NIST-FDA Workshop: Standards for Pathogen Detection via Next-Generation Sequencing

October 27-28, 2015
NIST Gaithersburg, MD

Report

Organizers: Scott Jackson[1], Heike Sichtig[2], Brittany Goldberg[2], Chelsie Geyer[2], Jason Kralj[1]

[1] National Institute of Standards and Technology, Gaithersburg, MD USA
[2] Food and Drug Administration, Silver Spring, MD USA

Purpose of the Workshop

The purpose of this NIST-FDA workshop was to seek input on defining reference materials, reference data and reference methods for assessing analytical sensitivity, specificity, and relative performance of NGS-based pathogen detection devices/assays. These reference materials will be critical in addressing the challenges associated with mixed pathogen detection in complex samples (i.e. clinical, environmental) using shotgun metagenomic sequencing and targeted resequencing (culture-independent diagnostics) approaches.

This workshop targeted the primary users/adopters of these standards; including clinicians, the biosecurity community, public health, industry, academic and government laboratories. We invited subject matter experts from these areas to provide their perspectives on the needs for standards in this field. Also included were break-out sessions that allowed attendees to discuss, and reach consensus on, the specific characteristics of the proposed standards.

Agenda

Tuesday, October 27, 2015 (Red Auditorium)

8:30am	Arrive Main Gate
9:00am — 9:25am	Opening Remarks – Scott Jackson, Laurie Locascio, Heike Sichtig
9:25am — 10:00am	Charles Chiu – University of California San Francisco School of Medicine
10:00am — 10:35am	Joseph Campos - Children's National Medical Center
10:35am — 10:50am	*Coffee Break*
10:50am — 11:25pm	John Besser - CDC
11:25pm — 12:00pm	Heike Sichtig – FDA
12:00pm— 1:00pm	*Lunch NIST Cafeteria*
1:00pm — 1:30pm	Poster Session
1:30pm — 2:00pm	Timothy Minogue – USAMRIID
2:00pm — 2:30pm	Tom Slezak – Lawrence Livermore National Lab
2:30pm — 3:00pm	Patrick Chain – Los Alamos National Lab
3:00pm — 3:15pm	*Coffee Break*
3:15pm — 3:45 pm	William Klimke – NIH-NCBI
4:00pm — 4:30pm	Justin Zook – NIST
4:30pm — 5:00pm	Wrap-Up and Logistics for Next Day

Wednesday, October 28, 2015 (Lecture Rooms A and B/Portrait Room)

9:00am — 9:15am	Opening Remarks
9:15am — 9:45am	Scott Jackson and Jason Kralj - Prototype NIST Mixed Pathogen RM
9:45am — 10:30am	Break-out Session 1
10:45am — 11:30am	Break-Out Session 2
11:30am — 1:00pm	*Lunch – NIST Cafeteria*
1:00pm — 2:00pm	Poster Session
2:00pm — 3:00pm	Report from Breakouts
3:00pm — 3:30pm	Closing Remarks

Introduction and Overview

A pathogen is broadly defined as a microorganism that is able to cause disease in a multicellular eukaryotic host (e.g. plants and animals). Pathogens come in the form of viruses, bacteria, fungi, protozoa, parasites and prions (http://www.bseinfo.org/prions.aspx). Due to extensive genetic diversity that exists within most classes of pathogens, it is impossible to precisely define the number of different pathogenic species that have been observed to cause disease to date.

According to the CDC, 23.6 million Americans visited a clinic as a result of an infectious disease in 2010 (CDC, 2010). In 2014, infectious diseases were associated with an economic burden of over $120 billion in the U.S. alone (Research America, 2015). Collectively, *in vitro* diagnostic tests for infectious diseases account for over one billion dollars in health care costs annually.

Our need to rapidly and reliably detect and identify disease-causing microbial organisms (pathogens) dates back to the origins of germ theory in the mid-19th century. In a paper read to the French Academy of Science in April 1878, Louis Pasteur stated "*To demonstrate experimentally that a microscopic organism actually is the cause of a disease and the agent of contagion, I know no other way, in the present state of Science, than to subject the microbe to the method of cultivation out of the body.*" Pasteur's comments still apply today in our era of modern medicine where pathogens are routinely isolated and characterized in the clinical setting. A large number of laboratory (*in vitro*) tools have been developed over the past decades for the purpose of pathogen identification in order to better inform healthcare providers with regards to the diagnosis, treatment and monitoring of infectious diseases. These diagnostics play critical roles in promoting public health, by guiding appropriate therapy to minimize antimicrobial resistance, identifying disease outbreaks, addressing potential biothreats and insuring the safety of our food and water supply.

