NIST Special Publication 1222

Report on the NIST/DHS/FDA Workshop: Standards for Pathogen Detection for Biosurveillance and Clinical Applications

August 14-15, 2017

Scott Jackson
Nancy J. Lin
Jason Kralj

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1222

National Institute of Standards and Technology
U.S. Department of Commerce

NIST Special Publication 1222

Report on the NIST/DHS/FDA Workshop: Standards for Pathogen Detection for Biosurveillance and Clinical Applications

August 14-15, 2017

Scott Jackson
Nancy J. Lin
Jason Kralj
Biosystems and Biomaterials Division
Material Measurement Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1222

April 2018

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology Special Publication 1222
Natl. Inst. Stand. Technol. Spec. Publ. 1222, 44 pages (April 2018)
CODEN: NSPUE2

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1222

Workshop Organizers

Scott Jackson
Nancy J. Lin
Jason Kralj

National Institute of Standards and Technology
Gaithersburg, MD USA

Workshop Sponsors

National Institute of Standards and Technology
Department of Homeland Security, Science and Technology Directorate
United States Food and Drug Administration

Corporate Sponsors

ATCC
CosmosID
Illumina
Microbiologics, Inc.
Zymo Research Corp.

Abstract

The 2017 NIST/DHS/FDA Workshop: Standards for Pathogen Detection for Biosurveillance and Clinical Applications was held at the National Institute of Standards and Technology (Gaithersburg, MD) on August 14-15, 2017. The workshop brought together subject matter experts from the clinical and biosurveillance communities to discuss standards to support the use of next generation sequencing for pathogen detection. The workshop consisted of invited and contributed oral presentations, poster sessions, a panel discussion, and break-out sessions. In general, the participants highlighted the need for continued development and utilization of cell- and DNA-based reference materials, standard operating procedures and protocols, reference data, defined metrics, and interlaboratory studies to assess measurement biases and overcome measurement and bioinformatic challenges and regulatory hurdles. This workshop was the third NIST-hosted workshop on sequencing for pathogen detection/identification since 2014. Pathogen detection is relevant to a number of programmatic priorities within the NIST Material Measurement Laboratory including microbial metrology, precision medicine, biomanufacturing, biodefense/forensics, food safety, antibiotic resistance, and water quality. Thus, it is our intention at NIST to continue to host this workshop theme approximately every two years.

Key words

Biothreat; clinical diagnostics; next generation sequencing; pathogen detection; reference materials; standards development.

Table of Contents

List of Figures

"*For the morbid matter of cholera having the property of reproducing its own kind, must necessarily have some sort of structure, most likely that of a cell. It is no objection to this view that the structure of the cholera poison cannot be recognized by the microscope, for the matter of smallpox and of chancre can only be recognized by their effects, and not by their physical properties*." — John Snow (1855)

1. Introduction

It has been nearly 150 years since scientists first demonstrated conclusively that microorganisms are agents of disease. With this discovery, germ theory became the predominant theory to explain disease transmission, quickly ousting the widely accepted miasmatic theory that described a noxious form of "bad air" emanating from rotting organic matter. This transformation sparked the "golden age" of microbiology during which time many microbes were identified as the causative agents of disease.

Since these early days, our ability to prevent, control, and understand the transmission of disease has been predicated on our ability to rapidly detect and identify pathogens in the host or environment. Detection and identification techniques have evolved over the past century but have largely been based on three fundamental approaches:

1) culture- and microscopy-based techniques for phenotypic and physical indicators,
2) antigen-based immunological techniques, and
3) DNA-based molecular genetic techniques.

Today, these methods are used to varying degrees across different application spaces.

DNA-based techniques in particular have lately seen rapid advances due to improvements in both DNA sequencing and computational tools, yielding a vast and diverse set of methods for pathogen detection. Today, the term "metagenomics" is used to describe the analysis of a collection of genomes (usually microbes) from an often complex sample using the combination of DNA sequencing and bioinformatics tools to identify constituents of the population. This approach holds tremendous potential, as these DNA-based tools can inform a wide spectrum of applications such as infectious disease, antimicrobial resistance, biosurveillance, and biomanufacturing. However, the *lack of consensus* among the myriad of sequencing and analysis tools underscores a real problem in translating this technology to the applied setting. To address this challenge, standards are needed to enable systematic characterization of the underlying processes, ultimately to increase confidence for decision-making in this high-stakes arena.

2. Workshop Overview

The purpose of this workshop was to present state-of-the-art pathogen detection technologies, primarily related to next-generation sequencing (NGS), and discuss ongoing and potential future efforts toward relevant standards, with an emphasis on needs in the clinical diagnostic and biothreat detection stakeholder communities. The workshop targeted expected primary users and adopters of these standards, including clinicians, the biodefense community, public health, industry, academia, and government laboratories. We invited subject matter experts from these areas to provide their perspectives on the needs for standards in this field. Break-

out sessions and a panel discussion were included to allow attendees to discuss specific topics identified previously as challenging to the field. In addition, over 35 posters were contributed by attendees to highlight their latest efforts in pathogen detection. NIST scientists also presented ongoing efforts focused on standards for pathogen detection. These efforts include a whole-cell (yeast) reference material developed in collaboration with DHS, a mixed pathogen DNA reference material developed in collaboration with FDA, and preliminary results of an ongoing metagenomic interlaboratory study hosted by NIST, CosmosID, and the Association of Biomolecular Resource Facilities (ABRF). The agenda included time for specific feedback on how these and other reference materials and standards can be implemented and expanded to address critical needs within the clinical and biosurveillance communities, as well as in environmental, public health, and biomanufacturing applications.

2.1. Workshop Participants

Workshop attendees (listed in Appendix A) included subject matter experts from a number of sectors. Of the 174 people registered for the workshop, approximately 142 were in attendance including 11 foreign nationals. The largest percentage of workshop registrants were from government (non-military) agencies, followed by industry, academia, non-profit organizations, and the military (Fig. 1).

