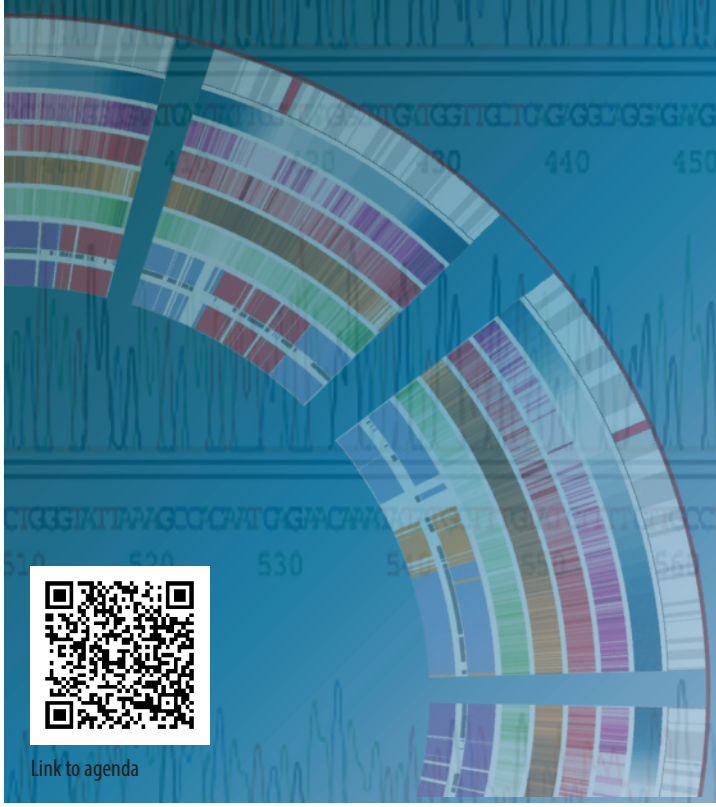
11th Annual Sequencing, Finishing, Analysis in the Future Meeting

Santa Fe, New Mexico
June 1-3, 2016



Link to agenda

Los Alamos
NATIONAL LABORATORY
EST. 1943

TABLE OF CONTENTS

Agenda Overview 3

June 1st Agenda 5

Speaker Presentations 7

Poster Presentations / Meet and Greet Party 35

June 2nd Agenda 99

Speaker Presentations 101

Happy Hour at Cowgirl with Map 127

June 3rd Agenda 129

Speaker Presentations 131

Attendees 153

Maps of Santa Fe, NM 159

2016 SFAF Sponsors 161

The 2016 "Sequencing, Finishing, and Analysis in the Future" Organizing Committee

* Chris Detter, Ph.D., Chief Science Advisor, MRIGlobal
* Johar Ali, Ph.D., Research Director, AA Ontario, Canada
* Patrick Chain, Bioinformatics/Metagenomics Team Leader, LANL
* Michael Fitzgerald, Microbial Special Projects Manager, Broad Institute
* Bob Fulton, M.S., Director of Project Development & Management, WashU
* Darren Grafham, Lab Manager, Children's Hospital, Sheffield, UK
* Alla Lapidus, Ph.D., Director, Centre for Algorithmic Biotechnology, SPbU; Russia
* Donna Muzny, M.Sc., Director of Operations, Baylor College of Medicine
* David Bruce, M.Sc. Project and Program Manager of Genomic Sciences, LANL
* Shannon Johnson, Ph.D., Project Manager, LANL

AGENDA OVERVIEW

WEDNESDAY, 1ST JUNE

07:30 - 08:30	Breakfast (Sponsored by New England Biolabs)
08:30 - 08:45	Welcome Introduction
08:45 - 09:30	Dr. Gilmore, Keynote Address (Sponsored by Advanced Analytic)
09:30 - 10:00	Dr. Earl, Invited Speaker
10:00 - 10:40	Oral Session 1 Pathogen Detection, Diagnostics and AMR
10:40 - 11:00	Coffee Break (Sponsored by LabCyte)
11:00 - 13:00	Oral Session 2 : Pathogen Detection, Diagnostics and AMR
12:40 - 14:00	Lunch Break (Sponsored by Promega)
14:00 - 15:40	Oral Session 3: Pathogen Detection, Diagnostics and Wetlab Applications
15:40 - 16:00	Coffee Break (Sponsored by Bioo Scientific)
16:00 - 16:40	Sequencing Vendor - Panel Discussion (Sponsored by Swift Bio)
16:40 - 18:25	Tech Time Talks 1
18:30 - 21:30	Meet and Greet Poster Session (Food and Drinks will be served, Sponsored by Roche)
	18:30 - 20:00 Poster Session 1a NM Room (1st floor)
	18:30 - 20:00 Poster Session 2a; Mezzanine (2nd Floor)
	20:00 - 21:30 Poster Session 1b; NM Room (1st floor)
	20:00 - 21:30 Poster Session 2; Mezzanine (2nd Floor)

THURSDAY, 2ND JUNE

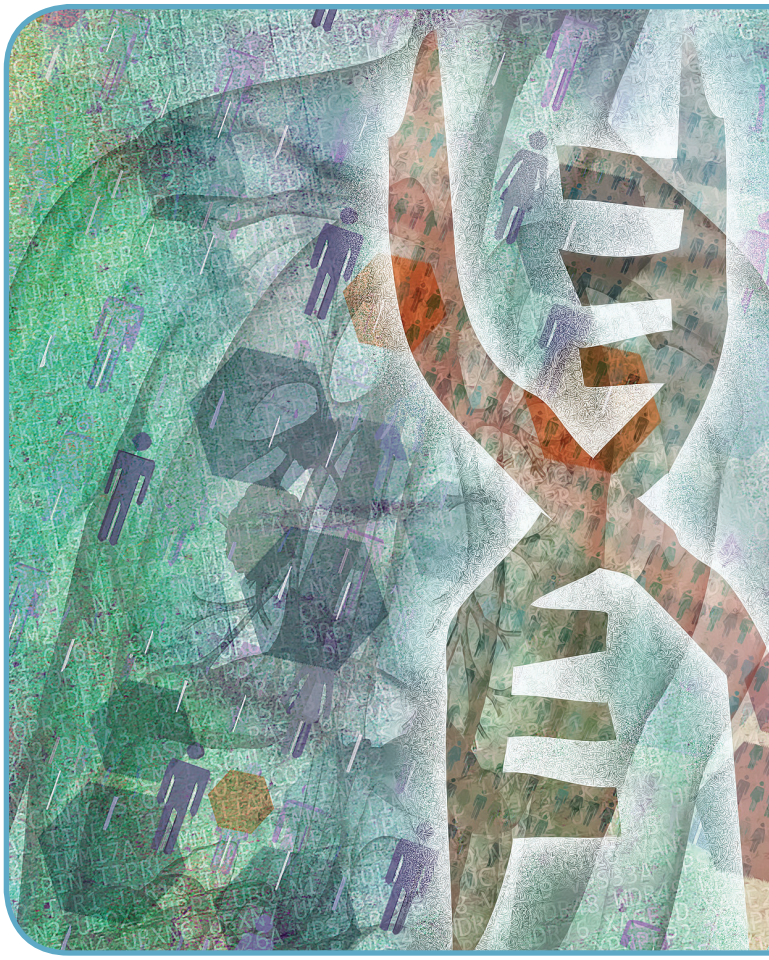
07:30 - 08:30	Breakfast (Sponsored by New England Biolabs)
08:30 - 08:45	Welcome Introduction and Opening Remarks
08:45 - 09:30	Dr. Pevzner, Keynote Address (Sponsored by Kapa Biosystems)
09:30 - 10:00	Dr. Borodovsky, Invited Speaker
10:00 - 10:20	Oral Session 4: Bioinformatics - Assembly and Analysis
10:20 - 10:40	Coffee Break (Sponsored by 10x Genomics)
10:40 - 12:00	Oral Session 5: Bioinformatics - Assembly and Analysis
12:00 - 13:20	Lunch Break (Sponsored by MRIGlobal)
13:20 - 15:30	Oral Session 6: Forensics (Human and Microbial)
15:30 - 15:50	Coffee Break (Sponsored by Becton Dickinson)
15:50 - 17:45	Tech Time Talks 2
18:30 - 20:30	Happy Hour(s) at Cowgirl Café, Sponsored by Illumina

FRIDAY, 3RD JUNE

07:30 - 08:30	Breakfast (Sponsored by New England Biolabs)
08:30 - 08:45	Opening Remarks
08:45 - 09:30	Dr. Petrosino, Keynote Address (Sponsored by Qiagen)
09:30 - 10:50	Oral Session 7: Metagenomics
10:50 - 11:10	Coffee Break (Sponsored by Dovetail)
11:10 - 11:40	Mr. Lakey, Invited Speaker
11:40 - 13:00	Oral Session 8: Plants and the Fruit Fly
13:00 - 14:00	Lunch Break (Sponsored by Dovetail)
14:00 - 15:40	Oral Session 9: Human Genomics
15:40 - 16:00	Coffee Break (Sponsored by Promega)
16:00 - 17:20	Oral Session 10: Human Clinical Dx and Analysis Pipelines
17:20 - 17:30	Closing Remarks

xGen® Exome Research Panel

- **Return consistent results** with probes manufactured to GMP standards
- **Achieve uniform exome coverage** using fewer sequencing reads
- **Generate reliable data** after a short hybridization



See for **yourself** at
www.idtdna.com/exome

IDT®
INTEGRATED DNA TECHNOLOGIES

custom oligos • qPCR • next generation sequencing • RNAi • genes & gene fragments • CRISPR genome editing

WEDNESDAY, 1ST JUNE

07:30 - 08:30 Breakfast (Sponsored by New England Biolabs)

08:30 - 08:45 Welcome Introduction

08:45 - 09:30 Dr. Gilmore, Keynote Address (Sponsored by Advanced Analytic)

09:30 - 10:00 Dr. Earl, Invited Speaker

10:00 - 10:40 Oral Session 1 Pathogen Detection, Diagnostics and AMR
Chaired by: Michael Fitzgerald and Bob Fulton

OS-1.01 :: Drug resistance diagnosis in Tuberculosis patients using targeted NGS
OS-1.02 :: Genomic sequencing and analysis of Neisseria gonorrhoeae clinical isolates to characterize antimicrobial resistance in Rio de Janeiro, Brazil

10:40 - 11:00 Coffee Break (Sponsored by LabCyte)

11:00 - 13:00 Oral Session 2 : Pathogen Detection, Diagnostics and AMR
Chaired by: Patrick Chain and Donna Muzny

OS-2.01 :: Human gut antimicrobial resistance: A comparison of microarray, targeted sequencing and deep metagenomics sequencing
OS-2.02 :: The NCBI Pathogen Detection Pipeline for Foodborne and Clinical Bacterial Pathogens
OS-2.03 :: Predictive Pathogen Biology: Genome-Based Prediction of Pathogenic Potential and Countermeasures Targets
OS-2.04 :: Genome wide characterization of Enterotoxigenic Escherichia coli serogroup O6 strains from multiple outbreaks and surveillance between 1975-2014
OS-2.05 :: Evaluating the potential of applying a metagenomics analysis for infectious disease clinical diagnostics and hotspot surveillance of vulnerable populations in challenging settings

12:40 - 14:00 Lunch Break (Sponsored by Promega)

14:00 - 15:40 Oral Session 3: Pathogen Detection, Diagnostics and Wetlab Applications
Chaired by: Kenny Yeh and Johar Ali

OS-3.01 :: SPIDR-WEB: an NGS biotechnology platform for diagnostic and transcriptomic applications
OS-3.02 :: Strand Specific RNA-sequencing using terminal breathing of RNA-cDNA duplexes for library synthesis
OS-3.03 :: The northward dissemination of a novel clade of West Nile Virus is associated with 2012 disease outbreak in North Texas
OS-3.04 :: Extensive Genetic Heterogeneity of Circulating Respiratory Syncytial Virus Strains Revealed by Targeted, Next Generation Whole Genome RNA Sequencing
OS-3.05 :: Implementation of sequencing platform in Gabon: Needs and Challenges for an efficient epidemiological surveillance strategy

15:40 - 16:00 Coffee Break (Sponsored by Bioo Scientific)

16:00 - 16:40 Sequencing Panel Discussion: M Appel (Roche), J Preston (Illumina) & S Turner (PacBio)
Chaired by: Bob Fulton and Patrick Chain

16:40 - 18:25 Tech Time Talks 1
Chaired by: Kenny Yeh and Alla Lapidus

TT-1.01 :: Purification Strategies, Quantitation, and QC Measurements for Predicting Downstream NGS Success with FFPE and Circulating Cell-Free DNA Plasma Samples
TT-1.02 :: Reducing input amounts and streamlining workflows in NGS library preparation
TT-1.03 :: Efficient Miniaturized Sequencing Library Preparation with Acoustic Liquid Handling
TT-1.04 :: Development of Amplicon Panels for Testing of Inherited Mendelian Diseases
TT-1.05 :: A Fast and Reliable Amplicon-based NGS Strategy for Screening of Germline and Somatic Mutations in BRCA1 and BRCA2 Genes
TT-1.06 :: High Performance, Streamlined Methods for RNA-Seq and Target Enrichment

18:30 - 21:30 Meet and Greet Poster Session (Food and Drinks will be served; Sponsored by Roche)

18:30 - 20:00 Poster Session 1a; NM Room (1st floor)
18:30 - 20:00 Poster Session 2a; Mezzanine (2nd Floor)

20:00 - 21:30 Poster Session 1b; NM Room (1st floor)
20:00 - 21:30 Poster Session 2b; Mezzanine (2nd Floor)

MULTIDRUG RESISTANT ENTEROCOCCI: EVOLUTION FROM PALEOZOIC BEGINNINGS TO LEADING HOSPITAL PATHOGEN IS WRITTEN IN THEIR GENOMES

Wednesday, 1st June 8:45 La Fonda Ballroom Keynote Address (KN-1)
Sponsored by Advanced Analytic

Dr. Michael Gilmore
Harvard School of Medicine

Enterococci are among the most widely distributed core components of gut flora in animals from invertebrates and insects to mammals. This led us to speculate that an ancestral Enterococcus colonized the last common ancestor, and was vertically disseminated as new host species evolved. Despite being numerically minor constituents of the gut microbiota, enterococci emerged among the vanguard of multidrug resistant hospital adapted pathogens. Interestingly this happened twice: in Enterococcus faecalis, and in the distantly related species E. faecium. This raises two questions: 1) What are the core properties of enterococci that make them nearly universal components of gut consortia of such a diverse range of animals? and 2) Why, among the great diversity of gut microbes, did enterococci repeatedly emerge to become leading causes of multidrug resistant hospital acquired infection? With antibiotic resistance now a leading global public health threat, there is a compelling need to understand the underlying biology and genetics that led to their hospital adaptation.

To determine the core traits of enterococci that both enable them to inhabit animals with diverse gut physiologies and diets, and predisposed them to adapt and proliferate in the modern hospital ecology, we selected 25 enterococcal species representing all major phylogenetic branches of the genus. We examined them in detail for phenotype, genotype, and where possible, correlated that with host association. We further compared these traits to those of both commensal and multidrug resistant strains of the most common human associated species, E. faecalis and E. faecium. We found that the enterococci acquired the ability to withstand episodic desiccation and starvation, among other stressors, and that speciation is largely driven by changing carbohydrates available in the gut of new hosts. Calibration of divergence indicates that enterococci arose commensurate with the terrestrialization of animals, and parallels their radiation, including gaps as occurred during the Permian Extinction. In adapting to cycles of deposition on land, the enterococci acquired traits that positioned them well for survival and adaptation to the modern hospital environment.

Speaker's biographical sketch
Michael S. Gilmore, PhD is currently the Sir William Osler Professor of Ophthalmology, and Microbiology and Immunobiology, Harvard Medical School. He serves on the steering committees of the Harvard Microbial Sciences Initiative, and the Infectious Disease Initiative of the Broad Institute of MIT and Harvard. As Principal Investigator of the Harvard-wide Program on Antibiotic Resistance, his research focuses on the evolution and development of multidrug resistant strains of enterococci, staphylococci, and streptococci, and the development of new therapeutic approaches. He is past chair of the NIH Bacterial Pathogenesis Study Section, the Gordon Conference on Microbial Adhesion and Signal Transduction, ASM Division D and the ARVO IM Section. He is founder and organizer of the international ASM Conference on Enterococci series, Editor in Chief of the public access book, Enterococci: From Commensals to Leading Causes of Drug Resistant Infection. Mike started his academic career in 1984 at the University of Oklahoma Health Sciences Center, where he rose through the ranks to Vice President for Research. He also held the MG McCool professorship and the George Lynn Cross chair. In 2004 he moved to Harvard Medical School to become the CL Schepens Professor of Ophthalmology, President and CEO of the Schepens Eye Research Institute, and Marie and DeWalt Ankeny Director of Research. In 2010, he moved his laboratories to their current location on the Massachusetts General Hospital campus of Harvard Medical School, in the Massachusetts Eye and Ear Infirmary. He continues to serve on numerous advisory boards and committees for public and private organizations, mainly focused on drug discovery, antibiotic resistance, and bacterial pathogenesis.

REWINDING THE CLOCK ON THE EVOLUTION OF DRUG RESISTANT TUBERCULOSIS

Wednesday, 1st June 9:30 La Fonda Ballroom Invited Speaker (IS-1)

Dr. Ashlee Earl
Broad Institute of MIT & Harvard

The spread of drug resistant TB (DRTB) is a major threat to global TB control, and the reason why universal drug susceptibility testing (DST) is a key component of the WHO "End TB" strategy. The able to rapidly detect complete drug resistance profiles in TB patients represents the most urgent need in TB treatment, but no scalable solution currently exists. All of the efforts to expand phenotypic methods for DST for case management and surveillance are hindered by cost, and by the need for complex laboratory infrastructure including TB culture facilities.

We have developed a rapid and sensitive amplicon next generation sequencing (NGS) solution for producing comprehensive genotypic resistant profiles for Mycobacterium tuberculosis (Mtb) populations direct from patient samples. In addition to improvements in patient care, an amplicon approach also has the potential to accelerate broadly implementable "culture-free" surveillance systems to quantify the worldwide distribution of TB drug resistance. In order for NGS to be successfully integrated into patient care and surveillance at a global scale we envision the need for a standardized and validated sequencing pipeline including an automated DNA extraction front-end, and a standardized, validated analysis, interpretation and archiving system on the back-end. The Relational Sequencing TB Data Platform (ReSeqTB) is a repository that is currently being developed for cataloging phenotypic, genotypic, and metadata associated with clinical DRTB isolates. Integrating our targeted sequencing approach into this repository, utilizing the vetted and standardized analysis pipeline, storage, and sharing platform of ReSeqTB, will help to develop and validate a simple, standardized NGS platform for the detection and interpretation of drug resistance in Mtb from sequences obtained directly from patient sputum.

Speaker's biographical sketch
Ashlee M. Earl is a research scientist and group leader for the Bacterial Genomics Group at the Broad Institute of MIT and Harvard. Within the Broad Institute's Genomic Center for Infectious Diseases, Earl is working to understand the relationship between microbes and human health including how multi-drug resistant pathogens emerge and spread.

Earl coordinated much of the Broad's research in the Human Microbiome Project and now leads a team of computational biologists to develop and utilize an array of 'omics analytical approaches to dissect bacterial and host contributions to several infectious diseases. She has organized and led dozens of local and international collaborations to bring genomic approaches to the study of tuberculosis, urinary tract infections, and hospital-acquired infections caused by the enterococci and carbapenem-resistant Enterobacteriaceae.

DRUG RESISTANCE DIAGNOSIS IN TUBERCULOSIS PATIENTS USING TARGETED NGS

Wednesday, 1st June 10:00 La Fonda Ballroom Talk (OS-1.01)

Rebecca Colman[1], Julia Anderson[2], Darrin Lemmer[3], Valeriu Crudu[4], David Dolinger[5], Claudia Denkinger[5], David Engelthaler[2], Timothy Rodwell[1]

[1]University of California, San Diego, [2]Translational Genomics Research Institute, [3]TGen North, [4]Phthisiopneumology Institute (PPI), [5]Foundation for Innovative New Diagnostics

The spread of drug resistant TB (DRTB) is a major threat to global TB control, and the reason why universal drug susceptibility testing (DST) is a key component of the WHO "End TB" strategy. The able to rapidly detect complete drug resistance profiles in TB patients represents the most urgent need in TB treatment, but no scalable solution currently exists. All of the efforts to expand phenotypic methods for DST for case management and surveillance are hindered by cost, and by the need for complex laboratory infrastructure including TB culture facilities.

We have developed a rapid and sensitive amplicon next generation sequencing (NGS) solution for producing comprehensive genotypic resistant profiles for Mycobacterium tuberculosis (Mtb) populations direct from patient samples. In addition to improvements in patient care, an amplicon approach also has the potential to accelerate broadly implementable "culture-free" surveillance systems to quantify the worldwide distribution of TB drug resistance. In order for NGS to be successfully integrated into patient care and surveillance at a global scale we envision the need for a standardized and validated sequencing pipeline including an automated DNA extraction front-end, and a standardized, validated analysis, interpretation and archiving system on the back-end. The Relational Sequencing TB Data Platform (ReSeqTB) is a repository that is currently being developed for cataloging phenotypic, genotypic, and metadata associated with clinical DRTB isolates. Integrating our targeted sequencing approach into this repository, utilizing the vetted and standardized analysis pipeline, storage, and sharing platform of ReSeqTB, will help to develop and validate a simple, standardized NGS platform for the detection and interpretation of drug resistance in Mtb from sequences obtained directly from patient sputum.

GENOMIC SEQUENCING AND ANALYSIS OF NEISSERIA GONORRHOEAE CLINICAL ISOLATES TO CHARACTERIZE ANTIMICROBIAL RESISTANCE IN RIO DE JANEIRO, BRAZIL

Wednesday, 1st June 10:20 La Fonda Ballroom Talk (OS-1.02)

A. Jeanine Abrams[1], Ana Paula Ramalho[2], David Trees[1]

[1]Centers for Disease Control and Prevention, [2]Universidade Federal do Rio de Janeiro

Neisseria gonorrhoeae, the etiological agent responsible for the sexually transmitted infection gonorrhea, is the second most common notifiable infection in the United States, and approximately 106 million new gonorrhea cases were globally estimated in 2008. Antimicrobial resistance to former first-line antibiotics (e.g., penicillins, tetracyclines, fluoroquinolones, and cephalosporins) in N. gonorrhoeae have been facilitated by both plasmid- and chromosome-mediated mechanis This acquired resistance has led to the current CDC treatment recommendation for uncomplicated gonorrhea, which outlines the dual-use of ceftriaxone with either azithromycin or doxycycline.

This study utilized genomic sequencing and analysis to characterize the rates and phylogenetic patterns associated with resistance to penicillin, tetracycline, ciprofloxacin, azithromycin, cefixime, and ceftriaxone in isolates from Rio de Janeiro, Brazil. We examined the genomes of 117 gonococcal isolates that were collected from public and private healthcare clinics between 2006 and 2015 in Rio de Janeiro. In addition to genomic data, we examined phenotypic data (antimicrobial susceptibility profiles) to further clarify the detected resistance patterns. The results indicated relatively low levels of reduced susceptibility to cefixime and azithromycin compared to the levels observed for the other antibiotics, and multi-drug resistance was detected in several samples. Moreover, two samples exhibited reduced susceptibility to all of the tested antibiotics, with the exception of ceftriaxone. These results will not only further our understanding of the evolution of decreased susceptibility to a variety of antibiotics in Rio de Janeiro, but will also potentially inform the development of local and global treatment guidelines.

COFFEE BREAK

Sponsored by LabCyte

10:40 – 11:00

HUMAN GUT ANTIMICROBIAL RESISTANCE: A COMPARISON OF MICROARRAY, TARGETED SEQUENCING AND DEEP METAGENOMICS SEQUENCING

Wednesday, 1st June 11:00 La Fonda Ballroom Talk (OS-2.01)

Tom Slezak[1], Tom Brettin[2], Dionysios Antonopoulos[2], Sarah Owens[2], Kenneth Frey[3], Shea Gardner[1], Jonathan Allen[1], Sam Minot[4], Nick Greenfield[4], Nisha Mulakken[5], Rongsu Qi[5], Chengya Liang[5], Gary Vora[3], Gary An[6]

[1]Lawrence Livermore National Laboratory, [2]Argonne National Laboratory, [3]Naval Medical Research Center, [4]One Codex, [5]Thermo Fisher, [6]University of Chicago

Patients in Intensive care units (ICUs) are particularly vulnerable to infections from antimicrobial resistant (AMR) organisms. These patients often receive multiple courses of broad-spectrum antibiotics and given the role of fecal auto-contamination in nosocomial infections we posit that characterizing the AMR determinants in their gut microbiomes can inform the timely construction of antibiotic regimens to limit the emergence of clinically significant AMR organisms. We report on an early-stage pilot program that evaluated the ability of 3 technologies (microarray, targeted sequencing, and deep metagenomics whole-genome shotgun (WGS) sequencing) to detect a key set of over 500 AMR genes in fecal microbiome samples obtained from three long-term ICU patients and three healthy volunteers. Microbiome structure and diversity was also characterized via 16S rRNA-based amplicon sequencing. We used the set of AMR genes detected by the Antimicrobial Resistance Determinant Microarray (ARDM) developed by the Naval Research Laboratory (NRL) and developed an equivalent AmpliSeq® targeted amplification panel (1,354 total amplicons) to test these samples. Deep WGS (approximately 200 million reads for each healthy subject and 60 million reads for each ICU patient) was performed as a comparison using the Illumina HiSeq and as a reference dataset for non-targeted sequences. The results demonstrated that the targeted sequencing consistently detected more AMR genes than WGS, in addition to being faster and less expensive (multiple targeted samples were run on a single Thermo Fisher Ion PGM or S5 run, compared to using a full HiSeq lane per samples in the WGS set). A cloud-based analysis and reporting package was prototyped and applied to both the targeted and WGS sequence data. We will discuss the sequencing and array results in detail and make a case for the clinical utility of targeted sequencing for known genes present in complex tissue/body fluid samples, both to impact the treatment of individual patients and reduce the emergence of epidemiological foci of AMR within the hospital.

THE NCBI PATHOGEN DETECTION PIPELINE FOR FOODBORNE AND CLINICAL BACTERIAL PATHOGENS

Wednesday, 1st June 11:20 La Fonda Ballroom Talk (OS-2.02)

William Klimke[1], Michael Feldgarden[1], Dan Haft[1], Arjun Prasad[1], Martin Shumway[1], Michael Dicuccio[1], Richa Agarwala[1]

[1]National Center for Biotechnology Information

The promise of using whole genome sequencing in public health laboratories for the analysis of bacterial pathogens has become a reality. In the United States the federal public health and regulatory agencies have formed an official collaboration along with NIH with the stated goal of coordinating the development of a network for the surveillance, detection and investigation of outbreaks and research into pathogens causing enteric illnesses transmitted by food and other routes. A successful pilot project for Listeria monocytogenes was started in 2013 where every Listeria collected in the United States, whether clinical or environmental, was sequenced and the data deposited at NCBI and made available to the public. The outcome of the pilot project were: 1) reductions in the case count per outbreak, 2) an increase in the number of clusters detected, 3) and more outbreaks solved. The federal agencies are now moving towards sequencing all four major bacterial pathogens responsible for foodborne illnesses: Campylobacter, STEC, Listeria, and Salmonella. The program currently exists with federal, state, industrial and academic partners with next generation sequencing machines producing genomic sequences in real time and submitting the data to NCBI. The NCBI Pathogen Detection system processes the raw data into a final analysis set for the public health labs. Once the data are deposited the system does an initial QC check on the data and then identifies the nearest genome in the reference set using a k-mer approach. Once a reference is identified a combined assembly approach is taken with both reference-guided and de novo assemblers to produce a final combined assembly. The assemblies are then placed into a phylogenetic context, first every assembly is compared to all others for each organism group using pairwise k-mer distances and FastME, and then a more fine-grained phylogeny is produced using SNPs. The end result is a single k-mer tree of all isolates for each organism group, and then many sub clusters of closely related isolates based on SNP distances. Clusters are reported daily whenever new data arrives and made available to the public health agencies. In addition, the assembled genomes are annotated using NCBI's PGAP 3.0 annotation system, and antimicrobial genes/proteins are listed using a newly built module within that pipeline based on a reference set of antimicrobial resistant proteins and HM The latter is part of the National Strategy for Combatting Antibiotic Resistant Bacteria (CARB) project that was initiated by the President and NCBI is putting together the National Database of Antibiotic Re- sistant Organisms. These two new parts of the analysis pipeline, 1) SNP clustering, 2) antimicrobial resistance gene/protein annotation will be discussed.

PREDICTIVE PATHOGEN BIOLOGY: GENOME-BASED PREDICTION OF PATHOGENIC POTENTIAL AND COUNTERMEASURES TARGETS

Wednesday, 1st June 11:40 La Fonda Ballroom Talk (OS-2.03)

Debjit Ray
Sandia National Labs

Horizontal gene transfer (HGT) and recombination leads to the emergence of bacterial antibiotic resistance and pathogenic traits. Genetic changes range from acquisition of a large plasmid to insertion of transposon into a regulatory gene. HGT events can be identified by comparing a large number of fully sequenced genomes across a species or genus, define the phylogenetic range of HGT, and find potential sources of new resistance genes. In-depth comparative phylogenomics can also identify subtle genome or plasmid structural changes or mutations associated with phenotypic changes. Comparative phylogenomics requires that accurately sequenced, complete and properly annotated genomes of the organism. Due to dramatic advances in "short read" sequencing technology, the raw sequence coverage needed for sequencing a bacterial genome now can be obtained in a couple of days for a few dollars sequencing costs, starting with only a few nanograms of genomic DNA. Assembling closed genomes requires additional mate-pair reads or "long read" sequencing data to accompany short-read paired-end data. To bring down the cost and time required of producing assembled genomes and annotating genome features that inform drug resistance and pathogenicity, we are analyzing the performance for genome assembly of data from the Illumina NextSeq, which has faster throughput than the Illumina HiSeq (~1-2 days versus ~1 week), and shorter reads (150bp paired-end versus 300bp paired end) but higher capacity (150-400M reads per run versus ~5-15M) compared to the Illumina MiSeq. Bioinformatics improvements are also needed to make rapid, routine production of complete genomes a reality. Modern assemblers such as SPAdes 3.6.0 running on a standard Linux blade are capable in a few hours of converting mixes of reads from different library preps into high-quality assemblies with only a few gaps. Remaining breaks in scaffolds are generally due to repeats (e.g., rRNA genes) are addressed by our software for gap closure techniques, that avoid custom PCR or targeted sequencing.

Our goal is to improve the understanding of emergence of pathogenesis using sequencing, comparative genomics, and machine learning analysis of ~1000 pathogen genomes. Machine learning algorithms will be used to digest the diverse features (change in virulence genes, recombination, horizontal gene transfer, patient diagnostics). Temporal data and evolutionary models can thus determine whether the origin of a particular isolate is likely to have been from the environment (could it have evolved from previous isolates). It can be useful for comparing differences in virulence along or across the tree. More intriguing, it can test whether there is a direction to virulence strength. This would open new avenues in the prediction of uncharacterized clinical bugs and multidrug resistance evolution and pathogen emergence.

GENOME WIDE CHARACTERIZATION OF ENTEROTOXIGENIC ESCHERICHIA COLI SEROGROUP O6 STRAINS FROM MULTIPLE OUTBREAKS AND SURVEILLANCE BETWEEN 1975-2014

Wednesday, 1st June 12:00 La Fonda Ballroom Talk (OS-2.04)

Vaishnavi Pattabiraman, Lee Katz, Ashley Sabol, Eija Trees

Centers for Disease Control and Prevention

Enterotoxigenic Escherichia coli (ETEC) are an important cause of diarrhea in children under the age of five in developing countries and are the leading bacterial agent of traveler's diarrhea in persons traveling to these countries. ETEC strains secrete heat-labile (LT) and/or heat-stable (ST) enterotoxins that induce diarrhea by causing water and electrolyte imbalance. In this study, we have characterized 17 ETEC serogroup O6 isolates from multiple outbreaks from 1975 to 2014 by whole genome sequencing (WGS) and present the findings on virulence factors, plasmids and high-quality single nucleotide polymorphisms (hqSNPs) analysis. Total genomic DNA was extracted using the DNeasy blood and tissue kit (Qiagen). DNA libraries were prepared using the Nextera XT kit (Illumina Inc.) and sequencing was performed on the MiSeq using the 2 X 150-bp chemistry. The raw reads were trimmed, quality control was performed using PRINSEQ and assembled with SPAdes, and contigs were uploaded into plasmid finder and virulence finder tools on the Center for Genomic Epidemiology website (https://cge.cbs.dtu.dk). Whole genome hqSNP analysis was performed with Lyve-SET version 1.1.4e (http://github.com/lskatz/Lyve-SET) on the raw reads which were cleaned with Computational Genomics Pipeline (CGP). SNPs were called with Varscan, and Lyve-SET was run with the following options: 20x minimum coverage, 95 % read support, and clustered SNPS in less than 5 base pairs were filtered. The draft genome of ETEC O6 strain 2011EL- 1370-2 (USA, 2011) was used as the reference strain. All 17 ETEC genomes carried one or both of the multidrug resistance plasmids pRSB107 and pHN7A8 of IncFII group that infer resistance to penicillins, cephalosporins and tetracycline among others. In addition, 7 genomes carried an IncI1 group plasmid eliciting multidrug resistance; F5524 carried multidrug resistance plasmids of IncI2 and IncP groups. pCoo plasmid of IncFII group which encodes the prototype ETEC colonization factor was detected in 13/17 genomes. Heat-labile enterotoxin A subunit (ltcA) and heat-stable enterotoxin 1b (st1b) were found in all 17 genomes. Longus type IV pilus or CS-21 (lngA) were found in 12/17 genomes. Results from the hqSNP analysis indicated overall high diversity among the serogroup O6 strains. Within clonal outbreaks, isolates differed from each other by less than 10 hqSNPs. However, two of the outbreaks appeared to be polyclonal. Three isolates that were presumed to belong to different, though temporarily associated, outbreaks were highly related with each other indicating a possible common source for the outbreaks or an occurrence of a common sequence type. In summary, the ETEC strains characterized in this study carried key virulence factors and multidrug resistance plasmids that contributed to their pathogenicity and marginal benefits from antimicrobial drug therapy. Whole genome hqSNP analysis appears to be a useful tool for outbreak cluster detection and source tracking of ETEC though polyclonal outbreaks will always prove challenging. Further evaluation of hqSNP analysis and other WGS-based methods such as whole genome multi-locus sequence typing (wgMLST) is needed before WGS can be implemented for routine surveillance of ETEC.

EVALUATING THE POTENTIAL OF APPLYING A METAGENOMICS ANALYSIS FOR INFECTIOUS DISEASE CLINICAL DIAGNOSTICS AND HOTSPOT SURVEILLANCE OF VULNERABLE POPULATIONS IN CHALLENGING SETTINGS.

Wednesday, 1st June 12:20 La Fonda Ballroom Talk (OS-2.05)

Oksana Lukjancenko[1], Katrine Joensen[2], Rene Hendriksen[1], Patrick Munk[1], Frank Aarestrup[1]

[1]Technical University of Denmark, National Food Institute, [2]Statens Serum Institute

World population faces the frequent appearance and rapid spread of infectious diseases. Diarrhea is a major global disease burden estimated to encompass 1.7 billion cases each year. Diarrhea is typically a symptom of a gastrointestinal infection, but may also be a symptom of several medical conditions or a result of drug treatment, e.g. antibiotic-associated diarrhea. Diarrhea of different infectious origin (bacterial, viral or parasitic) cannot be distinguished based on history or clinical observations and thus rapid analyses are important, since specific treatment as well as patient care depends on the pathogen. In addition, rapid and accurate diagnostics, characterization and comparison of pathogens are essential to identify both nosocomial and food borne outbreaks.

The current diagnostics involve numerous procedures and typically only result in identification of a minority of diarrhea-causing microbial agents, while often the pathogen is not identified in time to guide clinical management, and in many clinical cases is never identified. With next-generation sequencing (NGS) becoming cheaper it has huge potential in routine diagnostics and infection surveillance. NGS has already been used in clinical settings for elucidating bacterial outbreaks, and it has been proposed for real-time typing and surveillance of pathogens. NGS technology can be applied directly to clinical or sewage samples, potentially advancing diagnostics and leading to even more rapid diagnostic results.

Here, we evaluated the potential of NGS-based diagnostics through 1) direct sequencing of fecal samples from patients with diarrhea; and 2) sequencing of sewage samples, taken in the infor- mal settlement, Kibera in Nairobi, Kenya. Species distribution was determined with MGmapper (http://cge.cbs.dtu.dk/services/MGmapper/) and NGS-based diagnostic prediction was performed based on relative abundance of pathogenic bacteria, viruses, and parasites; and detection of bacterial pathogen-specific virulence genes. NGS-based diagnostic results were compared to conventional findings for clinical diarrheal samples; and epidemiological data for sewage samples.

The NGS-based approach enabled pathogen detection comparable to the conventional diagnostics, and the approach has potential to be extended for detection of all pathogens. Pathogen prediction in sewage samples showed correspondence to the epidemiological data.

LUNCH

Sponsored by Promega

12:40 – 14:00

SPIDR-WEB: AN NGS BIOTECHNOLOGY PLATFORM FOR DIAGNOSTIC AND TRANSCRIPTOMIC APPLICATIONS

Wednesday, 1st June 14:00 La Fonda Ballroom Talk (OS-3.01)

Momochilo Vuyisich
Los Alamos National Laboratory

We are transforming the field of infectious disease diagnostics with the development of the Sample Prep for Infectious Disease Recognition With EDGE Bioinformatics (SPIDR-WEB). SPIDR-WEB is a sample-to-result biotechnology platform that enables efficient use of next generation sequencing (NGS) for pathogen detection in clinical samples.

NGS has become a powerful tool for detection and characterization of both known and emerging pathogens. The main advantage of NGS is its non-biased approach that identifies all organisms in a sample. This is in contrast to traditional molecular assays that force us to look for a set of specific pathogens. In most clinical samples, the relative abundance of pathogen nucleic acids (DNA or RNA) is vanishingly small. Therefore, vast amounts of sequence data must be generated and analyzed to identify rare pathogen sequences. SPIDR-WEB is a sample-to-result process that relies on efficient laboratory and in silico steps.

Clinical samples mostly comprise non-informative host RNAs or abundant housekeeping gene transcripts. SPIDR-WEB incorporates removal of non-informative RNAs (RNR), thereby enriching all other RNAs, including those from pathogens. This step enables either higher sensitivity and specificity, or less expensive and faster sequencing. Our custom EDGE bioinformatics data analysis platform provides rapid read classification at all taxonomic levels, and reliably detects all organisms present in a sample. EDGE is an efficient process, as it uses databases with pre-computed signatures, instead of aligning sequencing reads to the entire Genbank. In addition to RNR and EDGE, SPIDR- WEB includes robust, inexpensive and rapid sample lysis, RNA extraction, and library preparation steps.

We want to implement SPIDR-WEB in both research and clinical settings to support a multitude of applications, such as discovery of novel mechanisms and biomarkers, study host-pathogen interactions, improve vaccines and therapeutics, and complement current diagnostic tools and help improve their utility.

We will describe SPIDR-WEB technology and show clinically-relevant results obtained from human blood, stool, respiratory, cerebrospinal fluid, urine, and other sample types.

STRAND SPECIFIC RNA-SEQUENCING USING TERMINAL BREATHING OF RNA-CDNA DUPLEXES FOR LIBRARY SYNTHESIS

Wednesday, 1st June 14:20 La Fonda Ballroom Talk (OS-3.02)

Brad Townsley[1], Mike Covington[1], Yasunori Ichihashi[2], Kristina Zumstein[3], Neelima Sinha[3]

[1]UC Davis/Amaryllis Nucleics, [2]RIKEN Center for Sustainable Resource Science, [3]UC Davis

Next Generation Sequencing (NGS) is driving rapid advancement in biological understanding and RNA-sequencing (RNA-seq) has become an indispensable tool for biology and medicine. There is a growing need for access to these technologies although preparation of NGS libraries remains a bottleneck to wider adoption. Here we report a novel method for the production of strand specific RNA-seq libraries utilizing the terminal breathing of double-stranded cDNA to capture and incorporate a sequencing adapter. Breath Adapter Directional sequencing (BrAD-seq) reduces sample handling and requires far fewer enzymatic steps than most available methods to produce high quality strand-specific RNA-seq libraries. The method we present is optimized for 3-prime Digital Gene Expression libraries and can easily extend to full transcript coverage shotgun type strand-specific libraries and is modularized to accommodate a diversity of RNA and DNA input materials. BrAD-seq offers a highly streamlined and inexpensive option for RNA-seq libraries.