Traditional Approaches for Pathogen Detection:

Traditionally, *in vitro* diagnostics of infectious diseases have been performed using culture-based techniques; sometimes in combination with antigen detection and/or serology-based tests. Today in the genomics era, DNA and RNA-based molecular assays have become more routine and are gaining a larger share of the *in vitro* diagnostic market space. A relatively simple DNA-based *in vitro* assay for diagnosing an infectious disease might take the form of a quantitative or qualitative PCR assay where the target for detection is a pathogen-specific gene or an antimicrobial resistance marker. While powerful in many regards, these targeted approaches are biased by the clinical experience of the ordering clinician, who must suspect the pathogen and then order the appropriate targeted diagnostic test. More recently, new diagnostics employing a multiplexed-by-design approach permit panels of pathogens to be assayed from the same sample, thus enabling the rapid diagnosis of multiple pathogens. Ultimately, all of the existing diagnostic resources require *a priori* knowledge of the potential targets, and commercial diagnostic testing has yet to be developed for emerging or rare pathogens.

Metagenomics Sequencing for Pathogen Detection:

Over the past decade, there has been a substantial rise in DNA sequencing throughput. At the time of this writing, state-of-the-art DNA sequencing instruments (so called Next-Generation Sequencing (NGS) or High-Throughput Sequencing (HTS) instruments) are capable of producing gigabases of sequence data in under 48 hours at a cost under $0.10 per megabase (https://www.genome.gov/27541954/dna-sequencing-costs/). This enormous throughput and low cost allows for modest-size laboratories to utilize from NGS technology and is responsible for the current level of research in the field of genomics. While whole genome sequencing of individual microorganisms/isolates was once the primary practice for microbial NGS technologies, the enormous increase in throughput has led to the adoption of more routine metagenomic sequencing approaches in which highly complex communities of microorganisms are sequenced in parallel.

Metagenomic sequence data obtained from complex samples (i.e. mixtures of microbes, usually with varying levels of constituents, and sometimes with additional DNA from a host organism) could provide a qualitative and quantitative understanding of the individual components of the original complex sample. Genus, species, and even strain-level taxonomic assignments of microorganisms, as well as their relative abundance, could potentially be readily obtained from complex samples by using metagenomic sequencing approaches. This ability to rapidly characterize and identify the entire (microbial) content of a complex sample provides a unique and novel strategy for pathogen detection and identification. Compared to the targeted approaches described above (e.g. PCR), metagenomic approaches are less biased in that they require no *a priori* knowledge of the sample contents. An added benefit is the ability to identify infections with a multitude of possible causative pathogens, such as pneumonia, or polymicrobial infections such as chronic wound infections. Finally, metagenomic sequence data could allow for the detection and identification of antibiotic resistance genes and virulence factors in complex samples; information that can be used to guide treatment options and improve antibiotic stewardship.

In the coming years, DNA sequencing technologies will continue to develop and improve. Advancements in NGS technologies will potentially offer higher sample throughput, greater speed and reduced running costs. As NGS technology matures, sample preparation, analysis and interpretation may evolve to allow for the streamlined diagnosis of a wide variety of pathogens, and provide health-care providers and patients with clinically-actionable results that could enhance patient care and overall public health.

Bioinformatics, Databases, and Data Interpretation:

Computational basics for NGS technology-based pathogen detection are i) robust bioinformatic tools and ii) comprehensive reference (pathogen) genome databases. Metagenomic data generated from a single complex clinical sample (e.g. stool) may consist of millions of individual sequence reads; often ranging from 50 to several thousands of bases in length. Depending on the sample type, often only a small fraction of these reads are derived from the target pathogens. Nevertheless, every read must be searched against a reference database of pathogen genomic sequences in order to identify sequence matches. As such, accurate pathogen detection requires a comprehensive and reliable database of microbial genomes, or at least pathogen-specific genetic signatures. To date, several microbial genome databases have been developed specifically for this purpose. For example, FDA-ARGOS (**FDA** d**A**tabase of **R**egulatory **G**rade micr**O**bial **S**equences) houses regulatory-grade microbial reference genomic sequences; JGI (Joint Genome Institute) Genome Portal includes access to various JGI genomic databases and NGS data analysis tools (http://jgi.doe.gov/data-and-tools/genome-portal/); HIV, HCV, HFV/Ebola databases from LANL (Lawrence Livermore National Laboratory) (http://www.hiv.lanl.gov/content/index) provides genomic sequences of many viruses; and Saccharomyces Genome Database provides sequencing information on the budding yeast *Saccharomyces cerevisiae* and NGS analysis tools (http://www.yeastgenome.org/).

Integral to this process is the availability of robust and efficient bioinformatic tools/algorithms that can analyze large amounts of sequencing data. Standard DNA sequence alignment tools like BLAST are powerful, but require massive computational resources in order to efficiently search through the large amounts of metagenomic sequence data generated from complex samples. As a result, a number of different bioinformatic tools/algorithms have been developed that allow more timely and effective search capabilities. To date, over 60 different custom bioinformatic pipelines have been developed that provide efficient and rapid sequence alignment capabilities. Some commonly-used tools include Kraken (Wood & Salzberg 2014), LMAT (Allen 2013), EDGE (Chain 2016), MetaPhlAn (Segata 2012), and Pathoscope (Hong 2014) among many others.