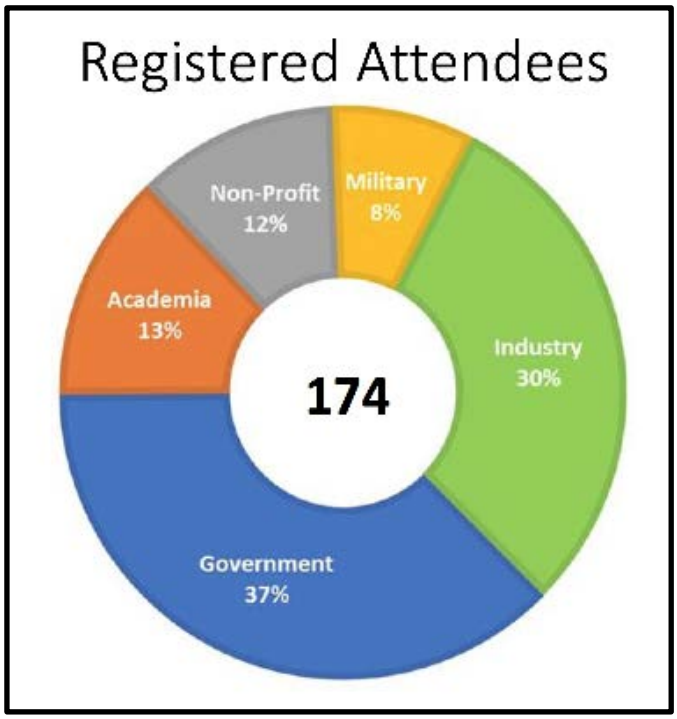

Fig. 1. Affiliations of the 174 registered workshop attendees.

The workshop was also webcast live to reach a broader audience. Online viewers emailed questions in real-time to NISTpathogen@nist.gov during the workshop, with moderators and

speakers addressing those questions as time permitted. Webcast statistics reported that the livestream was viewed by 59 of the 69 registered viewers. An archival recording of the webcast is available.

Social media has become a powerful tool for disseminating information and highlighting recent findings, news, and workshop content. Considering this tool, we encouraged participants to use hashtag #NISTpathogen for any social media posts pertaining to the workshop and to ask questions in real-time via Twitter. Following the workshop, we purchased an analytics report from Twitter (TWEETREACH) describing the activity associated with hashtag #NISTpathogen from August 11, 2017 to August 18, 2017. During that eight-day period, there were 275 tweets from 143 contributors that ultimately reached 207,788 Twitter accounts. The exposure was estimated to be 1,883,890 impressions, as determined by Twitter (Fig. 2).

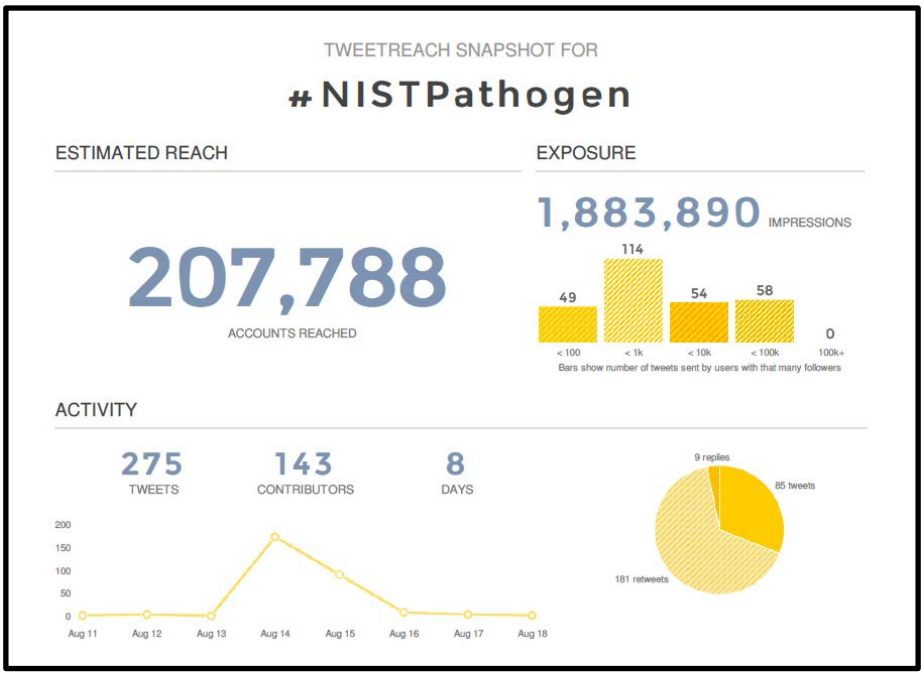

Fig. 2. Twitter Analytics Report for #NISTpathogen
Report covers August 11, 2017 to August 18, 2017.

3. Oral Presentations

The full agenda for the workshop is provided in Appendix B. The workshop began with a keynote address from Dr. Rita Colwell, followed by oral presentations from invited subject matter experts (abstracts provided in Appendix C). In addition, seven contributed poster presenters were invited to give a short lightning talk on their work (abstracts in Appendix D). Presentation slides are available on the workshop website, with permission from the speakers. Abstracts for contributed posters are provided in Appendix E.

4. Summary of Panel Discussion

Dr. Scott Jackson moderated a panel discussion on standards for pathogen detection. Panel members were Dr. Rita Colwell, Dr. Charles Chiu, Dr. Shanmuga Sozhamannan, Dr. Heike Sichtig, Dr. Robert Schlaberg, Dr. Jonathan Jacobs, and Ms. Alexa McIntyre. Highlights from the panel discussion are provided here, while the full panel discussion is available online for viewing.

In your own mind, what does a "standard for pathogen detection" look like?