THE NORTHWARD DISSEMINATION OF A NOVEL CLADE OF WEST NILE VIRUS IS ASSOCIATED WITH 2012 DISEASE OUTBREAK IN NORTH TEXAS

Wednesday, 1st June 14:40 La Fonda Ballroom Talk (OS-3.03)

Chukwuemika Aroh[1], Mary D'Anton[2], Beth Levine[1], Nan Yan[1], Edward Wakeland[1]

[1]UT Southwestern Medical Center, [2]Texas Department of State Health Services

West Nile Virus (WNV) is the leading cause of mosquito-borne encephalitis in the United States and thus is of significant public health concern. Since its arrival in the northeast region of United States in 1999, WNV has disseminated to the rest of the country and has led to frequent epidemics in birds, horses, and humans. For a long time it has been known that mid-central United States, from Texas to North Dakota, has the highest incidence of human West Nile disease, but the cause remains unclear. In this study we sequenced novel WNV isolates from environmental samples in Texas from 2012 to 2015 and analyzed these along with endemic WNV isolates in TX and other parts of the United States. Our results suggest that two lineages of WNV arrived in TX from northeastern United States between 2008 and 2012. These two lineages were associated with the most recent WNV epidemic in 2012 and 2013. Of these lineages, only the lineage characterized by a mutation in the NS2a region of the WNV genome spread into northern states in 2012, suggesting that this lineage may have readily infected migratory birds. Additionally, in 2013 this northward dissemination was again observed as 2012 strains from Houston-gulf area replaced mutant NS2a bearing strains in North TX area. These observations suggest a general northward dissemination of WNV strains in mid-central United States that may lead to the rapid dissemination of pathogenic viruses during an epidemic.

EXTENSIVE GENETIC HETEROGENEITY OF CIRCULATING RESPIRATORY SYNCYTIAL VIRUS STRAINS REVEALED BY TARGETED, NEXT GENERATION WHOLE GENOME RNA SEQUENCING

Wednesday, 1st June 15:00 La Fonda Ballroom Talk (OS-3.04)

Darrell Dinwiddie[1], Kurt Schwalm[1], Walter Dehority[1], Joshua Kennedy[2], Stephen Gross[3], Gary Schroth[3], Stephen Young[4]

[1]University of New Mexico Health Sciences Center, [2]University of Arkansas for Medical Sciences Arkansas Children's Hospital, [3]Illumina Inc, [4]TriCore Reference Laboratories

Background: Respiratory syncytial virus (RSV) is a major cause of childhood morbidity. Significant annual fluctuations in incidence and severity occur, yet the specific genetic variation that influences transmission, virulence, and pathogenesis for RSV is poorly understood.

Methods: We developed a hybridization-based method to target and enrich complete coding sequences of respiratory syncytial virus from clinical samples. The enriched samples undergo deep sequencing in a high-throughput, multiplexed, rapid manner on the Illumina MiSeq. A custom bioinformatic pipeline is used to determine specific viruses, construct nearly complete genome sequences, assess viral gene expression, detect genetic variation and conduct phylogenetic analysis.

Results: We have completed deep, next generation sequencing of 102 RSV positive samples across four infection seasons. Of the samples sequenced, we have generated complete or nearly complete genomes from 69 of 102 (67.6%) samples and have covered >50% of the genome from 29 of 102 samples (28.4%). Together, 96% of samples sequenced have >50% of their genomes sequenced. Alignment of sequencing reads to their appropriate prototype RSV A (NC_001803) and RSV B (AY353550) reference sequences produced an average of 410 variants (range 257-474) and 466 variants (range 418-510), respectively. The average number of non-synonymous variants was 79 for both RSV A and RSV B samples. Strain and phylogenetic analysis reveal the presence of at least 6 distinct strains of RSV co-circulating during the same infection season.

Conclusions: Taken together, our data reveal that current clinical RSV isolates differ from signifi- cantly for their reference sequences and suggest that genetic diversity of co-circulating RSV strains during the same infection season may be underappreciated. Evaluation of RSV by targeted, deep, next generation RNA sequencing provides important information about clinical viral isolates cur- rently not detected by clinical testing that may reveal genetic factors that impact clinical severity of illness and inform clinical management.

IMPLEMENTATION OF SEQUENCING PLATFORM IN GABON: NEEDS AND CHALLENGES FOR AN EFFICIENT EPIDEMIOLOGICAL SURVEILLANCE STRATEGY

Wednesday, 1st June 15:20 La Fonda Ballroom Talk (OS-3.05)

Nicolas Berthet[1], Ingrid Labouba[1], Andy Nkili Meyong[1], Patrick Chain[2], Momchilo Vuyisich[2], Tracy Erkkila[2], Eric Leroy[1]

[1]Centre international de Recherches médicales de Franceville, [2]Los Alamos National Laboratory

With the emergences of different viral pathogens in Africa such as Zika virus in Gabon in 2007 – 2010 (Grard et al:, PLoS Negl Trop Dis. 2014), Bas-congo rhabdovirus in Central Africa in 2012 (Grard et al:, PLoS Pathog. 2012), or more recently Ebolavirus in West Africa in 2014, the pathogen discovery field should become a priority in biology research for concerned countries as well as for entire African continent. There, providing required tools for detection, identification and characterization of potential emergent infectious pathogens becomes primordial to diagnose first, then adequately manage outbreaks and finally implement appropriate epidemiological surveillance.

The Centre International de Recherches Médicales de Franceville (CIRMF) located in south of Gabon focuses its research activities on human and animal infectious diseases. The main CIRMF's objectives are the epidemiological surveillance of human and/or zoonotic viral outbreaks mainly in Central Africa. This naturally includes not only diagnosis activities for adequate follow-up of population and neighbourhood; but also the identification of involved viral pathogens. CIRMF's equipment for specific and generic analyses based on PCR technologies has made it an efficient infectious diagnosis centre for the sub-region. Moreover, thanks to its installation of biosafety level 2, 3 and 4 the CIRMF hosts all required structures to support the treatment of high risk biological samples as well as different steps of in vitro pathogen isolation process. Finally the CIRMF was progressively equipped to answer to these public health needs.

Basically the CIRMF was locally able to detect and identify viral pathogens. However, further characterization by sequencing was only possible via subcontracting with European laboratories. This represented a considerable limit in terms of delay especially during outbreaks when time is so precious. In absolute, to locally treat samples from reception on site to sequencing data analyses would constitute an important gain of time and would improve epidemiological surveillance of all Central African sub-region.

Since November 2013, thanks to DTRA financial and LANL technical supports the CIRMF worked for installing and optimizing a sequencing platform to complete its research ability in pathogen discovery field. We suggest here to present the challenges met by CIRMF et al. to implement this platform from the purchase of the sequencer to date while it is workable even if further improvement are still required.

COFFEE BREAK

Sponsored by Bioo Scientific

BIOO SCIENTIFIC
MAXIMIZE SCIENCE FOR LIFE®
BIOO LIFE SCIENCE PRODUCTS

15:40 – 16:00

SEQUENCING PANEL ROUND TABLE

Sponsored by Swift Bio

16:00 – 17:00

Speaker 1: Maryke Appel (Roche Diagnostics Corporation)

The Roche Sequencing Story: Driving toward personalized medicine

Routine sequencing has the potential to revolutionize healthcare, especially in the areas of Oncology, Infectious Diseases and Prenatal Testing, and help realize the potential of personalized medicine. Roche Sequencing's long-term goals are to become the partner of choice for next-generation sequencing (NGS), and make NGS a routine clinical practice. To achieve this, we are working toward solutions that are truly transformative. Our core strategy comprises a phased approach to building simplified, end-to-end (sample in, result out) workflows. To this end, Roche has acquired best-in-class technologies across the whole workflow—from Sample Preparation through Sequencing, Analysis and Reporting.

In this presentation, we will provide a high-level overview of our vision, technology pipeline and focus on cell-free DNA as a key clinical sample type.

Speaker 2: Jeremy Preston (illumina)

Illumina sequencing technology updates and application in the microbial world

Illumina has been at the forefront of the genomics revolution with the development of a suite of sequencing technologies ranging from ultra-high throughput to low cost, bench top machines. This presentation will introduce Illumina's portfolio of sequencing systems with a focus on our new, low cost, easy to use bench top solutions and will summarize key findings in seminal publications using Illumina technology.

Speaker 3: Steve Turner (PacBio)

PacBio sequencing technology update and applications

PURIFICATION STRATEGIES, QUANTITATION, AND QC MEASUREMENTS FOR PREDICTING DOWNSTREAM NGS SUCCESS WITH FFPE AND CIRCULATING CELL-FREE DNA PLASMA SAMPLES

Wednesday, 1st June 17:00 La Fonda Ballroom Tech Talk (TT-1.01)

Doug Wieczorek, Spencer Herman, Curtis Knox, Jennifer Mook, Douglas Horejsh, Eric Vincent, Douglas Storts, Trista Schagat

Promega Corporation

Formalin fixed, paraffin embedded (FFPE) tumor tissue samples have long been an important source of genetic material for mutational analysis. However, the quality of DNA from FFPE samples is often highly variable, and the resulting degradation and crosslinking due to the fixation process can lead to issues with amplifiability and difficulty in NGS analysis. An alternative to FFPE is obtaining circulating cell-free DNA (ccfDNA) from plasma or other biological fluids. Collecting ccfDNA samples from plasma is non-invasive. Samples can be collected quickly and frequently and allows for the ability to monitor the presence or absence of mutations or DNA species over time. The drawbacks are that yields of ccfDNA are often very low and cell free DNA representing any specific cell population is typically present at low frequencies.

We have developed novel nucleic acid purification chemistries that improve upon current manual and automated methods for the purification of DNA from FFPE and plasma and demonstrate their use in NGS applications. DNA was purified from multiple FFPE tumor tissue types and matching plasma as well as normal FFPE tissue samples using multiple methods. DNA quantity and quality was measured by two separate strategies to study degradation levels of the nucleic acid obtained. Libraries were constructed using a commercially available 56 gene oncology panel for targeted NGS and sequencing quality was evaluated. The overall quality of the sequencing data correlates with measured quality metrics derived from a prototype DNA QC assay currently in development.

REDUCING INPUT AMOUNTS AND STREAMLINING WORKFLOWS IN NGS LIBRARY PREPARATION

Wednesday, 1st June 17:15 La Fonda Ballroom Tech Talk (TT-1.02)

Daniela Munafo

NewEngland Biolabs, Inc

The boundaries of NGS library construction are continually expanding, requiring higher performance with ever-decreasing input amounts, and often with sub-optimal quality of DNA and RNA. In order to achieve high quality libraries even with challenging samples, it is important to have high enzymatic efficiencies in all steps of library construction. It is also critical to achieve uniform genome or transcriptome coverage, regardless of GC content, input amount or nucleic acid quality. We continue to develop streamlined, automatable workflows in order to increase throughput, reduce opportunities for errors and to minimize sample loss. These new protocols include a novel enzymatic DNA fragmentation reagent and employ reformulated reagents at each library construction step, for high performance with even picogram (DNA) and low nanogram (RNA) input amounts, for a wide range of applications.

EFFICIENT MINIATURIZED SEQUENCING LIBRARY PREPARATION WITH ACOUSTIC LIQUID HANDLING

Wednesday, 1st June 17:30 La Fonda Ballroom Tech Talk (TT-1.03)

Austin Swafford

Labcyte

The Labcyte® Echo® Acoustic Liquid Handler revolutionizes liquid transfer by using acoustic energy for the non-contact transfer of fluids. Labcyte Echo systems automatically adjust for differences in surface tension and viscosity to achieve high precision and accuracy regardless for every transfer.

Increased efficiency and cost savings are realized from contamination-free miniaturized reactions (up to 100-fold for NGS), highly reproducible data, and the elimination of tip costs and washing fluids.

Upstream library preparation is no longer a significant bottleneck with the efficient utilization of sequencing capacity that is realized by nanoliter-scale normalization and pooling processes. The Labcyte AccessT Workstation seamlessly integrates the Echo with a full line of peripheral devices to automate complete multiplexed or partial workflow processes.

DEVELOPMENT OF AMPLICON PANELS FOR TESTING OF INHERITED MENDELIAN DISEASES

Wednesday, 1st June 17:45 La Fonda Ballroom Tech Talk (TT-1.04)

April Lewis, Radmila Hrdlickova, Jiri Nehyba, Carrie Firmani,
Dylan Fox, Colby Clear, Masoud Toloue

Bioo Scientific

Single gene disorders follow a pattern of Mendelian inheritance where a single mutation in one gene will cause disease. Objectives for screening of Mendelian diseases range from confirmation of clinical diagnosis and determination of appropriate therapies, to assessment of disease reoccurrence risk, carrier testing for at-risk family members, and prenatal diagnostics. Disease studies of this nature can be aided through the use of next generation sequencing (NGS). Custom amplicon-based assays for targeted NGS offer researchers an efficient solution for variant discovery, identification and characterization on genes related to particular diseases.

Bioo Scientific offers amplicon design and panel development as a custom service with the use of NEXTflex™ Amplicon Studio, proprietary primer design software. Compared to several other soft- ware programs, NEXTflex Amplicon Studio has well-defined primer selection criteria, and variable measurements are more accurate and precise. The software is more flexible in its ability to cater to different experimental requirements, allowing for the ability to offer custom services. Only the most optimally designed primer mates, as quantified through a novel scoring system, were introduced into these panels.

Bioo Scientific has also developed an innovative library preparation protocol for these assays, which embodies several new technical elements increasing efficiency and decreasing biases.

The proprietary software and innovative library preparation protocol were used in the production of 21 diagnostic gene panels for testing of predisposition to recurrent fever syndromes, cystic fibrosis, diabetes, obesity, breast and colon cancers, and a variety of other genetic disorders.

The panels have 100% target region coverage. Target design covers all coding sequences of canonical isoforms, as well as coding regions of a majority of alternative isoforms of 58 genes. Additionally, targets cover 5-50 bp (at least 25 bp in 56 genes) of flanking intronic sequences, several deep intronic regions, and promoter regions where known pathogenic mutations were reported. In total, gene targets make up 190.3 kb.

Primers were specifically designed for use with DNA from fresh or frozen samples to limit the number of amplicons, and therefore increase coverage uniformity and multiplexing capabilities of all panels. Consequently, amplicon inserts range from 60-251 bp. Total amplicon size was kept under 280 bp to allow for compatibility with the Illumina MiniSeq instrument. However, these panels work with all other Illumina and Ion Torrent platform Bioo Scientific provides complete kits with all reagents required for library preparation, including size selection beads and 384 Illumina-compatible barcodes if desired. The amplicon panels follow a rapid and user-friendly workflow, were validated by sequencing, and are optimized to deliver high coverage uniformity and target specificity.

Development of severe diseases and pathological conditions are often associated with inherited mutations in a single gene. Therefore, screening for these mutations is vital in determining treatments, assessing familial pedigrees for risk, and potentially aiding in preventative care. With distinct attention to primer design, target definition, and simplicity of library preparation, Bioo Scientific's amplicon panels offer a cost-effective and complete solution.

A FAST AND RELIABLE AMPLICON-BASED NGS STRATEGY FOR SCREENING OF GERMLINE AND SOMATIC MUTATIONS IN BRCA1 AND BRCA2 GENES

Wednesday, 1st June 17:55 La Fonda Ballroom Tech Talk (TT-1.05)

Jiri Nehyba, Radmila Hrdlickova, Josh Kinman, Carrie Firmani,
Dawn Obermoeller, April Lewis, Masoud Toloue

Bioo Scientific

Germline mutations in BRCA1 and BRCA2 genes are linked to hereditary breast and ovarian cancer. Individuals in affected families have heightened risk for breast and ovarian cancer and significant risk for prostate and pancreatic cancer. Somatic mutations in BRCA1 and BRCA2 also contribute to cancer development. Next generation sequencing (NGS) provides a convenient and simplified alternative to traditional Sanger sequencing to detect mutations in targeted BRCA genes. Sev- eral companies now offer BRCA NGS screening kits; however, there remains significant space for improvements in speed and cost efficiency.

Bioo Scientific has developed five high performance amplicon assays for detection of BRCA1 and BRCA2 mutations. These assays differ in the type of samples for which the panel was optimized (either fresh/frozen or DNA isolated from FFPE samples), sequencing platform compatibility (Illumina or Ion Torrent), primer design, and in PCR-based strategy for enrichment of the BRCA targets. Performance of these assays was evaluated for uniformity and depth of coverage, and reliability of mutation calling. Bioo Scientific's assays were then directly compared to competitor assays using fresh/frozen DNA. FFPE samples of varying levels of degradation were tested using the NEXTflex™ BRCA1 and BRCA2 Amplicon Panels.

All five Bioo Scientific assays for the detection of BRCA1 and BRCA2 germline and somatic mutations offer 100% target coverage, totaling 16.4 kb of the genome including all coding exons of both genes. Pathological mutations were correctly detected with these kits in six validated DNA samples from breast cancer patients (obtained from the HapMap project). In comparison with competitor kits, NEXTflex BRCA1 and BRCA2 Amplicon Panels can handle lower amounts of starting material with a shorter protocol for successful library preparation. Using two primer pools for fresh/frozen DNA samples enables the production of quality results with as little as 20 ng of DNA in only 4 hours. Library preparation for FFPE samples uses four pools, requires a minimum of 40 ng of input, and was tested to work with varying qualities of FFPE DNA preparations. Importantly, the NEXTflex BRCA1 and BRCA2 Amplicon Panels surpass other kits with regard to coverage uniformity and specificity, by reducing the amount of off-target reads and by ensuring equal amplicon performance. Bioo Scientific's assays provide >99% uniformity of amplicon reads for FFPE samples (100% for fresh/frozen DNA samples) and attain >97% specificity, defined as reads mapping within the desired target (>97.5% for fresh/frozen DNA samples). High uniformity and low off- target reads decrease sequencing costs, enabling a high level of multiplexing, especially when paired with NEXTflex™ 384 Illumina-compatible barcodes.

Because inherited and acquired mutations in BRCA1 and BRCA2 genes increase the risk of breast, ovarian, and several other types of cancer, the need to quickly and accurately detect these mutations is critical in research, preclinical, and clinical settings. The NEXTflex BRCA1 and BRCA2 Amplicon Panels are complete solutions, with high multiplexing capabilities, which bring efficiency, convenience, and speed to the detection of mutations associated with cancer in the BRCA1 and BRCA2 genes for several DNA sample types.

HIGH PERFORMANCE, STREAMLINED METHODS FOR RNA-SEQ AND TARGET ENRICHMENT

Wednesday, 1st June 18:05 La Fonda Ballroom Tech Talk (TT-1.06)

Maryke Appel

Kapa Biosystems

Roche Diagnostics

The expanding scope and application of next-generation sequencing in both research and clinical environments have been driving a demand for sample preparation methods that yield high-quality libraries, while supporting a higher degree of workflow automation and faster turnaround times. We have previously reported on advances in DNA library construction—most notably the incorporation of low-bias enzymatic fragmentation in a streamlined, single-tube workflow (KAPA HyperPlus). This method offers the speed and convenience of tagmentation-based protocols, but consistently outperforms the latter with respect to library yields, key sequencing metrics (coverage uniformity and depth), and flexibility across different sample types and experimental designs.

We have recently expanded our suite of streamlined sample preparation methods to include a complete, fully automatable workflow for target enrichment (HyperCap), as well as a novel, rapid workflow for the construction of stranded RNA-Seq libraries (KAPA RNA Hyper Prep).

The HyperCap workflow, co-developed by Kapa Biosystems and Roche Nimblegen, incorporates the KAPA Hyper Prep or HyperPlus chemistry in an application-specific approach to rapid target capture. General improvements to both the library construction and target capture portions of the workflow reduces turnaround time and eliminates steps that previously required user intervention and/or specialized equipment, and have therefore been difficult to automate. A series of optional improvements allows the end-user to further tailor the protocol to specific sample types, sequencing applications and operational objectives. In its most extreme form, the HyperCap workflow allows for the construction of high-quality, sequencing-ready libraries from input DNA in ~9 hours. A more conservative, 2-day workflow is recommend for challenging samples (e.g. low-input FFPE) and/or small capture panels. Examples from of both ends of this spectrum will be presented.

The KAPA RNA Hyper Prep workflow employs novel chemistries that allow for the combination of several steps in the construction of RNA-Seq libraries. As a result, sequencing-ready libraries can easily be prepared from total RNA in a standard 8-hour day, inclusive of RNA enrichment (mRNA Capture or ribosomal depletion). The protocol is also compatible with total RNA input for RNA capture applications. Higher library construction efficiency allows for successful library construction from lower RNA inputs, and higher success rates with FFPE samples. Data generated with Universal Human Reference RNA (UHR), ERCC spike-in controls and RNA isolated from fresh frozen and FFPE tissues will be presented.

Products are for life science research use only, not for use in diagnostic procedures.

POSTER PRESENTATIONS
&
MEET AND GREET PARTY

Sponsored by Roche Diagnostics

Enjoy!!!

Drink tickets (beer, wine, juice and sodas) provided
Use your yellow tickets

Poster Sessions

1a = 1st floor, 18:30 – 20:00 pm
2a = 2nd floor, 18:30 – 20:00 pm

1b = 1st floor, 20:00 – 21:30 pm
2b = 2nd floor, 20:00 – 21:30 pm

GENETIC CHARACTERIZATION OF STX2 PRODUCING ESCHERICHIA COLI (STEC) ISOLATED IN THE COUNTRY OF GEORGIA

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.01)

Tea Tevdoradze, Gvantsa Chanturia, Ekaterine Zhgenti, Giorgi Dzavashvili, Lia Tevzadze

National Center for Disease Control and Public Health, Georgia

Shiga toxin-producing Escherichia coli (STEC) infection is among the leading causes of bloody diarrheal disease complicated by hemolytic uremic syndrome (HUS) in the country of Georgia. Since 2009, many human cases of bloody diarrhea in Georgia complicated with HUS have been caused by E. coli strains positive for the Shiga toxin 2 gene (stx2).

In this study we characterized three stx2 producing E. coli strains isolated from three patients with HUS (in 2012, 2014, 2015 respectively). In previous studies isolates were characterized by confirmatory STEC (stx1, stx2, eae, ehxA), O104 (stx2, terD, rfbO104, fliC H4) and enteroaggregative (pCVD, AGGR) conventional multiplex PCR assays. STEC strains were genotyped by pulsed field gel electrophoresis (PFGE) and multilocus sequence typing (MLST). Draft whole-genome sequenc- ing (WGS) of isolates was performed using the next-generation sequencing (NGS) Illumina MiSeq platform. Obtained reads were further analyzed using the CLC Genomics Workbench (CLC Bio) software package.

In the present study, WGS-based SNP phylogenetic analysis was performed using the EDGE bioinformatics software; we also determined the virulence and the resistance genes profiles of three STEC strains based on the Center for Genomic Epidemiology database, Denmark (https://cge.cbs.dtu.dk). Two enteroaggregative 0104: H4 strains that belonged to sequence type ST-678 revealed a very similar set of virulence and resistance genes. The strain isolated in 2012 contained an additional resistance gene blaCTX-M-14 that was not present in the 2015 strain. The third non-O104 STEC strain ST-677, which could not be characterized with classical serotyping methods, was identified as E. coli O174:H21. Comparative sequence analysis performed against reference strains in established E. coli databases (https://cge.cbs.dtu.dk) indicated this strain lacked known resistance genes and encoded a limited range of virulence factors.

WGS-based SNP phylogenetic analysis revealed all 0104:H4 strains studied showed close affiliation with historical STEC isolates from Georgia (2009) and German (2011) outbreak strains. The third 0174:H21 isolate, lacking virulence genes, showed more genetic similarity with a single E. coli O152:H28 strain (SE11) isolated in Japan in 2008.

COMPARISON OF 12 FRANCISELLA TULARENSIS WHOLE GENOMES FROM THE COUNTRY OF GEORGIA

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.02)

Gvantsa Chanturia[1], Giorgi Dzavashvili[1], Jason Farlow[2]

[1]National Center for Disease Control and Public Health, Georgia
, [2]Farlow Scientific Consulting Company, LLC

The select agent Francisella tularensis is widely spread in the eastern part of Georgia. The National Center for Disease Control and Public Health of Georgia (NCDC) performs active surveillance of this pathogen by seasonal field sampling of vectors and amplifiers at two main foci. A collection of more than one hundred strains of F. tularensis currently exists at the NCDC- Lugar Center for Public Health Research.

The SNP and MLVA genotyping of archival strains was performed in the scope of previous studies in collaboration with Northern Arizona University (NAU) and Walter Reed Army Institute of Research (WRAIR). Whole genome SNP comparison of one F. tularensis strain from the NCDC collection and another F. tularensis reference genome revealed the Georgian strain as basal for most of the European clades. Country-specific SNPs were discovered and used for typing and phylogenetic analysis of the rest of the strains.

Twelve F. tularensis strains from the NCDC collection were selected for whole genome sequencing at WRAIR and Lugar Center using next generation sequencing platfor Bioinformatics tools available at the Lugar Center including CLC-Bio and EDGE were used for data processing and phylogenetic analysis. The F. tularensis SCHU S4 and Live Vaccine Strain (LVS) genomes were used for reference read mapping analyses. Close relatives to the Georgian clade reference strains were added from the NCBI database for phylogenetic analysis.

Phylogenetic analyses based on new SNPs and InDels among Georgian F. tularensis strains placed these isolates in separate branches that were not previously observed after PCR-based SNP typing. Whole-genome-based SNP analysis allowed the discovery of new diversity patterns inside the previously-established canonical SNP lineage (B.Br.013).

TINK: A NOVEL EUKARYOTIC EVIDENCE BASED PAN-TRANSCRIPTOME GENERATION PIPELINE

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.03)

Chandler Roe, Jason Travis, Nathan Hicks, Elizabeth Driebe, David Engelthaler, Paul Keim

TGen North

While next generation sequencing has become an increasingly easy laboratory procedure, eukaryotic genome annotation is still a challenging bioinformatic task. High-throughput mRNA sequencing (RNA-Seq) platforms allow for a variety of applications such as novel transcript and isoform discovery, expression estimate analysis, alternative splicing as well as exploration of non-model-organism transcriptomes. However, the required genome assembly and annotation is a complicated and time-consuming process that requires multiple steps and command line skills. Our pipeline, TINK, generates an evidence based pan-transcriptome reference to be used for RNA-Seq analysis. It provides a rapid, all encompassing, one-time analysis that allows for discovery of unique transcripts. This pipeline combines ab initio gene prediction using the program AUGUSTUS, protein homology prediction utilizing AAT and de novo RNASeq assemblies using both PASA and Trinity. These results are weighted and combined using EvidenceModeler to create individual genome annotations for each sequenced sample and further compiles, clusters and de-replicates these annotations to create a novel pan-transcriptome reference. We have used this technique to explore differential expression and identify novel transcripts from the fungal pathogen Cryptococcus gatti. This pathogen has been characterized into four types, I-IV, within which subgroups exist. In order to capture transcripts unique to one subtype of C. gattii as well as differing expression levels, multiple analyzes would need to be performed using a different reference each time, which is both computationally expensive and time consuming. TINK provided a reference to allow a single analysis on this data, greatly reducing time and resources.

A RAPID, CULTURE-FREE WHOLE GENOME ASSEMBLY APPROACH FOR CHARACTERIZING COMPLEX METAGENOMICS SAMPLES

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.04)

Paul Havlak[1], Brandon Rice[1], Brenden O'Connell[2], Christopher Troll[1], Ei Min[1], Jonathan Stites[1], Marco Blanchette[1], Margot Hartley[1], Nicholas Putnam[1], Robert Calef[1], Richard Green[1,2]

[1]Dovetail Genomics, LLC, [2]University of California-Santa Cruz

High-throughput sequencing allows genetic analysis of complex microbial communities that inhabit a wide variety of environments. Shotgun sequencing of environmental samples, which often contain microbes that are refractory to culturing in the lab, can reveal the genes and biochemical pathways present within the organisms in a given environment. However, high-quality de novo assembly of these highly complex datasets is generally considered to be intractable.

We have developed a powerful and efficient method for de novo genome assembly of read data from complex metagenomics datasets. Our approach involves efficient generation of connectivity data from these complex DNA samples. Adapting our Chicago Method, a technique that has worked well for de novo genome assembly of single organisms, we perform proximity ligation and sequencing of in vitro assembled chromatin from metagenomic DNA samples. When combined with shotgun sequence data, we are able to generate genome-scale assemblies of multiple organisms in metagenomics communities sampled directly from the human gut, soil, and stream water. This fast and simple approach enables a more complete understanding of microbial communities than that afforded by standard culturing techniques or shotgun sequencing alone.

DEVELOPMENT AND VALIDATION OF METAGENOMICS SEQUENCING PIPELINES FOR BIOSURVEILLANCE AND DIAGNOSTICS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.05)

Joe Russell[1], Jonathan Jacobs[1], Richard Winegar[1], Cynthia Zimmerman[1], J R Aspinwall[1], John Bagnoli[1], K Parker[1], J Stone[1], Brittany Campos[1], Tom Slezak[2], Patrick Chain[3], Karen Davenport[3], Po-E Li[3], Joseph Anderson[4], Kimberly Bishop Lilly[4], Kenneth Frey[4], Tim Postlethwaite[5], Tammy Spain[5], Jesper Jakobsen[6], Cecilie Boysen[6], Kristine Werking[1], Michael Cassler[1]

[1]MRIGlobal, [2]Lawrence Livermore National Laboratory, [3]Los Alamos National Laboratory, [4]Naval Medical Research Center, [5]Draper Laboratory, [6]Qiagen

Next generation sequencing (NGS) has the potential to allow unbiased detection and characterization of biothreat agents and emerging pathogens from a variety of clinical and environmental samples. This capability would greatly benefit multiple applications, including microbial forensics, biosurveillance, clinical detection and clinical diagnostics. However, current sample to sequence pipelines are complex, and there is a growing need for them to be simplified, standardized, and validated before results can be made comparable across multiple laboratories. Specific needs include standard reference materials, simplified sample and library preparation, trusted reference databases, robust bioinformatics pipelines, and clear regulatory pathway. MRIGlobal is leading a large team of vested organizations in developing and validating methods for accelerating the use of NGS as a powerful tool for the detection of infectious disease agents. Here we present the overall scope of the project and the trajectory of its development.

GENETIC STUDIES OF YERSINIA PESTIS STRAINS IN KAZAKHSTAN

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.06)

Berzhan Kurmanov, Elmira Zh. Begimbayeva, Dmitriy Berezovskiy, Kabysheva N.P.

M. Aikimbayev Kazakh Scientific Center for Quarantine and Zoonotic Diseases

Kazakhstan is the home of a most ancient, large and active plague foci that occupies almost the entire southern part of the country (more than 1 million sq. km.) and encompasses diverse ecological zones: desert, semi-desert, steppe, low-mountain, and high-mountain. As a result there are different types of major carriers: marmots, ground squirrels, gerbils, voles, mice, etc. Thus, the most typical carrier in the Central Asian desert focus is the great gerbil (Rhombomys opimus), whereas in the Tien Shan high-mountain focus the typical carrier is the gray marmot (Marmota baibacina). Such diversity in geographic range and hosts, offers the possibility that the strains of the plague pathogen – Yersinia pestis from different ecological zones may differ both phenotypically and genotypically.

Research works performed in Kazakhstan show changes in biochemical characteristics of plague microbe based on the capacity for fermentation of rhamnose, glycerol and arabinose, and nitrate reduction. Genetic polymorphism of strains was determined. Study of protein profile of Y. pestis strains showed that strains circulating in the plague natural foci of Kazakhstan had nine variants of protein profile. The genotypes of Y. pestis strains were studied by VNTR-analysis, which revealed 7 clusters of closely related strains.

Genetic characterization of Kazakhstani strains is described in national publications as well as in collaborative research works (Engelthaler D. M. et al 2000, Antolin M., et al 2003, Lowell J. L. et al 2007 etc.), where the strains from Kazakhstan are attributed to the four known biovars: Antiqua, Medievalis, Orientalis and Microtus.

Collaborative research with the Center for Microbial Genetics and Genomics of Northern Arizona University on genotyping of Kazakhstani strains by whole genome sequencing has been currently carried out. The study of genetic characteristics and genetic diversity of local Y. pestis strains has great prospects both for the diagnosis and establishing evolutionary ways of development. In addition, there is a need for a systematic meta-genome analysis to assess the impact of the carrier (host) in a focus on the properties of the pathogen.

For more in-depth analysis of the phylogenetic relationships of Kazakhstani Y.pestis strains it is necessary to conduct additional analysis of variable loci by the methods of Multilocus sequence typing (MLST), Melt Analysis of Mismatch Amplification Mutation Assays (Melt-MAMA), or by whole genome sequencing. The study of genetic characteristics and genetic diversity of local strains of plague microbe has great prospects both for the diagnosis and establishing evolutionary ways of development. In addition, there is a need for a systematic meta-genome analysis to assess the impact of the carrier (host) in a focus on the properties of the pathogen.

EVOLUTION OF SEQUENCING IN THE ARIC COHORT TO REVEAL THE GENETIC ARCHITECTURE OF COMPLEX TRAITS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.07)

Ginger Metcalf[1], Narayanan Veeraraghavan[1], Harsha Doddapaneni[1], Prof. Yi Han[1], Bing Yu[2], Simon White[1], William Salerno[1], Alanna Morrison[2], Xiaoming Liu[2], Andrew Carroll[3], Darren Ames[3], Donna Muzny[1], Prof. Richard Gibbs[1], Eric Boerwinkle[1]

[1]Human Genome Sequencing Center Baylor College of Medicine, [2]Human Genetics Center, University of Texas Health Science Center, [3]DNAnexus

The Atherosclerosis Risk in the Community (ARIC) study is a flagship project of the HGSC seeking to identify susceptibility genes underlying well-replicated GWAS findings for heart, lung, and blood diseases and their risk factors in combined analyses with the larger Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium.

Early efforts involved a multiplexed capture sequencing approach targeting a 2.2 Mb region of the genome. While informative, analyses were limited to the loci selected from the GWAS findings and thus biased towards common variants. As sequencing platforms improved, more comprehensive applications of whole exome sequencing and low pass (6X) genomes became affordable, followed by 30X genomes on the Illumina Hiseq X. To date, the HGSC has sequenced 15K whole exome and 8K whole genome samples spanning 5 longitudinal cohort studies with extensive phenotypic and clinical data.

These data have fueled considerable analytical work, forming the foundation for discovery of novel genes/loci influencing a number of chronic disease risk factors, but while an increased sample size increases the chance at discovery, it also introduces increased logistical challenges. The increasing footprint of genomic data creates challenges in data sharing, computational capability, and data storage. To address these complications we created an expansion of our local compute infrastructure as part of collaborations with DNAnexus and AWS. We built and validated our high throughput sequence processing and analysis framework into DNAnexus and staged the largest ever biomedical compute on the Amazon Cloud. Additionally a secure data commons was created whereby ARIC and CHARGE investigators from around the world could access the sequence data and run analysis tools on the cloud to achieve their individual study ai Analysis of 30X WGS data is underway in an effort to identify genomic regions influencing the human serum metabolome among ARIC members, with special emphasis on metabolites influencing risk of CVD. By focusing on proximal measures of physiologic processes we will optimize the size of a gene's effect relative to corresponding risk factor level or disease endpoint.

A NOVEL APPROACH FOR SELECTIVE ENRICHMENT OF GENE TARGETS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.08)

Andrew Barry[1], Daniel Kraushaar[2], Lynne Apone[1], Sarah Bowman[2], Kruti Patel[2], Noa Henig[1], Amy Emerman[2], Theodore Davis[1], Salvatore Russello[1], Cynthia Hendrickson[2]

[1]New England Biolabs, Inc., [2]Directed Genomics

Target enrichment of selected exonic regions for deep sequence analysis is a widely used practice for the discovery of novel variants, and identification and phenotypic association of known variants for a wide range of practical applications. Current available strategies for selective enrichment can be characterized as either hybridization-based enrichment, where long synthetic oligonucleotides are used to selectively capture regions of interest, or multiplexed amplicon-based, where pairs of short primer sequences leverage PCR to selectively amplify sequence targets. While hybridization-based methods have proven to be a tractable approach for large panels scaling to whole exome, the approach presents challenges in a relatively high sample input requirement, longer workflows, and inability to scale to very focused panels. In contrast, multiplexed amplicon approaches have proven valuable for small, highly focused panels, yet suffer from inherent challenges including the inability to scale content, loss of specificity associated with PCR duplication, and difficulties annealing primer pairs to already degraded materials.

The NEBNext DirectTM technology utilizes a novel approach to selectively enrich nucleic acid targets ranging from a single gene to several hundred genes, without sacrificing specificity. Further- more, intrinsic properties of the approach lend themselves to improved sensitivity and have proven amenable to challenging sample types including FFPE tissue and circulating tumor DNA (ctDNA). The result is a 1-day protocol that enables the preparation of sequence-ready libraries with high specificity, uniformity, and sensitivity for the discovery and identification of nucleic acid variants.

EXPLORING INTERSPECIES INTERACTIONS TO DISCOVER A NOVEL GUT BACTERIAL COCKTAIL TO TREAT ANTIBIOTIC-RESISTANT CLOSTRIDIUM DIFFICILE INFECTION

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.09)

Anand Kumar, Armand Dichosa, Shawn Starkenburg, Momchilo Vuyisich

Los Alamos National Laboratory

Gut microflora interactions play an important role in human health and disease. Deciphering these complex interactions is continually challenging, but is also an avenue for innovative diagnostics and therapeutics. Here, we propose to explore interspecies interactions and methods enabling the discovery of a specific gut bacterial cocktail to treat antibiotic-resistant Clostridium difficile infection (CDI). CDI is an emerging public health threat worldwide, and in the US alone, CDI is responsible for 0.5 million cases and 30,000 deaths annually. Currently, fecal transplants (FT) are the only therapy for antibiotic-resistant CDI, but this treatment method is not universally adopted due to unpredictable side effects. To develop a safe and reliable alternative treatment for CDI, we will identify optimal growth conditions and define baseline interactions in the normal healthy gut flora in vitro (Aim1); we will analyze C. difficile and gut microflora interactions to determine two specific groups of bacterial populations that either enhance or suppress C. difficile bacteria in vitro using established growth conditions (Aim 2); and we will test identified natural suppressor gut bacteria in a C. difficile pre-clinical animal model to demonstrate the therapeutic efficacy of a novel bacterial cocktail (Aim 3). Furthermore, any group of bacteria that enhances C. difficile growth could be useful to diagnose the severity and/or to track the progression of CDI in humans. Our previously established and reported High-throughput Screening of Cell-to-cell Interactions (HiSCI) in vitro technique will be utilized to understand interspecies interactions in the first and second ai

PERSONALIZING CYSTIC FIBROSIS: A LONGITUDINAL ANALYSIS OF THE MICROBIOME FROM CF PATIENTS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.10)

Josie Delisle[1], John Gillece[1], James Schupp[1], Jolene Bowers[1], Erin Kelley[1], Elizabeth Driebe[1], David Engelthaler[1], Cori Daines[2], Paul Keim[1,3]

[1]TGen, [2]University of Arizona, [3]Northern Arizona University

Cystic fibrosis (CF) is an incurable, genetic disorder that affects roughly 70,000 people worldwide and dramatically reduces the lifespan of the afflicted individual. Promising new drugs have significantly reduced symptoms for about 4% of the CF population, but unfortunately are expensive and do little for the other 96%. It affects many systems in the body, but the characteristic symptom that dramatically reduces quality of life is associated with the accumulation of mucus in the lungs. In the long term, this static mucoid environment gets colonized by a diverse community of microbes, viruses and fungi, therefore, becoming a reservoir for infection and requires chronic treatment with antibiotics. Generally, the choice of antibiotic is not targeted and in many cases is administered after the infection has advanced. In addition to long-term side effects of these drugs, eventually antibiotic resistant bacteria and fungi can take hold and complicate further treatment. In collaboration with the University of Arizona, we have initiated a study to track changes in the microbial and fungal community over time in patients with CF. The purpose of this study is to better understand the CF pulmonary microbiome over time and determine if emerging pathogens and antibiotic resistance can be detected prior to exacerbation. A primary goal of the study is to develop and evaluate personalized, targeted microbiome assays for rapid characterization of a patient's current pulmonary community. These prototype diagnostics could potentially be used to advise treatment that is personalized and timely. The study involves monthly sputa collections from up to 13 patients over the course of one year. The sputa samples were analyzed for 16S rDNA and Whole MetaGenome Sequence (WMGS) community analysis. In addition, the sputa samples were evaluated through our Next-Gen Antimicrobial Resistance Detection (N-GARD) assay, profiling the personalized microbial community in these samples through amplicon sequencing. Preliminary 16S analysis (samples from 2 to 6 time points for 6 patients) revealed high relative abundances of many of the common pathogens associated with CF, including Pseudomonas, Streptococcus, Staphylococcus, Haemophilus, and Enterobacteriaceae with variation among patients and among timepoints within patients. Analysis of metagenomic data through GOTTCHA and WGFAST revealed overlapping patterns with the 16S, but was able to resolve to species and strain level. N-GARD sequencing revealed two CF patients with MRSA (SCCmec type II) and another patient with MSSA. Other species targets such as for Achromobacter, Haemophilus and Streptococcus are currently under development. Analysis of the fungal community is ongoing.