Regulatory Oversight for NGS-Based Mixed Pathogen Detection Devices

Currently, the Division of Microbiology Devices within the Center for Devices and Radiological Health (CDRH) at the US FDA is trying to develop an adaptive regulatory approach for the evaluation of NGS-based microbial diagnostic assays. Assays developed with claims of testing mixed microbial samples present challenges as there is a lack of well-established validated physical and nucleic acid reference materials and genomic sequences. Clinical and scientific community leaders have emphasized the need for a mixed microbial reference material and protocols for developing and validating a microbial NGS-based *in vitro* diagnostic assay.

Initially, the stakeholders at this workshop proposed the development of standards for a mixed microbial genomic DNA reference material for the purpose of evaluating analytical performance metrics of microbial NGS-based diagnostic tests. This will not include testing of pre-analytical steps. Mixed sample reference material represents a significant advancement for the diagnostics community through enhancing the ability of timely proficiency testing of NGS-based diagnostics. Development and validation of genomic DNA reference materials can be timely and cost prohibitive for some laboratories. The stakeholders also proposed that this proficiency material can serve as a quality control option for an NGS-based diagnostic assay and it ensures that testing in the laboratory is being conducted in a reproducible and reliable manner. They also suggested that the use of an agnostic (metagenomics) detection system, coupled with a high quality reference database and a mixed sample microbial reference material, could help device developers, as well as advance epidemiological surveillance and assist in outbreak scenarios.

The multitude of potential uses of NGS for biothreat, surveillance, and everyday clinical purposes, as well as the existence of all required wet lab instrumentation and chemistries, underline the need for timely and accurate proficiency testing material. Mixed sample DNA reference materials have the potential to assist in the development, validation, and regulatory approval/clearance of NGS-based *in vitro* diagnostic devices.

Summary of Talks

On the first day of the workshop, we invited subject-matter experts to present their work and asked each speaker to follow-up with a summary (abstract) of their presentations. We've also made the presentation slides available (via hyperlinks below) with permission from the speakers.

Charles Chiu, M.D./Ph.D. – University of California, San Francisco School of Medicine

Associate Professor, Laboratory Medicine and Medicine / Infectious Diseases
Director, UCSF-Abbott Viral Diagnostics and Discovery Center
Associate Director, UCSF Clinical Microbiology Laboratory

Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis: the Critical Need for Reference Standards

Metagenomic next-generation sequencing (mNGS) is an emerging genomics tool to comprehensively detect all pathogens -- viruses, bacteria, fungi, and parasites, in a single assay. At UCSF, we have proven the utility of this approach in multiple clinical scenarios, including its use to save the life of a 14 year-old boy with a critical case of undiagnosed neuroleptospirosis (Wilson, et al., 2015, *New England Journal of Medicine*). This technology has been enabled by the vast increases in sequencing capacity over the past several years and the availability of rapid computational pipelines such as SURPI ("Sequence-based Ultra-Rapid Pathogen Identification") for analysis of NGS data. Critical to the utility of mNGS approach for infectious disease diagnosis will be clinical validation of the test in a CLIA-certified laboratory and eventual FDA regulatory approval. Key challenges that will thus need to be addressed include (1) generation of accurate reference materials and controls, (2) customization of research-based computational pipelines for clinical use, (3) development of user-friendly graphical interfaces for data analysis (4) data storage and HIPAA compliance issues, (5) generation of interpretative clinical reports, and (6) integration of sequence data into the patient electronic medical record (EMR). Of particular importance will be the critical need for established "wet" laboratory standards (samples and/or DNA) as well as comprehensive and accurate microbial reference databases. As a "proof-of-principle" demonstration project, we will launch in March of 2016 a new multi-center, prospective study of mNGS for precision diagnosis of acute infectious diseases across 3 medical centers: UCSF, UCLA, and UC Davis. We will initially focus on analysis of cerebrospinal fluid for diagnosis of meningitis and/or encephalitis using a CLIA-certified mNGS assay, with the results used clinically to guide patient care and management. A dedicated precision medicine consult team will evaluate the impact of mNGS on health care costs, diagnostic yield, and patient outcomes.
https://www.slideshare.net/secret/m1wirQf09qLoao

Joseph Campos, Ph.D., D(ABMM), F(AAM) – Children's National Medical Center

Interim Chief, Division of Laboratory Medicine
Director, Microbiology Laboratory, Molecular Diagnostics Laboratory, and Laboratory Informatics

Is NGS Ready for Use as a Clinical Diagnostics Tool?

https://www.slideshare.net/secret/ntGakslghzuPmA

John Besser, Ph.D. – Centers for Disease Control and Prevention

Deputy Chief, Enteric Diseases Laboratory Branch,
Division of Foodborne, Waterborne, and Environmental Diseases,
National Center for Emerging and Zoonotic Infectious Diseases (NCEZID)

