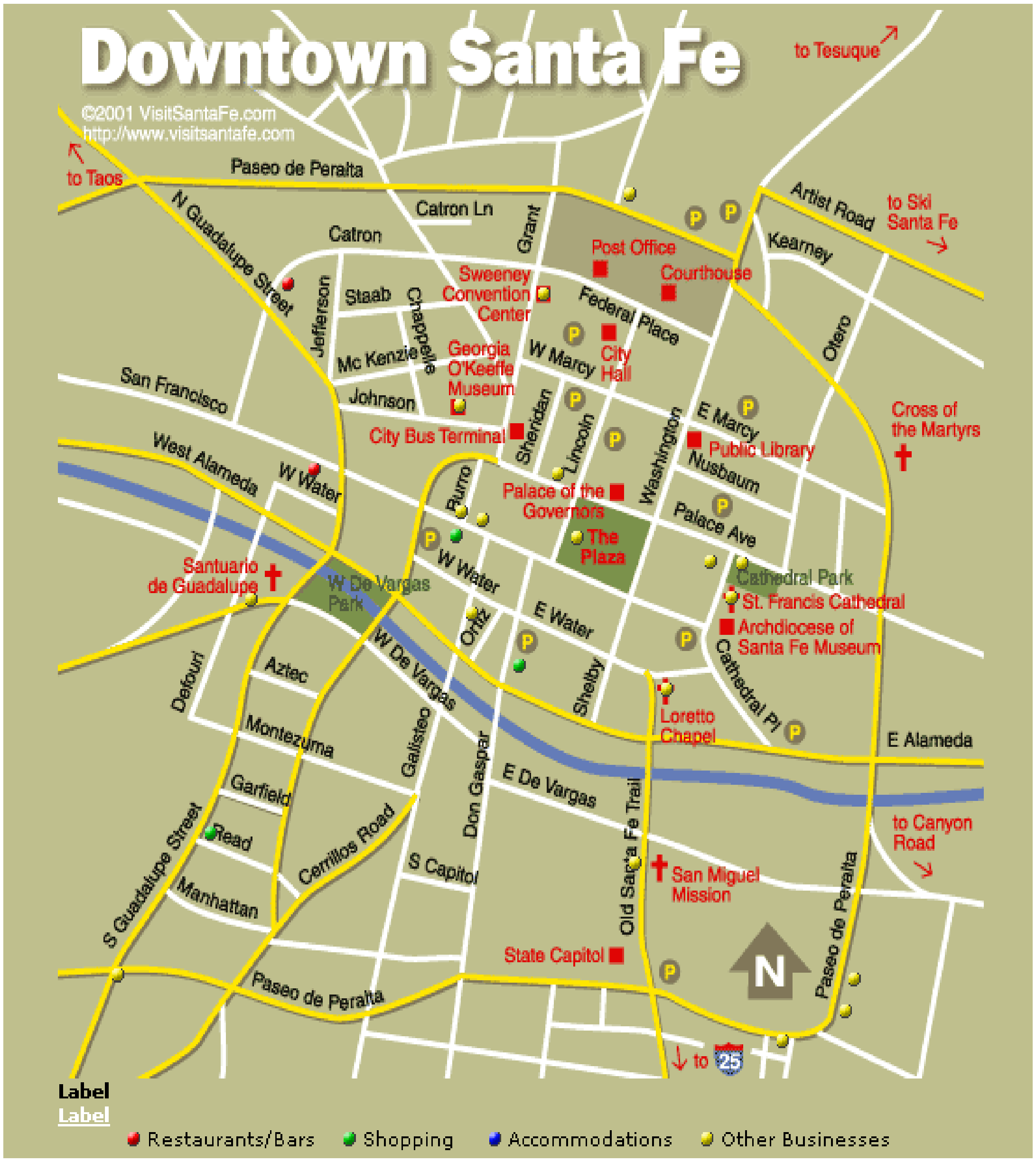
FF #	First	Last Name	Affiliation	email	Abstract
FF0245	Peter	Pesenti	Defense Threat Reduction Agency	Peter.Pesenti@dtra.mil	No
FF0193	Lori	Peterson	Caldera Pharmaceuticals	peterson@cpsci.com	No
FF0033	Caleb	Phillips	Texas Tech University	caleb.phillips@ttu.edu	No
FF0116	Amy	Powell	Sandia National Laboratories	ajpowel@sandia.gov	No
FF0306	Jennifer	Price	Los Alamos National Laboratory (LANL)	jprice@lanl.gov	No
FF0214	Bharath	Prithiviraj	University of Colorado, Boulder	Bharath.Prithiviraj@Colorado.EDU	No
FF0219	Daniela	Puiu	Johns Hopkins University	dpuiu@hotmail.com	**Yes**
FF0126	Kashef	Qaadri	Biomatters, Inc.	kashef@biomatters.com	**Yes**
FF0117	Thiru	Ramaraj	National Center for Genome Resources (NCGR)	tr@ncgr.org	No
FF0168	Krista	Reitenga	Los Alamos National Laboratory (LANL)	reitenga@lanl.gov	No
FF0048	Scott	Remine	JPEO	scott.remine@dtra.mil	No
FF0058	Michael	Requa	BioNano Genomics, Inc	MRequa@bionanogenomics.com	No
FF0310	Gary	Resnick	Los Alamos National Laboratory (LANL)	gary.resnick2@gmail.com	No
FF0102	Ernie	Retzel	National Center for Genome Resources (NCGR)	efr@ncgr.org	**Yes**
FF0031	Michael	Rey	Novozymes, Inc.	MWR@novozymes.com	No
FF0007	Michael	Rhodes	Life Technologies	Michael.Rhodes@lifetech.com	No
FF0072	Darrell	Ricke	MIT Lincoln Laboratory	darrell.ricke@ll.mit.edu	No
FF0035	Bryce	Ricken	Sandia National Laboratories	bricken@sandia.gov	No
FF0089	James	Robertson	FBI Laboratory, Quantico	James.Robertson2@ic.fbi.gov	No
FF0238	George	Rosenberg	University of New Mexico	ghrose@unm.edu	No
FF0096	Nicole	Rosenzweig	Edgewood Chemical Biological Center (ECBC)	carolyn.n.rosenzweig.civ@mail.mil	No
FF0208	Matthew	Ross	Baylor College of Medicine	mcross@bcm.edu	**Yes**