EDGE V2.0: AN UPDATED OF THE EMPOWERING THE DEVELOPMENT OF GENOMICS EXPERTISE BIOINFORMATICS PLATFORM

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.11)

Chien-Chi Lo, Po-E Li, Yan Xu, Sanaa Ahmed, Shihai Feng, Karen Davenport, Patrick Chain

Los Alamos National Laboratory

Continued advancements in sequencing technologies have fueled the development of new sequencing applications and promise to flood current public and private databases with raw data for genomes and metagenomes. A number of factors prevent the seamless and easy use of these data, including the breadth of project goals that often require highly specialized data processing, the wide array of tools that individually perform only small fractions of any given analysis, the large number of software and hardware dependencies these tools require to function, and the detailed expertise required to perform and interpret these analyses. In 2015, to address many of these issues, we have developed an intuitive web-based environment with a wide assortment of integrated and cutting-edge bioinformatics tools1. These preconfigured workflows provide even novice next-generation sequencing users with the ability to perform a number of complex analyses (e.g. metagenome analyses) with only a few mouse clicks, and, within the context of the same web environment, allow users to visualize and further interrogate the results. This bioinformatics platform is an initial attempt at Empowering the Development of Genomics Expertise (EDGE) in a wide range of applications for researchers who lack dedicated bioinformatics resources. Using the same strategy we have now expanded EDGE with (i) an amplicon data analysis module, (ii) a specialty genes detection module (for antibiotic resistance genes and virulence genes), (iii) UGE cluster submission support, (iv) MEGAHIT metagenome assembler, (v) a secure user management system update, (vi) new GOTTCHA database integration, (vii) third- party tools upgrade, (viii) and more interactive features for the result page. The improved EDGE v2.0 widens the usage for the NGS data analysis.

1. http://biorxiv.org/content/early/2016/02/21/040477

MOLECULAR DIVERSITY OF BRUCELLA STRAINS FROM THE COUNTRY OF GEORGIA

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.12)

Keti Sidamonidze[1], Kevin Drees[2], Jeffrey Foster[2], Ekaterine Zhgenti[1], Tea Tevdoradze[1], Nino Trapaidze[1], Gvantsa Chanturia[1], Paata Imnadze[1], Mikeljon Nikolich[3]

[1]National Center for Disease Control and Public Health, Georgia, [2]University of New Hampshire, [3]Walter Reed Army Institute of Research

Brucellosis is a globally important zoonotic disease that is endemic in the country of Georgia, where it causes substantial human morbidity and significant agricultural economic losses. Because of its high infectivity, Brucella abortus and B. melitensis are classified as Category B biological threat agents. Lack of genetic resolution with available methods has made it challenging to understand its evolutionary history and determine the spread of this pathogen across the globe. Whole genome sequencing (WGS) allows for a deeper understanding of phylogenetic relationships among bacterial strains.

In order to assess genetic variation among Brucella spp. circulating in Georgia, 15 Brucella strains – B. melitensis (n=5) and B. abortus (n=10), were whole genome sequenced. These strains were chosen as representatives of major genetic clusters, previously determined by MLVA-15 as part of the Defense Threat Reduction Agency's Cooperative Biological Research project GG-17. Whole genome assemblies from these strains were aligned to the reference genomes for each species, B. abortus-2308 and B.melitensis-16M, respectively, for SNP discovery. A phylogenetic comparison of Georgian Brucella whole genome sequences to a worldwide collection of genomes showed that Georgian strains of B. abortus largely form a unique clade basal to the most common radiation of strains from biovars 1, 2, and 4, and are most similar to strains from Central Asia.

Georgian B. melitensis isolates are less distinct and appear to mostly fall into the East Mediterranean lineage, but in select cases, also group with isolates found worldwide. Based on these WGS data, 15 Brucella strains were chosen for MLST analysis using the online database: BrucellaBase: Genome Information Resource. This study revealed that Brucella strains belong to sequence type 2 and 8 (ST2 and ST8). This panel will allow the screening of not only additional archival isolates, but also newly isolated Brucella strains in Georgia and the Caucasus region, thus allowing for a rapid assessment of their global phylogenetic context.

SEQUENCING STRATEGY FOR THE WHOLE GENOME OF SHEEPPOX VIRUS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.13)

Vitaliy Strochkov, Yerbol Burashev, Nurlan Sandybayev Sandybayev

Research Institute for Biological Safety Problems, MES Republic of Kazakhstan

Sheeppox virus (SPPV) is a widespread and economically significant pathogen considered endemic in central Asia and member of the genus Capripoxvirus. Capripoxvirus genomic DNA are double-stranded with lengths of around 150 kbp, with SPPV, goatpox virus (GTPV), and lumpy skin disease virus species generally sharing at least 96% nucleotide identity. Strains of SPPV and GTPV share at least 147 genes.

A previous strategy of shotgun sequencing of poxviruses is incomplete cleavage of genomic DNA with restriction endonucleases, e.g. Tsp509I. Fragments ranging in size from 1 to 2.5 kbp were randomly cloned into a plasmid and sequenced. By using this strategy, complete nucleotide sequences for several strains of sheep pox and goat poxvirus was defined by our group (JVI, 2002, 76(12):6054-6061).

We have adopted directed sequencing methods to rapidly obtain complete genome sequences from capripoxviruses. PCR arrays yielding overlapping PCR amplicons are being used to generate genomically tiled Sanger sequence data. Initial primer candidates were generated against the A strain genomic sequence, and targeting tiled amplicons of 1200 bp and overlapping by 500 bp. Specific primer selection was conducted to match annealing temperatures and to best match conserved regions relative to other capripoxviruses to reduce PCR failures resulting from sequence diversity. Primers were manufactured in 96-well plates and arranged to group primer sets with similar annealing temperatures. This enables generation of specific sequencing templates with redundancy at all sites using amplification.

DNA sequencing is performed by dideoxy sequencing method with use of termination dideoxy nucleotide (Sanger's method) on an automatic 16-capillary Genetic Analyzer 3130xl, Applied Biosyste As a polymer for capillaries POP-7 is used. Production of termination DNA outputs is carried out by the method of cyclic sequencing. These long read (up to approximately 1kbp) products are of particular use in sequence assembly, which is conducted de novo and by mapping against genomic reference genomes using both commercial and open source software.

With the use of direct sequencing of PCR products, we obtained nucleotide sequence of complete genomes for three sheeppox virus strains from different geographic regions using PCR and Sanger sequencing technologies currently in place at RIBSP. The work was performed under a study for Comparative genomics especially dangerous poxviruses (orthopox and capripox) in Kazakhstan.

USING NOVEL METHODS TO ASSEMBLY PLANT MITOCHONDRIAL GENOMES: AN EXAMPLE FROM THE MIMOSOID LEGUME GENUS LEUCAENA

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.14)

Sealtiel Ortega Rodriguez, Donovan Bailey

New Mexico State University

Plant mitochondria host structurally complex low gene content genomes known for high rates intra-cellular recombination, sequence loss, and sequence acquisition. As a result of these complexities very few such genomes have been published. This is even true for the ecologically and economically important legumes, where mitochondrial genomes are available for just fives species of closely related papilionoids. To diversify the representation of available legume mitochondrial genomes, we generated a draft sequence assembly for a member of the subfamily Mimosoideae, Leucaena trichandra. Mimosoid legumes are best represented throughout the tropics, where they are often critical species in natural ecosystems while also providing nearby human populations with a variety of important resources. Using 60X PacBio sequence data with 14.5kb average read length we employed a novel iterative reference seeded approach to de novo assemble the L. trichandra mitochondrial genome data from among a complex mix of nuclear and cpDNA reads. We discuss what we have learned about this iterative assembly approach along with the preliminary mitofy annotations of the draft genome.

FIRST COMPLETE GENOMIC SEQUENCE OF CLOSTRIDIUM SEPTICUM

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.15)

Michael E. Holder[1], Nadim J. Ajami[1], Bryan P. Roxas[2], Michael J. G. Mallozzi[3], Dorothy E. Lewis[4], Gayatri Vedantam[5], Joseph F. Petrosino[1]

[1]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, [2]University of Arizona, School of Animal & Comparative Biomedical Sciences, Tucson, AZ, [3]University of Arizona, School of Animal & Comparative Biomedical Sciences, Tucson, AZ The Gut Check Foundation, Tucson, AZ, [4]The Gut Check Foundation, Tucson, AZ The University of Texas Health Science Center, Department of Internal Medicine, Houston, TX, [5]University of Arizona, School of Animal & Comparative Biomedical Sciences, Tucson, AZ The Bio5 Research Institute, University of Arizona, Tucson, AZ

Clostridium septicum is an opportunistic pathogen responsible for about 2000 cases of deadly infections every year in the US. Gangrenous wound infections occur as well as gastrointestinal infections in those immunosuppressed during malignancies involving the bowel. These cancer-associated infections often present with little to no symptomology different from the common side effects of chemotherapy treatment, the most diagnostic of which are tachycardia and pain without identifiable cause. Currently, a rapid diagnostic tool for C. septicum does not exist and molecular biological research on this organism has been hampered by lack of genomic sequence, genetic tools, and the relatively low infection rate. Here we report a full genome assembly of Clostridium septicum using SPAdes and sequencing data from PacBio and Illumina 10Kb Nextera Mate Pair libraries. Two principle contigs, a 3.4 Mb circular chromosome and a 5Kb circular plasmid, were produced. Consensus error correction of the circularized sequences was performed using Quiver and the final result annotated with NMPDR RAST. This complete genomic reference revealed a high degree of homology with sequences from other Clostridia members. However, two regions appear to be unique to this species. One of these regions (50kb) bears the hemolytic aerolysin-like alpha toxin required for virulence. The second region (92kb) encodes a predicted novel, multi-domain, secreted 206 kDa protein of unknown function. Additionally, homologues of antibiotic resistance genes are observed in the plasmid. Annotation of the genome revealed a large number of genes dedicated to fermentation (59), cell wall and capsule biosynthesis (99) (eight of which are associated with exopolysaccharide biosynthesis), motility (48), sporulation and dormancy (76), oxidative stress (31), and virulence and disease (60) suggesting that a large proportion of the genome is dedicated to adaptation to the host during infection. Publication of the genome will help spur further advances in the diagnosis and treatment of this understudied pathogen.

A NEW NGS LIBRARY PREPARATION METHOD FOR TRANSCRIPTOME PROFILING WITH ENHANCED SENSITIVITY OF TRANSCRIPT DETECTION

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.16)

Daniela Munafo Erbay Yigit Deyra Rodriguez Mehmet Karaca Keerthana Krishnan Pingfang Liu Lynne Apone Vaishnavi Panchapakesa Laurie Mazzola Joanna Bybee Danielle Rivizzigno Fiona Stewart Eileen Dimalanta Theodore Davis

New England Biolabs, Inc.

RNA-seq (RNA sequencing) is a transcriptome-profiling method that uses next generation sequencing. It is widely used for genome-wide expression analysis as well as detection of mutations, fusion transcripts, alternative splicing, and post-transcriptional modifications. RNA-seq is becoming increasingly common in molecular diagnostics; providing better insights into how altered transcripts impact the biological pathways and the molecular mechanisms associated with disease progression. The successful adoption of RNA-seq into the molecular diagnostics will depend on the library preparation techniques that require low input RNA, and can capture the entire molecular repertoire within a sample without sequence bias.

Here, we present a high efficiency method for strand-specific RNA-seq that retains information about which strand of DNA is transcribed. Determining the polarity of RNA transcripts is important for the correct annotation of novel genes, identification of antisense transcripts with potential regulatory roles, and for correct determination of gene expression levels in the presence of antisense transcripts. This method is based on the labeling and excision of the second strand cDNA, and it is compatible with both poly A-tail enriched and ribosome-depleted RNA. Our results show this improved method generates significantly higher library yields that enable use of lower amounts of input RNA. More- over, our new method results in increased sensitivity and specificity, especially for low-abundance transcripts, reduced PCR duplicates and sequence bias, delivering high quality strand-specific data. This streamlined protocol is also amenable to large-scale library construction and automation.

DETECTION AND GENETIC CHARACTERIZATION OF BURKHOLDERIA PSEUDOMALLEI AND CLOSELY RELATED SPECIES DIRECTLY FROM SOIL USING A CUSTOM AMPLICON SEQUENCING ASSAY.

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.17)

James Schupp[1], Jason Sahl[2], Rebecca Colman[1], Jordan Buchaggen[1], Josie Delisle[3], John Gillece[1], Adam Vazquez[2], Joseph Gennarelli[2], Carina Hall[2], Joseph Busch[2], Vanessa Theobald[3], Glenda Harrington[3], Mirjam Kaestli[3], Mark Mayo[3], Bart Currie[3], David Engelthaler[1], Paul Keim[1,2], David Wagner[2]

[1]TGen, [2]Northern Arizona University, [3]Menzies School of Health Research

Rapid detection and characterization of clinical and forensic materials suspected of containing Burkholderia pseudomallei, a public health and potential bioterrorism agent endemic to Southeast Asia and Northern Australia, would be of enormous benefit to epidemiological and forensic investigations. Current methodologies, such as real time PCR, allow rapid detection but only limited characterization. High Throughput Sequencing (HTS) of multiple informative genetic loci can provide efficient, rapid detection and differentiation from near neighbor species, as well as fine scale genetic characterization. We have developed a 67 locus amplicon sequencing system that results in 1) detection of B. pseudomallei; 2) differentiation from B. mallei and near neighbor species; 3) potential detection of strain mixtures; 4) differentiation within B. pseudomallei; and 5) virulence gene characterization (10 vir genes), within 24-48 hours. The system couples highly multiplexed amplification reactions with a universal amplicon indexing system, resulting in efficient multilocus amplicon sequencing from potentially hundreds of samples in a single Illumina MiSeq sequencing run. We have detected B. pseudomallei in soils from Northern Australia, down to near single genome copy, as well as detection of near neighbor species, such as B. ubonensis. In analyzing soils from AZ, which show a very different bacterial community, no detection of B. pseudomallei and low level detection of other near neighbor species. We demonstrate detection of Burkholderia species mixtures as well as B. pseudomallei strain mixtures, utilizing variation within the targeted species specific, MLST and virulence loci. In samples containing a single predominant strain of B. pseudomallei, we also demonstrate strain level phylogenetic classification using the Whole Genome Focused Array SNP Typing (WG-FAST) pipeline directly from the soil sample.

INCREASED RATES OF SPONTANEOUS DUPLICATIONS AND DELETIONS UNDER HEAVY METAL EXPOSURE IN DAPHNIA MUTATION ACCUMULATION LINES

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.18)

Frédéric Chain, Jullien Flynn, James Bull[1], Melania Cristescu

McGill University

Heavy metals are pervasive environmental pollutants and toxic contaminants that are of major con- cern for ecosystem and human health. Metals can induce oxidative and DNA damage, promoting mutations through DNA repair and replication. Errors occurring during these processes are particularly important mechanisms generating copy number variations (CNVs) deletions, duplications and insertions but little is known about the long-term effects of metals on genome-wide mutations such as CNVs. Genomics approaches allow for nucleotide-resolution detection of CNVs, although these methods are limited by the quality and architecture of the reference genome. We analyzed CNVs among 42 whole genomes of clonal Daphnia mutation accumulation lines, all seeded by the same progenitor ancestor, but maintained in conditions either with or without metals at ecologically relevant concentrations. After propagating lines for an average of 100 generations, we found higher basal rates of both gene deletions and gene duplications in lines exposed to a mixture of nickel and copper compared to controls. However, our CNV rate estimates are limited to unique genomic regions to reduce the bias from reads mapping to multiple locations, which can be due to the mis-assembly of divergent alleles in diploid genomes. Our results show that metal exposure can increase the incidence of large-scale heritable mutations, but that complex genomes and assemblies limit the extent to which we can perform truly “genome-wide” analyses.

BAC SUDOKU SEQUENCING STRATEGY FOR IN SILICO SCREENING OF LARGE-INSERT SOIL METAGENOMIC LIBRARIES

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.19)

Scott Monsma[1], Jinglie Zhou[2], Alinne Pereira[2], Blaine Pfeifer[3], Timothy Bugni[4], Scott R. Santos[2], Megan Niebauer[1], Erin Ferguson[1], Ronald Godiska[1], Chengcang Wu[5], David Mead[1], Mark R. Liles[2]

[1]Lucigen Corporation, [2]Auburn University, [3]State University of New York at Buffalo, [4]University of Wisconsin-Madison, [5]Intact Genomics Inc

Soil microorganisms express diverse bioactive natural products; however, the majority of soil microbes are recalcitrant to cultivation. We are using a metagenomic approach to bypass cultivation and directly capture the DNA from diverse microbial genomes in natural environments such as soils. A metagenomic library from an agricultural soil (Cullars Rotation, Auburn, AL) was constructed in a broad host-range BAC vector that contained 19,200 clones with an average insert size of 110kb. Screening was accomplished using strategy termed BAC Sudoku sequencing, wherein a pooling strategy is used to multiplex the results, while still providing the ID of individual clones. BAC clones were sequenced in pools (row, column, and plate) using indexed primers and paired end reads on an Illumina HiSeq. Contigs were assembled for each pool and screened for secondary metabolite gene clusters using antiSMASH, resulting in identification of >1000 novel PKS/NRPS pathway-containing clones. The cloned pathways are very divergent from known pathways, with the GC content varying from 41 to 76% and the amino acid identity of the KS domains ranging from 32 to 83% to the best matching BLAST hit. Expression of these PKS pathway-containing clones in E. coli strain BTRA (engineered for polyketide precursor expression) has resulted in multiple clones with evidence for heterologous expression of a cloned PKS pathway. These results indicate a high degree of unique sequence space has been recovered from large-insert metagenomic clones and a subset of these clones are capable of being heterologously expressed to produce secondary metabolites, thereby expanding our available resources for natural product discovery.

ROLE OF PIRNAS IN THE ABSENCE OF ACTIVE TRANSPOSABLE ELEMENTS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.20)

Michael Vandewege[1], Roy Platt[2], David Ray[2], Federico Hoffmann[1]

[1]Mississippi State University, [2]Texas Tech University

Transposable elements (TEs) have the ability to mobilize throughout a genome and make up sizable proportions of mammalian genomes. A class of small RNAs, PIWI interacting RNAs (piRNAs), are part of a cellular pathway that protects genomes against the expression and proliferation of TEs. Post-transcriptional regulation of TE expression involves the guidance of PIWI proteins with endonuclease activity to TE transcripts via piRNAs that share sequence complementarity with TEs. In most mammals Long INterspersed Elements (LINEs) and Short INterspersed Elements (SINEs), are the dominant mobilizing TEs. However, in a rare instance, LINE and SINE elements are no longer mobilizing in the 13 lined ground squirrel (Ictidomys tridecemlineatus). This extinction occurred approximately 5 million years ago, and it is likely that piRNAs and PIWIs are no longer required to protect the genome against the expression of these elements. To determine the role of piRNAs in the absence of TEs, we sequenced and analyzed the piRNA repertoire of the ground squirrel at different life stages between birth and adulthood. The life stages coincide with the development of testis and expression of discrete classes of piRNAs. For direct comparisons, we also sequenced the piRNA repertoire of a juvenile and adult rabbit with active LINE and SINE elements. Among the rabbit piRNAs, there was a clear enrichment of LINE-like sequences. By contrast, there was no TE enrichment among any of the squirrel piRNA repertoires. We found evidence that piRNAs were derived from different locations of the genome among the squirrel juvenile and adult life stages, suggesting the presence different piRNA biogenesis pathways during testis development. There was evidence of residual LINE expression and piRNA directed cleavage of LINE elements in the squirrel, although it is unclear whether or not LINE mRNA is haphazardly derived from immobile LINE insertions. It is apparent that piRNAs are still produced through typical pathways in ground squirrel testis, although it is not apparent they are actively defending the genome. It is possible that piRNAs are continuously produced and used in a preventive manner.

RECENT ADDITIONS TO SPADES FAMILY OF TOOLS FOR GENOME ASSEMBLY AND ANALYSIS

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.21)

Alla Lapidus,. Dmitry Antipov, Anton Bankevich, Elena Bushmanova, Alexey Gurevich, Anton Korobeynikov, Alla Mikheenko, Dmitry Meleshko, Sergey Nurk, Andrey Prjibelski, Yana Safonova, Pavel Pevzner

Saint Petersburg State University

Despite its central role in genomics, accurate de novo genome assembly remains challenging. Moreover, the proliferation of new sequencing and sample-preparation technologies introduces additional levels of complications. In 2015, the SPAdes genome assembler (Bankevich et al., 2012), that was originally conceived as a scalable and easy-to-modify platform, was gradually extended into a family of SPAdes tools aimed at various sequencing technologies and applications. In addition to the constantly updated SPAdes assembler itself, it now includes:

- metaSPAdes assembler for metagenomics data (Nurk et al., 2015)
- rnaSPAdes: de novo RNA-seq data assembler (Prjibelsky et al., submitted)
- plasmidSPAdes: assembly of plasmids from the whole genome sequencing data (Antipov et al., submitted)
- dipSPAdes tool for assembly of highly polymorphic genomes (Safonova et al., 2015)
- exSPAnder module for repeat resolution that enables efficient utilization of mate-pair libraries and even mate-pairs only assemblies with NexteraMP libraries (Prjibelsky et al., 2014, Vasilinetc et al., 2015)
- hybridSPAdes tool for hybrid assembly of accurate short reads with long error-prone reads, such as Pacific Biosciences and Oxford Nanopore reads (Antipov et al., 2015)
- truSPAdes tool for assembling Illumina's barcoded True Synthetic Long Reads (Bankevich and Pevzner, 2015)

We will provide an overview of the SPAdes family of tools and benchmark them against state-of-the-art assembly tools using the QUAST family of assembly evaluation tools:

- QUAST tool for the quality assessment of genomics assemblers (Gurevich et al., 2013)
- metaQUAST tool for the quality assessment of metagenomics assemblers (Mikheenko et al., 2015)
- rnaQUAST tool for the quality assessment of RNA-Seq metagenomics assemblers (Bushmanova et al., 2015)

A DRAFT GENOME OF THE RESURRECTION LYCOPHYTE SELAGINELLA ARIZONICA SELAGINELLACEAE

Wednesday, 1st June 18:30 La Fonda NM Room (1st floor) Poster (PS-1a.22)

Anthony Baniaga,. ThiruvaranganRamaraj[2], Tara Hall[1]. Nils Arrigo[3], Michael Barker[1]

[1]University of Arizona, Department of Ecology & Evolutionary Biology, [2]National Center for Genome Resources (NCGR), [3]University of Lausanne, Department of Ecology & Evolution

Selaginella spp. Selaginellaceae have the potential to become an interesting model system to understand local adaptation and the evolutionary genetics associated with whole genome duplication and hybridization. They possess some of the smallest known genomes (1Cx = 81-182 Mbp) of all vascular land plants, and have drought-resistant resurrection phenotypes that permit survival for months in a desiccated dormant state and the ability to return to full photosynthetic capacity following soil moisture availability. With interest in the ecological and evolutionary importance of their resurrection phenotypes, and taking advantage of the small genome sizes in Selaginella, we generated high coverage PacBio (~80x), and paired-end and mate-pair Illumina libraries (~140x). We then performed a hybrid PacBio-Illumina assembly to generate a de novo assembly for the resurrection fern Selaginella arizonica. We present our de novo draft genome of S. arizonica and highlight genomic attributes associated with the resurrection phenotype.

INITIAL EVALUATION OF THE EARLY ACCESS STR KIT V1 FOR THE ION TORRENT™ PGM™

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.01)

Kelly A Meiklejohn[1], Sharon Wootton[2], Joseph Chang[2], Chien-wei Chang[2], Robert E Legace[2], Narasimhan Rajagopalan[2], James Robertson[1]

[1]FBI Laboratory, [2]Thermofisher

Forensic laboratories traditionally utilize commercial kits for STR genotyping, which simultaneously amplify the core CODIS STR loci and tag them with fluorescent dyes for separation by capillary electrophoresis (CE). Given the known limitations of CE typing, including the sensitivity issues with degraded DNA and the ability to deconvolute complex mixtures, massively parallel sequencing (MPS) is currently being evaluated as an additional tool for STR typing. In this study, we evaluated the Early Access STR Kit v1 for the Ion Torrent™ PGM™ using both commercially available pure native DNAs (n, 5) and forensic type DNA samples (i.e. blood, saliva and sexual fluid; n, 6). To assess reproducibility, libraries were prepared in triplicate for each sample using 1 ng of DNA as input (total n, 33). Initially we examined the performance of the Early Access STR Kit v1 using the data obtained from the five commercially available pure native DNA samples based on: 1) the depth of coverage (DoC) at each locus, and 2) the allele coverage ratio (ACR) for heterozygote loci. The average DoC among the 25 STR loci included in the panel was quite varied, ranging from 1775 – 8219X. In all loci except D13S317 and D9S2157, the ACRs were above the acceptable value of 0.5. We only observed a few instances of discordance between the PGM™ genotypes and those obtained using the traditional CE approach, indicating that reliable STR profiles can be obtained for both pure native and forensic type DNA samples. It is likely that with further advancements and modifications of the panel and associated analysis software prior to commercial release, instances of discordance will be infrequent. Subsequent evaluations should be focused on assessing the ability of the panel to detect mixtures at a range of ratios, along with the sensitivity limits.

DEVELOPING A PROTOCOL FOR RELIABLY OBTAINING DNA BARCODE DATA FROM BIOLOGICAL FRAGMENTS ISOLATED FROM FORENSIC-LIKE SAMPLES

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.02)

Kelly A Meiklejohn[1], James Robertson[2]

[1]FBI Laboratory/Visiting Scientist Program, [2]FBI Laboratory

Soil is often collected from a subject's tire, flatbed, shoes and shovel during an investigation of a crime. Studies documenting the biological diversity in such soil samples using MPS have primarily focused on a metagenomic approach, in which the operational taxonomic units (OTUs) are identified, not the individual species present. Given that biological organisms only inhabit specific ecosystems or habitats, identifying the individual species present in a soil sample could assist with geoattribution.

DNA barcoding and the associated public databases (e.g. Barcode of Life Database BOLD and GenBank) can provide a reliable molecular approach to species-level identification. While DNA barcode data has been obtained from pristine DNA extracted from freshly obtained specimens, no reports are available on DNA extracted from environmentally obtained samples, which may have fragmented DNA. The primary goal of our research was to develop a protocol to obtain DNA barcode data from forensic-like samples. To reach this goal we adapted previously published protocols for pristine DNA to compromised DNA, isolated from individual biological fragments found in soil samples. After completing this, we plan to develop a multiplex protocol for DNA barcoding of several biological specimens simultaneously on a MPS platform.

In late fall, soil samples were collected from 11 locations in the continental US. Where possible, 10 insect and 10 plant fragments were isolated from each sample for downstream extraction and amplification of the appropriate DNA barcoding regions (total n, 213). As expected, the quantity of extracted DNA was low (on average <15 ng/ul) and the purity was poor (presence of RNA, phenolic and aromatic contaminants). Given this, the amplification of the DNA barcoding regions from these extracts was not straight-forward; we identified PCR inhibition in the plant extracts and had to implement the use a high-fidelity polymerase for insect PCRs to ensure clean Sanger sequence data. We will discuss the developed protocol, strategies implemented to ensure routine PCR success in degraded samples and next steps needed to transition the protocol onto a MPS platform for multiple barcode analysis.

TOWARD A QUANTITATIVE, HIGH-THROUGHPUT METHOD FOR FORENSIC MICROSATELLITE (STR) SEQUENCING

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.03)

Melissa Scheible, Seth Faith

NC State University, Forensic Sciences Institute

Forensic science is poised to adopt new methods in DNA analysis utilizing next-generation sequencing (NGS) to obtain finer resolution and higher bandwidth in genetic analysis. To date, NGS workflows for forensic short tandem repeat (STR) sequencing do not afford a strict quantitative analysis (e.g., input output), which would be beneficial for mixture and low copy analysis. Forensic samples in NGS workflows are routinely normalized for library input quantities and molar library concentrations prior to sequencing. Furthermore, the current NGS methods are laborious, having numerous steps for introduction of operator error. Here, we present developmental research in optimizing a stream- lined forensic NGS workflow for STR sequencing that is non-normalized for quantitative analysis and automated for precision and accuracy. The workflow first amplifies STR loci with a balanced multiplex PCR reaction (PowerSeq, Promega Corp.). The post-PCR product is purified with magnetic beads, and the entire product is used for NGS library construction with a high efficiency adaptor ligation protocol (HyperPrep kit, KAPA Biosystems) operationalized on an Eppendorf 5075tc liquid handling workstation. The method does not employ a second PCR reaction for library enrichment and up to 96 samples are directly pooled without library normalization. Sequencing is conducted with Illumina MiSeq and data are analyzed using the Altius cloud-computing tool. We present data to associate gDNA input to sequence data output via measurements at checkpoints throughout the workflow: input DNA, post-PCR product, post-ligation library, pre-sequencing library, post- sequencing secondary and tertiary analyses. This approach will provide guidance for validation of NGS systems in forensic laboratories.

DETECTION AND CHARACTERIZATION OF BRAZILIAN GONOCOCCAL CLINICAL ISOLATES WITH REDUCED SUSCEPTIBILITY TO CEPHALOSPORIN ANTIBIOTICS

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.04)

Jack Cartee[1], Sean Lucking[1], Ana Paula Ramalho[2], A. Jeanine Abrams[3], David Trees[1]

[1]Centers for Disease Control and Prevention, [2]Universidade Federal do Rio de Janeiro, [3]Cen

Neisseria gonorrhoeae is the etiological agent responsible for the sexually transmitted infection gonorrhea. In 2008, the World Health Organization estimated 106 million new gonorrhea cases worldwide. The emergence of gonococcal resistance to former first-line antibiotics led to the current CDC treatment recommendation for uncomplicated gonorrhea, which outlines the dual-use of ceftriaxone with either azithromycin or doxycycline. However, the appearance of the mosaic form of the penA gene resulted in significant decreases in ceftriaxone susceptibility. Moreover, a consequence of the mosaic form of penA is the occurrence of treatment failures with various cephalosporins.

This study aimed to detect the presence of mosaic penA alleles in gonococcal isolates from Rio de Janeiro, Brazil, and to determine if the mosaic alleles were associated with reduced susceptibility to cephalosporins. We utilized genomic sequencing and phylogenetic analyses to examine the genomes of 117 gonococcal isolates that were collected at public and private healthcare clinics between 2006 and 2015 in Rio de Janeiro. Of these, seven samples with mosaic penA alleles exhibited reduced susceptibility to cefixime. These results will further aid our understanding of the evolution of decreased susceptibility to cephalosporins in N. gonorrhoeae.

MIDDLE EAST RESPIRATORY SYNDROME CORONAVIRUS (MERS-COV)

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.05)

Amanda Mercer[1], Cheryl D Gleasner[1], Tracy Erkkila[1], Shannon Johnson[1], Saied A Jaradat[2], Hazem Haddad[2], Helen Hong Cui[1]

[1]Los Alamos National Laboratory, [2]Princess Haya Biotechnology Center/JUST

Middle East respiratory syndrome coronavirus (MERS-CoV) is a severe acute respiratory syndrome (SARS)-like virus that affects the respiratory system of patients. Most patients develop severe acute respiratory illness symptoms including: fever, cough, shortness of breath, acute pneumonia, and acute renal failure. In June 2012, the first case of MERS-CoV was reported in a 60-year-old man from Jeddah, Saudi Arabia. As of July 12th, 2015, there have been a total of 1045 laboratory confirmed cases of MERS-CoV infection. About 44% of the confirmed cases have resulted in fatalities. All cases have been linked to countries around the Arabian Peninsula. Like all other coronaviruses that affect humans, MERS-CoV is assumed to have a zoonotic origin. Bats are believed to be the ultimate reservoir of MERS-CoV, but camels also appear to play a role in the transmission of the virus. Humans can be infected by exposure to air, or by consuming infected camel milk or meat. This zoonotic pathogen is not considered to have pandemic potential because it does not spread easily between humans. MERS mainly spreads in hospital settings and it has been observed that close contact with an infected patient has the potential to transmit the virus. In June 2015, there was an outbreak of MERS-CoV in South Korea. This was the largest disease outbreak outside of the Middle East, infecting 186 individuals, including 36 deaths. The recent South Korean outbreak brings greater urgency to studying this virus and its mechanisms of transmission and pathogenesis.

The collaborating institutions obtained clinical samples in Jordan from CoV infected patients of different clinical outcomes. We sequenced three coronaviral strains isolated from the sample samples, studied the phylogenic relationships, and compared with other reported MERS-CoV genomic sequence data. We will present the phylogenic correlation of these strains with other known strains, and related pathogenicity and epidemiological analyses.

GENETIC VARIATION BETWEEN THREE STRAINS OF CHLORELLA SOROKINIANA

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.06)

Blake Hovde, Yuliya Kunde, Karen Davenport, Shawn Starkenburg

Los Alamos National Laboratory, Los Alamos, NM

The freshwater chlorophyte genus Chlorella is an algal production strain of high interest in biotechnology applications. Here we present the genome sequences and gene annotations of three strains of Chlorella sorokiniana and present the results of a comparative analysis of gene content and genome structure between these strains (DOE 1412, UTEX 1228 and 1230). Genome sequencing and assembly was performed using the Pacific Biosciences long read sequencing platform alone or with combinations of Illumina or Opgen optical mapping platfor Though classified as the same species, we report a significant disparity of gene content, with each of the strains containing a large complement of strain specific genes (~40-100 genes per strain). Additionally, sequence identity at the nucleotide level is significantly different as well. Large numbers of genome rearrangements are also seen between the three strains and genome size varies by up to 3 Mb. Genome comparisons were made by using Symap and SynMap, and gene level comparative analyses were performed by using a combination of BLAST homology searches and using the Markov Cluster algorithm (MCL). Analysis of these unique genes, as well as genes shared between two of the three strains, may provide evidence to distinguish evolutionary history of these organisms and why growth conditions vary between strains.

PHYLOGENETIC ANALYSIS OF THE O-ANTIGEN BIOSYNTHESIS GENES IN VIBRIO CHOLERAE

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.07)

Daniel Negrón[1],Bruce Goodwin[2], Michael Smith[2], Shanmuga Sozhamannan[3]

[1]Noblis, Inc., [2]Defense Biological Product Assurance Office,
[3]The Tauri Group, LLC and Defense Biological Product Assurance Office, JPEO

The lipopolysaccharide (LPS) of Vibrio cholerae is a virulence factor involved in host-pathogen interactions. In particular, the O-antigen constituent of the LPS exhibits diverse genetic organization and is useful for classifying Vibrio strains and serogroups. Consequently, this provides valuable information for research into the ongoing pandemic. Our previous study developed a simple and effective bioinformatics pipeline to analyze the wb* gene cluster involved in O-antigen biosynthesis. The pipeline successfully extracted these regions from publicly available, whole genome sequencing data and generated a bootstrapped, maximum likelihood phylogenetic tree. This follow-up study compares the wb* region against the genomic backbone using phylogenic methods. Results from this study facilitate the identification and analysis of horizontal gene transfer (HGT) events, particularly those involving epidemic and non-epidemic strains.

THE CAPPABLE-SEQ APPROACH FOR ANALYSIS OF MICROBIAL TRANSCRIPTOMES

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2.08)

Ira Schildkraut, Laurence Ettwiller, John Buswell, Erbay Yigit, Bo Yan

New England Biolabs, Inc.

The initiating nucleotide found at the 5' end of primary transcripts has a distinctive triphosphorylated end that distinguishes these transcripts from all other RNA species. Recognizing this distinction is key to deconvoluting the primary transcriptome from the plethora of processed transcripts that confound analysis of the transcriptome. The currently available methods do not use targeted enrichment for the 5 end of primary transcripts, but rather attempt to deplete non-targeted RNA. We developed a method, Cappable-seq, for directly enriching for the 5' end of primary transcripts and enabling determination of transcription start sites at single base resolution. This is achieved by enzymatically modifying the 5 triphosphorylated end of RNA with a selectable tag. We first applied Cappable-seq to E. coli, achieving up to 50 fold enrichment of primary transcripts and identifying an unprecedented 16539 transcription start sites (TSS) genome-wide at single base resolution. We also applied Cappable-seq to a mouse cecum sample, mapping reads to the bacterial genomes contained in NCBI databases. We identified significant TSS signatures (>300 TSS) in many species and further analyzed representative members of 4 different phyla. Reads mapping to rRNA and transfer RNA (tRNA) represented less than 10 % of mappable reads indicating that Cappable-seq depletes processed transcripts such as rRNA and tRNA from microbiome total RNA. Cappable-seq captures the 5 end of primary transcripts enabling robust TSS determination in bacteria and microbiomes. Furthermore, Cappable-seq can reduce the complexity of the transcriptomes to quantifiable tags enabling digital profiling of gene expression in any microbiome.

WHOLE GENOME APPROACH TO MICROBIAL FORENSICS OF PLANT PATHOGENS

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.09)

Jacqueline Fletcher[1], Trenna Blagden[1], Ulrich Melcher[1], Brittany Campos[2], David Lin[3], Kumar Hari[3], Richard Winegar[2]

[1]National Institute for Microbial Forensics & Food and Agricultural Biosecurity, Oklahoma State University, [2]MRIGlobal, [3]cBio, Inc

A number of bacteria, viruses, and fungi pose serious health concerns to humans, threaten the U.S. economy, food and energy supplies, and/or the environment. The potential for pathogen use as biological weapons has been demonstrated, and the U.S. needs strong capabilities to respond to incidents involving such activity. Use of next-generation sequencing (NGS) for applications in forensic investigation enhances microbial detection, strain matching and discrimination capabilities. The technology also allows for detection of all microbes in complex samples, detection of previously unknown organisms, and identification of specific signatures or genes.

A multidisciplinary and multi-institutional team, led by MRIGlobal.,Inc. in cooperation with the Department of Homeland Security, assessed the genome sequencing status and then performed whole genome sequencing and assembly for a collection of high threat human, animal and plant pathogens, including U.S. select agents,

Genome sequence assessment surveys and surveillance plans were completed for the plant pathogenic bacteria Ralstonia solanacearum, Rathayibacter toxicus, Xylella fastidiosa, and Xan- thomonas oryzae, fungi Coniothyrium glycines, Puccinia graminis, Puccinia striiformis, Puccinia triticina, and Sybchytrium endobioticum, and oomycetes Peronosclerospora philippinensis and Sclerophthora rayssiae. Information and data gaps in existing genetic databases were used to set priorities for project collections and genome sequencing efforts.

Nucleic acid purified from the plant pathogens Ralstonia solanacearum (four strains), Rathayibacter toxicus (two strains), Xylella fastidiosa subsp. pauca (six strains), Xanthomonas oryzae (pv. USA) and its near neighbor X. campestris pv. leersiae, and Coniothyrium glycines were obtained from expert collaborators.

Illumina sequence libraries were prepared using Nextera XT. Whole genome 2×300 paired-end sequencing was performed using the Illumina MiSeq instrument. Reads were filtered and trimmed using Trimmomatic. The iMetAMOS pipeline was used to optimize de novo assembly and perform quality checks. Elements of the pipeline include FastQC, SPAdes, IDBA, KmerGenie, and QUAST. Resulting assemblies were polished using Pilon; with Samtools and BLAST used to remove low coverage and contaminating contigs. Final assemblies were annotated through the NCBI Prokaryotic Genome Automatic Annotation Pipeline.

An important outcome of this work was the creation of a NCBI-hosted genomic database for selected microbial biothreat plant pathogens of priority interest to the United States. The project generated provided genomic information necessary to assist forensic investigators in genotyping, phylogenetic analysis, and sample matching of evidentiary materials.

DISPLAYING METAGENOME SEQUENCING METADATA USING AN R SHINY APPLICATION

Wednesday, 1st June 18:30 La Fonda Mezzanine (2nd Floor) Poster (PS-2a.10)

Chris Stubben, Patrick Chain

Los Alamos National Laboratory

Over 200,000 metagenome sequencing runs have been deposited in the Short Read Archive since 2015 and in March 2016 over 35 runs are submitted every hour. The FASTQ files from each run are associated with a large amount of metadata describing experiments, samples and studies. We created an R Shiny app that connects to the REST API at the European Nucleotide Archive and retrieves metadata from studies, samples and experiments and allows users to filter, map and summarize metagenome metadata. Nearly two-thirds of metagenome samples since 2015 include locations and half of the remaining samples have place names that are geocoded using the Google maps API. The mean time between collection date and public release is 3.2 years and this delay has only declined slightly over the past 4 years. The app updates daily and displays studies and maps samples released over the last two weeks by default.