Culture-Independent Public Health Methods

Background: The next generation sequencing (NGS) and bioinformatics revolutions are driving evolution of foodborne disease surveillance. The usefulness of prospective, real-time whole genome sequencing (WGS) for Listeria monocytogenes outbreak detection and investigation has been demonstrated by a joint CDC, FDA, USDA, and NCBI pilot surveillance project. Significantly more clusters have been detected, outbreaks identified, and sources identified than in the pre-WGS era, and the median size of outbreaks has shrunk as outbreaks are detected when they are small. Technological advancements are also occurring in the field of clinical

diagnostic microbiology that will likely impact surveillance efforts in a variety of ways. A new generation of nucleic acid-based culture-independent diagnostic test (CIDT) panel tests makes it possible to simultaneously detect multiple bacterial, viral, and parasitic diarrheal disease agents in just a few hours. These tests are conducted directly on patient specimens, and yield results in a timeframe which will likely improve patient management. Unfortunately, these tests do not result in the production of bacterial or viral isolates, which are the raw material of WGS-based surveillance. CDC is pursuing multiple strategies to maintain laboratory-based surveillance in the CIDT era. One long-term strategy is to develop assays for high-level sequence-based characterization of microbes that are themselves culture-independent, much like CIDTs used for patient diagnosis. Scientific Approaches: Three broad development approaches are being explored by CDC. An amplicon sequencing approach is being explored that can potentially be implemented with existing WGS technology. Bacterial genomes are being mined for various targets that can be used for strain typing and determination of virulence and antimicrobial resistance potential. A variety of approaches are being taken to address phasing, which is the alignment of alleles belonging to the target organism. Shotgun metagenomics, which has been successfully applied for diagnosis of serious invasive infections, is not yet practical for routine surveillance using current technology. Several approaches are being taken address the major hurdles to practical metagenomics, which include signal-to-noise limitations and phasing issues. Finally, a single-cell sorting and sequencing technology is being explored through an industry collaboration. Evaluation of these approaches, and other research activities in the rapidly expanding world of metagenomics, requires standardized substrates for understanding assay dynamics and for inter-laboratory comparisons.
https://www.slideshare.net/secret/18rFZSjrSpDvse

Heike Sichtig, Ph.D. – US Food and Drug Administration

Regulatory Scientist and Principal Investigator of FDA-ARGOS

Regulatory Perspective on Infectious Disease NGS Dx Devices
Focus: Mixed Sample Reference Materials

Lack of clinical diagnostics for Infectious Diseases can have devastating consequences for public and global health as highlighted by the recent Ebola outbreak in West Africa. Supporting infrastructure needs to be developed to equip public and global health care providers with tools necessary to combat rising threats such as emerging pathogens and antimicrobial resistance. Microbial diagnostic devices based on Next Generation Sequencing (NGS) technologies are on the cusp of making it into clinical laboratories and hospitals.

The presentation outlined studies to evaluate the use of NGS-based devices as an aid in Infectious Disease diagnostics, and to gain a better understanding of potential NGS clinical implementation strategies. Focus was on the possible approaches to validation studies and data for the evaluation of infectious disease NGS-based diagnostics for potential regulatory clearance/approval, and the use of sequence outputs from infectious disease NGS-based devices to evaluate performance. Efforts towards generating an initial set of high quality, regulatory-grade microbial genomic reference sequences through the FDA-ARGOS project in collaboration with NCBI were also discussed. Our vision is a high quality ‘regulatory-grade’ microbial reference database that contains qualified sequence data for use by developers and clinical end users. The information contained in the presentation concerning possible approaches for validation were suggested approaches open for feedback.
https://www.slideshare.net/secret/A7cmDZnCvIcKhA

Timothy Minogue, Ph.D. – United States Army Medical Research Infectious Diseases

Chief, Molecular Diagnostics Department
Diagnostic Systems Division

Diagnostic Development for Biothreats: Use Cases for Reference Materials Critical for appropriate responses in any biothreat/public health challenge is surety in diagnostic accuracy, flexibility in response and alacrity in identification of etiologic agents. For the DoD, all medical decisions also require regulatory compliant use of diagnostic applications for adjudication of treatment. With this requirement comes absolute necessity of having curated well defined standards that impact diagnostic development, platform evaluation and proficiency testing. Current efforts within our lab focus on developing new molecular diagnostic applications while providing a clear path towards regulatory compliance. Through execution of this pipeline, we show how curated standards augment and define these efforts in diagnostic development.