Panel responses focused on the need to support and enable results that are accurate, reliable, reproducible, readable, and easily conveyable to non-experts. These types of results could be supported through relevant standards related to all aspects of the process from sample to answer. It was recognized that this goal for standards is far-reaching and will take time to achieve.

Is NGS going to make all other pathogen detection tools obsolete in the near future? Or will other (targeted) methods co-exist?

The panel indicated that typically an ensemble of methods outperforms a single one, thus older tools will likely remain in the toolbox for the foreseeable future. In the meantime, NGS hasn't been fully explored, and there are opportunities for additional tools including functional measurements of both the host and the microbe. Other points included expanding the focus of pathogen detection to consider polymicrobial infections, and the need for lower cost methods relative to NGS, particularly for biodefense.

What are the biggest challenges (gaps) that are preventing widespread use of NGS for pathogen detection in the clinic and in biosurveillance?

Answers included cost, expertise, time to actionable results, and dealing with different manufacturers for each step in the process. If a single technology producer made an end-to-end product solution, costs would come down dramatically.

Panelists also said that clinical application of microbial metagenomics requires advanced and improved clinical workflows, which are challenging to implement for a number of reasons, including

- quality control and quality assurance process controls are needed,
- it can take three to six months to train a certified technician,
- control materials are needed for proficiency testing,
- there is not an infectious disease-related billing code for sequencing, and
- transferring improvements in this rapidly developing field to the clinic requires retraining certified technicians and revalidating process controls, which presents a significant barrier in terms of expense and time consumption.

Dr. Colwell offered, “I think NGS is going through the same slow-motion accommodation and acceptance. It’ll get there, but it’s taking longer, I think, because the hurdles are higher and steeper than they were 20-30 years ago. Not a bad thing, but it is slow.

“I wish there was some way that FDA were able to somehow approve NGS for the difficult…infections. Clearly, PCR takes care of the single questions… Chronic sufferers, then, really do need something to help them. And it’s usually because…it’s the anaerobes or very particular pathogens.”

Dr. Sichtig said that the FDA is open to entertaining pre-submissions to discuss these areas and the possibility of alternative pathways for approval.

How should a reference material be used to validate a pathogen detection measurement?

Untargeted NGS can be performed in a validated way, such as by validating for each pathogen individually, and then reporting the full list of findings.

5. Summary of Breakout Sessions

Four breakout sessions were scheduled for the second day of the workshop to discuss and brainstorm specific questions related to standards development for pathogen detection. The four topics were Biosurveillance Community Needs, Benchmarking Bioinformatic Tools, Quantitating Whole Cell RMs, and Interlaboratory Study Design. There were two separate break-out session times, thereby allowing each participant to contribute to two different topics. Summaries of the discussions follow.

5.1. Topic 1: Biosurveillance Community Needs

The biosurveillance breakout session focused on addressing questions related to standards for the biosurveillance community, with a focus on biothreat detection.

1. What unique challenges does the biosurveillance community face regarding pathogen detection (as compared to the clinical community)?
2. What standards are needed to address these challenges?
3. Considering what has been presented at this workshop regarding standards development activities (e.g., DNA-based materials, whole cells, spiked materials, etc.), which efforts are sufficient? Which efforts should be expanded? What new efforts are needed?

5.1.1. Question 1: Unique Challenges for Biosurveillance

Participants indicated that pathogen detection for biosurveillance applications differs from clinical pathogen detection, in that it

- includes a wide range of background matrices and substrates, resulting in highly variable sampling;
- supports high stakes decisions where both false positives and false negatives can have serious repercussions for large groups of people ranging from panic and disruption to loss of many lives;
- encompasses decontamination efforts;
- does not have regulatory oversight or monitoring;
- does not usually have a clinical phenotype to suggest a target organism;
- has a need for continuous sampling that is tracked over time; and
- has rare positive events making it difficult to develop a process and understand what is needed to detect, monitor, and respond.

5.1.2. Questions 2 and 3: Standards Needs and Efforts

When participants were asked about the standards needs for the biosurveillance community, the answers mirrored those heard throughout the entire workshop, including materials, data, methods, and best practices to cover the entire detection process, where each standard has a very defined scope. Other specific comments included the need to include viruses, specific matrices of interest, expanded cell-based standards, operationally relevant materials, and safe mimics of BSL-4 agents. Also mentioned were standards for false negatives, continuous data collection, and sample collection.

Regarding the type of detection that would be most appropriate for biosurveillance, the participants thought that the initial focus should be targeted amplicon sequencing of agents of interest. However, it was recognized that this approach will not allow for detection of unanticipated, new, or modified organisms, thus shotgun metagenomic approaches to detect unknowns are also needed. Some participants thought it would be ideal to develop tools that could do both at the same time – detect known biothreat agents while also screening for other potential threat agents.

To detect specific agents of interest, development of bioinformatic tools that can make a presence/absence call with a reported uncertainty is needed. This type of approach will likely involve both sequencing-related metrics, such as genome coverage, confidence scores, etc., as well as biological metrics such as level of confidence in the detection of specific targets, e.g., toxin genes.

5.2. Topic 2: Benchmarking Bioinformatic Tools

One of the major topics of discussion during the workshop was metagenomic data analysis tools. At the time of this workshop, there were over 70 different metagenomic data analysis ("profiling") tools available. We asked this break-out group to discuss four topics:

1. Based on current benchmarking studies, it seems that many tools make trade-offs with respect to sensitivity and specificity. How might this be overcome?

2. Can we assign a confidence score? How is it calculated? What format should the output be? (binary, %, etc.)

3. When generating *in silico* raw data (fasta/fastq) for benchmarking, what needs to be considered in the reference data design?

4. Are current databases adequate? What's missing? Can/should we develop a standard database that is used across all profiling tools?

5.2.1. Question 1: Trade-offs in Sensitivity and Specificity

The overarching feedback was that the application would drive tool selection, and finding the right balance for sensitivity/specificity would be a major decision point in that selection.