FF0090	Anne	Ruffing	Sandia National Laboratories	aruffin@sandia.gov	**Yes**
FF0173	Surya	Saha	Cornell University	ss2489@cornell.edu	**Yes**
FF0187	Sakina	Saif	The Broad Institute	ssaif@broadinstitute.org	**Yes**
FF0288	Lisa	Salow	RainDance Technologies	salowl@raindancetech.com	No
FF0124	Joe	Salvatore	CLC bio	jsalvatore@clcbio.com	No
FF0037	Scott	Sammons	Centers for Disease Control and Prevention	zno6@cdc.gov	No
FF0232	Joshua	Santarpia	Sandia National Laboratories	jsantar@sandia.gov	No
FF0275	Ori	Sargsyan	Los Alamos National Laboratory (LANL)	sargsyan@lanl.gov	No
FF0248	Melissa	Scheible	Armed Forces DNA Identification Laboratory	melissa.scheible@us.army.mil	**Yes**
FF0220	Faye	Schilkey	National Center for Genome Resources (NCGR)	fds@ncgr.org	**Yes**
FF0034	Matt	Scholz	Los Alamos National Laboratory (LANL)	mscholz@lanl.gov	**Yes**
FF0206	James	Schupp	Translational Genomics Research Institute	jschupp@tgen.org	No
FF0142	Jerrod	Schwartz	University of Washington	jschwar@u.washington.edu	**Yes**
FF0036	Robert	Settlage	Virginia Bioinformatics Institute, Virginia Tech	rsettlage@vbi.vt.edu	No
FF0211	David	Sexton	Baylor College of Medicine	dsexton@bcm.edu	**Yes**
FF0308	Hong	Shen	Los Alamos National Laboratory (LANL)	xshen@lanl.gov	No
FF0091	Lou	Sherman	Purdue University	lsherman@purdue.edu	**Yes**
FF0301	Martin	Simonsen	CLC bio	msimonsen@clcbio.com	No
FF0146	Gary	Simpson	University of New Mexico	garyl.simpson@comcast.net	No
FF0139	Kanwar	Singh	Joint Genome Institute (JGI)	ksingh@lbl.gov	No
FF0003	Martina	Siwek	Critical Reagents Program Biosurveillance, CBMS	martina.siwek.ctr@us.army.mil	No
FF0276	Helena	Skar	Los Alamos National Laboratory (LANL)	skar@lanl.gov	No
FF0171	Tom	Slezak	Lawrence Livermore National Laboratory (LLNL)	Slezak1@llnl.gov	No