Tom Slezak – Lawrence Livermore National Laboratory

Associate Program Leader

Beyond Human: The need for Environmental Standard Reference Material(s)
The need for standard reference materials for the detection of pathogenic organisms in human clinical samples is well-established as a necessary step on the path towards eventual validation of NextGen Sequencing (NGS) as a clinical diagnostic. Somewhat less obvious are the needs for other standard reference materials where the background matrix is not simply human DNA. Gut microbiome, aerosols, water, food items, and other environmental samples (e.g., soils, indoor and outdoor swipes/swabs, etc.) all have roles in aspects of human health, food safety, biodefense, law enforcement and other fields where accurate comparisons of different NGS technologies are needed using samples that are relevant to the different mission needs. These all involve complex backgrounds which run into an immediate problem: what is the gold standard technology by which ground truth of the realistic complex backgrounds can be determined? We note that there is no perfect answer, but suggest that there is a feasible approach if one is willing to consider an evolving ground truth. Ultra-deep NGS (e.g., several billion reads of a HiSeq-class platform) of each batch of an environmental standard reference material can be generated and analyzed, with re-analysis possible at any time to take advantage of new genomic sequence knowledge. This approach would permit fair comparisons of NGS systems provided that all used the same batch of standard reference material. It is hoped that the Federal sponsors involved with NGS will find a way to jointly fund a solution to fill this important resource gap.
https://www.slideshare.net/secret/M5yOy7npA6hAIx

Patrick Chain, Ph.D. – Los Alamos National Laboratory

Genome Sciences, Bioscience Division

Pathogen detection via sequencing: Hurdles to realizing the promise of NGS
While most workflows tackling Next Generation Sequencing (NGS)-based pathogen detection appear straightforward, there are a number of issues that need to be addressed such that it can be routinely used in any (clinical) setting. These issues range from the differences between

sample types and confounding matrices recalcitrant to nucleic acid extraction, to the amount of host or commensal nucleic acids that may introduce noise into the sample, to the anticipated pathogen load within those specific samples. These front-end concerns will dictate the amount, and therefore cost of sequencing required to recover pathogen sequences capable of being detected. The downstream detection involves using a search strategy/algorithm to query the input sequences against a database. In this presentation, I will discuss the challenges facing bioinformatic pathogen detection, the algorithmic issues when trying to identify known as well as evolved or newly emerging pathogens rapidly, when compared with a database that is biased, incomplete and rife with errors (contamination, sequence errors, taxonomy inconsistencies, etc.). Other issues such as what can be interpreted as a true hit, how many hits are sufficient to make a call, how to eliminate the possibility of horizontal gene transfer to negatively affect a call, and how to separate relative abundance from probability of presence, are all topics of interest to the field and remain open questions.

William Klimke, Ph.D. – NCBI/NLM/NIH/DHHS

Senior Scientist

Moving Towards Standards for Next Generation Sequencing of Bacterial Pathogens

Public health is increasingly using next generation sequencing in bacterial pathogen analysis. The need for standards and quality metrics in order to trust the analysis results is of paramount importance. The National Center for Biotechnology Information is working towards improving the quality of bacterial genomes that are submitted to the public archives through a number of initiatives: 1) NCBI taxonomists have recently held a workshop aimed at improving the taxonomic assignment of genomes in GenBank, 2) the Interagency Collaboration on Genomics and Food Safety (Gen-FS) was formed between USDA-FSIS, FDA, CDC, and NCBI to address a number of issues including the formation of a working group for standards and quality metrics, 3) examination of discrepancies in SNPs when comparing closely related bacterial genomes. The taxonomy workshop led to the adoption of the use of Average Nucleotide Identity (ANI) along with the type strain sequences and genomes in GenBank to correct taxonomic misidentifications. NCBI has proposed a structured comment block to be put on GenBank records when misidentifications are corrected that would record the original scientific name that was submitted, the correction made by the NCBI ANI process, and the evidence for the change. The GenFS standards working group has assembled three benchmark datasets consisting of genome sequences as well as metadata from known outbreaks in order to test different SNP calling pipelines. Comparison of the SNP calls and phylogenetic trees from pipelines developed at the different federal agencies has uncovered differences that resulted in improvements in all pipelines. NCBI has been exhaustively examining the results of our Pathogen Detection pipeline with respect to SNP calling for closely related genomes sequences submitted from the collaborators in GenFS to the short read archive (SRA) as well as external submitters depositing pre-assembled genomes into GenBank in order to assist outbreak and trace back investigation for the food safety agencies. A number of subtle errors in the assembled genome consensus sequences have been uncovered that were easily filtered from downstream analysis when the short read data was deposited along with the genome. Therefore, one immediate and important recommendation is to always require the raw sequencing data deposited in SRA in order to independently measure the assembly quality. All of these efforts will lead to improved bacterial genome quality in the public archives.

https://www.slideshare.net/secret/i5AykfSX2JlQks

Justin Zook, Ph.D. – National Institute of Standards and Technology

Biomedical Engineer

Genome in a Bottle: You've sequenced. How well did you do?