As an example, tools for discovery of novel organisms have different criteria than those for diagnostic systems in a hospital. For the former, one is likely tolerant of false positives in the hope of finding the nearest neighbor, and therefore, an eventual classification. In the latter, improper and/or overtreatment carries additional risks to the patient.

It was not clear that a "universal" level of sensitivity/specificity would be possible even with fixed coverage (# of reads) for each sample.

5.2.2. Question 2: Confidence Scores

There was no clear consensus on the topic of confidence score. While possibilities were proposed, it was brought up that some tools may be doing this already. Another point raised the concern that confidence scores might not be comparable across platforms, as different tools may use different calculations to obtain the scores. Others noted that confidence scores may be unnecessary in discovery applications, as there is no confidence assumed or expected.

Some raised the issue of false positives via "shadow" detection from near neighbors, and how those are not handled in some systems. The resulting approaches could be too simple or too complex for the application at hand, and comparison of scoring methods would need additional vetting.

5.2.3. Question 3: *In silico* Data

There was general agreement that existing *in silico* data does a poor job representing real data. For instance, contamination is not generally represented, error profiles from different platforms vary, and coverage profiles may not present enough variability. In turn, synthetic sequences are easy to identify, even within a background of real data.
Still, there seems to be some momentum to develop these tools in a user-friendly, cloud-based design that would enable one to select organisms, abundances, and platforms to generate datasets for benchmarking tools. If designed correctly and characterized

thoroughly, these *in silico* data can provide "ground truth" as reference data with quantitative and qualitative attributes of high confidence.

5.2.4. Question 4: Database Adequacy

The answer to this question regarding databases seemed to be no, databases are not adequate. It seemed universally accepted that databases pose a significant challenge to accurate metagenomics analyses. Discussion points included the following.

- It is difficult to account for constituents of a database.
- Version control could help track/update changes.
- Updates needed to happen rapidly or be easily implemented.
- A large number of near neighbors gave redundancy but did not improve coverage of all genomes.
- Specific-use databases may find application for other use cases.

The concept of a single, all-encompassing database was not recommended, as various applications may require differing levels and amounts of information, and one large database might make the system less portable to end users.

5.3. Topic 3: Quantitating Whole Cell RMs

Whole cell RMs are recognized as having the capability to challenge the entire pathogen detection process, including DNA extraction. This group discussed appropriate formats for whole cell RMs, critical properties to be quantified for a whole cell RM, and how to consider addressing these properties.

1. What format is most appropriate for a whole cell reference material? Factors to consider include:
    a. pure strains, strain mixtures, in a relevant matrix, etc.
    b. live, inactivated, etc. (need to consider biosafety level?)
    c. minimum number of strains to include
2. What are the three most critical properties to be quantified for a whole cell reference material? What efforts should be started to address these needs?

5.3.1. Question 1: Format

Regarding what a whole cell RM should look like, at a minimum, it should be a stable material where effects of storage conditions on the number of cells and the expected DNA and RNA yields should be well-characterized and understood. The species included will likely be governed by the application and might consist of a set of individual species that could be mixed as desired by the end user, or a set of species mixed at a pre-determined ratio. Maintaining the species as physically separate samples provides more flexibility to the end user but also leaves room for more uncertainty in the final mixture concentrations due to pipetting and mixing errors. The set should include Gram-positive and Gram-negative

microbes, spore-formers, and microbes with varying lysis efficiency. Participants also discussed the need to include BSL-1 and BSL-2 organisms to enable broader use. For instance, BSL-2 strains may be more appropriate for clinical applications where BSL-1 strains may be sufficient for other applications.

The material format, such as preserved or inactivated, should be amenable to challenging both DNA and RNA measurements. All participants recognized the importance of verifying cell integrity after cell preservation or inactivation. A variety of methods to stabilize or inactivate the cells were discussed, including heat, radiation, glycerol, ethanol, and lyophilization, but a preferred method did not emerge. Each method has advantages and disadvantages, and the optimal method will likely vary with application. One concern with preserved or inactivated RMs was the likelihood of changes in the RNA, DNA, proteins, available epitopes, lipids, etc., to the point that the RM no longer presents the relevant characteristics of the starting material, such as difficulty in lysis.

The background matrix is another important aspect for a cell-based RM, yet it is highly dependent on the application. Participants indicated that an optional matrix could be included with the cells, a separate standard matrix could be provided for various applications, or each user could spike the cell RM into their application-specific matrix, ideally where there is low background within that matrix.

5.3.2. Question 2: Quantification of Cell RMs

Several methods to count microbial cells are available, yet it is challenging to determine and compare accuracy due to lack of relevant, cell-based reference materials. Cell RMs quantified for total cell number would find application in calibrating, validating, and comparing counting methods. In the meantime, multiple orthogonal methods could be evaluated with lesser-characterized control materials to assess proportionality, linearity, and relative differences among the methods.

The group proposed a potential starting point: a small number of well-characterized, well-documented microbial species available as both live and inactivated (fixed) cells for comparison of formats. It was also noted that the applicability of the cell material would increase if coordinating DNA-based RMs were available to match the strains found in the whole cell RMs.

5.4. Topic 4: Interlaboratory Study Design

The group was tasked with the following:

Design an interlaboratory study using a well-characterized mixed microbial gDNA reference material to determine sensitivity (e.g., limit of detection) and specificity (e.g., differentiating unrelated organisms including near neighbors) of sequencing-based pathogen detection. Factors that were discussed included:

1. How many strains should there be? How should they be mixed? How many mixtures are needed? How many replicates?
2. How/where should data be stored and analyzed? Who should analyze the data?
3. How should results be summarized and presented?
4. What are the three most important things to consider when designing an interlaboratory study?

For this purpose, the sensitivity (*Sen*) and specificity (*Spec*) are defined thusly:

$$Sen = \frac{TP}{TP+FN} \quad (1)$$

$$Spec = \frac{TN}{TN+FP} \quad (2)$$

Where *TP, TN, FP,* and *FN* are abbreviations for true/false positive/negative. Other metrics such as positive predictive value (PPV, or precision) and accuracy could be computed, as no additional information would be required.