IMPACTS OF FEEDING IN A RESISTANT TREE ON THE ASIAN LONGHORNED BEETLE AND ITS GUT MICROBIAL COMMUNITY

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.01)

Erin Scully

USDA-ARS Center for Grain and Animal Health Research

The Asian longhorned beetle (ALB; Anoplophora glabripennis) is an invasive, wood-boring pest capable of thriving in the heartwood of over 47 tree species worldwide where it faces a number of nutritional challenges, including digestion of lignocellulose and hemicellulose and acquisition of nitrogen, essential amino acids and nutrients that are present in low abundances in woody tissue. Through transcriptome sequencing, we have previously demonstrated that this insect possesses a rich repertoire of metabolic machinery that enables it to degrade major hardwood polysaccharides, such as cellulose, xylan, and pectin; detoxify host plant defensive compounds; recycle essential nutrients; and efficiently acquire protein and nitrogen from woody tissue or microbes that inhabit the gut. Furthermore, through metagenome and metatranscriptome sequencing efforts, we have also demonstrated that the taxonomically diverse gut microbiota encode diverse suites of genes that complement and augment ALB's endogenous physiological capacities, including the abilities to convert xylose sugars into compounds that can be directly utilized by ALB for the synthesis of fatty acids and amino acids, fix atmospheric nitrogen and recycle nitrogenous waste products, and synthesize several essential amino acids and nutrients that are present in low abundances in woody tissue. Further, it also encodes genes with the capacity to degrade large aromatic compounds and may collaborate with ALB to facilitate digestion of the lignin biopolymer. Thus, the metabolic potential of the gut community encodes an extensive suite of enzymes, which has been hypothesized to contribute to ALB's broad host range. Despite its broad host range, there are several tree species which display considerable resistance to ALB and other wood-boring pests, but the mechanisms underlying this resistance have not yet been characterized. In this study, we investigate the impacts of feeding in a resistant poplar tree on ALB and its gut microbiota, revealing that feeding in this resistant host causes substantial disruptions to the gut bacterial and fungal communities, interferes with the expression of beetle genes with predicted roles in detoxification and interactions with gut microbes, and reduces the abundances of proteins with key roles in digestion.

TOWARDS DEVELOPMENT OF A STANDARD INPUT FOR DETECTION OF MICROBES BY NEXT GENERATION SEQUENCING

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.02)

Rachel Spurbeck, Richard Chou, Nick Fackler, Jazmine Quinn

Battelle Memorial Institute

The utilization of Next Generation Sequencing (NGS) for bacterial pathogen detection is a powerful technique for biosurveillance purposes. However, the field suffers from a lack of standards for validation of NGS pipelines necessary to determine the limit of detection. Currently, there are several different methods for DNA extraction, library preparation, sequencing, and bioinformatics from which a researcher can pick and choose, without a means for determining which is best for their application. Presented here is the initial development of a standard for microbial detection by NGS. This standard is a quantified mixture of bacterial cells, which when sequenced using our chosen method, produce an expected number of reads mapping to the genomes of the bacteria present in the standard mix. To develop this standard, pure cultures of Staphylococcus aureus, Pseudomonas aeruginosa, and Escherichia coli were quantified by spectrophotometry (A600) and spread plates. Pure cultures were then diluted in a titration series from 1010 CFU to 102 CFU, DNA was extracted, and PCR-free libraries were prepared. Prior to quantification and sequencing on the Illumina MiSeq, the libraries were pooled and concentrated. Similarly, two organism mixtures (P. aeruginosa + E. coli, P. aeruginosa + S. aureus, E. coli + P. aeruginosa, and S. aureus + P. aeruginosa) were prepared from the quantified pure cultures with one organism being held constant at 108 CFU and the other organism in a titration series from 108 to 102 CFU. DNA was then extracted from the mixtures, PCR-free libraries prepared, and sequenced. The number of reads that mapped to each genome in the pure culture library titrations or in the mixed samples were quantified. For each of the bacteria in pure culture, read counts plateaued near 105 reads between 106 and 102 CFU. Thus, for individual or low DNA samples, each bacteria was detectable, but not quantifiable below 106 CFU. For TSB negative culture control, read counts were under 100 reads mapping to any of the queried genomes. For mixed cultures, where there was more background DNA to enable efficient reactions during library preparation, the plateau effect below 106 was not observed, consistent with the hypothesis that the presence of a threshold of DNA is necessary for efficient library preparation. With a background present, the coefficient of determination (R2) is 0.9667 with a P-value of 0.000421 for a titration of E. coli in a P. aeruginosa background, demonstrating that there is a correlation between the number of bacteria in a sample and the number of reads mapping to that organism. Future work is necessary to develop the standard further, by including more organisms, and testing the mix in different backgrounds. Using a known mix of organisms in a known quantity will enable true comparison and evaluation of bacterial detection systems and workflows, which is necessary for the general acceptance of pathogen detection results.

ASSESSING THE SINGLE NUCLEOTIDE POLYMORPHISM (SNP) VARIABILITY OF OUTBREAK STRAINS DURING IN VITRO PASSAGE

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.03)

Ashley Sabol

Centers for Disease Control and Prevention

Foodborne bacterial pathogens are responsible for approximately 9.4 million illnesses each year. The mode in which these organisms are passed from environmental or food source to humans can be varied. Furthermore, the rate at which these pathogens change can be identified through molecular subtyping of strains that belong to the same outbreak or source or by passaging them several times in vitro or in vivo. It is important to gain insight on the observed strain variability that may occur due to laboratory manipulations such as culture passages especially after recovery from a patient – in order to accurately distinguish genetic differences (unrelated strains) from background variation resulting from laboratory manipulations. This is particularly important for high resolution methods, such as whole genome sequencing (WGS) which is currently being adopted by the US public health laboratories for routine surveillance and outbreak investigations of foodborne bacterial pathogens.

Four outbreak-associated strains across 4 species Salmonella enterica serovar Newport, Shiga toxin-producing Escherichia coli (STEC) serogroup O157:H7, Vibrio cholerae, and Listeria monocytogenes – were passed 20 times in vitro. Three colony picks were selected at passes 1, 5, 10, 15, & 20 and streaked to blood agar plates. After incubation for 16-24 hours at 37°C, DNA extractions were performed for each pick, DNA libraries were prepared using the NexteraXT kit (Illumina Inc.) and sequenced paired-end on the Illumina MiSeq (500 cycles). The high quality single nucleotide polymorphism (hqSNP) calls were determined using Lyve-Set v1.1.4e (http://github.com/lskatz/Lyve-SET) at 10x coverage, 75% frequency for Listeria and Vibrio, and 20x coverage, 95% frequency for STEC and Salmonella. SNPs clustered closer than 5 bp were filtered.

Over the course of 20 passes, anywhere from 0-6 hqSNPs were observed among the 15 colony picks sequenced for each strain. The higher number of hqSNP differences for each strain were observed among picks from the later passes (i.e. passes 10-20). The least number of SNP differences between picks were observed among the STEC colony picks with 0-1 SNPs observed. The most SNP differences for a single pick were identified for Listeria pass 15, where 4-6 SNP differences were counted when compared to all other picks for that strain.

This study has provided preliminary evidence that WGS data, as measured by hqSNP analysis, remains stable despite of in vitro manipulations and can hence be used to assess genetic relatedness among foodborne bacterial strains during outbreak investigations. Future plans for this study are to expand it to include Campylobacter spp. and an additional Salmonella enterica serotype, as well as exploring different analysis approaches such as whole genome multi-locus sequence typing (wgMLST).

COMPARISON AND CONSOLIDATION OF VIRULENCE FACTOR DATABASES FOR NEW MODULES IN EDGE BIOINFORMATICS

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.04)

Logan Voegtly[1], Chienchi Lo[2], Po-E Li[2], Joseph Anderson[3], Karen Davenport[2], Kimberly Bishop Lilly[1], Theron Hamilton[4], Patrick Chain[2]

[1]Naval Medical Research Center; Henry M. Jackson Foundation, [2]Los Alamos National Laboratory, Los Alamos, NM, [3]Naval Medical Research Center; Defense Threat Reduction Agency, [4]Naval Medical Research CenterMD

Next generation sequencing (NGS) technology allows for not only rapid detection and identification of pathogens from a variety of sample types, but also rapid genetic characterization. Rapid characterization of pathogenic bacteria and their virulence factors from an isolate or a metagenomic sample is critical for threat detection and biodefense. A key component to virulence factors identification is a well curated database. With new virulence factors being discovered on a regular basis it is necessary to use the most up to date and well documented database possible. The EDGE Bioinformatic software package represents a collaborative effort between Los Alamos National Laboratory and Naval Medical Research Center to provide a user-friendly bioinformatic software package that enables rapid NGS data analysis, even when resources are limited. To support the development of a virulence factor identification module in EDGE, we have performed comparisons of the existing virulence factors databases (VFDB, PATRIC_VF, MvirDB, and Victors) to determine which databases are most up-to-date, most accurate and inclusive, and contain useful metadata about the virulence factors. We have condensed two databases, VFDB and PATRIC_VF, into one database, removing redundant entries at the strain level. This database will be used by a new module in the EDGE Bioinformatics software platform to annotate called genes as virulence factors within an isolate. With the use of software like ShortBRED the database will also be used to generate a profile of the virulence factors that are detected in a metagenomic sample. In an alternative approach, this database will also be used with the STRING protein-protein network database as input for a novel bioinformatic approach to determine the probability of sample virulence.

HIGH THROUGHPUT ANALYSIS OF NUCLEIC ACIDS FROM SMALL TO LARGE FRAGMENTS

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.05)

Denise Warzak, Jolita Uthe, Kit-Sum Wong, Steve Siembieda, Jon Hagopian

Advanced Analytical Technologies, Inc.

Quantitative and qualitative assessment of nucleic acids is an essential step in researching a variety of biological and biomedical processes. An important application of nucleic acid analysis is next-generation sequencing (NGS), which necessitates the use of high quality samples. Obtaining reliable measurements of sample integrity is a challenge in and of itself, as there is some variation in standards between different types of nucleic acids. For example, the amount of microRNA present in a sample is crucial when preparing small RNA libraries for NGS. Similarly, measurement of the integrity of un/sheared genomic DNA is important for next-generation long read sequencing. In short, NGS library profiling in a cost and time efficient manner is essential when evaluating if samples are worth the cost of sequencing.

The Fragment Analyzer™ (Advanced Analytical Technologies) has proven to be an indispensable instrument for the reliable qualification and quantification of nucleic acids. When employed in an NGS pipeline, the Fragment Analyzer is capable of assessing all sample types throughout library construction, from the initial assessment of sample integrity through final qualification. It is a flexible, high-throughput platform that employs different kits to analyze a variety of nucleic acids, in a wide range of concentrations and sizes: from a few picograms to several hundred nanograms; microRNA to total RNA; and amplicon/PCR fragments and large DNA fragments to genomic DNA. The versatility of the Fragment Analyzer allows for the efficient qualification and quantification of traditionally challenging nucleic acids for NGS and other downstream applications. In this presentation, we highlight the latest Fragment Analyzer applications to demonstrate the versatility and reliability of the instrument for DNA and RNA analysis.

EVALUATION OF AVERAGE NUCLEOTIDE IDENTITY USING MUMMER (ANI-M) AND RPOB GENE PHYLOGENY FOR IDENTIFICATION OF VIBRIONACEAE BY WHOLE GENOME SEQUENCE ANALYSIS

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.06)

Monica Santovenia[1], Maryann Turnsek[2], Lee Katz[2], Grant Williams[1], Jonathan Jackson[1], Cheryl Tarr[2]

[1]IHRC/CDC, [2]Centers for Disease Control and Prevention

A database that integrates multiple methods for analysis of whole genome sequence (WGS) data for identification of enteric pathogens including Vibrionaceae is being developed at CDC using BioNumerics v7.5 (Applied Maths) as a software platform. We evaluated for possible inclusion two methods for species identification: Average Nucleotide Identity using MUMmer (ANI-m), which provides a pairwise similarity between two genomes; and rpoB gene phylogeny, which places isolate sequences into a larger phylogenetic context with other Vibrio species.

ANI-m and rpoB gene phylogeny were evaluated using >80 Vibrionaceae genome assemblies, representing 15 clinically relevant Vibrio species. Genomic DNA was extracted using the ArchivePureTM DNA Cell/Tissue Kit (5 PRIMETM), and sequencing was performed on the Illumina MiSeq and Pacific Biosciences Single Molecule, Real Time (SMRT) sequencer platfor ANI-m was calculated using in-house developed scripts on a high performance computer. The rpoB gene sequences were aligned and phylogenetic analysis performed in MEGA v5 using the Neighbor-joining algorithm.

The phylogeny based on rpoB gene-sequence variation grouped isolates of each species into clusters that were clearly delineated from other Vibrio species. Overall genome similarity based on ANI-m was generally 95% for members within a species for the clinically-relevant Vibrionaceae. The rpoB phylogeny provides more information about the affinities of different species (e.g. V. navarrensis is closely related to V. vulnificus) but requires expertise in interpreting gene trees. ANI-m provides a more simplistic estimator for delineating species boundaries, but cannot guide additional testing if a query sequence does not match any genomes in a comparative database.

These approaches show promise for identification of Vibrionaceae, and further evaluation and validation on a larger set of genomes is currently ongoing. Also, rMLST, which analyzes variation in the 53 genes encoding the bacterial ribosome protein subunits (rps genes), will be evaluated for possible integration into the database.

THE IMPACT OF ENZYMATIC FRAGMENTATION AND LIMITED LIBRARY AMPLIFICATION ON DATA QUALITY FOR HUMAN WHOLE-GENOME LIBRARIES SEQUENCED ON THE ILLUMINA HISEQ X

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.07)

Maryke Appel[1], Beverley Van Rooyen[1], Jacqueline Meyer[1], Penny Smorenburg[1], Lisa Cook[2], Catrina Fronick[2], Vincent Magrini[2], Robert Fulton[2]

[1]Roche Diagnostics, [2]Washington University

Illumina's HiSeq X platform is making the "$1,000 genome" and large-scale human whole-genome sequencing (WGS) a reality. Robust, high-throughput pipelines for the production of high-quality human genomic shotgun libraries will be needed to sustain the predicted global upsurge in human WGS research. PCR-free library construction workflows are accepted to be the gold standard for these projects, but pose several challenges—including low and variable library yields from limited amounts of input DNA, and library stability issues.

In this study we optimized library construction parameters for libraries to be sequenced on the HiSeq X; using the KAPA Hyper Prep Kit (which requires Covaris shearing), or with the KAPA HyperPlus Kit with integrated enzymatic fragmentation. Different size selection and cleanup strategies were evaluated to achieve optimal library fragment sizes (350 370 bp) and yields (2 nM); and minimize adapter-dimer carry-over.

To address many of the pain points associated with high-throughput PCR-free workflows, we performed an amplification titration (0, 2, 4 or 6 cycles) of NA12878 libraries prepared with optimized KAPA Hyper Prep and KAPA HyperPlus protocols, and sequenced these libraries to ~30X cov- erage (1 library per lane equivalent). The impact of enzymatic fragmentation and low levels of PCR amplification on library QC metrics, library complexity, coverage uniformity and variant call statistics—and the implications for high-throughput library construction pipelines for human WGS research will be discussed.

Products are for life science research use only, not for use in diagnostic procedures.

A ROBUST, STREAMLINED, ENZYMATIC DNA FRAGMENTATION AND NGS LIBRARY CONSTRUCTION METHOD

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.08)

Fiona Stewart Lynne Apone Pingfang Liu Vaishnavi Panchapakesa Christine Sumner Christine Rozzi Deyra Rodriguez Karen Duggan Keerthana Krishnan Bradley Langhorst Joanna Bybee Laurie Mazzola Danielle Rivizzigno Barton Slatko Eileen Dimalanta Theodore Davis

New England Biolabs, Inc.

Sample preparation is quickly becoming a bottleneck in the Next generation sequencing (NGS) pipeline. While NGS libraries can be multiplexed, combined and sequenced together on a single lane, or chip, generating individual libraries is tedious and time consuming. In addition, the multiple steps required to construct a library provide numerous opportunities for errors and sample loss.

In order to increase the throughput of library construction, reduce errors and sample loss, we have developed a robust, flexible, enzymatic fragmentation method. Fragmentation can be combined with end repair and dA-tailing in a single step, or performed independently, to allow use in any NGS platform and optimization prior to library construction. This method is compatible with a broad range of DNA inputs and insert sizes. Libraries generated using this enzymatic fragmentation method with 5 ng and 100 ng of intact DNA show no significant difference in coverage uniformity or distribution from libraries generated with mechanically sheared DNA. Likewise, libraries generated to contain 200 bp and 900 bp inserts show no significant difference in sequence quality from each other or those generated with mechanically sheared DNA. Interestingly and importantly, enzymatic fragmentation generates libraries of substantially higher yields (2-3 fold) than those generated using mechanically fragmented DNA.

The ability to generate high quality NGS libraries from intact DNA without the need for numerous cleanup and liquid transfer steps will substantially reduce the time, cost and errors associated with library construction. In addition, these advances will permit greater use and adoption of NGS technologies into many scientific arenas.

QUALITY ASSESSMENT AND VALIDATION CRITERIA – TOWARDS THE DEFINITION OF TABLE 1.

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.09)

Dominika Borek, Maciej Puzio, Zbyszek Otwinowski

UT Southwestern Medical Center

Although next-generation sequencing provides the means to study properties of nucleic acids on unprecedented scales, concise measures for assessing the confidence of results from NGS experiments are lacking. There are tens to hundreds of statistical indicators available right now which separately provide information about: (1) the quality of the sequencing library and the material that was used to generate it, (2) the performance of the equipment, (3) potential biases in the results, and other important sequencing-related features. However, the average consumer of NGS technology is rarely in a position to efficiently integrate all of this information. This leads to decisions regarding whether an experiment was successful and whether the results are trustworthy frequently being arbi- trary. Comparative and meta-analyses are the area most affected by this; differences are attributed to biological phenomena when they frequently originated from differences in the experimental ap- proach. The lack of transparent validation criteria leads not only to the incorrect or sub-optimal interpretation of results but also to expensive over-sequencing.

We have developed alignment-free metrics that provide transparent and comprehensive validation of NGS experiment results and define the so-called Table 1, which concisely summarizes the quality of an experiment and data analysis so that NGS users and reviewers of publications and grant applications can quickly and yet with high certainty asses the quality of a particular NGS experiment. Our approach is based on data mining of sequencing reads, which includes analysis of overdispersion properties. This is followed by the analysis of residuals to detect whether our models of the experiment are sufficiently complete.

This approach provides partitioning of uncertainty into components related to error sources and estimates the magnitude of each error source. Together, these directly assess the quality of NGS experiments and contribute to the validation of NGS results.

NEXTSEQ V2 VS HISEQ RAPID RUN NGS DATA QUALITY COMPARISON IN THE CONTEXT OF DOWNSTREAM ANALYSIS FOR PUBLIC HEALTH SURVEILLANCE OF FOODBORNE BACTERIAL PATHOGENS

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.10)

Andrew Huang[1], Rebecca Lindsey[1], Blake Dinsmore[1], Jeremy Peirce[2], Charlotte Steininger[1], Peyton Smith[1], Lisley Garcia Toledo[1], Vikrant Dutta[1], Janet Pruckler[1], Kelly Hoon[2], Collette Fitzgerald[1], Heather Carleton[1]

[1]Centers for Disease Control and Prevention, [2]Illumina Inc

Introduction

Over the last four years, the decreasing cost and ease of use of benchtop sequencers has driven adoption of next-generation sequencing by local, state, regional, and federal public health laborato- ries for use in the identification and surveillance of foodborne pathogens. At the same time, there are a variety of sequencing devices and sequencing chemistries on the market. We had previously observed that there were differences in data quality when the same samples were sequenced using MiSeq v2 compared to using NextSeq v1 chemistries, resulting in significant differences in SNP and allele calling downstream. Because these differences could impact downstream bioinformatic analy- ses and outbreak cluster detection, we need to ensure that the same samples sequenced on different sequencers used by public health laboratories perform equivalently in downstream analyses. To this end, we have prepared a standard set of sequencing libraries, run them on HiSeq Rapid Run and NextSeq v2 sequencing chemistries, and compared the basic data quality, as well as wgMLST and hqSNP-based phylogenetic analyses between these sequencers.

Methods

Sequencing libraries were generated from genomic DNA extracted from 32 different strains each of Campylobacter, Shiga toxin-producing Escherichia coli, and Salmonella, using Covaris shearing and the NEB NEXT Ultra chemistry. These libraries were then aliquoted and run on HiSeq Rapid Run 2x250 and NextSeq v2 2x150 sequencing chemistry. To ensure sequencing effort uniformity, the sequencing read sets corresponding to each strain were trimmed to the same read length and down- sampled to a common coverage level prior to comparison. Basic quality analysis was conducted using R and QUAST (http://bioinf.spbau.ru/quast). We then built hqSNP-based phylogenies from the combined sets of data for each species using Lyve-SET pipeline (https://github.com/lskatz/lyve- SET), as well as wgMLST-based phylogenies for each species using Applied Maths' Bionumerics version 7.5.

Results

We noted that both sequencing chemistries had similar quality profiles with an advantage for HiSeq Rapid Run when constructing longer contigs. Despite these differences, the wgMLST analysis from Bionumerics showed no differences in allele calling between HiSeq Rapid Run sequenced samples and NextSeq v2 sequenced samples across any of the 32 samples in each of the three species. While the hqSNP analysis from Lyve-SET showed no SNP differences between HiSeq Rapid Run sequenced sam- ples and NextSeq v2 sequenced samples across any of the 32 samples each in STEC and Salmonella.

And only 5 SNP differences in one sample of Campylobacter when comparing HiSeq Rapid Run sequencing and NextSeq v2 sequencing, while the other 31 samples of Campylobacter showed no SNP differences between HiSeq Rapid Run sequencing and NextSeq v2 sequencing.

Conclusion

Despite small differences in basic data quality scores, similar outcomes were obtained when compar- ing sequencing from HiSeq Rapid Run and NextSeq v2 using phylogenetic analyses with hqSNPs or wgMLST.

ALLELE-MINING RNA-SEQ DATA OF COTTON (GOSSYPIUM HIRSUTUM L.) ROOTS

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.11)

Daojun Yuan[1], Alex Freeman[1], Christopher Hanson[1], Aaron Sharp[1], Sara Greenfiled[1], Lori Hinze[2], Richard Percy[2], Joshua A Udall[1]

[1]Brigham Young University, [2]USDA-ARS South Plains Agricultural Research Center

Cotton genomic resources and nucleotide polymorphism information is useful to understand germplasm diversity, crop domestication, and crop improvement. There are many publicly available resources of cotton; however most resources of gene expression are from fiber tissue and roots have been less studied. In this study, we report individual transcriptomes of 108 G. hirsutum accessions of root tissue, including 41 improved (domesticated) and 67 wild accessions from the USDA Germplasm collection. More than 3.0 billion clean reads were generated. Of the 37,505 genes in the diploid cotton G. raimondii reference genome, 61.4% were expressed in root tissue. Of these, 1,648 and 1,487 genes in improved and wild accessions respectively which were differentially expressed between the two subgenomes. The RNA-seq data was mined for alleles that were polymorphic among the diverse germplasm. The AT-genome had 234,266 SNPs in 25,993 genes, while the DT-genome had 169,123 SNPs in 25,836 genes (5x coverage). After a stricter filtering (<50% missing; 5% MAF), 31,513 and 23,585 high-confidence SNPs were used for population genomics analyses of AT and DT-genomes, respectively. Germplasm diversity, population structure and domestication sweeps were analyzed in each genome of polyploid cotton. These genic SNPs from root tissue will complement with the expanding genomic resources of cotton and provides a valuable resource to future genetic analyses and breeding progra

COMPARISON OF STANDARD CULTURE AND SINGLE-COLONY DNA EXTRACTION METHODS FOR WHOLE-GENOME SEQUENCING OF CAMPYLOBACTER JEJUNI

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.12)

Kun Liu, Karen Jinneman

US FDA

Introduction

Campylobacter jejuni is the most common food-borne bacteria causing diarrhea in the US and is most often associated with poultry, raw milk, unpasteurized eggs, contaminated water or produce. The technical advancement and decrease in cost of whole-genome sequencing (WGS) have made WGS more widely available for rapid and accurate identification of pathogens such as C. jejuni from contaminated food. Since bacterial culture and DNA extraction are critical to WGS, we evaluated the effects of two available methods on WGS data quality and analysis.

Purpose

To compare the quality of C. jejuni genomes prepared from two extraction methods. The subculture method, currently used in FDA, requires more biomass and involves an automated extraction. In contrast, a more rapid protocol needs only a single colony and mechanically releases DNA, which shows potential to greatly reduce culture time and reagent cost. Here, DNA was prepared from both methods, data quality was compared and method applications were evaluated based on WGS results.

Methods

C. jejuni B9 strain was isolated by Pacific Laboratory Northwest in 1983. DNA was extracted using QIAcube (Qiagen Inc., Valencia, CA) from subculture 1 or single-colony 2. DNA concentrations were determined by Qubit HS kit (Thermo Fisher Scientific Corp., Waltham, MA). Libraries were prepared with Nextera XT Kit (Illumina Inc., San Diego, CA). WGS experiments were performed on MiSeq (Illumina Inc., San Diego CA). Genomes were assembled with CLC genomics workbench (Qiagen Inc.). WGS data were evaluated using SpeciesFinder for 16S typing 3 and KmerFinder for speciation 3, 4. Genome annotations were performed with Rapid Annotation using Subsystem Technology (RAST) 5-7.

Results

The subculture method took 48 hours to grow C. jejuni and 2 hours on extraction, whereas the single-colony method took less than 30 minutes to finish all steps. The former yielded higher DNA concentration than the latter (11.4 vs. 0.5 ng/ul), but both were sufficient for library preparation which only requires 0.2 ng/ul.

Qualitatively, B9 was correctly identified to be C. jejuni based on WGS data from both extraction methods and highly homologous to reference strains NCTC-11168 and RM-1221. Quantitatively, de novo assembly from the subculture method yielded an N50 length of 188 kb from 63 contigs. In comparison, that from the single-colony method yielded an N50 length of 72 kb from 244 contigs. Average coverage rates were both greater than 270x. RAST annotations showed similar results. Thedraft genome from subculture was predicted to contain 1751 coding sequences, 315 subsystems, and 43 RNAs, while that from single-colony was predicted with 1753 coding sequences, 312 subsystems, and 43 RNAs.

Conclusions

We compared two DNA preparation methods on a C. jejuni field isolate and evaluated WGS data generated to assess applications for regulatory food safety science. For qualitative analysis following WGS, such as confirmation and speciation, the single-colony method is as good as subculture with a great time-saving advantage (49.5 hours shorter) and reduced cost ($4.60 less per sample). However, for deeper genomics analysis, DNA from subculture exhibits better quality based on N50 and contig counts.

EXAMINATION OF WHOLE GENOME MULTI-LOCUS SEQUENCE TYPING (WGMLST) FOR CAMPYLOBACTER SURVEILLANCE

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.13)

Lavin Joseph, Heather Carleton, Darlene Wagner, Michael Judd, Eija Trees, Vikrant Dutta, Janet Pruckler, Grant Williams, Kelley Hise, Collette Fitzgerald

Centers for Disease Control and Prevention

Background: Campylobacter species are a leading cause of bacterial foodborne illness in the United States. Pulsed-field gel electrophoresis (PFGE) is the current method used within PulseNet USA for surveillance and outbreak investigations. However, PulseNet USA is implementing whole genome sequencing (WGS) for surveillance of all the pathogens it tracks including Campylobacter. In this study, we examined the utility of whole genome multi-locus sequence typing (wgMLST) to cluster or differentiate outbreak-associated and sporadic Campylobacter isolates.

Methods: A BioNumerics 7.5 wgMLST database was developed in collaboration with domestic and international partners for Campylobacter surveillance. Fourteen public health laboratories (PHL) performed real-time WGS on 262 Campylobacter isolates (219 C. jejuni, 35 C. coli, 6 C. lari, 1 C. upsaliensis, and 1 C. fetus) isolates. In addition, 14 C. jejuni isolates from two temporarily associated outbreaks linked to a family reunion (two isolates) and church fundraiser (12 isolates) originating in the same state were sequenced by the local PHL. Sequencing was performed on the Illumina MiSeq using Nextera XT DNA libraries and 2x150 bp or 2x250 bp sequencing chemistry. Sequences with >20X coverage and average quality scores for R1 and R2 >30 were assembled using SPAdes and analyzed using wgMLST (2531 loci). For comparison, the C. jejuni sequences associated with the two outbreaks were also characterized with high quality single nucleotide polymorphism (hqSNP) analysis using the LYVE-SET pipeline (github.com/lskatz/lyve-SET) as well as K-mer analysis using the NCBI Pathogen Detection Pipeline (http://www.ncbi.nlm.nih.gov/pathogens/). PFGE patterns for all isolates were generated using the PulseNet Campylobacter protocol, analyzed in BioNumerics 6.6.10, and named in accordance with PulseNet naming guidelines.

Results: On average, 1547 (915-1752), 573 (250-737), 344 (305-473), 371, and 453 wgMLST loci were identified within C. jejuni, C. coli, C. lari, C. upsaliensis, and C. fetus sequences, respectively. C. jejuni and C. coli isolates presumed to be epidemiologically unrelated (n=20) formed 12 clusters containing indistinguishable SmaI/KpnI patterns; nine of these clusters contained isolates that were also indistinguishable or highly similar by wgMLST (0-5 allele differences). The isolates in the remaining clusters were further differentiated by wgMLST (21-98 allele differences). C. jejuni isolates from the two outbreaks clustered together by wgMLST (0-3 allele differences), hqSNP (0-3 SNP differences), and K-mer analysis even though these outbreaks clustered separately by PFGE.

Conclusion: The majority (82%) of the sporadic Campylobacter isolates indistinguishable by PFGE were not separated by wgMLST. The isolates within each cluster were from the same state and may be from patients exposed to a common source, whereas isolates differentiated by wgMLST could be epidemiologically unrelated. Interestingly, C. jejuni isolates from the two Campylobacter outbreaks clustered together by wgMLST, hqSNP, and K-mer analysis, indicating that they may have origi- nated from the same source that was not discovered during the epidemiological investigations. It is also possible the sequence type associated with these isolates is relatively common. Our study shows that wgMLST appears to be a promising and user-friendly high resolution tool for Campylobacter surveillance that can be implemented in local PHLs.

AN IMPROVED ASSEMBLY ALGORITHM FOR DE NOVO CIRCULAR GENOME RECONSTRUCTION

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.14)

Christian Olsen, Helen Shearman, Richard Moir, Matt Kearse, Simon Buxton, Matthew Cheung, Jonas Kuhn, Steven Stones Havas, Chris Duran

[1]Biomatters, Inc.

Circular chromosomes or genomes, such as viruses, bacteria, mitochondria and plasmids, are a common occurrence in nature, but despite the wide array of algorithms available for de novo assembly, the circularity of these DNA molecules is largely overlooked. There are some limitations of this oversight. Current NGS assembly algorithms assume a linear molecule and will result in a linearly represented genome, with a breakpoint in an arbitrary position. This is becoming increasingly problematic with increasing NGS read lengths resulting in fewer contig assemblies with less ambiguity. Additionally, long read sequencing technologies promise to provide sequences that may greatly span breakpoints at the expense of coverage resulting in a relatively large quanta of information loss.

Although there are currently methods for the re-circularisation of contigs post-assembly by identifying common trailing/leading sequence motifs, we present a more robust approach of circularising during the assembly process whilst still allowing the merging of similar and sub-contigs throughout the overlap-based approach. This method can also combat the issue of chimeric sequence due to contamination, a common problem when sequencing bacterial cultures, due to decreased likelihood of conserved contig ends due to timely circularisation.

We present results from both 28.6 million read Pan troglodytes illumina data set and 267,491 read Panthera leo persica (Asiatic Lion) mitochondrial NGS library produced using an Ion Torrent sequencing machine. These results are discussed and compared to some of the more popular linear assembly algorithms in common usage today.

Geneious R8 is the first bioinformatics software package to offer a circular de novo assembly method. The Geneious Circular de novo assembler is developed by Biomatters and may be found at http://www.geneious.com

PANGENOMICS: PERSISTENT HOMOLOGY FOR PAN-GENOME ANALYSIS

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.15)

Alan Cleary[1], Thiruvarangan Ramaraj[2], Joann Mudge[2], Brendan Mumey[1]

[1]Montana State University, [2]National Center for Genome Resources (NCGR)

The sequencing of multiple accessions per species has brought an opportunity for doing pan-genomic analyses, which can interrogate the unity and diversity within a species from a much broader perspective than looking at a single reference accession. Graph-based pan-genomic approaches have proven valuable but are subject to noise from artifacts which can change with parameters. True biological signal, however, will be more robust to changing parameters. In this work, we use persistent homology to quantify the significance of different features in pan-genomes, revealing true biological signal including evolutionary relationships and guiding further investigation.

Persistent homology is a method for computing topological features of a space at different spatial resolutions. More persistent features are detected over a wide range of resolutions and so are deemed more likely to represent true features of the underlying space. Since pan-genomes can be represented at the nucleotide resolution as de Bruijn graphs, we treat k-mer size as the spatial resolution. The features we track the persistence of are “bubbles” in the graph, which may represent heterozygous or homozygous variations, indels, repeats, or polymorphic sites, among other things.

By varying the k-mer size, we find which bubbles persist the longest within and among genomes. These data can be used to quantify the stability and significance of the features represented by the bubbles and can also define distance between different genomes within the population. Additionally, these data can be used to characterize interesting portions of the pan-genome when sketching a representative visualization, as the graph is typically too dense to draw verbatim. Persistent homology helps to get beyond the noise in the data to identify true branch points in the data, thereby characterizing repeats, identifying more accurate relationships between accessions, pinpointing regions of the genome where distances between accessions differ from general trends (possibly indicating regions of local adaptation), and, conversely, establishing regions of the genome which are syntenic.

USING MOLECULAR MODELS AND SEQUENCES TO UNDERSTAND NEW TECHNOLOGIES: CRISPR/CAS9 IN MOLECULE WORLD™

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.16)

Todd Smith, Sandra Porter

Digital World Biology

CRISPR/Cas9 is a hot topic. CRISPR/Cas9 is not only revolutionizing the field of gene editing, aspects of this system can be used to teach fundamental concepts in immunology and evolution. In terms of immunology, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and restriction/methylation systems protect bacteria from phage. CRISPR sequences are in essence an adaptive immune system for prokaryotes, storing information from previously encountered invaders.

Molecular models provide insights in the following ways. Models illustrate the interaction between CRISPR DNA and viral RNA sequences. They show how Cas9 and related proteins can form complexes to cleave viral RNA. Further, as these systems are widespread in bacteria, models can be used to predict and discover new syste

The excitement (and controversy) around CRISPR/Cas9 makes it a useful example for the classroom. Yet, there is a gap between this cutting edge research and educational use. Basic concepts about immunology, evolution, and inheritance of genetic information can be taught. For example, molecular models and sequence comparisons provide hands on ways for students to explore Cas9-related protein structures from bacteria to learn about differences between homology and convergent evolution. To facilitate classroom use, we have created a collection of CRISPR/Cas9 and related structures and have made this freely available at www.digitalworldbiology.com/dwb/structure-collections. This collection has been met with enthusiasm by biotechnology education programs (e.g. Madison College, WI) that incorporate CRISPR/Cas9 technologies in their curriculum.

AN EXTENDED CORE GENE MLST TARGET IDENTIFICATION AND SUBSET SELECTION PIPELINE FOR CULTURE-INDEPENDENT PATHOGEN SUBTYPING

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.117)

JoWilliams Newkirk[1], Eija Trees[2], John Besser[2], Heather Carleton[2]

[1]IHRC, [2]Centers for Disease Control and Prevention

While isolate-based whole genome sequencing is being rapidly integrated into the US public health surveillance system, isolate availability for surveillance continues to decline as a result of the adoption of culture-independent diagnostic tests by clinical laboratories. As affordable methods are not yet available for obtaining the same genome resolution directly from shotgun metagenomic sequencing of clinical samples, particularly microbially complex samples such as stool, alternative methods are needed to reliably capture genetic information relevant to pathogen subtyping. Targeted amplification and sequencing of informative genomic regions (i.e. multilocus sequence typing, MLST) is a well understood and robust typing method whose resolution is limited only by the number of sites used. Unfortunately, identifying large numbers of informative regions with conserved primer sites is labor intensive, particularly if hundreds or thousands of reference genomes are used for site selection.

To facilitate the rapid development of extended MLST schemes for targeted pathogen groups, we developed a custom pipeline leveraging widely used open source programs to identify potential MLST targets with conserved primers sites and to find subsets of those targets that recapitulate a reference phylogeny (user provided or generated by the pipeline from concatenated core genes). Our pipeline accepts whole genome annotation files from the targeted pathogen group in GenBank (.gbk) format. Core genes are identified by protein BLAST of all annotated open reading frames (ORFs) from a single genome against the ORFs from all GenBank files submitted to the pipeline. Hits are filtered to retain only single copy ORFs which occur in all submitted genomes and are 50% similar across 50% of the query length. Hits found in multiple putative single copy ORF groups are also discarded. The nucleotide sequences for these core ORFs are aligned in Muscle and trimmed to remove end gaps. Up to ten conserved primer pairs producing amplicons of ~250 bp are designed for each alignment in Primer3. The primer pairs and amplicons are filtered to retain only those that do not overlap and capture polymorphisms between input genomes. Users may either retain all passing amplicons or use one of two methods to select an optimized subset for typing. The concordance of the subtyping provided by the selected amplicons to the reference phylogeny can be assessed using a variety metrics, including those that compare the resulting trees (e.g. Kendall-Colijn metric) and those that compare cluster membership (e.g. adjusted Wallace coefficient).

Scripts for the pipeline were written in Python 2.7 and R, and management is provided by bpipe with support for both standard multicore machines and cluster environments. We demonstrate the utility of this pipeline using a collection of 266 Salmonella bongori and enterica genomes representing 68 serotypes.

ANALYSIS OF MICROBIAL FUNCTIONS USING CLC MICROBIAL GENOMICS MODULE

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.18)

Marta Matvienko, Andreas Pedersen

Qiagen Bioinformatics

CLC Genomics Workbench is a user friendly and powerful application for the analysis of NGS data. The package can be extended with a few additional plugins that specifically support the analysis of microbial and metagenomics data.

Microbial Genome Finishing Module (MGFM) automates steps like scaffolding, contig joining, and the ordering of contigs relative to each other or to a closely related reference genome. It also allows for easy and rapid error correction and assembly of PacBio data.

Microbial Genomics Module (MGM) supports three major applications: analysis of microbial compositions using OTU clustering; typing and clustering of bacterial isolates with NGS-MLST, and whole metagenome assembly and its functional analysis.

MetaGeneMark plugin provides the gene and CDS annotations of de novo assembled contigs.

In this presentation, we will cover the whole metagenome-based analysis of functional profiles in drinking water using CLC Microbial Genomics Module and MetaGeneMark Plugin. For the data samples, we downloaded the publicly available NGS sequencing reads (Chao et al, 2013). These sequencing samples came from whole metagenome sequencing using Illumina reads. The microor- ganisms were collected from river water and treated drinking water. We de novo assembled each of the five sequencing samples using the Module's metagenome assembler. The drinking water samples contained significantly fewer contigs with much longer (about 8kb) N50 values than the river water assemblies (N50 of ~870 nt). This may suggest that the metagenome complexity of the drinking water is simpler than the complexity of raw river water.

The Gene and CDS annotation was performed using the MetaGeneMark plugin. For the functional annotations, GO database and Pfam2GO mappings were downloaded directly to Genomics Workbench from the Gene Ontology Consortium. Pfam domains were identified for ~30% of CDS in river samples, and for ~57% of CDS in treated water samples. GO terms were assigned for ~20% of CDS in the river water samples and 42% of CDS in drinking water samples.

To estimate the abundance of functional categories, we remapped the reads to annotated assemblies, and then we built the GO functional profile for each sample. A similar analysis was done using Pfam domains counts. For the comparative analysis, the abundance tables for GO categories and Pfam domains counts were converted to experimental tables. After applying the statistical analysis tools available in the Genomics Workbench, we were able to identify the functions that were eliminated or enriched in the drinking water as compared to the river water. The analyzed data can be simultaneously viewed as a table, heat map, scatter plot, and volcano plot. The domains and functions can be extracted from any of these views. The GO biological process "pathogenesis" term was reduced to zero counts in drinking water. Many other GO functions such as "chemotaxis", "conjugation", and "signal transduction" were significantly more abundant in drinking water. A similar comparative analysis was performed for the Pfam domains abundance.

The described tools provide the gateway to sophisticated functional analysis, and empower users at any level of bioinformatics experience.

AVERAGE NUCLEOTIDE IDENTITY: A FAST WHOLE GENOME SEQUENCE-BASED METHOD FOR SPECIES IDENTIFICATION OF ESCHERICHIA COLI, E. ALBERTII AND E. FERGUSONII

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.19)

Sung Im, Heather Carleton, Lee Katz, Andrew Huang, Rebecca Lindsey

Centers for Disease Control and Prevention

In the context of national reference laboratory strain testing and foodborne pathogen surveillance it is important to quickly and accurately identify the genus and species of unknown bacterial whole genome sequences (WGS). Average nucleotide Identity (ANI) is an in-silico method that calculates the average similarity between two WGS by aligning and comparing the nucleotide bases between two genome assemblies. Two common algorithms are used to compute the ANI between two genomes: NCBI's BLASTn and MUMmer's DNAdiff. While the BLASTn method is computationally more expensive, it provides higher specificity as each WGS comparison is reverse analyzed to provide a two-way reciprocal best hit similarity value. The MUMmer based method calculates the ANI more rapidly using preselected high quality reference genomes for comparison.