Clinical laboratories, research laboratories and technology developers all need DNA samples with reliably known genotypes in order to help validate and improve their methods. The National Institute of Standards and Technology formed the Genome in a Bottle Consortium (genomeinabottle.org), which has been developing Reference Materials with high-accuracy whole genome sequences to support these efforts. Our pilot reference material is based on Coriell sample NA12878 and was released in May 2015 as NIST RM 8398 (tinyurl.com/giabpilot). To minimize bias and improve accuracy, 11 whole-genome and 3 exome data sets produced using 5 different technologies were integrated using a systematic arbitration method. The Genome in a Bottle Analysis Team is adapting these methods and developing new methods to characterize 2 families, one Asian and one Ashkenazi Jewish from the Personal Genome Project, which are consented for public release of sequencing and phenotype data. We have generated a larger and even more diverse data set on these samples. We are analyzing these data to provide an accurate assessment of not just small variants but also large structural variants (SVs) in both "easy" regions of the genome and in some "hard" repetitive regions. We have also made all of the input data sources publicly available for download, analysis, and publication. We combined the strengths of each of our input datasets to develop a comprehensive and accurate benchmark call set. Many challenges exist in comparing to our benchmark calls, and thus we are working with the Global Alliance for Genomics and Health to develop standardized methods, performance metrics, and software to assist in its use.

https://www.slideshare.net/secret/vFuQ395PeVZJ4U

Scott Jackson, Ph.D. & Jason Kralj, Ph.D. – NIST

Genome Scale Measurements

A Prototype Mixed Pathogen Reference Material

In this work, 4 pathogens (S. enterica, S. aureus, P. aeruginosa, and C. sporogenes) were mixed in two ways—one equigenomic (keeping the number of each pathogen genome constant), and one 10-fold dilution series. To better mimic clinical samples, a human background was chosen, representing approximately 1000-fold more genomic content than all microbial sources. The DNA were sourced from NIST as either reference materials (NIST RM 8398), or potential RMs that have been thoroughly sequenced and vetted. 4 taxonomic identification tools were employed to analyze the data, including 2 external vendors and 2 open-source programs (Kraken and MetaPhlAn). The results from each platform generally agreed with respect to identifying the genus (4/4 pathogens) and species (3/4 pathogens, missing C. sporogenes) for the equigenomic sample. The abundance in terms of number of reads attributed of each pathogen also were in good agreement, except for C. sporogenes, as most of the tools did not have that particular organism in their database, leaving identification to rely on sequence information from other Clostridium species. For the dilution series, only the 2 most abundance species were correctly identified, indicating that the limits of detection become challenging on these systems for read counts below a few hundred (here, approximately 0.1 % of the total sample). Tools incorporating viral genomes also found human herpesvirus 4 (e.g. Epstein-Barr) used to immortalize the RM 8398 cell line in amounts that exceeded the other pathogen genomes. Given the wide discrepancies in tool outputs, it was clear that standards would benefit developers and end-users alike in understanding and gaining confidence of the

combined sequencer and post-processing analyses, and that carefully-curated databases will be necessary to ensure proper identification for clinical analyses.
https://www.slideshare.net/secret/sghGHVcQpPrOYI

Summary of Breakout Groups

Breakout groups were formed to discuss and brainstorm the composition, quantification, characterization, and bioinformatics associated with the mixed microbial reference material. The following sections are the results of those discussions. Prior to the breakout sessions, attendees decided by consensus to focus the reference material(s) towards clinical samples, as this type of isolate likely has the broadest range of applications. Meeting attendees agreed that nucleic acids (DNA and RNA) should be available for utilization, noting the potential to deploy whole organism-based materials at some later date.

Group 1 (Composition)

There is an ecosystem of standards that the community needs to address. Standards could potentially be used during the pre-analytical process, sequencing library preparation, sequencing, bioinformatics, and metadata collection. This group discussed:

1. Standard clinical matrix (blood, urine, stool)
2. Standard cells
3. Standard genomic material
4. Standards data sets

Participants from this group agreed that the initial focus of the reference material should target the limit of detection and precision of identification. Looking forward, the goal may be to develop a suite of genomic DNA (gDNA), RNA, and whole cells in a variety of matrices. However, this task is beyond the scope of the workshop at this point.

One potential incarnation would be a collection of single-isolate gDNA samples, utilizing a protocol for how to mix the samples. Having a "sample background” would be desirable, and one idea that was suggested was to utilize the ERCC cDNA. However, it wasn’t clear how best to develop that concept and therefore, this approach will be reconsidered in the future. For the organisms themselves, G+C content, DNA/RNA, genome size, near neighbors, repetitive content, and extremophiles would ideally be represented. Non-pathogenic organisms are still valuable to include in the NGS reference standard, as this material assesses the tool performance rather than specific pathogens. Background DNA extracted form specific samples might be helpful, though finding large enough quantities would likely be a limiting factor, and was considered impractical at this time.

Group 2 (Quantification)

Participants in this breakout group discussed formulation of the reference material. The group agreed that procedures for taking the individual strains and mixing them appropriately would be needed. What was clear was the need to challenge the sequencing process and bioinformatics. Limit of detection-type analyses were considered high priority, and those would be facilitated through dilution series (e.g. $\log_{10}$) experiments. Near-neighbor discrimination was also discussed. For example, material with a pathogenic strain in a background of otherwise commensal microbes would have value both from a technical standpoint as well as LOD-type analysis. When matrices were discussed, group members stated that for the purpose of quantifying the DNA this aspect should be kept separate.