Two groups, each consisting of approximately 15 people, answered parts of these questions. Generally, no consensus was reached as far as a single high priority interlaboratory study that was easily defined. Instead, general guidelines and topics were suggested. Instead of a larger (more participants) interlaboratory study, the group recommended employing a set of smaller "pilot" interlaboratory studies to:

- identify necessary requirements and best practices for these studies,
- address very focused questions,
- quickly probe for experimental factors that have large, significant effects using a low statistical power, and
- maintain focus and leadership.

5.4.1. Question 1: Strains

More is not always better. The participants thought it is preferable to emphasize a quality study, where controlled work is performed using only a few well-characterized strains rather than trying to work with many strains (and an overly complex sample). Experimental factors to be evaluated should be defined in the context of the number of strains in the study.

Ideas discussed for testing library preparation included: (1) having two other laboratories sequence each laboratory's library, and (2) having NIST prepare libraries to be sequenced for a consistent, high quality material.

The use of technical replicates should be considered to test consistency within a laboratory. There was no strong opinion about the number of replicate samples per strain; rather, it was agreed that statistical analysis should dictate this number.

5.4.2. Question 2: Data Storage and Analysis

The consensus was that data should be stored in a centralized repository that is open access. Generally, the group was in favor of each laboratory analyzing its own data from the study as well as a subset of data from other participating laboratories. However, in most situations, all the raw sequence data should also be analyzed using the same bioinformatic tool (or tools) to understand variability associated with upstream steps in the measurement process.

5.4.3. Question 3: Presentation of Results

There was no consensus on how to summarize and present the results other than they should be defined before any study is undertaken. It was generally agreed that there were too many unique characteristics of each analysis tool to present the output from them as such; instead, the focus should be on answering performance questions, to the extent any tool is fit for purpose.

5.4.4. Question 4: Overall Study Design

There were a few individuals who weighed in on this question based on previous experience.

1. Think about how to implement negative controls; they are critical to reduce and characterize false positive (FP) rates.
2. Collect a lot of metadata. It is one of the only ways to ensure consistency between methods and pinpoint potential problems. Though it is probably considered onerous by most, it can help to recognize potential sources of variability.
3. Take laboratory background samples.
4. Have quality controls to rule out gross-level mistakes.
5. Validate procedures and materials before sending to participating groups.

Based on this feedback, the prudent approach is to address specific questions with a carefully designed study and a small set of participants. This approach has several logistical benefits, including the ability to coordinate and communicate quickly and efficiently. Because of the low number of participants, only large effects would be statistically significant. To analyze these effects in more detail, interlaboratory studies with larger cohorts would be needed.

NIST said they would query workshop participants for help with specific proposed questions or requests for participating in these pilot studies, as appropriate.

6. Overall Outcome and Future Directions

Overall, workshop attendees highlighted a critical need for standards to increase confidence in NGS results and support high-stakes decision-making in pathogen detection for clinical and biosurveillance applications.

Whole cell reference materials represent one clear area of need for this community, as they are critically important for validating analytical methods for pathogen detection. These reference materials should be quantitative and qualitative; that is, the material should be validated for cell count (enumeration) and the taxonomy/strain/genome should be available.

The role of genomic DNA-based reference materials focused on characterization of those materials via interlaboratory studies. DNA reference materials could be used to prepare mixed samples for interlaboratory studies to characterize devices for sensitivity and specificity of NGS-based microbial detection. Viral genomic materials are also needed.

In addition to the need for cell- and DNA-based reference materials, the need for *in silico* datasets was also evident. *In silico* data should be generated using only whole genome sequencing data from high quality genome assemblies. *In silico* studies can be performed in parallel with wet lab experiments, and indeed may help identify input parameters needed to improve predictive models. These studies would also complement wet lab work by improving the understanding of and confidence in the underlying bioinformatic tools used to assess sample composition. NIST will elicit input on to how best proceed with distribution and curation for *in silico* data analyses.

NIST, DHS S&T, and FDA will review the input and feedback gathered during the workshop, including the break-out sessions, to refine current and design future projects for standards development. Communications will continue via email, and regular updates will be posted to https://microbialstandards.org/.

6.1. NIST Future Directions

Development of standards for pathogen detection via NGS remains a priority for the NIST Biosystems and Biomaterials Division. As such, we will continue to engage stakeholder communities to prioritize standards needs and ensure relevance and community-adoption of resultant standards. Results from this workshop have informed ongoing programs at NIST, as described below. Workshops will continue to be conducted on a biennial basis.

NIST efforts to develop a reference material based on yeast (*Saccharomyces cerevisiae*) cells are now transitioning to efforts to quantify bacteria. Bacteria are typically smaller with more complex morphologies as compared to yeast cells. Stakeholder input for this effort is needed, particularly regarding the role of whole cell materials to support evaluation of pathogen detection workflows from sample to result as well as appropriate strains to include in a cell-based reference material to support biothreat detection. The consensus of participants was that the list of strains included in the NIST mixed pathogen DNA reference material is also relevant for a mixed cell RM to support biosurveillance. Inclusion of biothreat agents is not currently needed.

NIST is currently developing a mixed pathogen DNA reference material. The material is expected to be completely manufactured by the Summer of 2018. In addition to validating the material at NIST, it is our intention to crowd-source the characterization of this material. Data from multiple laboratories will contribute to the material characterization as well as provide a preliminary understanding of lab-to-lab reproducibility. To this end, we are offering the material as a "Research Material" (or candidate reference material) to colleagues and stakeholders who are willing and able to sequence the material in their laboratory. In exchange for the free material, we ask that the data be shared with NIST (ideally via a NCBI Biosample). To date, dozens of requests for the material have been received. We expect to ship the material before the end of 2018. We are also considering the addition of viral sequences, as we now have guidance on developing viral standards while still complying with requirements of the Federal Select Agent Program. However, specific community needs and the corresponding measurements and standards to address these needs must first be clearly articulated.