In a previous ANI analysis, 170 strains from 5 Escherichia species (E. coli/Shigella, E. albertii, E. fergusonii, E. hermanii & E. vulneris) were compared in an all-against-all pairwise fashion using the BLASTn algorithm (ANIb). This experiment determined that an ANI value of >95% is required to correctly identify an unknown Escherichia genome to its species. The results of this anal- ysis were used to select three complete Escherichia reference strains for the MUMmer based ANI (ANIm) method: Escherichia albertii KF1 (NZ_CP007025.1), Escherichia fergusonii ATCC_35469 (NC_011740.1) and Escherichia coli 08-4006. The accuracy of ANIm was tested with WGS from 100 Escherichia and 100 additional non-Escherichia enteric bacteria previously characterized by traditional methods. Additionally, a down-sampling experiment was conducted to test the coverage depth (from 40x 1x) at which species identification using ANIm became unreliable. For each Escherichia strain tested, reads were systematically removed from the original fastq sequence files. The down sampled read pairs were assembled and re-analyzed against the reference strains.

The initial ANIb analysis, averaging a run time of 4 minutes per comparison, yielded a defined genomic space from which Escherichia strains representing their respective genera were selected. Using the selected reference strains ANIm analysis of 100 Escherichia WGS correctly identified 81 E. coli, 12 E. albertii and 7 E. fergusonii. The 100 non-Escherichia WGS all tested negative for the Escherichia genus while testing positive to their respective organisms. The down-sampling experiment revealed that ANIm was accurate to <5x coverage.

ANIm of WGS was 100% accurate for the 200 strains, even at less than 5x genome coverage. This analysis pipeline allows for the implementation of a fast, reliable species level identification method capable of scaling to high volume reference laboratory strain testing and foodborne disease outbreak detection.

MINING FREQUENT SUBPATHS IN PAN-GENOMES

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.20)

Alan Cleary[1], Thiruvarangan Ramaraj[2], Joann Mudge[2], Brendan Mumey[1]

[1]Montana State University, [2]National Center for Genome Resources (NCGR)

In an era where next generation sequencing has enabled genome sequencing and assembly of multiple accessions per species, the need for pan-genomic algorithms to mine biological information from multiple genomes has become critical. As a pan-genome can be represented at the nucleotide resolution by a de Bruijn graph or at the gene resolution with a directed gene graph, we have implemented an approximate frequent subpaths algorithm to identify syntenic stretches between genomes.

Similar to approximate frequent itemset mining algorithms, our algorithm bounds the minimum number of supporting paths (transactions), as well as the amount of error allowed in each supporting path and the fraction of support for each node (item). Additionally, our algorithm bounds the length of divergent sequences inserted into supporting paths. By exploiting the structure of the graph, our algorithm achieves an efficient running time.

Biological applications of this algorithm include the identification of genes, the inference of phylogenies, the recognition of signatures of local adaptation, and gene clustering, among other things. Additionally, these data can be used to characterize interesting portions of the pan-genome when sketching a representative visualization, as the graph is typically too dense to draw verbatim. This algorithm could be used, for example, to identify core genes for a species, differences in assemblies run with different algorithms, and regions of the genome responding to selection pressures, such as genomic signatures of climate adaptation that can be used to predict future responses to climate change or genomic regions driving specialization such as those driving industrial specialization of yeast.

SYNERCLUST, A TRULY SCALABLE ORTHOLOG CLUSTERING TOOL

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.21)

Christophe Georgescu[1], Alison D Griggs[1], Aviv Regev[1], Ilan Wapinski[2], Brian J Haas[1], Ashlee Earl[1]

[1]Broad Institute, [2]enEvolv

Accurate ortholog identification is a vital component of comparative genomic studies. Popular sequence similarity based approaches, such as OrthoMCL, struggle to cluster or- thologs when there are high rates of paralogs, and although phylogeneticbased methods handle paralogs, they are not sufficiently fast or scalable to work on large sets of whole genomes. Fur- thermore, most approaches do not take synteny into account, which means information useful for distinguishing paralogs is unused. Synergy, originally developed to work on eukaryotic species, uses a hybrid approach to resolve ortholog clusters, relying upon sequence similarity, synteny and phy- logeny. Here, we present Synerclust, a tool that takes the fundamentals of Synergy and adds a number of improvements that retain Synergy's high accuracy, but makes it amenable to ortholog clustering of hundreds to thousands of whole genome data sets, representing either eukaryotic or prokaryotic species. SynerClust bypasses the all vs all Blast requirement inherent to other cluster- ing tools by selecting and comparing cluster representatives at each node in an input species tree. Working from tip to root, SynerClust solves and keeps track of orthology relationships, ultimately providing the most parsimonious solution that takes into account gene gains and losses, common- place in prokaryotes. We have also optimized SynerClust for memory usage and made it amenable for running on many different compute infrastructures.

EXPLORATION OF READ DEPTH AND LENGTH FOR STRUCTURAL VARIATION DETECTION

Wednesday, 1st June 20:00 La Fonda NM Room (1st floor) Poster (PS-1b.22)

Adam English, Jesse Farek, Donna Muzny, William Salerno,
Eric Bowerwinkle, Richard Gibbs

Baylor College of Medicine

Long-read sequencing (>1 kbp) offers more complete genomic information when compared to short-read sequencing (~100 bp), but the accuracy and relatively high cost-per-base limits the practicality of long reads as the sole data source in high-throughput whole-genome sequencing projects. An alternate, more cost-effective strategy is to combine data types, which has been effectively implemented by de novo assembly tools including pacbioToCA and PBJelly. Here we illustrate how SV detection varies with different combinations of sequencing technologies, methods, and coverages.

We first create calls from a haploid cell-line CHM1-tert from PBHoney (PMID: 24915764) from 40x PacBio coverage, 134x/400 bp Illumina data and an independently derived set of PacBio SVs through Parliament (PMID: 25886820), a consolidation SV discovery tool, to generate ~25,000 variant loci, ~9,000 of which are supported by short- and long-read hybrid assembly.

Next, using lower per-data type coverage, we explore SV detection when applied to the diploid human HS1011 using 20x PacBio coverage (i.e., 10x per haploid genome), and multiple coverages and insert sizes of Illumina paired-end sequencing as well as other technologies including aCGH and BioNano Irys optical mapping. These combinations show that PacBio data for evaluation expands the hybrid assembled variants by 42% and PBHoney's PacBio discovery by an additional 46%.

Finally, we evaluate the added value of long-read data of an Ashkenazim trio with ~30x coverage for each parent and ~60x proband coverage. We find a Mendelian consistency rate of 90% for parental homozygous calls and 75% for proband homozygous calls.

By exploring coverage titration points, we have quantified the impact on SV detection of specific combinations of short- and long-read data. Together, these experiments suggest that robust SV detection from whole-genome data can be achieved with hybrid read data at notably low coverages.

EXTENDED VIRAL EXPLORATION OF HEMORRHAGIC FEVER SYNDROMES WITH NEGATIVE EBOLA DIAGNOSIS

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.01)

Ingrid Labouba[1], Andy Nkili Meyong[1], Patrick Chain[2], Maganga Gael[1], Momchilo Vuyisich[2], Tracy Erkkila[2], Eric Leroy[1], Nicolas Berthet[1]

[1]Centre international de recherches médicales de Franceville (CIRMF GABON), [2]Los Alamos National Laboratory

Background:

While in West Africa the hugest Ebola virus disease (EVD) outbreak was in progress, on 26 August 2014 the WHO reported another one in the North-West of the Democratic Republic of Congo (DRC), in the district of Bouende (Equateur province) located at 1200km of North of Kinshasa. Since the virus was discovered in 1976, this was the 7th in this country. On 7 October 2014, a total of 69 cases (3 suspected, 28 probable, 38 confirmed) and 49 deaths (21 males, 28 females) have been recorded. 33 Blood samples collected from those patients were sent to the CIRMF for diagnosis confirmation and complementary investigation. Based on qPCR specific diagnosis, 7/33 cases were found Ebola Zaire positive (Maganga et al., NJEM 2014). 26/33 remaining samples were found negative for Ebolaviruses (Zaire, Bundibyo, Sudan species) and Marburg virus specific diagnosis as well as for generic Pan Filoviruses PCR. These then constitute a cohort of cases with an unknown etiology that require a deeper exploration. Here we suggest to investigate the presence of potential infectious causal agent able to be associated with hemorrhagic fever syndrome (HFS) in those patients found negative for any filoviruses.

Methods:

Filovirus-negative whole blood samples were stored at -80°C until their total RNA extraction initiated under high security conditions in BSL-4 and completed in BSL-3 laboratories. After a ribosomal RNA removal, remaining RNA will be used for preparing DNA libraries directly and after a whole genome amplification (WTA) step to increase chances of viral pathogen detection by high throughput sequencing. In parallel, we will proceed to the isolation of potential pathogens by cell culture or mouse brain inoculation. Positive generated isolate or inoculate will also be treated for high throughput sequencing. Subsequent analyses will be performed by bioinformatician using CLC and EDGE softwares.

Expected results and perspectives:

Here we aim to identify the other potential infectious causal agents, different from Ebola and Marburg viruses that stroke during 2014 DRC outbreak. This project also aims to implement a standardized procedure for an extended and complete exploration of clinical syndromes with unknown etiologies. Currently dedicated to this small cohort of HFS patients, it will be used in the long term on entire CIRMF's biobank and more.

Funding Sources:

Defense Threat Reduction Agency

Agence Nationale de la Recherche FRANCE

EVALUATION AND OPTIMIZATION OF THE ILLUMINA NEXTSEQ 500 SYSTEM FOR NEXT GENERATION SEQUENCING (NGS)-BASED SURVEILLANCE OF INFLUENZA

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.02)

Shoshona Le, Thomas Stark, Samuel Shepard, Elizabeth Neuhaus, David E Wentworth, John Barnes[1]

Centers for Disease Control and Prevention

Background-

The CDC Influenza Division (ID) currently utilizes Next Generation Sequencing (NGS) techniques to conduct genetic surveillance of circulating influenza virus strains. As a World Health Organization (WHO) Collaborating Center, ID receives between 8,000 and 10,000 specimen submissions per year for surveillance of influenza. ID has implemented a high throughput pipeline to meet this demand that is based on processing 96 influenza samples/controls per run across two Illumina MiSeq instruments, to generate all of the sequencing data used for influenza genetic characterization. For the Northern Hemisphere influenza 2015-2016 season (October-April), ID generated complete influenza genomes of over 3,500 viruses using this NGS pipeline, totaling more than 30,000 genome segments available for influenza analysis. During the peak of the influenza season, the high volume of influenza samples resulted in 12 MiSeq runs during a single month, with a large number of specimens still awaiting processing. The Illumina NextSeq 500 was procured in an effort to mitigate this issue, as it has 10-15 times more data output of the current MiSeq in roughly the same run time. Because the NextSeq instrument differs significantly from the MiSeq instrument in the method of measuring fluorescence signals from the flowcell, assessments were conducted to determine whether the NextSeq would have any adverse effects on the influenza genetic data upon integration into the ID pipeline.

Methods-

To compare the data quality between the MiSeq and the NextSeq 500 instruments, libraries that had been prepared and previously run on the Miseq were re-sequenced on the NextSeq, utilizing a loading workflow that was modified according to Illumina specifications. Pools of 96 samples that had been barcoded using an Illumina single index kit were run, then expanded to quad-pools of 384 samples that had been barcoded utilizing Illumina Nextera A, B, C, and D 96-index kits. Quality of the data and quantity of the reads were compared between the instruments, as well as consensus sequences, subpopulations, and coverage of the influenza genomes sequenced. Further comparisons between NextSeq v1 and NextSeq v2 chemistry were also examined.

Conclusions-

The NextSeq 500 System offers a cost efficient NGS-based strategy for increasing the capacity of influenza NGS pipelines, and likely other viral pipelines, without sacrificing data quality or quantity. Its optimizable workflow, capability to provide a 10-fold increase in data generation, and capacity to quadruple sample number per sequencing run, yet improve coverage enables high throughput sequencing to meet the demand for global surveillance of the influenza A and B viruses at less than half the cost of multiple MiSeq runs.

BACTERIAL PATHOGEN NEXT-GENERATION SEQUENCING DATA TRIMMING, CORRECTION, AND SNPS DISCOVERY

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.03)

Darlene Wagner[1], Lee Katz[2], Eija Trees[2], Heather Carleton[2]

[1]IHRC, [2]Centers for Disease Control and Prevention

Background: Next-generation sequencing (NGS) allows rapid in-house sequencing of bacterial strains implicated in local or multi-state outbreaks. Single-nucleotide polymorphisms (SNPs) analyses incorporating NGS reads can aid outbreak cluster identification. This study evaluated effects of read trimming and correction (healing) on SNPs analysis quality.

Methods: NGS reads representing outbreak clusters from Salmonella enterica serovar Bareilly, Shiga toxin-producing Escherichia coli (STEC) serogroup O157, and Campylobacter jejuni, were cleaned using nine healing methods. Methods used to implement healing included prinseq, fastx_trimmer, BayesHammer, BayesHammer with fastx_trimmer, Musket, Quake, Blue, CG-Pipeline with quality trimming, and CG-Pipeline with quality cutoff masking. Forward-read (R1) and reverse-read (R2) errors were estimated through base-call qualities (Phred) and ambiguous nucleotide (N) counts. SNPs discovery through Lyve-SET (https://github.com/lskatz/lyve-SET) was assessed by counting informative aligned positions. Possible false positive/negative SNPs were inferred by counting SNPs shared across results of the different healing methods.

Results: R1 and R2 reads from Salmonella ser. Bareilly averaged 300,000 ambiguous nucleo-tide reads while R2 reads exhibited average Phred scores as low as 25.8 (median 29.6). Un-healed Bareilly reads yielded 56 SNP positions through Lyve-SET. BayesHammer, an edit-distance-based method, and kmers-based methods, Quake and Blue, raised Phred scores in the Bareilly set above
30.0. Blue, along with Musket, another kmers method, increased Bareilly SNP positions to 122 and 123, respectively, with 1 potential false positive site each. BayesHammer, fastx_trimmer, and BayesHammer with fastx_trimmer increased Bareilly SNP positions to 94, 99, and 110, respectively, without false positives. The STEC O157 outbreak clus-ter R2 reads exhibited median quality of 29.3 with up to 1.07x106 reads containing ambiguous nucleotides. Unhealed O157 reads yielded 62 SNP positions, which increased to 66 positions with no false positives after healing through Quake. R1 and R2 reads of the Campylobacter cluster had quality scores well above 30.0 but with an average of 11,000 ambiguous-nucleotide reads in R2. Unhealed Campylobacter reads yielded 137 SNP positions while fastx_trimmer, Blue, and BayesHammer with fastx_trimmer increased SNPs to 150, 152, and 153, respectively, with no inferred false positives. Across all three organism/outbreak sets, prinseq failed to in-crease SNP counts, while the CG-Pipeline-based methods reduced SNP counts in the Bareilly and Campylobacter sets. Musket, Quake, and Blue, consistently increased SNP counts, but each produced possible false positive/negative SNP sites in at least one organism set.

Conclusions: An optimal read trimming or cleaning method should increase the number of SNP positions without adding false positives, thus enhancing outbreak phylogeny. This study has shown that Musket, Blue, and Quake increase numbers of SNP positions for NGS reads with high ambiguous base-calls and Phred scores < 30.0. Yet, the kmers-based methods occasionally introduced SNPs not shared across methods within the organism set, indicating possible false positive/negative positions. BayesHammer, fastx_trimmer, and the two methods combined all increased counts of SNPs which were reproducible across more than one healing method. For future studies, outbreak sets with validated SNP sites will be used for more accurate assessment of false discovery or false exclusion of SNPs.

IDENTIFYING STRUCTURAL VARIATION, COMPONENT ISSUES AND OTHER SEQUENCE ARTEFACTS BY INTEGRATING LONG RANGE GENOME MAPS IN A WEB-BASED GENOME BROWSER

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.04)

William Chow, Kerstin Howe

Wellcome Trust Sanger Institute

The web-based genome evaluation browser gEVAL (geval.sanger.ac.uk) has been used extensively to curate and improve assemblies from reference consensus standards to novel draft genomes. However despite the wealth of evidential datasets available to aid in improving assemblies to gold standard such as clone ends, genetic markers or transcript models, there are still difficult regions in the genome that can't be sequenced, finished and resolved.

The single molecule genome mapping technology developed by Bionano Genomics, provides unbiased physical maps indicative of the genomic structure. By comparing these maps to an assembly, it can aid in both long range assembly correction as well as structural variant detection.

gEVAL has integrated and aligned genome map datasets against the various human, mouse and zebrafish assemblies presently available in the browser. Whilst we also provide trackhubs (ngs.sanger.ac.uk/production/grit/track_hub/hub.txt), gEVAL's representation of the aligned data was developed to allow users to easily identify regions of discordance between map(s) and assembly. This has been helpful in quickly sizing sequence gaps and resolving problematic regions such as those caused by repetitive elements.

gEVAL hosts in-house generated maps, as well as datasets for 4 human individuals (one trio plus one individual) from the Genome In A Bottle initiative (https://sites.stanford.edu/abms/giab). This aids not only the improvement of the human reference genome, but also provides a resource to help in identifying potential structural variation.

VISUAL ANALYTIC TOOL FOR INVESTIGATING UNEXPLAINED RESPIRATORY DISEASE OUTBREAKS THROUGH THE USE OF TARGETED AMPLICON RE-SEQUENCING

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.05)

Shatavia Morrison, Heta Desai, Bernard Wolff, Alvaro Benitez, Maureen Diaz, Jonas Winchell

Centers for Disease Control and Prevention

The infectious etiology of respiratory outbreaks remains unidentified in ~50% of Unexplained Respiratory Diseases Outbreaks (URDO) investigated by CDC, even when traditional multi-pathogen detection methods, such as real-time PCR are used. The current URDO panel provides the basic level of detection (presence vs. absence) for a known set of respiratory etiologic agents. By transforming the foundation of this approach to a next generation sequencing-based (NGS) method, we are able to generate a more comprehensive data set for all targeted agents that may be present in clinical specimens, along with the ability to identify novel or rare pathogens. A continued challenge in the "–omics" assays is the data analysis step. Existing metagenomics visualization tools are either too convoluted for user interpretation, non-interactive, or lack the ability to display parent-child taxonomy relationships.

We present an interactive visual analytic tool that allows for analyses of multiple metagenomics datasets simultaneously in order to identify the etiologic agent(s) during URDO outbreaks. By using this tool with in-house mock clinical samples, we were able to detect the seeded organisms. These datasets were generated with the Illumina platform. After sequence read data cleansing, Kraken was used to assign taxonomic labels using a k-mer based approach. The Kraken output was used as the input into the URDO visualization. This visualization has two components: (i) the back-end infrastructure and (ii) the front-end web-visualization component. The back-end consists of a MySQL database on a MAMP web server and PHP scripts. The MySQL database contains the parent-child relationships for all the bacteria and virus taxonomy lineages classified in NCBI and PHP scripts that allow for the communication between the back-end and front-end visualization. The front-end visualization was built with HTML5 and D3.js. The D3.js is a data-driven framework that allows for easy manipulation of text data into visual glyphs. It allows the user to upload multiple files and generate independent visual interpretations for each dataset as well as generate a cumulative view of all samples to allow for inter-level comparisons. The web page navigation is based on taxonomy classification hierarchical schema. This allows the user to select at which taxonomy level they would like to begin their analysis. The visualization is divided into two visual panes for easy interpretation of parent-child relationships. Using this approach we are able to display many- to-many comparisons. The size of the bubbles represent the number of reads associated with each organism. The user is able to hover over the bubble glyph to retrieve data such as the percentage of reads represented in the dataset. The summative view allows the user to highlight a specific classification unit and its corresponding units will highlight across samples to make it easy for the user to identify any specific classification group of interest.

This tool is useful for exploring metagenomics datasets. The framework joins scalable technologies together to make it modular to incorporate additional characterization features such as antibiotic resistance features, serogrouping, or receptor use. Ultimately, this tool will vastly improve URDO investigations.

DETERMINATION AND CHARACTERIZATION OF CLINICALLY OBTAINED VIRAL STRAINS VIA NEXT GENERATION SEQUENCING AND POST SEQUENCING BIOINFORMATIC ANALYSIS

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.06)

Brianna Mulligan[1], Walter Dehority[1], Kurt Schwalm[1], Stephen Young[2], Darrell Dinwiddie[1]

[1]University of New Mexico, [2]TriCore Reference Laboratories, Albuquerque, New Mexico

Biological resource centers (BRCs) serve contemporary life sciences by collecting, archiving, updating, and integrating a variety of research data. Researchers on the individual and institutional level can then freely access that information through user-friendly interfaces with computational analysis tools as an essential resource. In 2006, the World Data Center for Microorganisms had over 500 BRCs registered, but currently there are only 32 BRCs dedicated solely to virology. The amount of web-accessible information available to virologists is significantly less than other fields of study, and furthermore, often there is inefficient linkage between these databases and to larger databases such as Genbank or PubMed. Accordingly, we sought to develop tools and a database that will simultaneously integrate virus sequence data to identify viral strains, and characterize and annotate genomic variation while enabling researchers to connect with additional available analytical tools, and integrate information to pertinent BRCs in rapid and high throughput manner. We are currently developing our toolset and database utilizing complete and nearly complete genomes obtained using next generation sequencing of samples from 102 patients in New Mexico with respiratory syncytial virus infections. Implemented as a web interface, our database intends to accomplish this integration to BRCs by a knuckles-and-nodes approach which will accommodate expansion to 30 additional respiratory viruses. These tools include viral strain differentiation through application of the Needleman–Wunsch algorithm, identity scoring, and weighting of least variable segments within the aligned sequences of the collected viruses. Our database will enable researchers and clinicians to conduct rapid and efficient genomic and epidemiologic examination of clinical viral strains, which can provide critical insight into the pathogenesis of infection and will ultimately lead to an improved clinical understanding of the disparate clinical outcomes seen in acute pediatric respiratory viral infections.

USAGE OF SEQUENCE INDEPENDENT SINGLE PRIMER AMPLIFICATION NEXT GENERATION SEQUENCING (SISPA NGS) FOR WHOLE GENOME SEQUENCING OF HANTAAN VIRUS

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.07

Daesang Lee[1], Dong Hyun Song[1], Won-keun Kim[2], Se Hun Gu[1], Jeong-Ah Kim[2], Seung-Ho Lee[2], Jin Sun No[2], Prof. Jin-won Song[2], Seong Tae Jeong[1]

[1]Agency for Defense Development, [2]Korea University

Hantavirus, a family of bunyaviridae, is a single-stranded, negative sense RNA virus containing tripartite genomes. Hantavirus infection causes hemorrhagic fever with renal syndrome (HFRS) and hantavirus pulmonary syndrome (HPS) with the mortality rate of 1-36%. The outbreak or endemic infection of hantavirus are a critical threat for world public health because of the lack of effective prophylactic and therapeutic strategies. In addition, according to the Center for Disease Control and Prevention (CDC), hantavirus belongs to Category C of the bioterrorism agent categories, suggesting it is a possible biological agent to threat the national security because of easy production, rapid transmission and less understandable pathogenesis. Recently, the usage of NGS provides a potential tool to acquire whole viral genome sequence for the advance of diagnostics and vaccines. Here, based of Sequence Independent Single Primer Amplification Next Generation Sequencing (SISPA NGS), we have attempted whole genome sequencing of Hantaan virus (HTNV), the prototype of hantavirus and a etiology of HFRS, from endemic areas in the Republic of Korea (ROK). In this study, SISPA NGS applied to the whole genome sequencing and the detection of HTNV tripartite genomes from the isolates. The results showed that whole genome sequences of seven HTNV strains was recovered by the NGS and validated by phylogenetic analysis with hantavirus. Thus, SISPA NGS will be a robust tool for the accurate genomic-based diagnosis and whole genome sequencing of hantaviruses in natural reservoirs as well as HFRS and HPS patients.

GENOME SEQUENCE OF RIFT VALLEY FEVER VIRUS ISOLATES FROM THE 2008-2010 DISEASE OUTBREAK IN SOUTH AFRICA

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.08)

Prof. Phelix Majiwa[1], Rachel Maluleke[2], Alison Lubisi[2], George Michuki[3], Maanda Phosiwa[4], Phemelo Kegakilwe[5], Steve Kemp[3]

[1]ARC Onderstepoort Veterinary Institute and University of Pretoria, Faculty of Veterinary Science, [2]ARC-OVI/University of Pretoria, [3]ILRI, [4]ARC-OVI, [5]Veterinary Services, Northern Cape

Changing climate and increasing pressure to produce food sufficient for the increasing human population lead to higher frequency of contacts between humans and animals particularly in the developing countries. This leads to greater likelihood of disease spill-over from animals to human, for it is known that a majority of infectious disease of man in the developing countries originate in animals (Foresight, Infectious Diseases: Preparing for the Future (Office of Science and Innovation, London, 2006)). One such disease is Rift Valley fever (RVF), transmitted from infected animals to man by mosquitoes. Humans become infected by Rift Valley fever virus also through contaminative contact with tissues of infected animals. The disease causes major loses in livestock and has significant negative impact on the livelihoods of livestock keepers. The virus responsible for this disease has a potential for dual use for inhumane purposes.

Rift Valley fever is a re-emerging zoonotic disease. Recent outbreaks (2008, 2009 and 2010) in South Africa occurred unexpectedly and with increased frequency. As of August 2010, there were 232 human cases, with 26 confirmed deaths.

In order to obtain comprehensive information on the genetic composition of the RVF viruses (RVFVs) circulating in South Africa, genome sequence analyses was undertaken on RVFVs isolated from samples collected over time from animals at discrete foci of the outbreaks, with emphasis on isolates from 2008-2010 outbreaks. The genome sequences of these viruses were compared with those of the viruses from earlier outbreaks in South Africa and elsewhere.

Representative virus isolates from the 2008-2010 RVF outbreak cluster into two distinct clades, one similar to clade C and the other to clade H previously described by Grobbelaar et al (Grobbelaar AA, Weyer J, Leman PA, Kemp A, Paweska JT, Swanepoel R. Molecular epidemiology of Rift Valley fever virus. Emerg Infect Dis. 17:2270-6, 2011). The isolates from outbreaks of 1955, 1974 and 1975 all which fall into a third cluster designated Z.

A majority (87.5%) of 2008 outbreak isolates are in Clade C (Grobbelaar et al +2011), which contains only one isolate from the 2009 outbreak. A majority of the 2009 (33%) and 2010 (67%) outbreak isolates clustered in clade H.

Rift Valley Fever viral genome sequences determined by Bird et al (J Virol. 81:2805-16, 2007), included in the analyses for comparison, did not cluster with the 2008-2010 isolates from the disease outbreaks; instead, they clustered with the isolates from 1955, 1974 and 1975 South African outbreaks. The data will be presented and discussed in the context of phylogenetic relationships among the isolates, and implicit recombination, if any, in the genes encoding glycoproteins Gc and Gn, which have a role in protective host immune response.

ANALYSIS OF MITOCHONDRIAL GENOME ISOFORMS IN LETTUCE

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.09)

Alexander Kozik, Jie Li, Dean Lavelle, Sebastian Reyes Chin Wo, Richard Michelmore

Genome Center, University of California, Davis, CA

Mitochondria are integral organelles of most eukaryotic cells that have distinct genomes from the main nuclear one. Plant mitochondrial genomes are much larger than mitochondrial genomes of animals and fungi. Frequently, plant mitochondrial genomes also possess fragments of chloroplast genomes. Interchange between mitochondrial and nuclear genome segments are common phenomena in plants. Moreover, mitochondrial genome of a particular plant species may consist of several structural isoforms that coexist simultaneously. These isoforms are represented by the shuffling of large basic DNA sequence blocks separated by short repeats. The complexity of plant mitochondrial genomes consequently makes determining their sequence and structure challenging. There are fewer annotated plant mitochondrial genomes at NCBI GenBank than completed animal mitochondrial sequences. Only approximately 100 mitochondrial genomes are available for vascular plant species http://www.ncbi.nlm.nih.gov/genome/browse/?report=5. As a part of Lettuce Genome Sequencing Project https://lgr.genomecenter.ucdavis.edu/ we have assembled and characterized the lettuce mitochondrial genome. We extracted mitochondrial components from the Illumina and PacBio libraries used to assemble the lettuce nuclear genome and assembled them into a complete mitochondrial genome using a multistep approach. Initial assembly with CLC Genomics Workbench, Velvet, and FALCON generated multiple contigs that appeared to be the basic building blocks of several isoforms. The contact frequency for contig terminals with Illumina mate libraries, and contig bridging with PacBio reads were used to assemble the contigs into circular graphs. The total unique sequence of the lettuce mitochondrial genome is ~ 300 Kb. Multiple circular isoforms of the lettuce mitochondrial genome were identified and annotated.

THE EXOME COVERAGE AND IDENTIFICATION (EXCID) REPORT: A GENE SURVEY TOOL FOR CLINICAL SEQUENCING APPLICATIONS

Wednesday, 1st June 20:00 La Fonda Mezzanine (2nd Floor) Poster (PS-2b.10)

Rashesh Sanghvi, Kimberly Walker, Qiaoyan Wang, Harsha Doddapaneni, Yi Han, Huyen Dinh, Eric Boerwinkle, Donna Muzny, Richard Gibbs

Baylor College of Medicine

The Exome Coverage and Identification (ExCID) Report was developed at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) to represent gene, transcript and exon sequence coverage for samples analyzed with panel sequencing, WES and WGS. Since March 2013, the report has been used to analyze more than 22,888 WES, 9,428 regional captures and 32,660 WGS research samples for the BCM-HGSC and more than 23,647 WES/Carrier/CAP-CLIA samples at Baylor Miraca Genetics Laboratories (BMGL).

ExCID is a one-stop tool for providing annotations, gathering mapping metrics such as %reads mapped and %duplicates reads, coverage metrics such as average coverage and bases 20X, gene% coverages for OMIM genes and reporting exact bases that are poorly covered based on user-defined coverage threshold. ExCID provides gene, transcript and exon annotations to the input BED regions from RefSeq, VEGA, CCDS and MIRbase. These external databases are downloaded with the installation of ExCID and can be updated using the scripts provided in the installation.

The user can visualize coverage over the region of interest using the UCSC tracks (WGS and BigBed) and the OMIM Gene % coverages bar chart. Results are reported as boxplots and other 'coverage tracks' that can be visualized in popular browsers such as the Integrative Genome Viewer (IGV) and the UCSC Genome Browser. ExCID output was utilized by BMGL website to provide clinicians easy access to in-depth information on clinical design aspects and coverage.

In addition, ExCID's batch ability makes it possible to compare data from hundreds of samples to reveal trends in the performance of large scale sequencing projects and asses the consistency of sequencing strategies. This can be used to determine the common problematic regions across samples leading to design improvements and unique solutions. Analyzing clinical WES samples led to technical developments such as capture reagent spike-in (i.e. panel killer) and improving coverage of nearly all clinically targeted genes (N=3,200 at 100% coverage). Furthermore, ExCID is used for evaluating the efficiency of in-house pipelines to meet requirements and set thresholds for various coverage metrics. ExCID's ability to asses the percentage of gene covered is routinely used both clinical and research setting to samples with irregular performance. ExCID has also been used by the Clinical Sequencing Exploratory Research (CSER) Sequencing Standards working group, a consortium of ten clinical sequencing centers across the country, to determine the regions of the genome that remain difficult to sequence, regardless of the method used. Establishment of robust WGS pipeline metrics are essential for broad applications in both the research and clinical setting.

THURSDAY, 2ND JUNE

07:30 - 08:30 Breakfast (Sponsored by New England Biolabs)

08:30 - 08:45 Welcome Introduction and Opening Remarks

08:45 - 09:30 Dr. Pevzner, Keynote Address (Sponsored by Kapa Biosystems)

09:30 - 10:00 Dr. Borodovsky, Invited Speaker

10:00 - 10:20 Oral Session 4: Bioinformatics - Assembly and Analysis
Chaired by: Bob Fulton and Patrick Chain
OS-4.01 :: Estimating the effects of repeats on assembly contiguity

10:20 - 10:40 Coffee Break (Sponsored by 10x Genomics)

10:40 - 12:00 Oral Session 5: Bioinformatics - Assembly and Analysis
Chaired by: Donna Muzny and Alla Lapidus
OS-5.01 :: Nucleotid.es - objective benchmarking of bioinformatics software
OS-5.02 :: Automation and Management of Influenza Next-Generation Sequencing Data Analysis and Quality Control Procedures
OS-5.03 :: Distinguishing Outbreaks of Botulism Using a Reference-based SNP Analysis and High Quality Reference Genome Sequences of Clostridium botulinum
OS-5.04 :: Importance of WGS for identifying environmental sources of Legionella pneumophila during outbreak investigations of Legionnaire's Disease

12:00 - 13:20 Lunch Break (Sponsored by MRIGlobal)

13:20 - 15:30 Oral Session 6: Forensics (Human and Microbial)
Chaired by: Minh Nguyen and Michael Fitzgerald
OS-6.01 :: Targeted CRISPR/Cas9 DNA fragmentation and selective primer sequencing enables massively parallel microsatellite analysis
OS-6.02 :: Advancing Molecular Diagnostics in Sudden Unexplained Death – a New York City Experience
OS-6.03 :: Assessment of Single Nucleotide Polymorphism (SNP) Genotyping Chemistries and Comparison of Phenotypic and Ancestry Prediction Tools for Forensic Investigative Leads
OS-6.04 :: RIGEL: Analysis System for Microbial Attribution, Bioforensics and Biosurveillance
OS-6.05 :: Massively Parallel Sequencing Technologies for Expanded DNA Identification Capabilities at the Federal Bureau of Investigation Laboratory
OS-6.06 :: The Effect of Storage Time and Temperature on the Recovered Microbiome from Pubic Hairs
OS-6.07 :: National Institute of Justice Funding Opportunities

15:30 - 15:50 Coffee Break (Sponsored by Becton Dickinson)

15:50 - 17:45 Tech Time Talks 2
Chaired by: Johar Ali and Kenny Yeh
TT-2.01 :: Direct determination of genome sequences
TT-2.02 :: Single-Molecule Mapping with Solid-State Detectors
TT-2.03 :: One Codex: Accurate, robust, and sensitive tools for applied microbial metagenomics
TT-2.04 :: A powerful method for comprehensive structural variation detection with short reads
TT-2.05 :: Analysis of The Functional Contents of Microbial Communities Using a Novel QIAGEN Bioinformatics Pipeline
TT-2.06 :: Geneious: a bioinformatics platform for biologists
TT-2.07 :: Artemisinin Drug Resistance Workflow for Plasmodium falciparum
TT-2.08 :: Towards building complete genome assemblies using BioNano Next-Generation Mapping technology
TT-2.09 :: De Novo Assembly and Structural Variation Detection of Human Genomes using Single Molecule Next-Generation Mapping and SV Call Validation by Inheritance and Orthogonal Measurements

18:30 - 20:30 Happy Hour(s) at Cowgirl Café, Sponsored by Illumina

ANTIBIOTICS DISCOVERY: FROM GENOME SEQUENCING TO GENOME MINING TO SPECTRAL NETWORKS

Thursday, 2nd June 8:45 La Fonda Ballroom Keynote Address (KN-2)
Sponsored by Kapa Biosystems

Dr. Pavel Pevzner

University of California, San Diego

Genomics studies revealed numerous antibiotics-encoding genes across a wide range of bacterial and fungal species, including various species in the human microbiome. However, little is known about the hundreds of secondary metabolites (including antibiotics) produced by microorganisms in the gut, despite the fact that humans are chronically exposed to them. Deep exploration of this meta-antibiome critically depends on a transition from the current one-off process of antibiotics analysis to a high-throughput antibiotics sequencing. I will discuss recent advances in computational antibiotics discovery that span bioinformatics techniques ranging from genome sequencing to genome mining to spectral networks.

Speaker's biographical sketch

Dr. Pevzner is Ronald R. Taylor Distinguished Professor of Computer Science and Director of the NIH Technology Center for Computational Mass Spectrometry at University of California, San Diego. He holds Ph.D. (1988) from Moscow Institute of Physics and Technology, Russia. He was named Howard Hughes Medical Institute Professor in 2006. He was elected the ACM Fellow (2010) for "contribution to algorithms for genome rearrangements, DNA sequencing, and proteomics" and ISCB Fellow (2012). He was awarded a Honoris Causa (2011) from Simon Fraser University in Vancouver. In 2015, he founded the Center for Algorithmic Biotechnology at Saint Petersburg State University, Russia. Dr. Pevzner has authored textbooks "Computational Molecular Biology: An Algorithmic Approach" in 2000, "Introduction to Bioinformatics Algorithms" in 2004 (with Neal Jones), and "Bioinformatics Algorithms: An Active Learning Approach in 2014 (with Phillip Compeau). His latest textbook has become the basis of Bioinformatics Specialization at Coursera, a series of Massive Online Open Courses with over 230,000 students enrolled in the last 2 years.

ALGORITHMIC SOLUTIONS FOR HIGH ACCURACY GENE FINDING IN JUST ASSEMBLED NGS GENOMES

Thursday, 2nd June 9:30 La Fonda Ballroom Invited Speaker (IS-2)

Mark Borodovsky

Georgia Institute of Technology

Gene prediction and annotation plays central role in genomics. However, in spite of much attention, open problems still exist and stimulate a search for new algorithmic solutions in all categories of gene finding.

Prokaryotic genes can be identified with higher average accuracy than eukaryotic ones. Nevertheless, the error rate is not negligible and largely species-specific. Most errors are made in prediction of genes located in genomic regions with atypical G+C composition. I will talk about our efforts to improve GeneMarkS, a self-training tool used in many genome projects. The new tool, GeneMarkS-2 (Tang et al., submitted), uses local G+C-specific heuristic models to make initial predictions of atypical genes that serve as 'external' evidence in subsequent self-training iterations. Unlike the current GeneMarkS the new tool makes adjustments of the model structure within the self-training process. In multiple tests we have demonstrated that the new tool is favorably compared to the existing gene finders.

We also report progress in developing tools for structural annotation of eukaryotic genomes. We have constantly updated the self-training ab initio gene prediction tool, GeneMark-ES. Recently, it was extended to fully automated GeneMark-ET (Lomsadze et al., 2014) that integrates information on mapped RNA-Seq reads as well as extension to GeneMark-EP (Lomsadze et al., in preparation) that uses initial self-training and gene prediction to generate external evidence in terms of genomic footprints of homologous proteins.

For *ab initio* gene prediction in fungal genomes we have developed fungi specific self-training methods. The constantly updated fungal version of GeneMark-ES has been used in a number of DOE JGI and Broad Institute fungal genome sequencing projects since 2007.

Our metagenomic gene finder, MetaGeneMark (Zhu et al., 2010) employed in IMG/M for metagenome annotation and conventionally used for analysis of bacterial and arhaeal sequences was further developed to predict genes in fungal metagenomes.

Finally, we describe BRAKER1 (Hoff et al., 2015), a pipeline for unsupervised RNA-Seq-based genome annotation that combines advantages of GeneMark-ET and AUGUSTUS. We observed that BRAKER1 was more accurate than MAKER2 (Holt et al., 2011) when it is using assembled RNA-Seq as sole source of extrinsic evidence. BRAKER1 does not require pre-trained parameters or a separate manually curated training step.

All the tools described above can be applied for analysis of just assembled NGS genomes.

Speaker's biographical sketch

Mark Borodovsky is a Regents' Professor at the Join Wallace H. Coulter Department of Biomedical Engineering of Georgia Institute of Technology and Emory University and Director of the Center for Bioinformatics and Computational Genomics at Georgia Tech. He is also a Chair of the Department of Bioinformatics at the Moscow Institute of Physics and Technology in Moscow, Russia.

Mr. Borodovsky is interested in promoting bioinformatics education. He is a Founder of the Georgia Tech Bioinformatics M.Sc. and Ph.D. Program, a Member of Educational Committee of the International Society of Computational Biology as well as organizer of a series of International Conferences in Bioinformatics at Georgia Tech started in 1997.

ESTIMATING THE EFFECTS OF REPEATS ON ASSEMBLY CONTIGUITY

Thursday, 2nd June 10:00 La Fonda Ballroom Talk (OS-4.01)

Shoudan Liang, Jason Chin

Pacific Biosciences of California

For a perfect assembler and at a high coverage, the contiguity of the assembly at a finite read length is limited by repetitive sequences. We study the limit imposed by repeat structures in plants, and contrast it to human, as the read length is increased. We started with assembled contigs from long reads and perform an all-against-all alignment. Non-unique regions of the contigs define repeats. We require each alignment to be longer than a minimum length, S. Repeats shorter than S will not align. Therefore, as the minimum overlap S is increased, we observed a decrease in the number of repeat regions. For example, for coffee genome, when the minimum allowed overlap increases from 500 to 5,000 bp, the number of distinct repetitive regions is reduced by more a factor of 10. This is partially due to long repeats being less abundant and partially because the short repeats are occurring in clusters that are seen as unique sequences in the alignment. We developed a method to separate these two effects. We show the tendency of repeats to cluster in several plant genomes. Clustered repeats are especially difficult to assemble from short reads because even when all short reads are identified to be from the same 100 kb region, they are still repetitive in the repeat-cluster.

A related method to estimate the repeats is by counting the abundance of two k-mers separated by a fixed distance. The distance between the k-mers is a proxy for the repeat length. This method has an advantage of potentially being directly applied to the long-read data before assembly. We compare the direct estimate from the read with the estimate from the contigs for several plant genomes.