FF0070	Todd	Smith	Perkin Elmer	Todd.Smith@PERKINELMER.COM	**Yes**
FF0092	Catherine	Smith	Centers for Disease Control and Prevention	cab2@cdc.gov	No
FF0225	Jenifer	Smith	Pennsylvania State University	jas1110@psu.edu	No
FF0295	Mike	Smith	JPEO - CRP	michael.aaron.smith1@us.army.mil	No
FF0097	Shanmuga	Sozhamannan	Naval Medical Research Center (NMRC)	Shanmuga.Sozhamannan@med.navy.mil	No
FF0309	Shawn	Starkenburg	Los Alamos National Laboratory (LANL)	shawns@lanl.gov	No
FF0249	Shannon	Steinfadt	Los Alamos National Laboratory (LANL)	shannon@lanl.gov	No
FF0268	Keven	Stevens	Integrated DNA Technologies	kstevens@idtdna.com	No
FF0039	Fiona	Stewart	New England Biolabs	stewart@neb.com	**Yes**
FF0052	Randy	Stratton	OpGen	rstratton@opgen.com	No
FF0302	Tod	Stuber	U.S. Dept. of Agriculture, APHIS	Tod.P.Stuber@aphis.usda.gov	No
FF0262	Hui	Sun	Joint Genome Institute (JGI)	hsun@lbl.gov	**Yes**
FF0100	Anitha	Sundararajan	National Center for Genome Resources (NCGR)	asundara@ncgr.org	No
FF0253	Sirisha	Sunkara	Joint Genome Institute (JGI)	ssunkara@lbl.gov	No
FF0120	Deacon	Sweeney	OpGen	dsweeney@opgen.com	**Yes**
FF0188	Sean	Sykes	The Broad Institute	ssykes@broadinstitute.org	**Yes**
FF0080	Cristina	Takacs-Vesbach	University of New Mexico	cvesbach@unm.edu	**Yes**
FF0040	Cheryl	Tarr	Centers for Disease Control and Prevention	crt6@cdc.gov	No
FF0167	Tootie	Tatum	Joint Genome Institute (JGI)	oltatum@lbl.gov	No
FF0267	Ken	Taylor	Integrated DNA Technologies	ktaylor@idtdna.com	No
FF0008	Hazuki	Teshima	Los Alamos National Laboratory (LANL)	hazuki@lanl.gov	No
FF0221	Glenn	Tesler	University of California, San Diego	gptesler@math.ucsd.edu	**Yes**
FF0133	Sterling	Thomas	Noblis	Sterling.Thomas@noblis.org	No
FF0103	Graham	Threadgill	Beckman Coulter Life Sciences	gjthreadgill@beckman.com	No