NIST, CosmosID, and ABRF co-organized an interlaboratory study (The Metagenomics MVP Challenge v1.0) using a metagenomic DNA mixture as the input material. This study was designed to be small and relatively short (completed within 12 months) in order to characterize measurement bias associated with shotgun library preparation techniques and NGS platforms. In addition, the study was designed to provide experience and insight into feasibility of interlaboratory studies. NIST will use experience gained from this study to guide future studies, including those for the NIST genomic DNA microbial reference material.

Acknowledgments

The workshop was sponsored by NIST, DHS Science and Technology Directorate, and FDA. Corporate sponsorship for the workshop was provided by ATCC, CosmosID, Illumina, Microbiologics Inc. and Zymo Research Corporation.

The organizers thank their colleagues Megan Cleveland, Sandra Da Silva, Kevin Kiesler, and Samantha Maragh for assistance with the breakout sessions, as well as Jonathan Jacobs, Shanmuga Sozhamannan, and Scott Tighe for moderating breakout sessions. We also thank the NIST Public Affairs Office for help with logistics, including Mary Lou Norris, Gladys Arrisueno, Crissy Robinson, and Karen Startsman. The organizers thank all presenters and participants for their contributions to workshop discussions.

The Department of Homeland Security (DHS) Science and Technology Directorate supported this work under the Interagency Agreement FTST-17-00017 with NIST.

Appendix A: Attendee List

Name	Organization
Nadim Ajami	*Baylor College of Medicine*
Beena Akolkar	*NIDDK*
Omayma Al-Awar	*Illumina*
Marc Allard	*US Food and Drug Administration/CFSAN ORS*
Kensey Amaya	*Contractor*
John Bagnoli	*MRIGlobal*
Michael Bazaco	*U.S. Food and Drug Administration*
Linda Beck	*DoD Naval Surface Warfare Center Dahlgren*
Brian Beck	*Microbiologics*
Kimberly Bishop-Lilly	*Naval Medical Research Center*
Levi Blue	*Canon US Life Sciences*
Philip Bocock	*Biolog*
Richard Boykin	*Nanostring Technologies*
Liliana Brown	*NIH/NIAID*
Lucy Burns	*Microbe Inotech Labs, Inc*
Juan Bustamante	*uBiome*
Rachel Campbell	*DHS OHA NBIC*
Patrick Chain	*Los Alamos National Laboratory*
Lindsey Chambers	*Canon U.S. Life Sciences*
Corey Chen	*Aperiomics*
Charles Chiu	*University of California, San Francisco*
Jessica Chopyk	*University of Maryland*
Steven Choquette	*NIST*
Shinhai Chu	*Illumina*
Jongsik Chun	*ChunLab/Seoul National Univ.*
Megan Cleveland	*NIST*
Rita Colwell	*CosmosID, Inc.*
Seth Commichaux	*US Food and Drug Administration/CFSAN*
Turner Conrad	*USAMRIID*
Colette Cote	*BioReliance*
Lorenzo D'Amico	*Baylor College of Medicine*
Sandra Da Silva	*NIST*
Manoj Dadlani	*CosmosID*
Tamar Dickerson	*FDA-MRIGlobal*
Linda Duffy	*NIH/NCCIH*
Robert Duncan	*US Food and Drug Administration/CBER*
Tara Eskandari	*NIST*
Kathrine Figueroa	*Walter Reed Army Institute of Research*
Carolyn Fisher	*US Food and Drug Administration/CBER*
Kenneth Frey	*NMRC-Frederick*
Christian Fung	*Walter Reed Army Institute of Research*
Jayanthi Gangiredla	*US Food and Drug Administration*
Solomon Gebru	*US Food and Drug Administration/GbF*
Karlis Graubics	*Cosmos ID*
Nick Greenfield	*One Codex*
Yan Guo	*Pacific Biosciences*
Nur Hasan	*CosmosID*
Manzour Hazbon	*ATCC*
Jordan Hitz	*Tetracore Inc*
Hayley Hogan	*Illumina*
Charles Hong	*DTRA JSTO*
Kelly Hoon	*Illumina*
Chung-Ying Huang	*NSTG*
Kyle Hubbard	*Harris*
Joshua Hyman	*University of Wisconsin*
Crystal Icenhour	*Aperiomics*
Scott Jackson	*NIST*
Jonathan Jacobs	*MRIGlobal*
Shannon Johnson	*Los Alamos National Laboratory*
Arifa Khan	*US Food and Drug Administration/CBER*
Kevin Kiesler	*NIST*
Katalin Kiss	*ATCC*
William Klimke	*NCBI*

Frank Kolakowski	*Tetracore, Inc.*
Jason Kralj	*NIST*
William Kramp	*HHS/ASPR/BARDA*
Claudia Lam	*US Food and Drug Administration*
John Lee	*BARDA*
Joseph Leonelli	*ATCC*
Jaclyn Levy	*IDSA*
Po-E Li	*Los Alamos National Laboratory*
Nancy Lin	*NIST*
Sabina Lindley	*US Food and Drug Administration-CFSAN*
Laurie Locascio	*NIST*
Cynthia Long	*ATCC*
Jennifer Lu	*Johns Hopkins University*
Maria Mayda	*ATCC*
Alexa McIntyre	*Weill Cornell Medicine*
David Melka	*US Food and Drug Administration-CFSAN*
Tim Mercer	*Altius Institute*
Heather Miller	*Johns Hopkins University*
Daniela Miller	*US Food and Drug Administration/ORISE*
John Miller	*US Food and Drug Administration/DOE*
Tim Minogue	*USAMRIID*
Dev Mittar	*ATCC*
Kelly Moffat	*CosmosID*
Robert Moss	*Illumina*
Michael Muchow	*NIST*
Vikram Munikoti	*BioWatch Program, DHS*
Todd Myers	*US Food and Drug Administration*
Daniel Nasko	*University of Maryland*
Yonas Nebiyeloul-Kifle	*DHS S&T*
Anumeet Nepaul	*Fraunhofer CESE*
Kristen O'Connor	*Defense Threat Reduction Agency*
Christian Olsen	*Pacific Biosciences*
Nathanael Olson	*NIST*