A third way of estimating repeat abundance from long reads is by performing an all-against-all alignment using about 1% of data. This, when compared to the expected alignment of an idealized genome of the same size that does not have any repeatitive regions, reveals excessive alignments related to repeats at different lengths. This can be helpful in choosing assembly parameters. The method we developed is available at https://github.com/pb-sliang/TAP.

COFFEE BREAK

Sponsored by 10x Genomics

10:20 – 10:40

NUCLEOTID.ES OBJECTIVE BENCHMARKING OF BIOINFORMATICS SOFTWARE

Thursday, 2nd June 10:40 La Fonda Ballroom Talk (OS-5.01)

Michael Barton

Joint Genome Institute

As novel technologies emerge, fueled by scientific questions, the number bioinformatics tools continues to grow. When publishing new bioinformatics software, the tool is typically compared against others' using test data. Naturally, researchers will select the results that portray their software in the best way, which can lead to a subjective comparison, sometimes even lacking reproducibility. Combining this with the current large numbers of software publications in bioinformatics it is difficult for researchers to objectively evaluate which software will work best in their analysis, since different tools are rarely tested on identical data and settings for running tools can be very custom.

To address these challenges, we developed the nucleotid.es project, which places bioinformatics tools in software containers that allow them to be used on any platform, and reproducibly benchmarked against reference sequencing data. This allows software to be evaluated simultaneously and without requiring manual installation or setting of parameters. The tools benchmarked in nucleotid.es are submitted by the bioinformatics community or created from the existing literature. This effectively crowd-sources software from anyone who wishes to participate. The latest bioinformatics research can thus constantly be evaluated against the current existing corpus and the result of this benchmarking then shared with the wider bioinformatics community as they are generated.

This talk will present nucleotid.es and the results of running this analysis on a range of genome assemblers.

AUTOMATION AND MANAGEMENT OF INFLUENZA NEXT-GENERATION SEQUENCING DATA ANALYSIS AND QUALITY CONTROL PROCEDURES

Thursday, 2nd June 11:00 La Fonda Ballroom Talk (OS-5.02)

Thomas Stark[1], Samuel Shepard[2], Sujatha Seenu[2], Elizabeth Neuhaus[2], John Barnes[2]

[1]Battelle Memorial Institute, Atlanta, GA,
[2]Influenza Division, Centers for Disease Control and Prevention,

The CDC Influenza Division recently transitioned to using next generation sequencing technologies for molecular surveillance and now routinely sequences original clinical specimens as part of its operational pipeline. A customized Clarity LIMS workflow supports these operations and serves an important role within this pipeline, particularly because it has been extensively configured to streamline high throughput sequencing tasks. This overall workflow includes automated queueing of available specimens from a central metadata store as well as auto-assignment of samples to a central library construction workflow. Experimental procedures are easily tracked and quality control metrics are readily produced for each phase of the overall process, which includes monitoring of in-progress instrumentation as well as review of post-run analysis results. Genome assembly of the LIMS-staged samples is automatically scheduled on grid computing infrastructure as soon as NGS runs complete. Downstream curation and aggregate quality control assessments are also triggered automatically within this framework, enabling access to interpretation-ready data as early as one hour following the completion of a 96-plex Illumina MiSeq run. Our custom-configured Clarity LIMS pipeline furnishes e-mail status alerts and provides summary reports on a per-run basis, two features that serve to notify submitters of completion or failure status in real time. This overall system has become a critical component of Influenza Genomics Team operations by minimizing clerical technical errors within the laboratory, by maintaining consistent records of laboratory metadata associated with these experimental procedures, by improving the timeliness of the availability of processed data through the automation of analysis, and by providing computational scrutiny of multiplexed data. Finally, this LIMS configuration and data management regime has been replicated in a cloud-computing environment in order to establish a network of procedurally identical Influenza Reference Center satellite sites, which serve to increase national surveillance capacity, reduce turn-around times for the availability of genetic data, and establish readiness for pandemic preparedness.

DISTINGUISHING OUTBREAKS OF BOTULISM USING A REFERENCE-BASED SNP ANALYSIS AND HIGH QUALITY REFERENCE GENOME SEQUENCES OF CLOSTRIDIUM BOTULINUM

Thursday, 2nd June 11:20 La Fonda Ballroom Talk (OS-5.03)

Brian Shirey[1], Shannon Johnson[2], Maliha Ishaq[1], Carolina Luquez[1], Karen Hill[2], Susan Maslanka[1]

[1]Centers for Disease Control and Prevention, [2]Los Alamos National Laboratory

Reference-based whole genome SNP analysis can be used to support laboratory investigations of botulism by providing the necessary resolution, speed, and accuracy to distinguish outbreak-associated strains from sporadic cases. The National Botulism Laboratory Team (NBLT) at CDC has partnered with Los Alamos National Laboratory (LANL) to generate additional high-quality, finished Clostridium spp. reference sequences, and develop a user-friendly, reproducible SNP-based workflow that utilizes open-source, freely-accessible bioinformatics tools.

Currently, there are 20 complete genome sequences of botulinum neurotoxin producing clostridia (BTPC) available in the NCBI database; however, many commonly-occurring BTPC strains lack representation in the database. These gaps in representation limit our ability to perform high-resolution, reference-based BTPC strain subtyping. The Clostridium botulinum PulseNet database of PFGE patterns was used, in part, to identify 10 commonly-occurring BTPC strains that lack sequence representation in public databases. The LANL Genomics Group sequenced (using PacBio and Illumina platforms), assembled, and annotated the genomes of these 10 strains to generate high-quality, complete genome sequences that will be publically available in the NCBI database.

NBLT used the Ion Torrent PGM to sequence an additional 100 BTPC strains to facilitate the evaluation of open-source bioinformatics tools for reference-based SNP genotyping. Through this effort, NBLT has successfully identified an analytical workflow to: 1) rapidly identify the most appropriate BTPC reference sequences based on average nucleotide identity thresholds, 2) construct accurate SNP phylogenies using the RealPhy v112 tool, and 3) identify a BoNT serotype-specific parameter regime for distinguishing BTPC strains based on estimates of evolutionary divergence. This method can be employed by users with minimal bioinformatics expertise to rapidly and accurately distinguish BTPC strains within a single botulism outbreak, or between a small number of outbreaks, without relying on high performance computing requirements. This workflow will facilitate labora- tory investigations of botulism by providing a reproducible, yet customizable subtyping tool with proven efficacy to distinguish outbreak strains.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

IMPORTANCE OF WGS FOR IDENTIFYING ENVIRONMENTAL SOURCES OF LEGIONELLA PNEUMOPHILA DURING OUTBREAK INVESTIGATIONS OF LEGIONNAIRE'S DISEASE

Thursday, 2nd June 11:40 La Fonda Ballroom Talk (OS-5.04)

Brian Raphael, Shatavia Morrison, Jeffrey Mercante, Natalia Kozak Muiznieks, Jonas Winchell

Centers for Disease Control and Prevention

Legionnaire's disease (LD) is a severe pneumonia caused by various Legionellae (most commonly L. pneumophila) which can colonize man-made water systems and cause infections in humans when aerosolized. Confirming the environmental sources associated with an outbreak is important for controlling disease transmission. We evaluated the utility of whole genome sequencing (WGS) com- pared to other available subtyping methods in order to better understand genetic relationships among strains isolated during LD outbreak investigations. We also assessed various WGS analysis methods to determine their ability to cluster outbreak-related isolates.

L. pneumophila is a genetically diverse species consisting of multiple serogroups (sg) and three known subspecies; subsp. pneumophila, subsp. fraseri, and subsp. pascullei. In order to more fully understand differences in gene content among these strains, our laboratory has generated three complete genome sequences of L. pneumophila subsp. pascullei strains isolated nearly 30 years apart from the same facility and compared these to other previously determined L. pneumophila subsp. sequences. Although these isolates switched from sg5 to sg1, we identified minimal changes in gene content other than variation in the LPS biosynthesis region.

L. pneumophila subtyping has relied on Sequence Based Typing (SBT) where the nucleotide se- quences of seven loci are utilized to generate a Sequence Type (ST). ST1 is the most common ST associated with sporadic LD cases. We demonstrated that outbreak-specific clades can be detected among draft genomes of ST1 isolates (N=50) representing different LD outbreaks, sporadic clinical isolates, and various environmental sources using core gene SNP analysis and whole genome MLST.

When sequences of isolates within a suspected outbreak cluster were examined, average nucleotide identity and distance estimation using the MinHash technique (Mash) proved to be the most rapid initial method of identifying outbreak-associated strains. Moreover, a comparison of isolates from 10 separate LD investigations occurring in New York State revealed that WGS analysis provided superior resolution of isolates compared to pulsed-field gel electrophoresis. Extraction of SBT loci sequences in silico was only partially successful due in large part to the presence of paralogous sequences of one allele (mompS).

Ideally, the genomes of Legionella spp. present in clinical specimens or environmental sources will need to be sequenced and analyzed without the time-consuming isolation of this organism in order to realize the full potential of WGS for rapidly informing public health decision making. As a first step toward this goal, we examined archived water samples obtained from 3 separate LD investigations. While large differences in the abundances of various bacterial taxa were observed using 16S rDNA amplicon sequencing, Legionella-specific sequences, when present, were consistently found to be <1% of the overall number of operational taxonomic units observed.

These data and workflows are useful for informing public health practices during investigations of Legionnaire's disease outbreaks and results should be interpreted in proper epidemiological context.

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

LUNCH

Sponsored by MRIGlobal

12:00 – 13:20

TARGETED CRISPR/CAS9 DNA FRAGMENTATION AND SELECTIVE PRIMER SEQUENCING ENABLES MASSIVELY PARALLEL MICROSATELLITE ANALYSIS

Thursday, 2nd June 13:20 La Fonda Ballroom Talk (OS-6.01)

Gi Won Shin, Susan M Grimes, Hojoon Lee, Billy Lau, Charlie Xia, Hanlee P Ji

Stanford University

Microsatellites are multi-allelic genetic markers, composed of short tandem repeats (STRs). They have unit motifs composed of mononucleotides, dinucleotides and large motifs up to hexamers. Next generation sequencing approaches and other methods for STR analysis rely on the analysis of a limited number of PCR amplicons, typically in the tens. To massively increase throughput and improve genotyping accuracy of microsatellites, we developed STR-Seq, a next generation sequencing technology that analyzes over 1,000 STRs in parallel. STR-Seq uses in vitro CRISPR/Cas9 fragmentation to produce specific DNA molecules that encompass the complete microsatellite sequence. Afterwards, STR-selective primers enable massively parallel, targeted sequencing of large STR sets. Amplification-free library preparation provides single molecule sequences without using unique molecular barcodes. Overall, STR-Seq's genotyping has high throughput, high accuracy and provides informative haplotypes. Using these new features, STR-Seq can identify a 0.1% minor genome fraction in a DNA mixture composed of different, unrelated samples.

ADVANCING MOLECULAR DIAGNOSTICS IN SUDDEN UNEXPLAINED DEATH A NEW YORK CITY EXPERIENCE

Thursday, 2nd June 13:40 La Fonda Ballroom Talk (OS-6.02)

Yingying Tang

New York City Office of Chief Medical Examiner (OCME)

Sudden Unexplained Death (SUD) in previously healthy individuals is a vexing challenge facing forensic pathologists and a critical issue in public health today. SUD is a diagnostic exclusion which is defined as an undetermined cause of natural death after scene investigations, autopsy, and the review of past medical records. SUD is a critical issue that is not routinely investigated at molecular level. Powered with next-generation sequencing technology, OCME evaluated the utility of molecular testing of a large number of genes in a well-characterized and demographically diverse SUD cohort. I will present the results from that study to demonstrate the effectiveness and feasibility of molecular autopsy in a medical examiner setting. Furthermore, I will highlight the importance of family studies in clinical evaluation of high risk family members and function studies in enhancing our understanding of variants of uncertain significance, both type of studies are conducted through multi-institutional and multi-disciplinary collaborations. The feasibility and utility of broad molecular autopsy provides the first needed step in personalized diagnostics and precision preventive medicine for families of victims of SUD.

ASSESSMENT OF SINGLE NUCLEOTIDE POLYMORPHISM (SNP) GENOTYPING CHEMISTRIES AND COMPARISON OF PHENOTYPIC AND ANCESTRY PREDICTION TOOLS FOR FORENSIC INVESTIGATIVE LEADS

Thursday, 2nd June 14:00 La Fonda Ballroom Talk (OS-6.03)

Lilly Moreno, Michelle Galusha, Jodi Irwin

Federal Bureau of Investigation

Next Generation Sequencing kits for forensic applications have recently been commercialized and allow for concurrent amplification of a variety of markers, including hundreds of forensically relevant single nucleotide polymorphisms (SNPs). These SNPs could be useful not only for identification purposes but also in generating investigative leads for a variety of scenarios. As a complement to the chemistries, commercial data analysis packages have been developed to impart weight to the resulting genotypes and provide ancestry and phenotypic predictions. Thus far, however, the consistency and robustness of the raw data from these commercial assays have not been thoroughly examined, and the utility and reliability of the software tools have not been systematically tested. With SNP data for ancestry and phenotype from 100 samples, we analyzed the performance of the commercial chemistries. We also compared the SNP genotypes, as well as the ancestry and phenotype predictions, from different chemistry and software combinations to better understand and characterize the various workflows. Finally, by comparing the software predictions to the truth data for all 100 samples, we assessed the practical viability of these markers for aiding forensic investigations. The results of these studies will be presented.

RIGEL: ANALYSIS SYSTEM FOR MICROBIAL ATTRIBUTION, BIOFORENSICS AND BIOSURVEILLANCE

Thursday, 2nd June 14:20 La Fonda Ballroom Talk (OS-6.04)

Willy Valdivia[1], Deepak Sheoran[1], Juergen Richt[2], Chester Mcdowell[2]

[1]Orion Integrated Biosciences Inc., [2]Kansas State University

The use of weapons of mass destruction and in particular biowarfare agents have produced different appreciations and responses by the international community. From the sixties to the nineties, security efforts focused on state players and the development of countermeasures against traditional bioweapons. However, this attention shifted as extremist groups and individuals used pathogens in acts of bioterrorism and dual use technologies including synthetic biology became available (1). It is argued that next generation high throughput DNA sequencing offers the possibility of typing biological samples for microbial infection diagnosis, attribution, bioforensics and biosurveillance. However, the sensitivity and accuracy thresholds of these analyses can be affected by the error rate of databases storing information of reference genomes. The ambiguity of metadata associated with specific strains and the contamination genomic material during sequencing can affect the confidence of attribution. Although several algorithms have been developed to identify pathogen-specific genomic signatures, the use of these libraries to determine the taxonomic composition of a given metagenomics sample is affected by the mutational landscape, the sample size, the number of DNA reads generated by each sample, the use of phylogenetic distances and statistical weights to establish a probability match. In order to address these challenges, here we summarize our work towards the development of an integrated and comprehensive genomic-based analytical strategy for known and unknown microbial taxonomic composition assessment. This architecture can establish accurately intra- and interspecies relationships among more than 380,000 known taxonomies. RIGEL generates motif fingerprints and genomic signatures and during this process; errors in reference entries are flagged, disambiguated, corrected and/or eliminated. Once this process is completed RIGEL- mtp can be deployed to scan genomic or metagenomics samples from any sequencing platform. We demonstrate the performance of our system using more than 200 genomes of Fransicella tularensis, Bacillus anthracis, Burkholderia spp. In addition, we summarize the analysis of more than 1000 clinical, biological and environmental metagenomic samples sequenced with different technologies. Our presentation summarizes the benchmarking of sensitivity, speed and accuracy in the discrimination of known vs. unknown microbial species and/or strains at resolution levels to support biosurveillance and bioforensics efforts. In addition, we summarize the performance of our technology during the 2015-2016 Zika outbreak in South America and the 2016 Elizabethkingia outbreak in the US.

1.W. A. Valdivia-Granda, Biosurveillance enterprise for operational awareness, a genomic-based approach for tracking pathogen virulence. Virulence 4, (Oct 23, 2013).

MASSIVELY PARALLEL SEQUENCING TECHNOLOGIES FOR EXPANDED DNA IDENTIFICATION CAPABILITIES AT THE FEDERAL BUREAU OF INVESTIGATION LABORATORY

Thursday, 2nd June 14:40 La Fonda Ballroom Talk (OS-6.05)

Jodi Irwin[1], Lilly Moreno[1], Michael Brandhagen[2], Michelle Galusha[1], Rebecca Just[2], Anthony Onorato[2], Thomas Callaghan[2]

[1]Federal Bureau of Investigation, [2]FBI Laboratory

Though Massively Parallel Sequencing (MPS) has transformed numerous genetic disciplines over the past decade, it is only within the past few years that evaluations of MPS for forensic application have been undertaken in earnest. Given the potential of MPS to not only increase the quantity and discriminatory power of genetic data but also improve the overall throughput of samples through the laboratory, the Federal Bureau of Investigation is evaluating MPS assays for future casework application. Long-term laboratory efforts are directed towards employing MPS as a common platform for testing of all markers of forensic interest. However, near-term efforts are directed specifically towards evaluating the technology for its utility in expanding existing institutional capabilities. Three areas of current interest are 1) mitochondrial DNA typing and the development of entire mitochondrial genome (mtGenome) data in particular and 2) highly challenging samples and the benefits of MPS for improved information recovery, and 3) no-subject crime scene samples and the value of ancestry and phenotype markers for developing investigative leads. Given the significant benefits that complete mtGenome data bring to the discriminatory power of mtDNA evidence, we are evaluating methods that efficiently yield robust mtGenome data from high quality specimens, as well as approaches that address the significant challenge of recovering entire mtGenome data from limited evidentiary material. MPS presents a relatively minor shift from currently employed methods and workflows for mtDNA typing in forensics, and thus mtDNA applications are the primary focus of our near-term MPS validation and implementation efforts. In line with these efforts to recover more, and more discriminatory, mtDNA information from the most limited evidentiary material, we are also assessing the general benefits of MPS typing for expanding the lower range of sample quality from which probative data can be recovered. The sensitivity of the MPS process to low quantities of DNA, the benefits of numerous marker systems in a single assay, the utility of STR sequence information when only a small number of markers are recovered are some of the topics being examined and characterized for this application. Finally, we are evaluating commercially available assays and software tools for predicting ancestry and phenotype to better understand the true practical utility of these tools in developing investigative leads. With a better practical understanding of these and other benefits of MPS, new approaches to highly challenging samples can be devised, the lower range of sample type and quality from which probative data may be recovered can likely be broadened and the overall number of cases that can benefit from DNA typing can be expanded. Here, we present an overview of these efforts.

THE EFFECT OF STORAGE TIME AND TEMPERATURE ON THE RECOVERED MICROBIOME FROM PUBIC HAIRS

Thursday, 2nd June 15:00 La Fonda Ballroom Talk (OS-6.06)

Diana Williams

Defense Forensic Science Center

The use of the microbial genetics in forensic science has been limited despite its potential application in linking individuals. Recent work on the human head and pubic hair microbiome has suggested that the pubic hair microbiome may be transferred between individuals during sexual intercourse and detected using high-throughput sequencing. Because this study was limited in scope, further studies must be performed to evaluate the efficacy and reliability of this analytical method for criminal investigations. One aspect that must be addressed involves the impact of storage on the microbiome of samples recovered for forensic testing. Forensic samples may be stored from days to years in a variety of conditions before analysis occurs. To test the effects of storage, pubic hair samples were collected from volunteers and stored at room temperature, refrigerated (4oC), and frozen (-20oC). The samples were subject to high-throughput sequencing at baseline, 1 week, 2 weeks, 4 weeks, and 6 weeks post-collection. The subsequent analysis of the sequence data showed that individual microbial profiles and the differences in gender were the greatest source of variation between samples. Within the samples collected from each individual, no statistically significant difference was observed while time or temperature varied. For short-term storage (< 6 weeks), the microbiome recovered was not significantly affected by the storage time or temperature, suggesting that investigators and crime labs could use already-existing evidence storage methods.

Disclaimer: The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the United States Department of the Army or United States Department of Defense.

NATIONAL INSTITUTE OF JUSTICE FUNDING OPPORTUNITIES

Thursday, 2nd June 15:20 La Fonda Ballroom Talk (OS-6.07)

Minh Nguyen

National Institute of Justice

The National Institute of Justice (NIJ) - the research, development and evaluation agency of the U.S. Department of Justice - is dedicated to improving knowledge and understanding of crime and justice issues through science. NIJ's Office of Investigative and Forensic Sciences supports this mission by sponsoring research to provide objective, independent, evidence-based knowledge and tools to meet the challenges of criminal justice, particularly at the State and local levels.

As DNA sequencing technologies advance and costs decrease, the interest in applying sequence analysis methods and workflows to forensic issues increases. NIJ anticipates continued interest in sequencing technologies for forensic applications and is interested in engaging the genomics community in examining forensically relevant research questions.

Ms. Nguyen will present an overview of NIJ's research and development portfolio, information on its funding cycle, and general information about R&D funding opportunities at NIJ.

COFFEE BREAK

Sponsored by Becton Dickinson

15:30 – 15:50

DIRECT DETERMINATION OF GENOME SEQUENCES

Thursday, 2nd June 15:50 La Fonda Ballroom Tech Talk (TT-2.01)

David Jaffe, Neil Weisenfeld, Vijay Kumar, Kamila Belhocine, Rajiv Bharadwaj, Deanna Church, Paul Hardenbol, Jill Herschleb, Chris Hindson, Yuan Li, Patrick Marks, Pranav Patel, Andrew Price, Michael Schnall Levin, Alex Wong, Indira Wu

10x Genomics, Inc

We introduce a new method for determining the genome sequence of an organism. Our method has the following key advantages:

1. Our starting material consists of 1 ng of high molecular weight DNA, as compared to typical requirements of 1,000-10,000 ng or more.
2. We create a single library, as compared to typical methods which require creation of multiple libraries and often multiple data types. This makes our process fundamentally more robust than typical methods.
3. Our costs are about ten times lower.
4. The entire process, including assembly, is not organism-specific. For example, there are no parameters to specify to the algorithm.
5. We produce a genome sequence that mirrors the actual chromosomes in the sample, as contrasted with prior methods for which a contig is a mélange of homologous sequences.

To accomplish this, we create a single 10X Genomics Linked-Read (Chromium Genome) library and sequence it on the HiSeq X instrument. The data type consists of barcoded pools of reads, each originating from several very long molecules, with each molecule represented by many reads, and shallowly covered. Our new turn-key software, the Supernova Assembler, exploits these pools, first to create local assemblies, that can capture difficult regions, and then to phase homologous chromosomes.

Using human genomes, we obtain contigs of size 100 kb, in scaffolds of size larger than 10 Mb, and composed of phase blocks of size 3-4 Mb. Notably, these phase block lengths greatly exceed those obtained from any other method, in spite of being constructed from dramatically less expensive data. We rigorously assess the accuracy of our assemblies, including by means of a HGP sample for which 340 Mb of finished sequence is available. We further demonstrate our method on a wide range of organisms (both animal and plant), obtained from diverse starting materials. For many purposes, our method supplants all prior methods for obtaining genome sequences by providing a direct and inexpensive path to the true sequence of the sample.

SINGLE-MOLECULE MAPPING WITH SOLID-STATE DETECTORS

Thursday, 2nd June 16:05 La Fonda Ballroom Tech Talk (TT-2.02)

John Oliver

Nabsys 2.0

Structural variants are difficult to detect with short read technologies because genomes are hard to assemble. Polymorphism, repeats, and sequencing bias can turn even small genomes into assembly nightmares. Maps constructed from long reads, however, can be used to inform the assembly process.

Nabsys 2.0 has developed a DNA mapping technology that utilizes a completely electronic single- molecule detection scheme to generate map information from single molecules that are hundreds of kilobases in length. Long- range information is preserved so structural rearrangements and duplications are easily identified. The advantages of electronic sensing are higher resolution, accuracy, and density of the resulting maps versus those provided by optical technologies. An additional benefit is that since the sensors are manufactured in silicon wafers, solid-state foundries can be leveraged to provide scalability.

Several human genomes have been mapped using the Nabsys platform. The data have been used to develop a systematic and automated method to call structural variants as small as 300 bp in size. Examples of structural variant calls, both the raw data and methodology for making the call will be presented.

ONE CODEX: ACCURATE, ROBUST, AND SENSITIVE TOOLS FOR APPLIED MICROBIAL METAGENOMICS

Thursday, 2nd June 16:20 La Fonda Ballroom Tech Talk (TT-2.03)

Sam Minot, Nick Greenfield

One Codex

Next-generation sequencing (NGS) can be a powerful tool for the analysis of microbial samples, detecting low-abundance organisms, characterizing pathogenicity, and profiling complex microbial mixtures. However, unlocking the full potential of NGS requires powerful, robust, and highly accurate computational analysis. One Codex provides a comprehensive data platform for microbial metagenomics that encompasses best-in-class metagenomics, secure data management, and novel tools for microbial characterization. With a comprehensive database of ~40,000 complete microbial genomes, One Codex enables users to detect the broadest swath of bacteria, viruses, fungi, archaea, and protists.

One Codex has created a set of applied bioinformatic tools that deliver A) sensitive detection of low-abundance pathogens for biodefense, B) reference-free genomic clustering for strain-level outbreak epidemiology and NGS-based microbial attribution, C) integrated analysis of NGS amplicon panels for targeted detection, and D) functional characterization of specific microbes for public health. Moreover, evaluation of read-level accuracy using 50 million simulation WGS reads from 10,639 microbial genomes showed that One Codex had the highest degree of sensitivity and specificity (AUC = 0.97, compared to 0.82-0.88 for other methods).

Although the per-base error rate of modern NGS technology is quite low, the large volume of data produced by NGS instruments has presented a challenge for the detection of low-abundance pathogens in the presence of near neighbor organisms. This challenge was seen most acutely in the spurious detection of B. anthracis in the New York City subway system in early 2015. One Codex worked with the PathoMAP group to design and implement an automated tool for the specific detection of B. anthracis while accounting for sequencing error and near neighbors (https://science.onecodex.com/bacillus-anthracis-panel/). This tool indicated that the NYC B. anthracis results were consistent with sequencing error, while maintaining a high degree of analytical sensitivity for low-level spike-in samples. The B. anthracis detection tool is run automatically on samples uploaded to the One Codex platform, and provides a model for high-confidence, high-sensitivity pathogen detection tools for applied metagenomics.

A POWERFUL METHOD FOR COMPREHENSIVE STRUCTURAL VARIATION DETECTION WITH SHORT READS

Wednesday, 1st June 16:35 La Fonda Ballroom Tech Talk (TT-2.04)

Nicholas Putnam[1], Sanjeev Balakrishnan[1], Jonathan Stites[1], Richard E. Green[1,2]

[1]Dovetail Genomics, LLC, [2]University of California-Santa Cruz

Characterization and understanding of large genomic structural variants (SVs) is of fundamental importance for human health. Such variation is common, accounts for a large portion of all genomic variation in humans, and has been implicated in a diverse array of genomic diseases including cancers and psychiatric disorders. Despite its relevance to human health, cost-effective methods for thorough characterization of SVs in human populations and individuals have remained elusive. Using in vitro proximity ligation, Dovetail Genomics has developed novel sequencing library preparation and analysis methods that bring significant power to bear on this problem. Applying these methods to a variety of samples, we have successfully detected and characterized both simple and complex structural variants across a wide range of size scales, including rearrangements flanked by tens of kilobases of low complexity sequence. Our method is rapid and affordable, enabling more effective research studies and the characterization of individual clinical cases.

ANALYSIS OF THE FUNCTIONAL CONTENTS OF MICROBIAL COMMUNITIES USING A NOVEL QIAGEN BIOINFORMATICS PIPELINE

Thursday, 2nd June 16:50 La Fonda Ballroom Tech Talk (TT-2.05)

Andreas Sand Pedersen, Francesco Strino, Aske Simon Christensen, Martin Bundgaard

Qiagen Bioinformatics

The field of microbial ecology is currently being revolutionized through next generation sequencing studies, revealing unprecedented detailed insight into environmental as well as host-associated microbiomes. But still, very few tools offer efficient and user-friendly analysis of the functional genomic contents of microbiomes.

We have implemented a unique pipeline of tools for functional and comparative analysis of microbiomes as a component in QIAGEN Bioinformatics' comprehensive solution for microbial genomics - CLC Microbial Genomics Module. First step in this pipeline produces high-quality metagenome contigs using a novel fast and highly memory efficient metagenome de novo assembler. The resulting contigs are subsequently CDS-annotated using the MetaGeneMark plugin for CLC Genomics Workbench, and CDSs are annotated with PFAM protein families, GO terms and/or top BLAST hits (using e.g. UniProt). Finally, the relative abundance of each identified functional genomic element is computed and presented to the user in a user-friendly manner.

The functional contents of individual samples can be studied in a tabular format or visualized as pie charts, while visual comparison of multiple samples is done using stacked bar charts, layered area charts and principal coordinate plots. Finally, statistical comparison of multiple microbiome samples can be done to back-up our observations. All visualisations and statistical analysis can be performed in the context of user-defined meta-information such as patient/control data.

We will present benchmarks of this pipeline, performed on both previously published mock community datasets and several recently published simulated microbiome datasets, designed for benchmarking of microbiome analysis solutions.

We demonstrate that QIAGEN Bioinformatics' new metagenome de novo assembler runs as fast as the fastest alternative (MegaHit) but uses only a fraction of the memory required by any other benchmarked de novo assembler. At the same time it produces contigs that are both longer than the contigs produced by other popular metagenome assemblers (e.g. MegaHit and IDBA_UD) and are of significantly higher quality than the contigs produced by any other benchmarked de novo assembler.

We furthermore demonstrate that the high-quality contigs produced by our metagenome de novo assembler enables taxonomic and functional profiling of very high quality.

And finally, we show that the complete analysis pipeline producing functional profiles from raw microbiome read datasets in five easy steps is able to reliably detect changes in the functional content of microbiomes with statistical significance.

GENEIOUS: A BIOINFORMATICS PLATFORM FOR BIOLOGISTS

Thursday, 2nd June 17:05 La Fonda Ballroom Tech Talk (TT-2.06)

Christian Olsen, Helen Shearman, Richard Moir, Matt Kearse,
Sidney Markowitz, Jonas Kuhn, Sebastian Dunn, Alex Cooper

Biomatters, Inc.

Biomatters' Geneious is a bioinformatics software platform which allows researchers the use of industry leading algorithms for their genomic and protein sequence analyses. Geneious offers a comprehensive suite of functions, including peer-reviewed tools that enable researchers to be more efficient with their bioinformatic workflows. Geneious is comprised of an extensive tool suite for next-generation sequence analysis, molecular cloning, chromatogram assembly, and phylogenetics.

This major version release includes tools for RNAseq read mapping and expression analysis, support for scaffolding with de novo assemblies, Golden Gate assembly, and an updated CRISPR tool. Researchers may also choose to extend Geneious with new plugins like Blast2GO (Gene Ontology enrichment), FreeBayes variant finder, Augustus (gene prediction), and BBTools (NGS data QC). Additional features include the 16S Biodiversity tool and Sequence Classifier plugin. The 16S Biodiversity tool identifies high-throughput 16S rRNA amplicons from environmental samples using the RDP database, and visualizes biodiversity as an interactive chart using a secure web viewer. The Sequence Classifier plugin taxonomically classifies an organic sample by how similar its DNA is to your own database of known sequences using a BLAST-like algorithm with multiple-loci and phylogenetic trees to assist with an unknown identification.

Geneious easily affords real-time dynamic interaction with sequence data and empowers biologists to produce stunning publication quality images to increase the visibility of their research. By utilizing Geneious, biologists can quickly streamline their sequence analysis workflows. This tech talk aims to present new features and benefits of the highly-integrated Geneious sequence analysis platform.

ARTEMISININ DRUG RESISTANCE WORKFLOW FOR PLASMODIUM FALCIPARUM

Thursday, 2nd June 17:15 La Fonda Ballroom Tech Talk (TT-2.07)

Christian Olsen[1], Eldin Talundzic[2], Helen Shearman[1]

[1]Biomatters, Inc., [2]Centers for Disease Control and Prevention

About 3.2 billion people are at risk of malaria. In 2015, there were ~214 million malaria cases and an estimated 438,000 malaria deaths. Young children, pregnant women and non-immune travellers from malaria-free areas are particularly vulnerable to the disease when they become infected. Falciparum malaria is the most lethal human malaria and is transmitted by the Anopheles mosquito. The primary methods for combating malaria are antimalarial treatments like artemisinin and control of the Anopheles mosquito vector.

Artemisinin is derived from the sweet wormwood plant Artemisia annua. Artemisinin based compounds are used with other classes of antimalarials to form artemisinin-based combination therapies (ACTs) which are, currently, the best available treatment for falciparum malaria worldwide. The recent emergence of artemisinin resistance in the Greater Mekong sub-region poses a major threat to the global effort to control falciparum malaria. Tracking the spread and evolution of artemisinin-resistant parasites is critical in aiding efforts to contain the spread of drug resistance.

This poster describes a workflow, which utilizes a Kelch-13 (K13) molecular marker, previously identified to be associated with artemisinin resistance. Kelch proteins are a group of proteins that contain multiple Kelch motifs that form a -propeller tertiary structure. Kelch-repeat -propellers are generally involved in protein-protein interactions, though the large diversity of domain architectures and limited sequence identity between kelch motifs make characterization of the kelch superfamily difficult.

A total of 417 patient samples from the year 2007, collected during malaria surveillance studies across ten provinces in Thailand, were genotyped for the K13 marker. The targeted gene is amplified with PCR and sequenced using Sanger sequencing. The sequences obtained from samples are compared to a reference sequence and the SNP combinations are examined. This pipeline is implemented in Geneious and was used to identify and track Artemisinin resistance in Thailand.

TOWARDS BUILDING COMPLETE GENOME ASSEMBLIES USING BIONANO NEXT-GENERATION MAPPING TECHNOLOGY

Thursday, 2nd June 17:25 La Fonda Ballroom Tech Talk (TT-2.08)

Andy Wing Chun Pang[1], Thomas Anantharaman[1], Xiang Zhou[1], Jian Wang[1], Joyce Lee[1], Evan Eichler[2], Tina Graves Lindsay[3], Alex Hastie[1], Han Cao[1]

[1]BioNano Genomics, [2]University of Washington, [3]Washington University

High-quality assemblies are important when trying to understand the biology of genomes. Current short-read assemblers are memory intensive and have difficulties in constructing contiguous assemblies; collecting deep coverage data by long-read technologies can be time-consuming and expensive. BioNano's Next-Generation Mapping (NGM) data complements short-read data by flagging and correcting inaccuracies and increasing contiguity, thereby reducing the need to collect high-coverage long-read data.

BioNano Genomics Irys® System utilizes high-molecular-weight DNA to construct physical genome maps. These maps can be used to reveal large structural variants, but can also be combined with sequencing assemblies to produce hybrid scaffolds of unprecedented lengths, with some spanning chromosomal ar In addition, when one aligns the sequence and BioNano assemblies, one can identify chimeric joins errors, which would appear as conflicting alignment junctions. Chimeric joins – two distal regions in the genome are incorrectly placed together by assembly algorithms may form when short reads, molecules, or paired-end inserts are unable to span across long DNA repeats.

We developed a new hybrid scaffold pipeline that detects and resolves these conflicting junctions between the sequence and the BioNano assemblies. At a conflicting junction, the pipeline uses BioNano's long molecules to determine which assembly has been constructed incorrectly. Specifically, it checks the chimeric quality scores surrounding the conflict junction on the genome map for any evidence of a misassembly. This score indicates the percentage of BioNano molecules aligned 55 kb to the left and to the right of a locus. If the junction on the genome map has low scores (less than 35%), then the genome map support would be considered relatively weak; hence, the pipeline would cut the genome map at the conflict, thus resolving the conflict. Conversely, if the genome map has high chimeric quality scores, then the sequence contig would be cut. Importantly, this automatic conflict-resolution function can be manually modified to enable users to have fine control in generating high quality and complete hybrid scaffolds.

We applied this hybrid scaffold pipeline on a haploid human genome (CHM1). The genome has been sequenced and genome mapped, and the assemblies' N50 values are 27.7 Mb and 3.9 Mb, respectively. The pipeline resolved 23 chimeric joins in the sequence assembly and three in the BioNano genome maps. Moreover, by combining the two refined assemblies, the ultra-long hybrid scaffolds resulted in a 58.4 Mb N50 value and 2.9 Gb in length.

This new hybrid scaffold functionality further enhances the construction of highly accurate and contiguous reference assemblies for complex plants and animal genomes using BioNano mapping technology.

DE NOVO ASSEMBLY AND STRUCTURAL VARIATION DETECTION OF HUMAN GENOMES USING SINGLE MOLECULE NEXT-GENERATION MAPPING AND SV CALL VALIDATION BY INHERITANCE AND ORTHOGONAL MEASUREMENTS.

Thursday, 2nd June 17:35 La Fonda Ballroom Tech Talk (TT-2.09)

Alex Hastie[1], Thomas Anantharaman[1], Tiffany Liang[1], Ernest Lam[1], Joyce Lee[1], Khoa Pham[1], Michael Saghbini[1], Ali Bashir[2], Han Cao[1]

[1]BioNano Genomics, [2]Mount Sinai School of Medicine

Structural variation detection is generally based on reference mapping to find discordant evidence. In order to detect structural variations comprehensively, a reference independent (de novo) approach is needed as it allows assembly of regions absent from the reference. Using Next-Generation Mapping (NGM) from BioNano Genomics), we produced high-resolution genome maps that were de novo assembled and preserved the long-range genomic information necessary for structural variation detection.

Here, the Genome in a Bottle (GIAB) reference trio of Ashkenazi Jewish descent (NA24385, NA24149, NA24143) has been de novo assembled using the Irys System. Structural variation analysis reveals insertions, inversions, and deletions, including large deletions in the UGT2B17 gene (involved in graft versus host disease, osteopathic health and testosterone and estradiol levels) in the mother and son. We compared structural variants found in the son (NA24385) by genome mapping to those found in his parents. Deletion and insertion calls, one kilobase and up, found in the son are also found in the parents at a rate of 90% and 92% respectively. We also use structural variation calls made with PacBio sequencing by reference mapping and local assembly approaches to validate BioNano's structural variation calls, resulting in up to 80% cross-validation rates for deletions of 2-5kb, while a much lower rate for larger deletions and all insertions by PacBio's methods. When considering PacBio SV calls, BioNano calls could validate all categories at a rate of 80-90%. Thus, structural variation calling by next generation mapping is a fast, inexpensive, robust and accurate method that can be orthogonally validated in size bins that other methodologies are able to interrogate as well as provide novel SV information where other technologies are unable to interrogate.

HAPPY HOUR(S) @ COWGIRL CAFÉ

Sponsored by illumina!!!

illumina

Cowgirl Cafe

505.982.2565 319 S. Guadalupe St Santa Fe, NM
See map on next page!

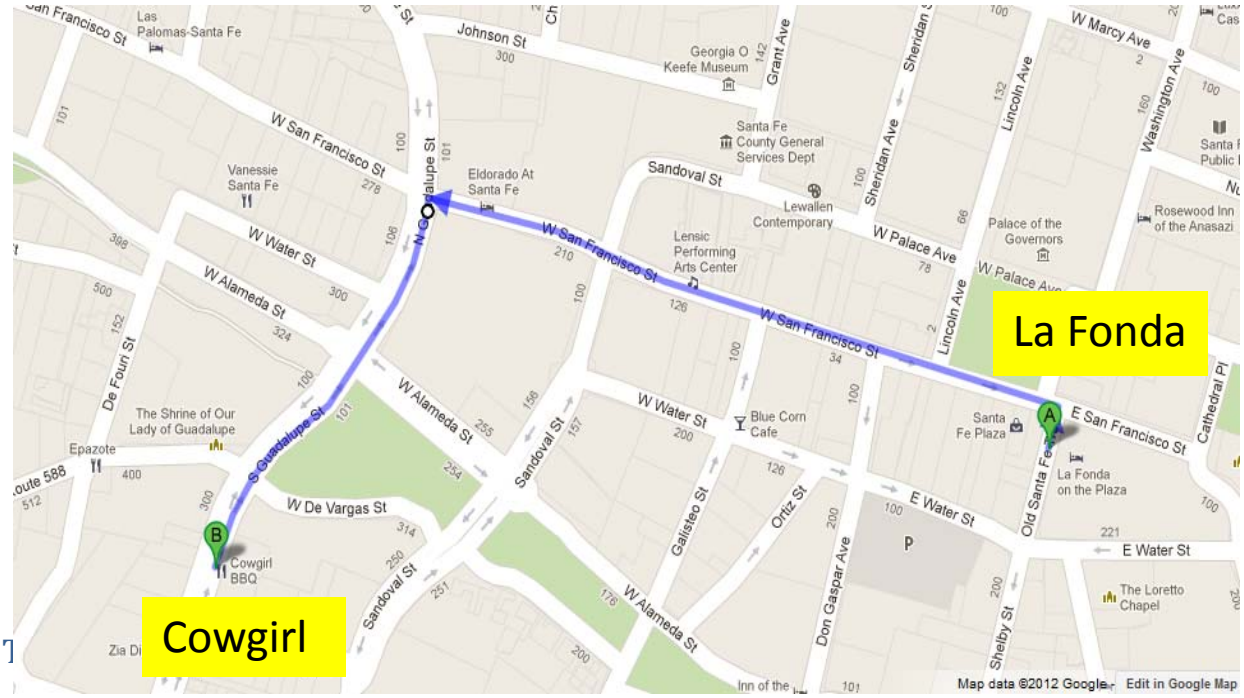
18:30pm – 20:30pm, June 2nd

Drink tickets (margaritas, beer, sodas) provided
Use your blue tickets

Enjoy!!!