2012 Attendees (by name)

FF #	First	Last Name	Affiliation	email	Abstract
FF0135	Matthew	Tobelmann	Defense Threat Reduction Agency	Matthew.Tobelmann@DTRA.Mil	No
FF0194	Nicole	Touchet	Caldera Pharmaceuticals	nltouchet@yahoo.com	No
FF0061	David	Trees	Centers for Disease Control and Prevention	dlt1@cdc.gov	No
FF0062	Eija	Trees	Centers for Disease Control and Prevention	eih9@cdc.gov	No
FF0109	Julien	Tremblay	Joint Genome Institute (JGI)	jtremblay@lbl.gov	**Yes**
FF0038	Steve	Turner	Pacific Biosciences	sturner@pacificbiosciences.com	**Yes**
FF0044	Maryann	Turnsek	Centers For Disease Control and Prevention	hud4@cdc.gov	No
FF0019	George	Vacek	Convey Computer Corporation	gvacek@conveycomputer.com	**Yes**
FF0136	Pete	Vallone	National Institute of Standards and Technology	peter.vallone@nist.gov	**Yes**
FF0236	David	Van Horn	University of New Mexico	vanhorn@unm.edu	No
FF0020	George	VanDegrift	Convey Computer Corporation	gvandegrift@conveycomputer.com	No
FF0056	Harper	VanSteenhouse	BioNano Genomics, Inc	hvansteenhouse@bionanogenomics.com	No
FF0137	Edward	Wack	MIT Lincoln Laboratory	wack@ll.mit.edu	No
FF0051	Trevor	Wagner	OpGen	twagner@opgen.com	**Yes**
FF0160	Ward	Wakeland	UT Southwestern Medical Center at Dallas	Edward.Wakeland@UTSouthwestern.edu	**Yes**
FF0186	Bruce	Walker	The Broad Institute	bruce@broadinstitute.org	**Yes**
FF0059	Ron	Walters	Pacific Northwest National Laboratory (PNNL)	ron@ron-walters.com	No
FF0041	Ian	Watson	Defense Threat Reduction Agency	Ian.Watson@dtra.mil	No
FF0032	George	Weinstock	The Genome Institute at Washington University	gweinsto@genome.wustl.edu	**Yes**
FF0162	Maggie	Werner-Washburne	University of New Mexico	maggieww@unm.edu	No
FF0174	Christian	Whitchurch	Defense Threat Reduction Agency	christian.whitchurch@dtra.mil	**Yes**
FF0175	Chris	Whitehouse	USAMRIID	chris.whitehouse@us.army.mil	No
FF0278	Richard	Winegar	MRIGlobal	rwinegar@mriglobal.org	No
FF0110	Aye	Wollam	Washington University School of Medicine	awollam@genome.wustl.edu	**Yes**
FF0128	Jo	Wood	Wellcome Trust Sanger Institute	jmdw@sanger.ac.uk	**Yes**
FF0279	Xun	Xu	Beijing Genome Institute (BGI)	xuxun@genomics.cn	**Yes**
FF0296	Robert	Yamamoto	FLIR	robert.yamamoto@flir.com	**Yes**
FF0066	Xing	Yang	BioNano Genomics, Inc	xyang@bionanogenomics.com	No
FF0170	Shuangye	Yin	The Broad Institute	shuangye@broadinstitute.org	**Yes**
FF0027	Malin	Young	Sandia National Laboratories	mmyoung@sandia.gov	No
FF0184	Sarah	Young	The Broad Institute	stowey@broadinstitute.org	**Yes**
FF0233	Brian	Young	Battelle Memorial Institute	youngb@battelle.org	No
FF0115	Ahmet	Zeytun	Los Alamos National Laboratory (LANL)	azeytun@lanl.gov	No
FF0017	Lucy	Zhang	Los Alamos National Laboratory (LANL)	xlz@lanl.gov	No
FF0297	Yilin	Zhang	Elim Biopharmaceuticals, Inc.	yilin12@gmail.com	No
FF0065	Justin	Zook	National Institute of Standards and Technology	justin.zook@nist.gov	**Yes**

Notes

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.