Andrea Ottesen	*US Food and Drug Administration*
Richard Pearlson	*Rap/Sat-NIST Committee for Social and Economic Resilience*
Jose Perez Donoso	*uBiome*
James Pettengill	*DHHS/FDA/CFSAN*
Casandra Philipson	*Naval Medical Research Center & DTRA*
Arthur Pightling	*US Food and Drug Administration/CFSAN*
Mihai Pop	*University of Maryland*
Anjan Purkayastha	*OPENBOX BIO*
Kashef Qaadri	*One Codex*
Tracy Radcliffe	*Biolog*
Sujatha Rashid	*ATCC*
Shashi Ratnayake	*NBACC*
Marco Riojas	*ATCC*
Gabriela Riscuta	*NCI*
Jon Ryan	*CosmosID*
Michael Salter	*AB Agri*
Saul Sarria	*US Food and Drug Administration*
Robert Schlaberg	*IDbyDNA Inc.*
Caitlin Sharpes	*US Army ECBC*
Heike Sichtig	*US Food and Drug Administration*
Stephanie Sincock	*HHS/ASPR/BARDA*
Alexei Slesarev	*BioReliance*
Daniel Sommer	*NBACC*
Shanmuga Sozhamannan	*Defense Biological Product Assurance Office*
Prem Sreenivasan	*Colgate Palmolive*
Becky Steffen	*NIST*
Robin Stombler	*Auburn Health Strategies*
Christina Strange	*Canon US Life Sciences*
Austin Swafford	*UCSD*
Shuiquan Tang	*Zymo Research*
Peter Thielen	*Johns Hopkins APL*
Scott Tighe	*Univ Of Vermont*
Ruth Timme	*US Food and Drug Administration*

Irina Tiper	*US Food and Drug Administration*
Thu-Thuy Tran	*US Food and Drug Administration*
Todd Treangen	*University of Maryland*
Stephen Turner	*University of Virginia*
Willy Valdivia	*Orion Integrated Biosciences*
Marc Van Eden	*Zymo Research Corp.*
Deanna Vella	*DNA Electronics*
Alamelu Venkatachalam	*Baylor College of Medicine*

Lakshmi Viswanathan	*BioReliance Corporation*
Alexandra Whale	*LGC*
Chris Whitehouse	*US Food and Drug Administration*
Ajoke Williams	*US Food and Drug Administration*
Joshua Wolfe	*JHUAPL*
Philip Wyatt	*Wyatt Technology Corp.*
David Yarmosh	*MRIGlobal*

Appendix B: Agenda

Monday, August 14, 2017

Time	Item
8:00 AM	Arrival, Registration, Poster Set-up, Exhibitor Set-up
9:00 AM	Welcome: **Scott Jackson** *NIST* Opening Remarks: **Laurie Locascio** *Director, Material Measurement Laboratory, NIST*
9:20 AM	**KEYNOTE ADDRESS: RITA R. COLWELL, Ph.D.,** *Distinguished Professor, University of Maryland College Park and Johns Hopkins University Bloomberg School of Public Health* Application of Next Generation Sequencing and Bioinformatics for Rapid and Accurate Pathogen Detection and Characterization
10:10 AM	**Jonathan Jacobs** *MRIGlobal* Culture-Independent Detection and Characterization of Infectious Agents Directly from Clinical and Environmental Sources: A Plea for Standards and Benchmarks
10:35 AM	Morning Break and Poster Viewing
11:05 AM	**Lightning Talks** (five minutes each) ● **Timothy Mercer** *Garvan Institute of Medical Research,* "Synthetic microbial communities provide internal reference standards for metagenome sequencing and analysis" ● **Marc Allard** *FDA CFSAN ORS,* "GenomeTrakr database: WGS network for foodborne pathogen traceback" ● **William Klimke** *NCBI,* "NCBI Pathogen Detection: Facilitating Traceback and Outbreak Investigation of Pathogen Genome Sequences in Real-Time Using an Automated SNP Clustering Analysis Pipeline" ● **Ruth Timme** *FDA,* "Bacterial benchmark datasets for comparison and validation of phylogenomic pipelines"
11:30 AM	**Robert Schlaberg** *University of Utah School of Medicine* Metagenomics-Based Detection of Respiratory Pathogens in Routine Practice
11:55 AM	**Vikram Munikoti** *DHS* A Standards-Driven Approach to the Deployment of BioWatch Detection Technologies
12:20 PM	**Alexa McIntyre** *Cornell University* Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers: Short and Long Reads
12:45 PM	Lunch (available for purchase at the NIST cafeteria)

1:50 PM	**Lightning Talks** (five minutes each) ● **Nadim Ajami** *Baylor College of Medicine*, "VirMAP, a viral metagenomics analysis platform for virus taxonomic classification" ● **James Pettengill** *FDA, CFSAN, OAO*, "Segal's Law, 16S rRNA gene sequencing, and the perils of foodborne pathogen detection within the American Gut Project" ● **Peter Thielen** *JHU Applied Physics Laboratory*, "Optimization of High Throughput Total RNA Sequencing for Acute Encephalitis Diagnostics"
2:10 PM	**Jason Kralj** *NIST* The Mixed Microbial DNA Reference Material for Pathogen Detection
2:35 PM	**Nancy Lin** *NIST* Yeast Cells as a Candidate Reference Material to Support Training in On-Site Biological Agent Sampling and Detection
3:00 PM	**Scott Tighe** *ABRF Metagenomics Research Group* Technical Strategies for Oxford Nanopore Sequencing and Rapid Pathogen Detection and Bio-surveillance
3:25 PM	Afternoon Break and Poster Viewing
3:55 PM	**Scott Jackson** *NIST* Initial Results from The Metagenomic MVP Challenge: An Interlab Study Designed to Assess Bias Associated with Library Preparation Methods and NGS Platforms
4:20 PM	Panel Discussion
5:20 PM	Closing Remarks
5:30 PM	Adjourn