Group 3 (Characterization)

The discussion focused on what was needed to characterize the different strains that could be used to create the reference material. Many of the standard analyses should be done such as G+C content, KK-PFGE, 16S, MALDI-TOF, and cell line ID. Nucleic acid impurities were noted as a potential problem, but highly-purified kits are now available that could help to mitigate this issue. Also, strain-specific RT-PCR assays/probes could be developed as a companion to the reference material. Length of fragment DNA was also a considered a component of characterization, although it would be challenging if the goal were long-read sequencing.

Group 4 (Bioinformatics)

This breakout group specified the benefit of early distribution of *in silico* materials (e.g. data sets from prospective reference materials). These data should approximately match the coming standards and potential clinical presentations. Highly accurate reference genomes for all pathogens are needed to create this reference material. Background data or metadata was another necessary element that should be provided with the reference material. This type of information should come from the laboratory, institution, or agency providing the organism. Releasing individual strains would also enable users to develop materials better representing their applications, and enable additional *in silico* experiments.

Developing a Mixed Pathogen Reference Material

The consensus outcome from the NIST-FDA Workshop on microbial NGS-based standards was to use a plug and play modular design such as a “96-well” plate (or another cassette of individual organisms) device that could allow for custom microbial mixtures, spike-in synthetic markers, and genetically engineered organisms. The customer can create their own mixtures and NIST/FDA would provide them with the standard operating procedure on how to perform the synthesis. There was consensus amongst workshop participants that the initial reference materials be composed of nucleic acid that has been derived from organisms having both clinical relevancy and diverse genome characteristics. Once these standards or version #1 have been established, there could be a transition to whole cells (version #2). Genomic material should be characterized by multiple methods such as qPCR, dPCR, and NGS. Also, effective strain characterization should involve “accepted” or “traditional” methodologies including PFGE, structural mapping, 16S sequencing, MALDI-TOF, or serotyping. This type of information should be included in the certification or package insert that comes with the standards.

Nucleic acid extraction methods used to obtain genomic material for the standards require characterization in order to evaluate extraction bias and efficiency. Ultra-high purity reagent kits are currently available and could be used for extracting the nucleic acid of interest. The background contamination must be thoroughly evaluated and subtracted from the microbial nucleic acid. However, it can be challenging to differentiate between what is a contaminant versus what is general background. The nucleic acid should be distributed in “real world” human clinical backgrounds including saliva, blood, urine, sputum, and stool. Standards should be quantitated using pre-analytical tools such as a Bioanalyzer®, Qubit®, or other established technology. The instrument used for quantitation should be reported out with the standard.

The organisms included in the reference mixtures should have genomes that cover a variety of characteristics. These qualities encompass high/low G+C content, large/small genome sizes, accessories elements, and repetitive content (e.g. homopolymer regions). Not only should pathogenic, virulent, and multidrug resistant organisms be included if possible as reference standard components, but near neighbors, commensals, and extreme organisms (i.e. slow growers, growth in cold/hot temperatures, growth in high salinity) are also important to include in the reference materials. Some workshop participants suggested that select agents should not be used because of restrictions and difficulties in shipping the organisms nationally and internationally. Instead, signature regions of biothreat agents could be created and used to develop the standards. In addition to bacterial isolates, viral sequences should be used but created as synthetic constructs. Acquisition of organisms that uphold the defined characteristics could come from workshop participants or FDA-ARGOS.

Participants suggested that reference mixtures have a defined volume and be mixed at clinically relevant ratios. Equigenomic mixtures or equal ratios of genomes were also discussed at the workshop. The use of spike-in controls that mimic ERCC ExFold RNA Spike-In Mixes, quality control samples, and well-characterized background samples need to be included in the design of the reference plate.

These standards must be compatible with available NGS platforms and therefore, prior to marketing and distribution, the standards must be evaluated on a variety of NGS systems and assessed with multiple replicates. In order to adapt to long read sequencing technology (e.g. Pacific Biosciences) these standards may need to be modified to be of high molecular weight. Whole cell standards may be necessary (or version #2).

Factors to Consider When Formulating an NGS-Based Reference Material

- Quantitation of genome copies (e.g. Qubit, NanoDrop, qPCR, ddPCR, RT-qPCR, Mass Spectrometry)
- Limit of detection (high, low samples)
- Phylogeny (i.e. Kingdom)
- Pathogen ratios
- Ease to grow and supply
- High and low G+C ratio
- Cell wall
- DNA/RNA
- Fragment size
- Stability
- Goal (e.g. quantitation)
- Gram negative/positive
- Matrix
- Interfering substances
- Biohazard, danger and legal issues
- Synthetic or naturally-sourced

The primary goal of this first material is to meet as many of these criteria/goals while minimizing the number of organisms to employ.