MAP TO COWGIRL

505.982.2565 319 S. Guadalupe St Santa Fe, NM



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.

FRIDAY, 3RD JUNE

07:30 - 08:30 Breakfast (Sponsored by New England Biolabs)

08:30 - 08:45 Opening Remarks

08:45 - 09:30 Dr. Petrosino, Keynote Address (Sponsored by Qiagen)

09:30 - 10:50 Oral Session 7: Metagenomics
Chaired by: Patrick Chain and Michael Fitzgerald
- OS-7.01 :: Sequencing the elephant in the room: rapid resolution of whole genome sequences from uncultured bacteria in the human vaginal microbiome
- OS-7.02 :: Task-driven growth of SPAdes applications
- OS-7.03 :: Assembling whole genomes from mixed microbial communities using Hi-C
- OS-7.04 :: Building Reference Materials for Mixed Microbial Detection with Next Generation Sequencing

10:50 - 11:10 Coffee Break (Sponsored by Dovetail Genomics)

11:10 - 11:40 Mr. Lakey, Invited Speaker

11:40 - 13:00 Oral Session 8: Plants and the Fruit Fly
Chaired by: Kenny Yeh and Johar Ali
- OS-8.01 :: PacBio, Dovetail, and BioNano technologies enable quality plant genome assemblies
- OS-8.02 :: A genome in the sugar beet field
- OS-8.03 :: Trait mapping and improvement of the melon fly (Bactrocera cucurbitae) genome
- OS-8.04 :: Characterizing chromosomal translocations associated with a genetic sexing system

13:00 - 14:00 Lunch Break (Sponsored by Dovetail Genomics)

14:00 - 15:40 Oral Session 9: Human Genomics
Chaired by: Donna Muzny and Bob Fulton
- OS-9.01 :: Improving genome analysis using Linked-Reads
- OS-9.02 :: A reference-agnostic and rapidly queryable NGS read data format allows for flexible analysis at scale
- OS-9.03 :: Reliable detection of Copy Number Variations based on data mining approaches
- OS-9.04 :: Robust genome-wide transcriptome enrichment sequencing for research and clinical platforms
- OS-9.05 :: Deep RNA sequencing of Brassica rapa RIL population for SNP discovery, genetic map construction, eQTL analysis, and genome improvement

15:40 - 16:00 Coffee Break (Sponsored by Promega)

16:00 - 17:00 Oral Session 10: Human Clinical Dx and Analysis Pipelines
Chaired by: Kenny Yeh and Michael Fitzgerald
- OS-10.01 :: ASAP: A Customizable Amplicon Sequencing Analysis Pipeline for High-Throughput Characterization of Complex Samples
- OS-10.02 :: Evaluation of HiSeq X Ten Performance: Towards Clinical Applications
- OS-10.03 :: YCGA Bioinformatics at Yale
- OS-10.04 :: Scalable and extensible next-generation sequence analysis pipeline management for over 50,000 whole-genome samples

17:00 - 17:10 Closing Remarks

METAGENOMIC APPLICATIOSN FOR MICROBIOME-RELATED STUDIES OF COMPLEX DISEASE

Friday, 3rd June 8:45 La Fonda Ballroom Keynote Address (KN-3)
Sponsored by Qiagen

Dr. Joseph Petrosino
Baylor College of Medicine

Nadim J. Ajami, Shelly A. Buffington, Matthew C. Wong, Daniel P. Smith, Ginger A. Metcalf, Donna M. Muzny, Richard A. Gibbs, Richard Lloyd, Beena Akolkar, Kendra Vehik, Jeffery P. Krischer, the TEDDY Study Group, Mauro Costa-Mattioli, and Joseph F. Petrosino

The incidence of complex disease, including immunity-related diseases (e.g. type 1 diabetes (T1D)) and neurological disorders (e.g. autism), has increased dramatically over the last 50 years while incidence of infectious diseases have declined. These trends cannot be explained by genetic factors alone, but suggest that the modern environment has changed leading to this increased risk. The link between our genetic blueprint, in utero exposures, and the development of our microbiome in early life sets our baseline health state. Increased gut permeability, intestinal inflammation and dietary impacts have all been observed in children with T1D and autism. Data support the hypothesis that an infectious trigger and/or microbial influence, perhaps even in utero, may be responsible for the onset of autism and the autoantibodies that ultimately lead to the decline to T1D.

We are exploring the comprehensive taxonomic and functional changes in the microbiome between birth and T1D onset in over 22,000 samples from 820 cases and controls (1:1 match) in the TEDDY international prospective cohort. Advanced analyses of 16S rRNA gene, and bacterial/viral metagenomic data have identified significant increased odds ratios for microbial community-based prediction of autoantibody emergence and T1D onset ($p \leq 0.05$). Integrated analyses of additional "-omics" data and extensive TEDDY metadata will begin to provide a complete perspective of the network of factors predisposing, and perhaps triggering, autoimmunity and T1D.

In a separate collaboration with Mauro Costa-Matioli (BCM), we have been examining the influence of the microbiome on a murine, maternal high-fat diet model for autism spectrum disorder (ASD). We find that social behavioral deficits, but not other ASD-like behaviors, associated with maternal high fat diet (MHFD)-induced obesity are mediated by alterations in the offspring gut microbiome. Moreover, oral treatment with a single, live commensal species corrects oxytocin levels and synaptic dysfunction in the ventral tegmental area and reverses social deficits associated with ASD in this model. If these effects translate to humans, this study suggests microbiome-based therapeutics could have a positive impact on ASD-related social dysfunction.

Speaker's biographical sketch
Joseph F. Petrosino, PhD, is an Associate Professor of Molecular Virology and Microbiology at Baylor College of Medicine and the Director of the Alkek Center for Metagenomics and Microbiome Research. He holds joint appointments in the Human Genome Sequencing Center, Department of Ophthalmology, and is a member of the Cell and Molecular Biology and Translational Biology and Molecular Medicine progra

Dr. Petrosino has authored 40 original papers. Among 14 published in 2012 are the June HMP flagship manuscripts in Nature, collaborative studies examining microbiome associations with Cystic Fibrosis, pregnancy, nutritional intervention in colitis, rotavirus infection, and the shaping of the microbiome from birth in murine syste He has been an invited to speak at numerous institutions and meetings nationally and internationally, and recently he has been named an American Society for Microbiology Distinguished Lecturer for 2012-2014.

SEQUENCING THE ELEPHANT IN THE ROOM: RAPID RESOLUTION OF WHOLE GENOME SEQUENCES FROM UNCULTURED BACTERIA IN THE HUMAN VAGINAL MICROBIOME

Friday, 3rd June 09:30 La Fonda Ballroom Talk (OS-7.01)

Laura Sycuro[1], Andrew Wiser[1], Josh Burton[2], Ivan Liachko[3], Jonathan Golob[1], Maitreya Dunham[2], Jay Shendure[2], David Fredricks[1]

[1]Fred Hutch Cancer Research Center, [2]University of Washington, [3]Phase Genomics

It is widely assumed that most species comprising the human microbiome have now been catalogued through 16S rRNA gene sequencing. However, many species remain uncultured and without additional genomic data from these organisms, our ability to consistently detect and functionally characterize them is handicapped all too often, they are the proverbial elephant in the room.

The human vaginal microbiome contains dozens of uncultured species, some of which are highly prevalent in women with bacterial vaginosis (BV) and increasingly implicated in preterm birth. We selected a single BV(+) vaginal sample with a high density of uncultured species for metagenomic sequencing and bacterial genome segregation using intracellular chromosome linkage analysis (Hi-C). Reconstruction of 16S rRNA genes from our shotgun reads (EMIRGE) predicted 27 species with relative abundance (RA) above 0.1% (total abundance = 99.5%), 11 of which are uncultured (total abundance = 72.3%). From a series of metagenomic assemblies (maximal assembly size = 59 Mb, n50 = 17 kb), we successfully segregated draft genome sequences for 23 of the predicted species, including all taxa >0.2% RA, 8 of the 11 uncultured species, and two distinct strains of Atopobium vaginae. Most of our Hi-C-derived genomes approached the quality of isolate genome sequences (median completeness and contamination of all 23 genomes assessed with CheckM = 85% and 1%, respectively).

Draft genomes were submitted to genome finishing and characterization pipelines, dubbed CoAlEScE and AnPhIRL, that we are developing for high throughput processing of bacterial genomes obtained from metagenomes. CoAlEScE was used to collapse redundant genomes clustered by Hi-C from different assemblies, resulting in significantly improved genome contiguity. Using assembled 16S rRNA and cpn60 marker genes, as well as whole genome comparative functionalities within AnPhIRL (based on pplacer and MiSi methodologies), we provisionally classified the 8 uncultured species as representatives of a novel phylum, 2 novel classes, 1 novel family, 1 novel genus, and 3 novel species. Since five of these species were moderately abundant (each 1-10% RA) and a sixth dominated the community (>50% RA), our findings underscore the extent to which uncultured and largely undescribed biological diversity within the vaginal microbiome could impact community function and reproductive health. Indeed, 3 of the uncultured species for which we successfully resolved a genome sequence, including both species representing novel classes, have never before been named or otherwise consistently identified across studies; nonetheless, all 3 were previously detected using full-length 16S amplification and sequencing in vaginal samples from women experiencing BV and/or preterm birth.

This work demonstrates the feasibility of rapidly populating reference genome databases with high quality, and in some cases strain-resolved genomes of uncultured bacteria using Hi-C linkage anal- ysis coupled with automated in silico genome finishing. Newly generated genomes of uncultured BV-associated bacteria that may contribute to pregnancy complications will provide linkages be- tween phylogeny and function that are needed to better understand the microbiome's role in human reproduction.

TASK-DRIVEN GROWTH OF SPADES APPLICATIONS

Friday, 3rd June 09:50 La Fonda Ballroom Talk (OS-7.02)

Anton Korobeynikov, Dmitry Antipov, Anton Bankevich, Elena Bushmanova, Alexey Gurevich, Alla Mikheenko, Dmitry Meleshko, Sergey Nurk, Andrey Prjibelski, Yana Safonova, Alla Lapidus, Pavel Pevzner

Saint Petersburg State University

Following the increased demand of the high throughput analytical software and pipelines SPAsed assembler was gradually extended into a family of SPAdes tools aimed at various sequencing technologies and applications.

Metagenomics is one of the fastest developing areas of microbial research with widely spread applications such as ecology, medicine, novel natural products discovery including antibiotics, anti-tumor and anti-cancer agents and more. The viability of these discoveries in the majority of cases relies on the ability to produce decent assembly of a metagenome. However, the problem of accurate de novo assembly of complex metagenomics datasets is far from being solved, thus stifling the biological discoveries.

The newly developed metaSPAdes is designed to fill this gap. It brings together new algorithmic ideas and proven solutions from the SPAdes toolkit to address the metagenomic assembly challenges.

Plasmids harbor biomedically important genes (such as genes involved in virulence and antibiotics resistance), however there are no specialized software tools for extracting and assembling plasmid data from whole genome sequencing projects. plasmidSPAdes is another problem-driven tool built on top of SPAdes platform to help assembling plasmids from whole genome sequencing data sets.

ASSEMBLING WHOLE GENOMES FROM MIXED MICROBIAL COMMUNITIES USING HI-C

Friday, 3rd June 10:10 La Fonda Ballroom Talk (OS-7.03)

Ivan Liachko[1], Joshua Burton[1], Laura Sycuro[2], Andrew Wiser[2], David Fredricks[2], Maitreya Dunham[1], Jay Shendure[1]

[1]University of Washington, [2]Fred Hutchinson Cancer Research Institute

Assembly of whole genomes from next-generation sequencing is inhibited by the lack of contiguity information in short-read sequencing. This limitation also impedes metagenome assembly, since one cannot tell which sequences originate from the same species within a population. We have overcome these bottlenecks by adapting a chromosome conformation capture technique (Hi-C) for the deconvolution of metagenomes and the scaffolding of de novo assemblies of individual genomes.

In modeling the 3D structure of a genome, chromosome conformation capture techniques such as Hi-C are used to measure long-range interactions of DNA molecules in physical space. These tools employ crosslinking of chromatin in intact cells followed by intra-molecular ligation, joining DNA fragments that were physically nearby at the time of crosslink. Subsequent deep sequencing of these DNA junctions generates a genome-wide contact probability map that allows the 3D modeling of genomic conformation within a cell. The strong enrichment in Hi-C signal between genetically neighboring loci allows the scaffolding of entire chromosomes from fragmented draft assemblies. Hi-C signal also preserves the cellular origin of each DNA fragment and its interacting partner, allowing for deconvolution and assembly of multi-chromosome genomes from a mixed population of organisms.

We have used Hi-C to scaffold whole genomes of animals, plants, fungi, as well as prokaryotes and archaea. We have also been able to use this data to annotate functional features of microbial genomes, such as centromeres in many fungal species. Additionally, we have applied our technology to diverse metagenomic populations such as craft beer, bacterial vaginosis infections, soil, and tree endophyte samples to discover and assemble the genomes of novel strains of known species as well as novel prokaryotes and eukaryotes.

The high quality of Hi-C-based assemblies allows the simultaneous closing of numerous unculturable genomes, placement of plasmids within host genomes, and microbial strain deconvolution in a way not possible with other methods.

BUILDING REFERENCE MATERIALS FOR MIXED MICROBIAL DETECTION WITH NEXT GENERATION SEQUENCING

Friday, 3rd June 10:30 La Fonda Ballroom Talk (OS-7.04)

Jason Kralj, Scott Jackson

National Institute of Standards and Technology

Myriad new sequencing and analytical technologies have emerged within the realm of metagenomic sequencing. This has promised to transform the way clinical and environmental samples are analyzed. Indeed, short (e.g. PGM Ion Torrent, MiSeq and HiSeq from Illumina) and long (Pacific Biosciences, Nanopore) read platforms have much to offer in terms of ability to detect the faintest traces of pathogens within a complex mixture of organisms. In addition, scores of post-sequencing analysis software packages offer a range of speed and sensitivity options to analyze the large data files, each with unique algorithms, databases, and interfaces. All this choice presents a serious question—does this particular analysis set work for my sample? While it is inconceivable that one could image the entire scope of experiments and sample types for microbial systems, there are a few basic analyses that would lend confidence to the abilities of these emerging tools.

We have endeavored to develop mixtures of microbial DNA with the purpose of identifying lim- its of detection, contamination, bias, noise, and general strengths/weaknesses of various shotgun metagenomics sequencing platfor Purified DNA represents an ideal condition that can be tightly controlled and analyzed, without other confounding pre-analytical variabilities such as extraction efficiency. Preliminary experiments using NIST reference materials (RMs) and prospective RMs have allowed us to make idealized mixtures of human, S. aureus, S. enterica subsp. enterica LT2,
P. aeruginosa, and C. sporogenes. Each well-defined mixture (one equigenomic, the other a dilution series) demonstrated the breadth of results from multiple analysis tools, and enable the end user to recognize the limits of these syste

The results indicated that at DNA "concentrations" approaching 1:10 000 (in a background of human DNA), or a few thousand total reads, detection was not challenging. However, accurate identification highly depended upon the underlying database used for each tool. Furthermore, some tools' filtering algorithms accurately discriminate accurate reads from reads containing significant numbers of sequencing errors, reducing the number of false positives. With this knowledge, we've proposed in silico analyses that will utilize reference data sets to inform new mixture designs that will provide benchmark analyses for the informatics tools. Meanwhile, we are expanding our suite of organisms to provide greater breadth of genomic characteristics (e.g. near neighbors, high/low G+C content, repetitive sequences) to stress test the sequencing pipeline. Together, these will enable developers to better ascertain the capabilities of their systems, and give regulatory agencies the tools and analyses needed to confidently evaluate the new tools in mixed microbial detection.

COFFEE BREAK

Sponsored by Dovetail Genomics

10:50 – 11:10

THE OIL PALM GENOME IN OUR PALMS: THE IMMEDIATE IMPACT OF THE GENOME SEQUENCE OF THE WORLD'S MOST IMPORTANT OIL CROP

Friday, 3rd June 11:10 La Fonda Ballroom Invited Speaker (IS-3)

Nathan Lakey
Orion Genomics, LLC

Nathan Lakey, Rajinder Singh, Meilina Ong-Abdullah, Eng-Ti Leslie Low, Rajanaidu Nookiah, Steven W Smith, Jared M Ordway, Robert A Martienssen and Ravigadevi Sambanthamurthi.

We recently reported 1.535 Gb of assembled sequence and transcriptome data from the African oil palm, Elaeis guineensis, and the interfertile S. American oil palm, Elaeis oleifera, which diverged in the new world. The oil palm sequence has already led to the discoveries of the VIR gene, the SHELL gene and the epigenetic cause of the mantled somaclonal abnormality. VIR is responsible for oil palm fruit color, and the identification of VIR mutations associated with the virescens fruit color phenotype will facilitate the development of elite breeding lines with a natural color indicator for fruit ripening. SHELL is responsible for increased oil yields in tenera hybrids through single gene heterosis. Statistical sampling of palms in independent planting sites throughout Malaysia suggests that a novel molecular precision agriculture approach involving SHELL gene testing and culling of non-tenera (low-performing) palms at the nursery stage before field planting will enable the exclusive planting of high yielding tenera pal This will result in a significant increase in oil yield from existing planted area and increasing wealth creation among the nation's poorest farmers. The genome sequence also enabled epigenome-wide association studies, which were used to find Karma, a transposon inserted into the MANTLED gene, whose methylation protects clonal planting materials from floral abnormalities. These discoveries, and more to come, are already helping to achieve sustainability for the most important oil crop worldwide.

Speaker's biographical sketch
Nathan D. Lakey, MBA, is Founding Principal, President & Chief Executive Officer of Orion Genomics, LLC. Lakey received a BA in Biochemistry from the University of Texas at Austin and an MBA from Washington University St. Louis. Lakey was regionally recognized with the top 40 under 40 award (2004 St. Louis), presented with the governor's top technology award (2005 Missouri) and currently serves as chairman of the Investment Advisory Committee, Biogenerator, and Chairman of Orion Biosains SDN BHD, and serves on the boards of Orion Genomics LLC, Missouri Baptist Hospital, Apse LLC, EpigenTX, INC, and YourBevCo LLC. Lakey has more than 25 years of experience in genomics, he was Director of DNA Sequencing at Millennium Pharmaceuticals, Inc., helped form Millennium Predictive Medicine, Millennium Biotherapeutics and Cereon Inc. Before joining Millennium, Lakey held various positions with Molecular Dynamics, Ambion Inc., and Harvard Medical School, Department of Genetics, in the laboratory of next generation sequencing pioneer, Professor George M. Church. Lakey holds numerous issued and pending patents in the US and around the world.

PACBIO, DOVETAIL, AND BIONANO TECHNOLOGIES ENABLE QUALITY PLANT GENOME ASSEMBLIES.

Friday, 3rd June 11:40 La Fonda Ballroom Talk (OS-8.01)

Thiruvarangan Ramaraj[1], Diego Fajardo[1], Karen Moll[1], Peng Zhou[2], Peter Tiffin[2], Jason Miller[3], Kevin Silverstein[2], Nevin Young[4], Joann Mudge[1]

[1]National Center for Genome Resources (NCGR), [2]University of Minnesota, [3]J. Craig Venter Institute, [4]University of New Mexico

Next generation sequencing and physical/optical mapping technologies have enabled interrogation of rapidly evolving genomic regions that generate a high frequency of tandem duplications. In plants, these regions are important because they control interactions with microbes. For Medicago truncatula, a nitrogen-fixing legume, interactions with microbial partners include not only plant defense, but symbiotic interactions with both nitrogen-fixing bacteria and fungal mycorrhizae, neither of which is present in the model plant, Arabidopsis thaliana. Here, we have used long reads (PacBio) to generate a de novo genome assembly of R108, an important Medicago truncatula accession because of its ability to be transformed. We have also generated a BioNano map and a Dovetail library for scaffolding. We've used the BioNano and Dovetail data to improve the continuity of the assembly, generating five different assemblies with different combinations and orderings of these long-range technologies. On these data, assembly connectivity increased with applications of BioNano or Dovetail and even further with their combination. The largest increase was produced having the Dovetail step precede the BioNano step.

A GENOME IN THE SUGAR BEET FIELD

Friday, 3rd June 12:00 La Fonda Ballroom Talk (OS-8.02)

Mitch McGrath[1], Belinda Townsend[2], Karen Davenport[3], Hajnalka Daligault[3], Shannon Johnson[3], Alex Hastie[4], Sven Bockland[4], Aude Darracq[5], Glenda Willems[5], Steve Barnes[5], Paul Galewski[6], Andy Funk[6], Jane Pulman[6], Tiffany Liu[6], Kevin Childs[6], Robert Bogden[7], Jon Wittendorp[7]

[1]USDA-ARS & Michigan State University, [2]Rothamsted Research, [3]Los Alamos National Laboratory, [4]BioNano Genomics, [5]SES Vanderhave, [6]Michigan State University, [7]Amplicon Express

Beets (Beta vulgaris) have been, and are, widely consumed food and fodder crops over the past three millennia, most recently for their luxurious production of the sweetener sucrose, and perhaps transitioning to a burgeoning energy and industrial crop that would hearken to its days fueling draft animals during pre-industrial times. As a member of the Carophyllales, a group of plant taxa known for their habitation in stressful and unusual environments, their unusual chemo-systematics, and as a sister eudicot lineage to the asterid and rosid clades, beets occupy a niche in crop production in cooler, northern temperate climates with remarkable ability to produce biomass and accumulate solutes on an annual cycle. The capacity of beets to meet growing needs is prodigious. Promises of productivity depend on the ability to engineer products, for which a schematic would be useful. Such a resource is available through the genome, which in this case, was assembled during 2015 from raw reads to contigs (PacBio, Illumina) to scaffolds (BioNano, Dovetail), resulting in 86 super-scaffolds in 566 Mb with 6% N's and collapsing to 9 pseudo-molecules plus 5 super-scaffolds (4.5 Mb) unassigned. Plant breeding only has two goals: Improvement in quality traits and protection of existing traits. This schematic is being explored for features of trait construction (accumulation of small molecules such as betalain pigments and sucrose), development, and genetic indicators of crop diversity (leafy chards vs roots of vegetable and industrial beets) as well as disease resistance traits (rhizomania "crazy root" disease, resistance gene analogs). Such a resource, tied to a host of inbred genetic populations which themselves are a novel and unique resource for discovering genes through traditional and new comparative approaches, now gives the ability to dissect the genetic architecture of beets, and the phenology of agronomic trait development in general, in a species where traditional genetic analyses have not been possible due to the out-crossing nature of its breeding system. Assembly of a high quality genome was relatively affordable from a deeply inbred individual. With a defined genetic architecture, variants detected at the population level, the level for which beet breeding has been most successful, are being identified with specific traits to gain insight into the biochemical and physiological processes contributing. Paired with gene expression changes across developmental transitions, specific genes responsible for bulk life-stage properties may be uncovered. Such phyto-centric information has never been available for beets, and plants in general, and strategies to exploit this information will likely still remain in the realm of particular crop idiosyncrasies. And beets do have many idiosyncrasies, which ideally, will be better interpreted and manipulated through close inspection of the beet genome.

TRAIT MAPPING AND IMPROVEMENT OF THE MELON FLY (BACTROCERA CUCURBITAE) GENOME

Friday, 3rd June 12:20 La Fonda Ballroom Talk (OS-8.03)

Sheina Sim[1], Scott Geib[2]

[1]University of Hawaii, Manoa, [2]USDA-ARS DKI US PBARC

The melon fruit fly Bactrocera cucurbitae (Coquillett), is a destructive agricultural pest and is the subject of strict quarantines that are enforced to prevent its establishment outside of its current geographic range. In addition to quarantine efforts, additional control measures are necessary for its eradication in the case of invasion to agriculturally rich areas. The sterile insect technique (SIT) has been effective in the control of medfly (Ceratitis capitata), and is part of a management strategy that regulatory agencies intend to expand to B. cucurbitae and other important pests.

A requirement of SIT is the availability of a genetic sexing strain (GSS) which enables the automation of sorting males from females so that only sterile males are released. In medfly, genetic sexing is based on pupal color (females have white pupae, males have wild-type brown pupae) and temperature sensitivity (females die at elevated temperatures, males can survive at elevated temperatures). Similarly, there exists a GSS for B. cucurbitae in which pupal color is also sexually dimorphic where females have a white pupal case and males have a wild-type brown pupal case, but its genetic basis is largely unknown. Genetic sexing by temperature sensitive lethal does not currently exist in Bactrocera. To facilitate the use of the B. cucurbitae GSS for SIT release, it is necessary to develop foundational tools for its biological, genetic, and genomic characterization to determine if the white pupae genes in B. cucurbitae and C. capitata are orthologs, and if it is possible to induce the tsl mutation in this species.

The first step in this is to identify the genetic basis of wp in B. cucurbitae. In this study, the whole B. cucurbitae genome was sequenced and assembled. Five mapping populations for this species were then sequenced and genotyped using a double digest restriction associated digest sequencing library. From this, a consensus linkage map for B. cucurbitae was generated and used to super-scaffold 69% of the draft assembly which includes 75% of annotated genes. White pupae was mapped to a few tightly linked loci found on one 8kb scaffold using QTL analysis. The locus with the highest LOD score was validated, showing consistency between genotype and phenotype. This data allows for the comparison of wp in melon fly with wp in medfly and the identification of the specific mutation causing wp.

CHARACTERIZING CHROMOSOMAL TRANSLOCATIONS ASSOCIATED WITH A GENETIC SEXING SYSTEM

Friday, 3rd June 12:40 La Fonda Ballroom Talk (OS-8.04)

Scott Geib[1], Sheina Sim[2]

[1]USDA-ARS DKI US PBARC, [2]University of Hawaii, Manoa

The Mediterranean fruit fly (medfly) is an important agricultural pest of many fruit and vegetable species. To protect the mainland United States from this pest, the sterile insect technique (SIT) is employed, involving release of tens of millions of sterile male medfly into the Los Angeles basin of California weekly. These flies have several mutations making them amenable to mass release. Due to a chromosomal translocation between the 5th chromosome and the male Y chromosome, females are homozygous recessive for both a temperature sensitive lethal mutation and a white pupal mutation, allowing straightforward separation of male and female flies and generation of male only release strains. Heterozygosity for these loci is maintained in males through the chromosomal translocation, linking wild-type phenotype with the sex chromosome. Utilizing several different sequenced based approaches, we compare the utility of these techniques to identifying the location of the Y-A translocation in this species. Structural variation was determined utilizing Hi-C (3C-seq) libraries, 10X genomics GemCode libraries and a high-density SNP based genetic maps derived from test crosses from both wild-type and translocated lab lines. In addition, scaffolding approaches utilizing these library types individually, as well as in combination with each other, are investigated for generating chromosome-scale genomic assemblies at a low cost.

LUNCH

Sponsored by Dovetail

13:00 – 14:00

IMPROVING GENOME ANALYSIS USING LINKED-READS

Friday, 3rd June 14:00 La Fonda Ballroom Talk (OS-9.01)

Deanna Church, Kristina Giorda, Cassandra Jabara, Sofia Kyriazopoulu Panagiotopoulou, Andrew Wei Xu, Heather Ordonez, Haynes Heaton, Mark Pratt, Patrick Marks, Paul Hardenbol, Adrian Fehr, Michael Schnall Levin

10x Genomics, Inc

High-throughput sequencing (HTS) has revolutionized genome analysis. Tens of thousands of genomes and hundreds of thousands of exomes have been analyzed globally allowing for new biological insights at both population and individual levels. Despite these advances, it has become increasingly clear that traditional methods are insufficient for providing a complete view of the genome. Paralogous sequences can often confound alignment, leaving biomedically important regions of the genome with low quality alignments and variant calls. Extracting information on large-scale events, includingcopy number variants (CNVs) and complex structural variants (SVs), is challenging using only short read data. Further, haplotype-level resolution in a single individual is not attainable using short read analysis. To address these problems, we have developed a technology that allows for the retention of long range information while retaining the power, accuracy, and scalability of short read sequencing technologies, producing a data type referred to as 'Linked-Reads' that enables a more complete analysis of a genome. At its core, haplotype-level dilution of long input molecules into over 1 million barcoded partitions allows for high-resolution reference-based analysis. We have demonstrated the ability to reconstruct individual haplotypes that span several megabases and have validated these haplotype reconstructions using trio sequencing data. Coupling Linked-Reads with novel algorithms that take advantage of these linkages allows for improved performance in regions of the genome that are typically inaccessible due to the presence of paralogous sequence. Validation of these variant calls has been challenging as they typically fall outside the Genome In a Bottle (GIAB) high confidence regions, but we have confirmed several hundred of these using orthogonal sequencing technologies. The power of the long range linkages also enables the improved detection of complex structural variants. In addition to identifying copy number variants (CNVs) we detect inter and intra-chromosomal events as well as more complex structural rearrangements. Linked-Read technology can be used in both a genome and targeted sequencing context, allowing access to a broader range of applications. The development of Linked-Reads is an important step in the evolution of genome analysis by allowing access to more of the genome, resolving complex variants and reconstructing long-range haplotypes.

A REFERENCE-AGNOSTIC AND RAPIDLY QUERYABLE NGS READ DATA FORMAT ALLOWS FOR FLEXIBLE ANALYSIS AT SCALE

Friday, 3rd June 14:20 La Fonda Ballroom Talk (OS-9.02)

Niranjan Shekar[1], William Salerno[2], Adam English[2], Adina Mangubat[1], Jeremy Bruestle[1], Eric Boerwinkle[3], Richard Gibbs[2]

[1]Spiral Genetics Inc, [2]Human Genome Sequencing Center Baylor College of Medicine, [3]University of Texas Health Science Center at Houston

In identifying the complement of genetic variants that are associated with complex disease, larger sample sizes increase power. Studies such as the Alzheimer's Disease Sequencing Project and the CHARGE Consortium where samples are collected from a range of centers show heterogeneous data, requiring informatics that can additively scale to thousands of samples and analytics that go beyond identifying small variants in NGS data. At scale, the challenge of evaluating SNPs, indels and SVs becomes the "N+1" problem of incrementally adding samples without having to perpetually reevaluate petabytes of population read data stored in BAM files.

The Biograph Analysis Format (BAF) is a method of indexing NGS data that extends the Burrows Wheeler Transform to allow for multiple paths, effectively creating a read overlap graph of the data. A BAF of HiSeq X 30x WGS data is 8.3 Gb, 95% smaller than the corresponding BAM. Generated from the BAM in 14 hours, the BAF can be queried up to 200,000 times a second. Multiple BAFs can be combined, which at scale results in a file size of approximately 3GB per individual. Because the BAF can be batched across individuals, query time grows less than linearly with the number of individuals.

For example, if 30,000 putative SV sites to be queried, SV-typing these sites across 10,000 HiSeq X WGS samples in BioGraph Analysis Format would require less than 30 TB of storage (for all the read data), 16 CPU hours, and 10 minutes (using 100 machines).

Here, we perform read over assembly to genotype 4,276 SVs larger than 80bp detected in at least one individual of the Ashkenazi Jewish Trio by Pindel. At 1,195 of these locations, there was at least one SV call in any one individual and all of these calls, except for 25 (2.1%) were consistent with mendelian inheritance. Further, read overlap assembly to genotype variants was performed at 3,935 locations where PBHoney called an SV with long read sequencing data on the same Trio. Of those, 1,327 locations had at least one genotype with all but 55 (4.1%) being consistent with mendelian inheritance.

Additionally, the data are reference-agnostic, so variants can be called against any reference or against the read graph of any other set of individuals, dramatically reducing the time for data harmonization. Further, information is divided such that the "read overlap graph" created from all the individuals is separate from the information indicating that path through the graph for each individual. This allows a search for a particular variation of interest directly from the read data remotely and rapidly, without the opportunity to reveal the exact individual(s) from that the variant originates.

Because the data are essentially a read overlap graph, it is possible to accurately characterize SVs by traversing the graph from a particular location or search for a particular sequence associated with the SV. So, fast querying of small files with reasonable compute requirements provides an N+1 solution for SVs.

RELIABLE DETECTION OF COPY NUMBER VARIATIONS BASED ON DATA MINING APPROACHES

Friday, 3rd June 14:40 La Fonda Ballroom Talk (OS-9.03)

Zbyszek Otwinowski, Maciej Puzio, Dominika Borek

UT Southwestern Medical Center

Copy number variations (CNVs) are important for understanding biology, and cancer development in particular. Despite significant progress in the development of tools for directly detecting CNVs in sequencing reads, no single tool detects all types of CNVs. This is because all tools analyze either explicitly or implicitly coverage variability in order to model its overdispersion, but they do not identify all sources of variations. There are many contributors to overdispersion other than CNVs, and these contributors vary significantly between experiments, resulting from biases in fragmentation and other steps of library preparation, systematic sequencing errors and artifacts of mapping.

We designed a method to separate these artifactual contributions from the biological signal, i.e. the CNVs. By means of data mining, we map the patterns of artifactual variability on a genome of interest and then we correct the coverage distribution for these patterns so that the resulting distribution can be used in efficient and reliable CNV detection. The reduction of overdispersion provides a very stringent validation criterion.

The method also has great potential to further the analysis of pan-genome CNVs variations and to determine whether particular CNVs represent a population or a private variant. For the latter, on average we expect different phenotypic effect.

ROBUST GENOME-WIDE TRANSCRIPTOME ENRICHMENT SEQUENCING FOR RESEARCH AND CLINICAL PLATFORMS

Friday, 3rd June 15:00 La Fonda Ballroom Talk (OS-9.04)

Harsha Doddapaneni[1], Prof. Jianhong Hu[1], Hsu Chao[2], Simon White[2], Tittu Matthew[2], Viktoriya Korchina[2], Caitlin Nessner[2], Sandra Lee[2], Donald W Parsons[3], David A Wheeler[2], Angshumoy Roy[4], Eric Boerwinkle[5], Donna Muzny[1], Richard Gibbs[2]

[1]Human Genome Sequencing Center Baylor College of Medicine, [2]Human Genome Sequencing Center, Baylor College of Medicine, [3]Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, [4]Department of Pathology & Immunology, Baylor College of Medicine; Department of Pediatrics, Texas Children's Hospital, [5]University of Texas Health Science Center at Houston

Transcriptome sequencing (RNA-Seq) together with whole genome sequencing (WGS) offers an in- tegrated informative dataset for functional characterization of human transcriptome. Generation of high-quality data is essential for success in RNA-Seq studies. However, standard RNA-Seq methods rely heavily on very high quality RNA samples, and the library preparation involves multiple enzy- matic manipulations as RNA is first converted to double-stranded cDNA and then into paired-end libraries. Therefore to increase RNA-Seq utility in routine clinical setting, these protocols have to overcome RNA quality constraints as well as need rapid preparation protocols.

Since August 2010, at BCM-HGSC we have generated transcriptome data for 4519 samples in support of different cancer, pharmacogenomics, vascular and other disease projects. These RNA-Seq libraries are strand-specific and poly(A)+ enriched and use automated library preparation workflow. To monitor sample and process variability, in addition to sequence metrics, we use RNA developed by the External RNA Controls Consortium (ERCC). Our primary RNA-Seq data analysis pipeline is built on STAR and Cuff-Links software and we assess performance of individual RNA-Seq libraries by 12 different metrics for library process consistency.

Since 2014, we have supported exome-capture transcriptome-seq by combining our strand-specific RNA-Seq protocol with whole exome sequencing protocol as tool to handle low quality RNA and RNA extracted from FFPE specimens. Total RNA (40 ng) from control samples as well as FFPE samples from cancer patients (2-3 RIN and DV 200 of >30%) sequenced using this protocol have shown that sensitivity of this approach for detecting fusions and somatic SNVs is comparable to that of standard poly-A+ enriched libraries.

Our recent efforts are focused on designing ways in which we can simplify the RNA-Seq protocol to reduce sample preparation time as well as increase sample throughput. In this regard, we have developed a RNA-Seq protocol using first strand cDNA as a template for preparing libraries using Accel-NGS 1S DNA Library Kit from Swift Biosciences. Libraries prepared using Universal Hu- man Reference (UHR) RNA and human Placenta RNA with this protocol generated 80-120 million reads/sample with high unique rates (89 % for UHR and 75 % for Placenta). Comparison of gene ex- pression values as Fragments Per Kilobase of transcript per Million mapped reads (FPKM) between 1S and RNA-Seq protocol gave a correlation of 0.97 for UHR sample and 0.97 for Placenta sample. Over all, this workflow eliminates the need to generate second cDNA synthesis and also reduces the library preparation time by half over standard poly-A+ enriched workflow. This protocol also allows to multiplex samples after ligating the molecular barcodes to the samples, but before PCR to increase sample throughput.

In all, these improvements to our RNA-Seq workflows will allows us to handle low quality RNA including RNA from FFPE specimens as well as result in fast turnaround times to have practical utility in both research and clinical settings.

DEEP RNA SEQUENCING OF BRASSICA RAPA RIL POPULATION FOR SNP DISCOVERY, GENETIC MAP CONSTRUCTION, EQTL ANALYSIS, AND GENOME IMPROVEMENT

Friday, 3rd June 15:20 La Fonda Ballroom Talk (OS-9.05)

Mike Covington[1], RJ Cody Markelz[2], Upendra Kumar Devisetty[3], Marcus Brock[4], Cynthia Weinig[4], Julin Maloof[2]

[1]University of California, Davis Amaryllis Nucleics, [2]University of California, Davis, [3]University of Arizona, [4]University of Wyoming

Here we describe deep RNA sequencing, high-density SNP discovery, and genetic map construction for a recombinant inbred line population derived from the Brassica rapa accessions R500 and IMB211. We demonstrate how this new resource is a significant improvement for QTL analysis over the current low-density genetic map. We also use the genotype data from the population to detect and remedy putative genome misassemblies and to assign scaffold sequences to their likely genomic locations. These improvements to the assembly represent 7.1-8.0% of the annotated Brassica rapa genome.

COFFEE BREAK

Sponsored by Promega

15:40 – 16:00

ASAP: A CUSTOMIZABLE AMPLICON SEQUENCING ANALYSIS PIPELINE FOR HIGH-THROUGHPUT CHARACTERIZATION OF COMPLEX SAMPLES

Friday, 3rd June 16:00 La Fonda Ballroom Talk (OS-10.01)

Darrin Lemmer[1], Jolene Bowers[1], Erin Kelley[1], Rebecca Colman[1], Matt Enright[1], Elizabeth Driebe[1], James Schupp[1], David Engelthaler[1], Paul Keim[2]

[1]TGen North, [2]TGen/Northern Arizona University

A novel technique, Universal Tail amplicon sequencing, allows for multiplexing numerous target amplicons for multiple bacterial samples together on the same sequencing run. Targeted, multiplexed, amplicon sequencing is useful for many applications, such as resistance gene detection, metagenomic sample characterization, biosurveillance, and forensics. For example, this technique is ideal for analyzing clinical samples, as tens to hundreds of different DNA-based assays can be run directly on each sample without having to culture bacterial isolates. Human DNA contamination is limited, so the pathogen signal is not masked as it would be for full metagenomic sequencing. Using this technique, we have sequenced more than 200 targets for 100 samples at up to 10,000x coverage on a single MiSeq run, resulting in massive amounts of data to analyze and interpret.

The Amplicon Sequencing Analysis Pipeline (ASAP) is a highly customizable, automated way to examine amplicon sequencing data. The important details of the amplicon targets are described in a text-based input file written in JavaScript Object Notation (JSON). This data includes the target name, genetic sequence (or sequences in the case of gene variant assays), any known SNPs or regions of interest (ROIs) within the target, and what the presence of this target or SNP signifies, clinically. This file can be hand-generated or created from an Excel spreadsheet using a provided template and Python script. The sequenced reads are processed by performing adapter, and optionally, quality trimming, and then aligned to the reference amplicon sequences extracted from the JSON file using one of several aligners. The resulting BAM files are analyzed with a custom Python script that combines the alignment data in the BAM file with the assay data in the JSON file and interprets the results. The output is an XML file with complete details for each assay against each sample. These details include number of reads aligning to each target, any SNPs found above a user-defined threshold, and the nucleotide distribution at each of these SNP positions. For ROI assays, the output includes the sequence distribution at each of the regions of interest both the DNA sequences and translated into amino acid sequences. Also, each assay target is assigned a significance if it meets the requirements laid out in the JSON file (i.e. a particular SNP or amino acid change is present) To make this output easier for the user to interpret, a number of XSLT stylesheets are provided for transforming the XML output into other, more readable formats, including Excel spreadsheets, web pages and PDF documents. Additionally, the use of XSLT stylesheets allows for multiple different views of the same data, from clinical summaries showing only the most important or relevant results to full researcher summaries containing all of the data. While designed for analyzing amplicons, ASAP works just as well for finding any gene targets, specific SNPs, or other biomarkers in whole genome sequencing data.