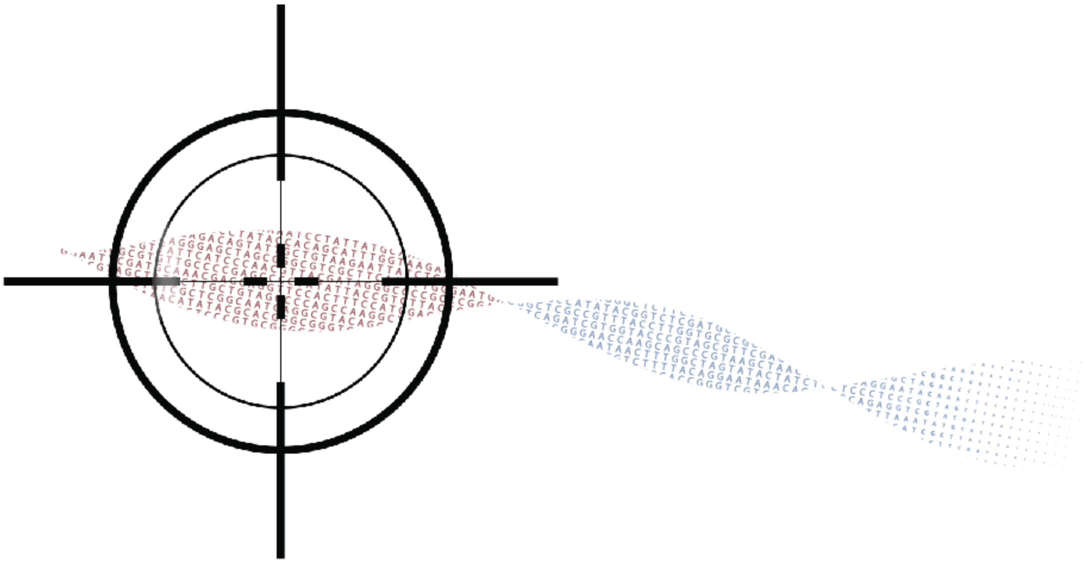
Statehood (1912 to present)

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.



Coming Soon xGen™ Lockdown™ Probes (BETA)

- Pool up to 2,000 individually synthesized probes into a single tube
- Develop your first custom panel within 10 days
- Optimize panels over time
- Mix and match panels for added flexibility
- Spike probes into exome sets to improve coverage performance

All probes are QC'd with ESI mass spectrometry and modified with 5' biotin. Pools ship within 7-10 business days. Probe tiling to customer specifications.

Available in three scales:

Product	Yield	Pool Container	Plate Container
Mini	2 pmole/probe	2 mL tube	NA
Standard	20 pmole/probe	4.5 mL tube	96 or 384 well plate
XL	200 pmole/probe	10 mL tube	96 or 384 well plate

xGen Lockdown Probes will launch as a beta product with limited support. Familiarity with target capture and in-solution hybridization will be required to use the product.

Sign up to be notified when the product is released
www.idtdna.com/pages/xgen

THE CUSTOM BIOLOGY COMPANY

WWW.IDTDNA.COM

IDT®
INTEGRATED DNA TECHNOLOGIES

“Sponsors”

Roche

http://www.roche-diagnostics.us/
Meet and Greet Party

http://www.lifetechnologies.com
Happy Hour x2

life technologies™

Caliper
a PerkinElmer company

http://www.caliperls.com/
Bags and Gifts

https://www.beckmancoulter.com/
Lunch

BECKMAN COULTER®

illumina®

http://www.illumina.com/
Lunch

“Sponsors”

http://www.neb.com
Fruit and Juice - Breakfast

NEW ENGLAND BioLabs Inc.

Agilent Technologies

http://www.home.agilent.com
Lunch

OpGen®

http://www.opgen.com/
Break

IDT®
INTEGRATED DNA TECHNOLOGIES

http://www.idtdna.com/
Meeting Guides

“Sponsors”

Roche

life technologies™

Caliper
a PerkinElmer company

BECKMAN COULTER®

illumina®

IDT®
INTEGRATED DNA TECHNOLOGIES

Thank You !!!