Top Organisms to Consider for Mixed Sample Reference Material

Naturally-Derived Materials

Mycobacterium avium complex - (extreme, slow-grower; source from ARGOS)
Listeria monocytogenes - (extreme; source from CDC –John Besser?)
Shigella sonnei - (virulence; source from ARGOS)
E. coli 0157:H7 - (virulence and near neighbor to Shigella, ESBL-KPC producer, Shiga toxin production, repetitive content; source from ARGOS)
Klebsiella pneumoniae - (MDR, plasmid; source from ARGOS)
Staphylococcus epidermidis - (commensal bacteria, near neighbor *Staphylococcus aureus*; source from ARGOS)
Bacteroides fragilis, Achromobacter xylosoxidans - (commensal bacteria; source from ARGOS)
Rhinovirus - (commensal virus, asymptomatic carriers; source?)
Neisseria sicca - (repetitive content, resistance; source from ARGOS)
Aeromonas hydrophilia, Streptomyces coelicolor - (high GC content; source *A. hydrophilia* from ARGOS, *S. coelicolor*?)
Plasmodium falciparum, Acinetobacter baumannii - (low GC content; source *A. baumannii* from ARGOS, *P. falciparum*?)
Burkholderia cepacia/cenocepacia - (large genome, high GC content; source from ARGOS)
Haemophilus influenzae - (small genome; source from ARGOS)
Streptomyces scabiei 87 22 – (large genome, high GC content, source?)
Staphylococcus saprophyticus – (low GC content, genus similarity to *S. aureus*)
Salmonella enterica subsp. Typhimurium LT2 – (moderate genome and GC content, NIST source)
Salmonella enterica subsp. Arizonae – (similar to *S. enterica* LT2 for discriminating similar strains)
Pseudomonas aeruginosa – (high GC content, moderate genome, NIST source)
Synechocystis sp. PCC 6803 – (moderate GC and genome, environmental species unlikely to be found in clinic or laboratory setting)
Bacillus thuringiensis – (low GC, environmental, *B. anthracis* simulant)
Vibrio furnissii – (moderate GC, genome size, relatively benign *Vibrio* strain with similarities to *Shigella*, *Yersinia*, *E. coli and Pseudomonas*)
Lentivirus simian immunodeficiency virus (SIV) – (small genome)
***Flavivirus* West Nile Virus** – (small genome, mosquito-vector spread)

Synthetic Constructs

Note: Natural materials in this category carry significant distribution challenges. Generated constructs would be inspired by known genome assemblies, but sequence will be effectively scrambled to ensure the material can be disseminated freely.
Ebola
***Filovirus* Lloviu cuevavirus** – (distant relative to Ebola and Marburg, non-human pathogen, unculturable)
Chikungunya
MERS
Hantavirus

Miscellaneous

FDA and NIST will take the opinions gathered from the breakout groups and formulate a prospective procedure for designing and generating the first mixed pathogen reference material for NGS-based assays. Communications will continue via email, until a website and/or online groups can be established.

In addition to the prospective RM, we will look into the possibilities of reference materials containing whole organisms to assist characterization of pre-analytical variability, and reference data that could be used for generating in silico data for experimental work on bioinformatics tool development. NIST will elicit help from other attendees for how best to proceed with distribution and curation on both of those topics.

Total attendance was estimated at 60. We limited registration to 100, and approximately 20 participants could not attend.

Conference Services were used to help with the logistics. Our point of contact was Mary Lou Norris, and she was assisted by Jami Schwartz and Crissy Robinson.
We tentatively planned the next workshop for next year in October.

Bibliography

CDC. (2010). *National Ambulatory Medical Care Survey*. Retrieved from http://www.cdc.gov/nchs/data/ahcd/namcs_summary/2010_namcs_web_tables.pdf

Research America. (2015). Retrieved from http://www.researchamerica.org/sites/default/files/InfectiousDiseaseFactSheet_0.pdf

FDA-ARGOS. http://www.ncbi.nlm.nih.gov/bioproject/231221

JGI Genome Portal. (http://jgi.doe.gov/data-and-tools/genome-portal/)

LANL HIV Databases. http://www.hiv.lanl.gov/content/index

Saccharomyces Genome Database. http://www.yeastgenome.org/

DE Wood & SL Salzberg (2014). *Genome Biology* **15**:R46. DOI: 10.1186/gb-2014-15-3-r46.

N Segata, et al. (2012). *Nature Methods* **9**: 811-814. DOI: 10.1038/nmeth.2066.

C Hong, et al. (2014). *Microbiome* **2**:33. DOI: 10.1186/2049-2618-2-33.

P.-E. Li, et al. (2016). *bioRxiv*. DOI**:** 10.1101/040477.

SK Ames, et al. (2013). *Bioinformatics* **29**:18. (2013). DOI: 10.1093/bioinformatics/btt389.

.