EVALUATION OF HISEQ X TEN PERFORMANCE: TOWARDS CLINICAL APPLICATIONS

Friday, 3rd June 16:20 La Fonda Ballroom Talk (OS-10.02)

Kimberly Walker[1], Rashesh Sanghvi[1], Qiaoyan Wang[1], Harsha Doddapaneni[1], Jianhong Hu[1], Adam English[1], William Salerno[1], Yi Han[1], Huyen Dinh[1], Eric Boerwinkle[2], Richard Gibbs[1], Donna Muzny[1]

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

High-throughput parallel nucleotide sequencing has revolutionized genomic research and reshaped applications in clinical health care. The HiSeq X Ten platform further expands these opportunities with unprecedented capacity. The Human Genome Sequencing Center (HGSC) at Baylor College of Medicine adopted the HiSeq X Ten system in the fall of 2014, with a view to eventual deployment in a CAP/CLIA environment.

To evaluate the instruments, we have analyzed more than 1,093 flowcells, representing >8,441 30X human genomes. These studies have included common disease cohorts, inherited cancers, mendelian disease cases as well as DNA from cell lines of lung and endometrial cancer. PCR-Free library methods (Illumina, Kapa Biosystems, and Swift Biosciences) have been evaluated and implemented for optimize coverage in GC-rich regions. Metrics related to coverage, sample integrity and variant representation were established to ensure high quality genome sequencing.

Based on our experience with the HiSeq X platform, we have implemented several standard metrics including >53% Pass Filter, >90% aligned bases, <3.0% error rate, >85% unique reads and >75% Q30 bases to achieve at least 90 GB unique aligned bases per lane. These are utilized for daily tracking of quality. Genome coverage metrics are also tracked to achieve 90% of genome covered at 20x and 95% at 10x with a minimum of 86 x 109 mapped, aligned bases with Q20 or higher. Additional metrics such as library insert size (mode and mean) per sample, duplicate reads, read 1 and read 2 error rates, % pair reads and mean quality scores are also monitored. Platform sensitivity and precision at ~30 coverage was determined to be 97.8% and 99.6% respectively using control sample NA12878. To ensure integrity in our production pipeline, we have implemented the SNPTrace assay by Fluidigm to confirm sample identity and VerifyBamID to detect sample contamination. Assessment of appropriate coverage benchmarks for clinically relevant genes and variants utilizing the OMIM gene list is in progress.

These evaluation efforts have provided valuable insight as to how sequencing depth and coverage uniformity impact the ability to accurately detect variants. Overall the platform has been consistent in performance. Recent data has shown stability in platform run-to-run yield and quality in more than 1,600 PCR-Free Kapa Hyper library samples achieving the high quality metrics described above. Establishment of robust PCR-Free WGS methods and associated pipeline metrics are essential for broad applications in both the research and clinical setting.

YCGA BIOINFORMATICS AT YALE

Friday, 3rd June 16:40 La Fonda Ballroom Talk (OS-10.03)

James Knight

Yale University

The Yale Center for Genome Analysis provides DNA sequencing services for Yale and external customers, with a focus on whole-exome, whole-genome, RNA-seq and ChIP-seq sequencing. The YCGA bioinformatics group provides bioinformatic analysis services both for custom projects and general tools made available to the Yale community, and to all. We describe recent developments and research in the YCGA bioinformatics group, including (1) an analysis of the newest exome kits from Nimblegen and IDT, which close the coverage gap with WGS and match its sensitivity for detecting coding region variants, (2) integration and support for both hg19 and hg38 reference analysis for GATK whole-exome and whole-genome analysis, (3) tertiary analysis tools for trios, families and cohorts of samples (annotation, denovo mutations, gene burden, kinship analysis, …), (4) combined variant/LOH/CNV analysis of FFPE tumor/normal exome datasets, and (5) workflow software for quickly scripting cluster-based, computational pipelines.

SCALABLE AND EXTENSIBLE NEXT-GENERATION SEQUENCE ANALYSIS PIPELINE MANAGEMENT FOR OVER 50,000 WHOLE-GENOME SAMPLES

Friday, 3rd June 17:00 La Fonda Ballroom Talk (OS-10.04)

Jesse Farek, Adam English, Daniel Hughes, William Salerno, Kimberly Walker, Donna Muzny, Richard Gibbs

[1]Human Genome Sequencing Center Baylor College of Medicine

The Baylor College of Medicine Human Genome Sequencing Center (HGSC) has recently added Illumina HiSeq X Ten sequencers to its sequencing fleet, which currently processes over 2,000 whole-genome samples per month. These samples originate from multiple projects and collaborators, including the Alzheimer's Disease Sequencing Project, the Trans-Omics for Precision Medicine Program, Baylor Miraca Genetics Laboratory, the Centers for Common Disease Genomics, the CHARGE Consortium, and the Center for Mendelian Genomics. In order for sequence analysis at the HGSC to scale to this increased workload, numerous improvements have been made to the efficiency and reliability of the center's sequence analysis infrastructure. HgV is the workflow management system for primary and secondary Illumina sequence analysis at the HGSC and features tiered XML pipeline protocols, job tracking, LIMS communication, verbose logging and stable reproducibility. HgV's protocol definition and LIMS communication infrastructure has been reworked for greater configurability so that pipeline protocol and LIMS parameters can be easily modified to accommodate different project requirements. Specifically, the HGSC has configured HgV use both local and cloud-based compute resources and to enforce CAP- and CLIA-compliant data handling for clinical pipelines. Secondary analysis programs have been rewritten for increased computational efficiency. The Atlas2 SNP and Indel variant callers (originally written in Ruby) have been rewritten and combined into a single C++ program that runs on average more than 50 times faster than Atlas2 and with improved variant calling quality and consistency. Two new custom reporting programs, SeqAnalyzer, which calculates FASTQ sequence metrics, and AlignStats, which calculates BAM alignment and coverage metrics, have been written to use significantly fewer computational resources than the existing programs they replace. Other areas of improvements to analysis workflows have also been investigated, including measuring the effects of local Indel realignment and base quality recalibration on variant call quality and researching efficient N+1 joint calling solutions for creating project level VCFs. These improvements have resulted in fast, extensible, and easily manageable analysis pipelines for human resequencing and other applications on the HiSeq X platform that have allowed the HGSC to concurrently support the heterogeneous analysis requirements of multiple large-scale sequencing projects. To date, HgV has managed the analysis of over 5,000 whole-genome samples and is expected to handle over 50,000 more samples in the near future.

Attendee Listing

First Name	Last Name	Company	Email
Audrey	Abrams McLean	Centers for Disease Control and Prevention	xxp7@cdc.gov
Omayma	Al-Awar	Illumina, Inc	oalawar@illumina.com
Johar	Ali	AA	ali.johar@gmail.com
Michael	Alonge	Driscoll's	michael.alonge@driscolls.com
Jayaleka	Amarasinghe	US Food and Drug Administration	Jayaleka.Amarasinghe@fda.hhs.gov
Mercedes	Ames	Biomax Informatics	merivames@gmail.com
Warren	Andrews	BioNano Genomics Inc.	wandrews@bionanogenomics.com
Elnaz Saberi	Ansari	Institute for Research in Fundamental Sciences	elnaz.saberiansari@gmail.com
Maryke	Appel	Kapa Biosystems	maryke.appel@kapabiosystems.com
Sereena	Aragon	Albuquerque Police Department	Snaragon@cabq.gov
Carlos	Arana	UT Southwestern Medical Center	carlos.arana@utsouthwestern.edu
Eugene	Arinaitwe	National Animal Disease Diagnostics and Epidemiology Centre	arieugene@yahoo.com
Aroh	Aroh	UT Southwestern Medical Center	Chukwuemika.aroh@utsouthwestern.edu
Jonathan	Arzt	Plum Island Animal Disease Center	Jonathan.Arzt@ARS.USDA.GOV
Jonathan	Atencio	Pacific Biosciences, Inc.	jatencio@pacificbiosciences.com
Jennifer	Ayres	Illumina, Inc	jayres@illumina.com
Giorgi	Babuadze	CBEP	gbabuadze@ncdc.ge
Donovan	Bailey	New Mexico State University	dbailey@nmsu.edu
Robert	Baker	Texas Tech University	Robert.Baker@ttu.edu
Sanjeev	Balakrishnan	Dovetail Genomics	jeev@dovetail-genomics.com
Anthony	Baniaga	University of Arizona	abaniaga@email.arizona.edu
Andrew	Barry	New England Biolabs	barrya@neb.com
Dhwani	Batra	Chenega/CDC	dhwani.govil@gmail.com
Elmira	Begimbayeva	Kazakh Scientific Center for Quarantine and Zoonotic Diseases	ebegimbayeva@kscqzd.kz
Elmira	Begimbayeva	Research Institute for Biological Safety Problems	ebegimbayeva@kscqzd.kz
Callum	Bell	National Center for Genome Resources	cjb@ncgr.org
Laura	Bell	Albuquerque Police Department	ljbell@cabq.gov
Lindsay	Benage	Los Alamos National Laboratory	lbenage@lanl.gov
Dmitriy	Berezovskiy	Kazakh Scientific Center for Quarantine and Zoonotic Diseases	dberezovsky@kscqzd.kz
Dmitryi	Berezovskyi	Research Institute for Biological Safety Problems	dberezovsky@kscqzd.kz
Nicolas	Berthet	Centre International de Recherches Médicales de Franceville (CIRMF)	nicolas.berthet@ird.fr
Jasbir	Bhangoo	Driscoll's	jasbir.bhangoo@driscolls.com
Joseph	Bogan	MRIGlobal	jbogan@mriglobal.org
Bonnie	Bond	AK PHL	bonniejbond@gmail.com
Dominika	Borek	UT Southwestern Medical Center	dominika@work.swmed.edu
Mark	Borodovsky	Georgia Institute of Technology	borodovsky@gatech.edu
Cecilie	Boysen	Qiagen	cboysen@mac.com
Cecilie	Boysen	Qiagen	cecilie.boysen@qiagen.com
Christopher	Bradburne	Johns Hopkins University, Applied Physics Lab	chris.bradburne@jhuapl.edu
Cathy	Branda	Sandia National Laboratories	cbranda@sandia.gov
Michael	Brandhagen	Federal Bureau of Investigation	Michael.Brandhagen@ic.fbi.gov
Barrett	Bready	Nabsys	bready@nabsys.com
David	Bruce	Los Alamos National Laboratory	dbruce@lanl.gov
Lijing	Bu	University of New Mexico	lijing@unm.edu
Yerbol	Burashev	Research Institute for Biological Safety Problems	yerbol.bur@gmail.com
Robert	Calef	Dovetail Genomics	robert@dovetail-genomics.com
Heather	Carleton	Centers for Disease Control and Prevention	hcarleton@cdc.gov
Traci	Carlson	FBI/ORISE	traci.carlson@ic.fbi.gov
John	Cartee	Centers for Disease Control and Prevention	yil5@cdc.gov
Michael	Cassler	MRIGlobal	mcassler@mriglobal.org
Frédéric	Chain	McGill University	frederic.chain@mail.mcgill.ca
Patrick	Chain	Los Alamos National Laboratory	pchain@lanl.gov
Gvantsa	Chanturia	NCDC	gvantsa.chanturia@ncdc.ge

First Name	Last Name	Company	Email
Olga	Chertkov	Los Alamos National Laboratory	ochrtkv@lanl.gov
William	Chow	Wellcome Trust Sanger Institute	wc2@sanger.ac.uk
Deanna	Church	10X Genomics	deanna.church@10xgenomics.com
Alan	Cleary	Montana State University	alan.cleary@msu.montana.edu
Rebecca	Colman	University of California, San Diego	beckyecolman@gmail.com
Rita	Colwell	U Maryland & Johns Hopkins School of Public Health	rcolwell@umiacs.umd.edu
Alex	Copeland	JGI	accopeland@lbl.gov
Mike	Covington	Amaryllis Nucleics, Inc.	mfcovington@gmail.com
Rachel	Creager	Defense Forensic Science Center	rachel.l.creager-allen.ctr@mail.mil
Helen	Cui	Los Alamos National Laboratory	hhcui@lanl.gov
Anna	Cunanan	Roche Diagnostics	anne.cunanan@roche.com
Kevin	Cupit	Front Range Community College	kcupit@gmail.com
Hajnalka	Daligault	Los Alamos National Laboratory	hajkis@lanl.gov
Paul	Daniel	Qiagen	Nonna.Druker@qiagen.com
Debanjana	Dasgupta	MRIGlobal	ddasgupta@mriglobal.org
David	Davenport	US Army Crime Lab	david.m.davenport30.civ@mail.mil
Karen	Davenport	Los Alamos National Laboratory	kwdavenport@lanl.gov
Josie	Delisle	TGen North	jdelisle@tgen.org
Chris	Detter	Los Alamos National Laboratory	cdetter@yahoo.com
Armand	Dichosa	Los Alamos National Laboratory	armand@lanl.gov
Todd	Dickinson	Dovetail Genomics	todd@dovetail-genomics.com
Sheila	Diepold	Tetracore	sdiepold@tetracore.com
Kariena	Dill	Dovetail Genomics, LLC	kariena@dovetail-genomics.com
Darrell	Dinwiddie	University of New Mexico	dldinwiddie@salud.unm.edu
Harsha V	Doddapaneni	Baylor College of Medicine	doddapan@bcm.edu
Norman	Doggett	Los Alamos National Laboratory	doggett@lanl.gov
Joseph	Donfack	FBI Laboratory	joseph.donfack@ic.fbi.gov
Qian	Dong	Qiagen	Nonna.Druker@qiagen.com
James	Dowling	Albuquerque Police Department	jdowling@cabq.gov
Russ	DuChene	Labcyte	russ.duchene@labcyte.com
Robert	Duncan	Texas Tech University	bigbitbucket@mac.com
Giorgi	Dzavashvili	NCDC	dzavashviligeorge@gmail.com
Ashlee	Earl	Broad Institute	aearl@broadinstitute.org
Jennifer	Elwell	Albuquerque Police Department	jelwell@cabq.gov
Adam	English	HGSC @ BCM	english@bcm.edu
Tracy	Erkkila	Los Alamos National Laboratory	terkkila@lanl.gov
Matthew	Ezewudo	Critical Path Institute	mezewudo@c-path.org
Diego	Fajardo	National Center for Genome Resources	dfajardo@ncgr.org
Jesse	Farek	Baylor College of Medicine	jesse.farek@bcm.edu
Andrew	Farmer	National Center for Genome Resources	adf@ncgr.org
Kevin	Fengler	Dupont Pioneer	kevin.a.fengler@pioneer.com
Haley	Fiske	Swift Biosciences	fiske@swiftbiosci.com
Michael	FitzGerald	Broad Institute	fitz@broadinstitute.org
Bob	Fulton	Washington University	bfulton@genome.wustl.edu
James	Gale	Tricore Reference Lab	James.Gale@tricore.org
Alfredo	Garcia	Government Scientific Source	agarcia@govsci.com
Daniel	Garcia	Centers for Disease Control and Prevention	int7@cdc.gov
Scott	Geib	USDA-ARS	scott.geib@ars.usda.gov
Christophe	Georgescu	Broad Institute of MIT and Harvard	cgeorges@broadinstitute.org
Michael	Gilmore	Harvard Medical School	michael_gilmore@meei.harvard.edu
Cheryl	Gleasner	Los Alamos National Laboratory	cdgle@lanl.gov
Rosalie	Gomez	Government Scientific Source	rgomez@govsci.com
Darren	Grafham	Children's Hospital	darrengrafham@gmail.com
Jeff	Gunter	US Department of Defense	jngunter@gmail.com
Jon	Hagopian	Advanced Analytical	jhagopian@aati-us.com
Jessica	Halpin	Centers for Disease Control and Prevention	JLHalpin@cdc.gov
John	Hanson	RTL Genomics	j.delton.hanson@researchandtesting.com
Tim	Harkins	Swift Biosciences	harkins@swiftbiosci.com
Alex	Hastie	BioNano Genomics Inc.	ahastie@bionanogenomics.com
John	Havens	Integrated DNA Technologies	jhavens@idtdna.com
Paul	Havlak	Dovetail Genomics	havlak@dovetail-genomics.com
Derek	Hofer	Department of Defense	dmhofer@verizon.net

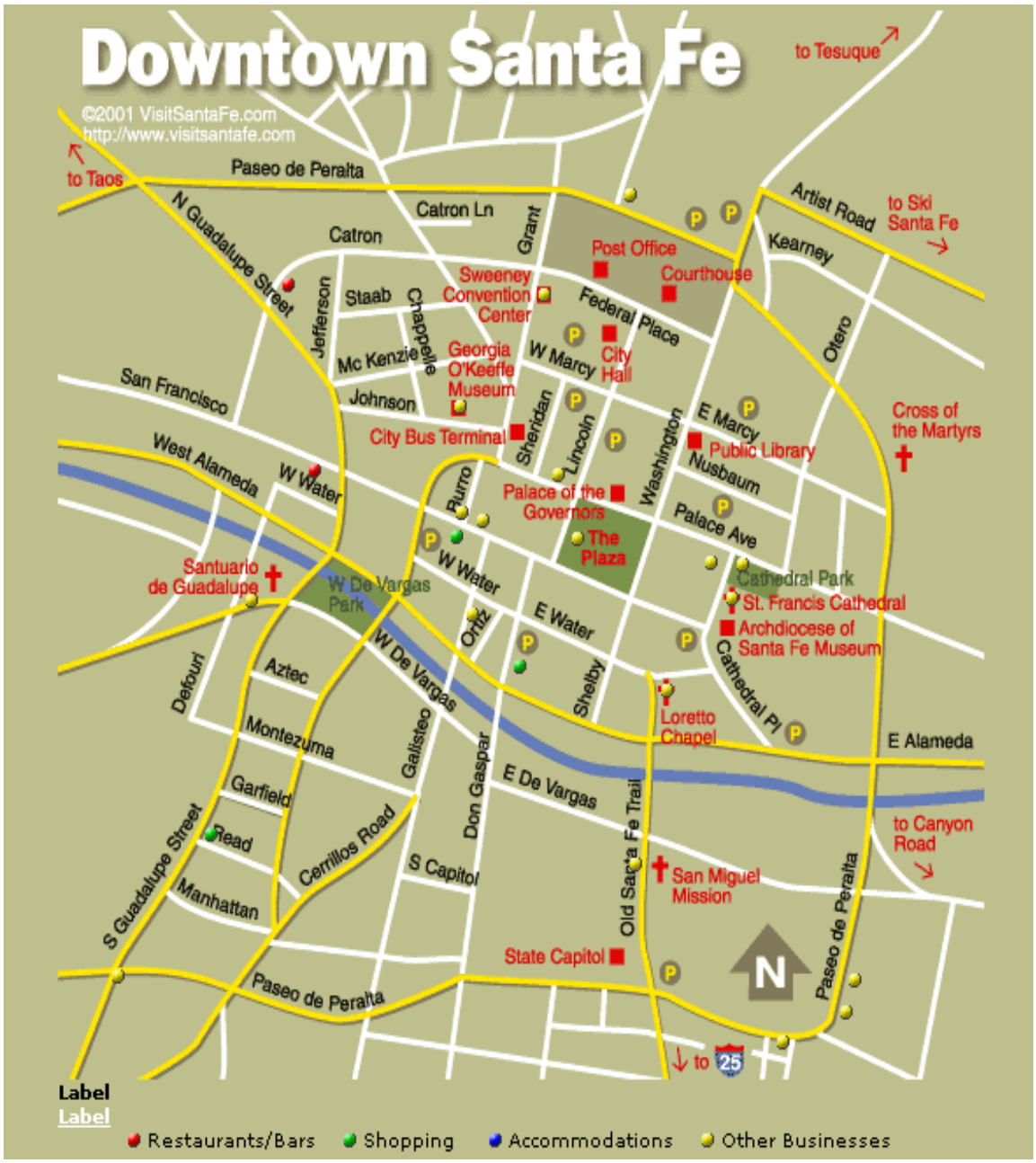
First Name	Last Name	Company	Email
Michael	Holder	Baylor College of Medicine	mholder@bcm.edu
Kelly	Hoon	Illumina, Inc	khoon@illumina.com
James	Horne	Los Alamos National Laboratory	j.horne308@gmail.com
Blake	Hovde	Los Alamos National Laboratory	hovdebt@lanl.gov
Bin	Hu	CSRA, CDC	hubin.keio@gmail.com
Andrew	Huang	Centers for Disease Control and Prevention	wwm8@cdc.gov
Previn	Hudetz	St. Pius X HS	previnhudetz@gmail.com
tim	hunkapiller	Discovery Biosciences	tim@discoverybio.com
Sung	Im	Centers for Disease Control and Prevention	wla9@cdc.gov
Emily	Innis	Los Alamos National Laboratory	einnis@lanl.gov
Jodi	Irwin	FBI	jodi.irwin@ic.fbi.gov
Omar	Ishak	Los Alamos National Laboratory	mishak@lanl.gov
Srinivas	Iyer	Los Alamos National Laboratory	siyer@lanl.gov
Scott	Jackson	NIST	scott.jackson@nist.gov
Val	Jackson-Hundley	Los Alamos National Laboratory	valjh@lanl.gov
Jacobs	Jacobs	MRIGlobal	jonathan.jacobs@gmail.com
David	Jaffe	Broad Institute	jaffe@broadinstitute.org
Anitha	Jayaprakash	Girihlet Inc	anitha@girihlet.com
Hanlee	ji	Stanford University	genomics_ji@stanford.edu
Nicole	Johnson	DOE Joint Genome Institute	nicolevjohnson@lbl.gov
Shannon	Johnson	Los Alamos National Laboratory	shannonj@lanl.gov
Lavin	Joseph	Centers for Disease Control and Prevention	GYU4@CDC.GOV
Bonaventure	Juma	Centers for Disease Control and Prevention	xwl2@cdc.gov
Ekaterina	Kalinkevich	Phytoengineering R&D Center	e.kalinkevich@phytoengineering.ru
Okumu	Kaluoch	US Food and Drug Administration	okumu.kaluoch@fda.hhs.gov
Laura	Kavanaugh	Syngenta Biotechology	laura.kavanaugh@syngenta.com
John	Kayiwa	Uganda Virus Research Institute	jkayiwa@uvri.go.ug
Jonathan	Kayondo	Uganda Virus Research Institute	jkayondo@gmail.com
Susan	Kerfua	National Livestock Resources Research Institute	kerfuas@gmail.com
Gladys	Kiggundu	National Animal Disease and Epidemiology Center	gladyskiggundu@yahoo.com
Luke	Kingry	Centers for Disease Control and Prevention	vtx8@cdc.gov
William	Klimke	NCBI/NLM/NIH/DHHS	klimke@ncbi.nlm.nih.gov
James	Knight	Yale University	j.knight@yale.edu
Kristen	Knipe	Centers for Disease Control and Prevention	wgg9@cdc.gov
Lars	Koenig	RTL Genomics	lars.koenig@researchandtesting.com
Frank	Kolakowski	Tetracore, Inc.	fkolakowski@tetracore.com
Anton	Korobeynikov	Saint Petersburg State University	anton@korobeynikov.info
Nato	Kotaria	NCDC	n.kotaria@ncdc.ge
Adam	Kotorashvili	NCDC	adam.kotorashvili@gmail.com
Alexander	Kozik	University of California, Davis	akozik@ATGC.ORG
John	Krebsbach	Albuquerque Police Department	jfkrebsbach@gmail.com
Anand	Kuamr	Los Alamos National Laboratory	dranandgundu@gmail.com
Ravi	Kumar	Novozymes	rvku@novozymes.com
Jochen	Kumm	InsightfulAI	jochen.kumm@gmail.com
Berzhan	Kurmanov	Kazakh Scientific Center for Quarantine and Zoonotic Diseases	berzhan@bk.ru
Ingrid	Labouba	Centre International de Recherches Médicales de Franceville (CIRMF)	ilabouba@gmail.com
Nathan	Lakey	Orion Genomics	lakey@oriongenomics.com
Ernest	Lam	BioNano Genomics Inc.	Elam@bionanogenomics.com
Alla	Lapidus	Saint Petersburg State University	yevalmi@gmail.com
Steve	Lasky	Advanced Analytical	slasky@aati-us.com
Haythem	Latif	GENEWIZ	haythem.latif@genewiz.com
Adrian	Lawsin	Centers for Disease Control and Prevention	kqj9@cdc.gov
Darrin	Lemmer	TGen North	dlemmer@tgen.org
April	Lewis	Bioo Scientific	aprill@biooscientific.com
Ivan	Liachko	University of Washington	ivanliachko@gmail.com
Shoudan	Liang	Pacific Biosciences, Inc.	sliang@pacificbiosciences.com
Sabina	Lindley	US Food and Drug Administration CFSAN	sabina.lindley@fda.hhs.gov
Carol	Lindsey	Los Alamos National Laboratory	ckcarr@lanl.gov
Anna	Lipzen	DOE Joint Genome Institute	alipzen@lbl.gov
kun connie	liu	US Food and Drug Administration	kun.liu@gmail.com

First Name	Last Name	Company	Email
Chienchi	Lo	Los Alamos National Laboratory	chienchi@lanl.gov
Chad	Locklear	Integrated DNA Technologies	Clocklear@idtdna.com
Vladimir	Loparev	Centers for Disease Control and Prevention	vnl0@cdc.gov
Sean	Lucking	Centers for Disease Control and Prevention	yim0@cdc.gov
Oksana	Lukjancenko	National Food Institute, Technical Univ. of Denmark	oklu@food.dtu.dk
Phelix	Majiwa	Agricultural Research Council	MajiwaP@arc.agric.za
Moabi Rachel	Maluleke	Agricultural Research Council	MalulekeR@arc.agric.za
Donna	Manogue	Albuquerque Police Department	dmanogue@cabq.gov
Cheriece	Margiotta	Los Alamos National Laboratory	CMargiotta@lanl.gov
Joel	Martin	DOE Joint Genome Institute	j_martin@lbl.gov
Rich	Masino	MRIGlobal	rmasino@mriglobal.org
Arne	Materna	Qiagen	Arne.Materna@qiagen.com
Marta	Matvienko	Consulting	matvienko@gmail.com
Marta	Matvienko	Qiagen	mmatvienko@clcbio.com
Peter	Maughan	Brigham Young University	Jeff_Maughan@byu.edu
Greg	Mayer	Texas Tech University	greg.mayer@ttu.edu
Carl	Mayers	DSTL	cnmayers@dstl.gov.uk
Kirsten	McCabe	Los Alamos National Laboratory	kjmccab@lanl.gov
Mitch	McGrath	USDA ARS	mitchmcg@msu.edu
Kim	McMurry	Los Alamos National Laboratory	kmcmurry@lanl.gov
Kelly	Meiklejohn	ORISE/FBI Laboratory	kelly.meiklejohn@ic.fbi.gov
Kevin	Meldrum	Illumina, Inc	kmeldrum@illumina.com
Michael	Mendrysa	Government Scientific Source	mmendrysa@govsci.com
Amanda	Mercer	Los Alamos National Laboratory	a.n.mercer@me.com
Ginger	Metcalf	Baylor College of Medicine	metcalf@bcm.edu
Thomas	Meyer	Defense Forensic Science Center	thomas.a.meyer37.civ@mail.mil
Tim	Minogue	USAMRIID	timothy.d.minogue.civ@mail.mil
Samuel	Minot	One Codex	sam@onecodex.com
Yimam Getaneh	Misganie	Ethiopian Public Health Institute	yimamgetaneh@gmail.com
Karen	Moll	National Center for Genome Resources	karen.moll@ncgr.org
Scott	Monsma	Lucigen Corp	smonsma@lucigen.com
Lilliana	Moreno	Federal Bureau of Investigation	lilliana.moreno@ic.fbi.gov
Shatavia	Morrison	Centers of Disease Control and Prevention	xxh5@cdc.gov
Joann	Mudge	National Center for Genome Resources	jm@ncgr.org
Teri	Mueller	Roche	teri.mueller@roche.com
Brianna	Mulligan	University of New Mexico	k11bm01@unm.edu
Brendan	Mumey	Montana State University	brendan.mumey@montana.edu
Daniela	Munafo	New England Biolabs, Inc	munafo@neb.com
William	Murphy	Texas A&M University	wmurphy@cvm.tamu.edu
Tim	Muruvanda	US Food and Drug Administration CFSAN	tim.muruvanda@fda.hhs.gov
Donna	Muzny	Baylor College of Medicine	donnam@bcm.edu
Madhugiri	Nageswara-Rao	New Mexico State University	mnrbhav@yahoo.com
Judy	Ney	Kapa Biosystems	judy.ney@kapabiosystems.com
Sock Hoon	Ng	DSO National Laboratories	ng_sock_hoon@dso.org.sg
Minh	Nguyen	National Institute of Justice	minh.nguyen@usdoj.gov
Ainsley	Nicholson	Centers for Disease Control and Prevention	agn0@cdc.gov
Andy	Nkili	Centre International de Recherches Médicales de Franceville (CIRMF)	andynkili@gmail.com
Laure	Nucci	Qiagen	Nonna.Druker@qiagen.com
Kristen	O'Connor	Defense Threat Reduction Agency	kristen.l.oconnor5.civ@mail.mil
John	Oliver	Nabsys	oliver@nabsys.com
Christian	Olsen	Biomatters	christian@biomatters.com
Andrea	Ottesen	US Food and Drug Administration CFSAN	andrea.ottesen@gmail.com
Zbyszek	Otwinowski	UT Southwestern	zbyszek@work.swmed.edu
Oliver	Oviedo	Los Alamos National Laboratory	oviedo@lanl.gov
David	Owuor	Centers for Disease Control and Prevention	cowuor@kemricdc.org
Clint	Paden	Centers for Disease Control and Prevention	fep2@cdc.gov
suman	pakala	University of Georgia	spakala@uga.edu
Andy	Pang	BioNano Genomics Inc.	Apang@bionanogenomics.com
Kyle	Parker	MRIGlobal	kparker@mriglobal.com
Vaishnavi	Pattabiraman	Centers for Disease Control and Prevention	vxe9@cdc.gov
Justin	Payne	US Food and Drug Administration CFSAN	justin.payne@fda.hhs.gov

First Name	Last Name	Company	Email
Andreas Sand	Pederson	Qiagen	Nonna.Druker@qiagen.com
Benjamin	Pelle	Los Alamos National Laboratory	bpelle@lanl.gov
Joseph	Petrosino	Baylor College of Medicine	joseph.petrosino@bcm.edu
Pavel	Pevzner	University of California, San Diego	ppevzner@ucsd.edu
Caleb	Phillips	Texas Tech University	caleb.phillips@ttu.edu
Robert	Player	Johns Hopkins Applied Physics Lab	robert.player@jhuapl.edu
Sandra	Porter	Digital World Biology	digitalbio@me.com
Nicholas	Putnam	Dovetail Genomics, LLC	nik@dovetail-genomics.com
Padmini	Ramachandran	US Food and Drug Administration	padmini.ramachandran@fda.hhs.gov
Thiruvarangan	Ramaraj	National Center for Genome Resources	tr@ncgr.org
Chris	Rampey	IDT	crampey@idtdna.com
Brian	Raphael	Centers for Disease Control and Prevention	BRaphael@cdc.gov
Debjit	Ray	Sandia National Labs	debray@sandia.gov
Linda	Ray	Illumina, Inc	lray@illumina.com
Eric	Rees	RTL Genomics	eric.rees@researchandtesting.com
Brandon	Rice	Dovetail Genomics, LLC	brandon@dovetail-genomics.com
Todd	Richmond	Roche NimbleGen	todd.richmond@roche.com
James	Robertson	FBI Laboratory / FBI Academy	james.m.robertson@ic.fbi.gov
Rosanna	Robertson	US Dept of Homeland Security	rosanna.robertson@associates.hq.dhs.gov
Allison	Rodriguez	US Food and Drug Administration	allison.rodriguez@fda.hhs.gov
Chandler	Roe	TGen North	croe@tgen.org
George	Rosenberg	University of New Mexico	ghrose@unm.edu
Lori	Rowe	Centers for Disease Control and Prevention	ioy0@cdc.gov
Denise	Ruffner	IBM	ruffner@us.ibm.com
Joseph	Russell	MRIGlobal	jrussell@mriglobal.org
Ashley	Sabol	Centers for Disease Control and Prevention	asabol@cdc.gov
Joseph	Salvatore	Qiagen	joe.salvatore@qiagen.com
Scott	Sammons	Centers for Disease Control and Prevention	zno6@cdc.gov
Nurlan	Sandybayev	Research Institute for Biological Safety Problems	nurlan.s@mail.ru
Rashesh	Sanghvi	Baylor College of Medicine	rashesh.sanghvi@bcm.edu
Monica	Santovenia	Centers for Disease Control and Prevention	MSantovenia@cdc.gov
Stephanie	Sarnese	Tetracore, Inc.	ssarnese@tetracore.com
Wendy	Schackwitz	DOE Joint Genome Institute	wsschackwitz@lbl.gov
Melissa	Scheible	North Carolina State University	mkscheib@ncsu.edu
Ira	Schildkraut	New England Biolabs	schildkraut@neb.com
Kelly	Schilling	National Center for Genome Resources	kschilling@ncgr.org
Matthew	Scholz	Vanderbilt University	matthew.b.scholz@vanderbilt.edu
Curt	Schuerman	Defense Forensic Science Center	curt.a.schuerman.civ@mail.mil
Edward	Schulak	ES Equities, LLC	edward@schulak.com
James	Schupp	TGen North	jschupp@tgen.com
Erin	Scully	USDA-ARS	Erin.Scully@ars.usda.gov
Johnny	Sena	National Center for Genome Resources	jsena@ncgr.org
Pavel	Senin	Los Alamos National Laboratory	psenin@lanl.gov
Brian	Sereni	GENEWIZ Inc.	brian.sereni@genewiz.com
Migun	Shakya	Dartmouth College	migun.shakya@dartmouth.edu
Helen	Shearman	Biomatters	helen@biomatters.com
Niranjan	Shekar	Spiral Genetics	niranjan@spiralgenetics.com
Sarah	Sheldon	Centers for Disease Control and Prevention	hso5@cdc.gov
Brian	Shirey	Centers for Disease Control and Prevention	TShirey@cdc.gov
Michael	Shoemaker	USA/SLAA	mvshoemaker@verizon.net
Heike	Sichtig	US Food and Drug Administration	Heike.Sichtig@fda.hhs.gov
Keti	Sidamonidze	NCDC	keti_sida@yahoo.com
Steve	Siembieda	Advanced Analytical	ssiembieda@aati-us.com
Sheina	Sim	University of Hawaii, Manoa	ssim8@hawaii.edu
Gary L	Simpson	Placitas Consulting Collaborative	garyl.simpson@comcast.net
Furman	Sizemore	Albuquerque Police Department	fsizemore@cabq.gov
Tom	Slezak	Lawrence Livermore National Laboratory	slezak@llnl.gov
Jason	Smith	Illumina, Inc	jsmith6@gmail.com
Todd	Smith	Digital World Biology	digitaltodd@me.com
Dr. Annette	Sobel	Texas Tech University	bigbitbucket@mac.com
Shanmuga	Sozhamannan	Defense Biological Product Assurance Office	shanmuga.sozhamannan.ctr@mail.mil
Rachel	Spurbeck	Battelle Memorial Institute	spurbeck@battelle.org

First Name	Last Name	Company	Email
Ganesh	Srinivasamoorthy	Centers for Disease Control and Prevention	sganesh02@yahoo.com
Shawn	Starkenburg	Los Alamos National Laboratory	shawns@lanl.gov
Fiona	Stewart	New England Biolabs	stewart@neb.com
Jonathan	Stites	Dovetail Genomics	jon@dovetail-genomics.com
Charles	Strittmatter	US Food and Drug Administration	cstritt@strittech.onmicrosoft.com
Vitaliy	Strochkov	Research Institute for Biological Safety Problems	vstrochkov@gmail.com
Michelle	Su Yen Wong	DSO National Laboratories	wsuyen@dso.org.sg
Shawn	Sullivan	Phase Genomics, Inc.	shawn@phasegenomics.com
Anitha	Sundararajan	National Center for Genome Resources	asundara@ncgr.org
Austin	Swafford	Labcyte	swafford.ade@gmail.com
Austin	Swafford	Labcyte	aswafford@labcyte.com
Laura	Sycuro	Fred Hutch Cancer Research Center	lsycuro@fredhutch.org
Kevin	Tang	Centers for Disease Control and Prevention	ktang@cdc.gov
Yingying	Tang	New York City Office of Chief Medical Examiner	ytang@ocme.nyc.gov
Tea	Tevdoradze	NCDC	t.tevdoradze@ncdc.ge
Masoud	Toloue	Bioo Scientific	mtoloue@biooscientific.com
Ryan	Toma	Los Alamos National Laboratory	ryan.toma@mail.com
Brad	Townsley	Amaryllis Nucleics, Inc.	bttownsley@gmail.com
Jason	Travis	TGen North	jtravis@tgen.org
Angie	Trujillo	Centers for Disease Control and Prevention	ATrujillo@cdc.gov
Brittany	Twibell	Los Alamos National Laboratory	btwibell@lanl.gov
Joshua	Udall	BYU	jaudall@gmail.com
George	Unc	Los Alamos National Laboratory	georgeunc@lanl.gov
Sahra	Uygun	Michigan State University	uygunsah@msu.edu
Willy	Valdivia	Orion Integrated Biosciences	Willy.valdivia@orionbio.com
Michael	Vandewege	Mississippi State University	mike.vandewege@gmail.com
Michelle	Vierra	Dovetail Genomics	michelle@dovetail-genomics.com
Eric	Vincent	Promega Corporation	eric.vincent@promega.com
Nicholas	Vlachos	ORISE/FBI Laboratory	ntvlachos@gmail.com
Logan	Voegtly	NMRC	loganvoegtly@me.com
Momchilo	Vuyisich	Los Alamos National Lab	vuyisich@lanl.gov
Darlene	Wagner	Centers for Disease Control and Prevention	ydn3@cdc.gov
Edward	Wakeland	UT Southwestern Medical Center	edward.wakeland@utsouthwestern.edu
Bruce	Walker	Applied Invention	bw@ai.co
Kimberly	Walker	Baylor College of Medicine	kimberly.walker@bcm.edu
Ron	Walters	Pacific Northwest National Laboratory	grwalters@frontier.com
Charles	Wang	US Food and Drug Administration CFSAN	Charles.wang@fda.hhs.gov
Haibin	Wang	Centers for Disease Control and Prevention	wanghaibin_2002@yahoo.com
Judson	Ward	Driscoll's	jud.ward@gmail.com
Jennifer	Watson	Albuquerque Police Department	jwatson@cabq.gov
Simon	Weller	DSTL	cnmayers@dstl.gov.uk
Doug	Wieczorek	Promega Corporation	Doug.Wieczorek@promega.com
Jeremy	Wilkinson	RTL Genomics	jeremy.wilkinson@researchandtesting.com
Alanna	Williams	Albuquerque Police Department	alwilliams@cabq.gov
Diana	Williams	Defense Forensic Science Center	diana.w.williams8.civ@mail.mil
A. Jo	Williams-Newkirk	IHRC	igy7@cdc.gov
Patti	Wills	Los Alamos National Laboratory	wills@lanl.gov
Andrew	Wiser	Phase Genomics, Inc.	andrew@phasegenomics.com
Kim	Woller	10X Genomics	kim@10xgenomics.com
stephen	wyatt	stephen m wyatt	stephenmwyatt@gmail.com
Zhenjiang	Xu	University of California, San Diego	zhx054@ucsd.edu
Elaine	Yeh	US Food and Drug Administration	elaine.yeh@fda.hhs.gov
Kenneth	Yeh	MRIGlobal	kyeh@mriglobal.org
Guohua (Karen)	Yin	USA	guohuayin1997@gmail.com
Tong	Yin	Thomson Reuters	tongyin@gmail.com
Diane	Yost	Los Alamos National Laboratory	dgyost@lanl.gov
Sarah	Young	Broad Institute of MIT and Harvard	jwineski@broadinstitute.org
Daojun	Yuan	Brigham Yong Unviversity	robertorun@gmail.com
Xiang	Zhou	BioNano Genomics Inc.	xiz407@gmail.com
Sulushash	Zhumabayeva	CH2M	Sulushash.Zhumabayeva@ch2m.com

Map of Santa Fe, NM



Reliable solutions for focused NGS

From predesigned and custom panels based on xGen® Lockdown Probes to adapter blockers, all of your target enrichment solutions are built upon individually synthesized and quality controlled oligos, ensuring you **consistently achieve greater sensitivity, higher throughput, and better uniformity** than delivered by array-synthesized probes.

"xGen Lockdown Probes are high quality reagents that work robustly and also provide the flexibility and scalability we need in our experiments."

Dr Jim Hughes
University of Oxford, UK

See for **yourself** at
www.idtdna.com/exome

IDT®
INTEGRATED DNA TECHNOLOGIES

custom oligos • qPCR • next generation sequencing • RNAi • genes & gene fragments • CRISPR genome editing

SFAF 2016 Sponsors

SFAF 2016 Sponsors

SFAF 2016 Sponsors