Tuesday, August 15, 2017

Time	Item
8:00 AM	Arrival/Registration
8:30 AM	Opening Remarks
8:40 AM	**Charles Chiu** *UCSF School of Medicine* Control Materials for Metagenomic Next-Generation Sequencing Assays
9:05 AM	**Heike Sichtig** *US Food and Drug Administration* FDA's Role in Building the ID NGS Diagnostic Toolkit
9:30 AM	**Shanmuga Sozhamannan** *DBPAO* Next Gen Sequencing Based Biodefense Assays: The Need for Standardized Alternate Reference Materials
9:55 AM	Instructions for Break-out Sessions
10:10 AM	Morning Break and Poster Viewing
10:40 AM	Break-out Session 1
11:30 AM	Move to next session
11:40 AM	Break-out Session 2
12:30 PM	Lunch (available for purchase at the NIST cafeteria)
1:30 PM	Reports from Break-out Sessions
3:15 PM	Closing Remarks
3:30 PM	Adjourn

Appendix C: Abstracts from Invited Talks

Keynote Address

Rita R. Colwell, Ph.D., Distinguished Professor, *University of Maryland College Park and Johns Hopkins University Bloomberg School of Public Health*

Application of Next Generation Sequencing and Bioinformatics for Rapid and Accurate Pathogen Detection and Characterization

Next generation sequencing (NGS) combined with high-resolution bioinformatics, offers a powerful method for detection, identification, and characterization of pathogenic microorganisms (bacteria, viruses, fungi, and parasites). This approach to diagnosis of infectious disease agents and infectious diseases offers accuracy, speed, and actionable information, the sequencing within a day or two and the bioinformatics analysis within minutes. We have applied this method in clinical studies, including retrospective case control studies comprising samples of known and unknown etiology, as well as samples from healthy individuals. The results are exciting and demonstrate microbiome analysis can be used to differentiate healthy, diseased, and asymptomatic carriers, including individuals in early stages of infection and disease.

Invited Speaker Abstracts

(in presentation order)

Jonathan Jacobs *MRIGlobal*

Culture-Independent Detection and Characterization of Infectious Agents Directly from Clinical and Environmental Sources: A Plea for Standards and Benchmarks

K Parker(1), JA Russell(1), B Campos(1), J Stone(1), J Bagnoli(1), D Yarmosh(1), M Torres(3), T Slezak(3), P Li(2), KW Davenport(2), PS Chain(2), R Winegar(1), JR Aspinwall(1), JL Jacobs(1)
(1) MRIGlobal, (2) Los Alamos National Labs, (3) Lawrence Livermore National Labs

Recent advances in sample preparation methods, next generation sequencing technologies, and advances in bioinformatics have collectively paved the way for a future where direct, culture-independent detection and characterization of pathogens is a common laboratory capability. Today, however, significant pitfalls and hurdles along the road to that destination. In this work, we present an integrated solution for universal pathogen detection and characterization that addresses each of the critical steps needed to fulfill the promise of culture-independent testing for infectious diseases. We have developed methods for sample prep, sequencing, bioinformatics analysis, and reporting for pathogens directly from blood and soil, using commercial-off-the-shelf (COTS) kits and technologies and free, easy to use software solutions. For human clinical blood samples, we tested and evaluated over 43 COTS kits for sample prep alone. In addition, we have developed a bioinformatics pipeline and an associated web-based graphical user interface (GUI), PanGIA, hand-in-hand with our laboratory development efforts and designed to run on low-cost commodity hardware commonly found in even modestly equip clinical laboratories. Herein, we present our end-to-end method for detection and characterization of Gram (-) and Gram (+) bacteria as well as RNA and DNA viruses, from human blood and environmental forensic swab samples. Our method relies on an optimized sample prep method that combines DNA/RNA isolation, incremental enzymatic based de-hosting, and a combined library prep culminating with sequencing on an Illumina MiSeq. The complete

sample to answer turnaround time is under 24 hours for up to 6 samples, and is capable of unbiased detection of pathogens as low as 1E3 cfu/pfu per ml titer levels from human blood, with similar performance for environmental swab samples. During the course of our development efforts, we have identified numerous sources of bias that can interfere with downstream interpretation of results, including differences in sample handling procedures, laboratory containment and (de)contamination protocols, choice of experimental controls, frequent reagent and kit contaminants, specificity issues from various bioinformatics analysis pipelines, and common challenges with database consistency and inclusivity.

Robert Schlaberg *University of Utah School of Medicine*
Metagenomics-Based Detection of Respiratory Pathogens in Routine Practice

Hypothesis-free pathogen detection by next-generation sequencing-based metagenomics can improve diagnostic yield compared to culture and PCR-based tests, especially in patients with complex healthcare needs. As these methods are being introduced in diagnostic laboratories, it is critical to develop quality control protocols and standards to ensure that complex laboratory and data analysis workflows perform as expected.

NGS-based tests have been in diagnostic use in other fields, including genetics and oncology, for several years. While lessons can be learned from these applications, many challenges pertaining to specimen processing, data analysis, and reference sequence databases are unique to the field of microbiology.

We have developed a metagenomics-based test for detection of respiratory pathogens using both, DNA and RNA-seq. This test makes extensive use of internal and external controls and was validated using both and virtual patient samples. Current challenges and examples of solutions will be discussed.

Vikram Munikoti *DHS, OHA, BioWatch*
A Standards-Driven Approach to the Deployment of BioWatch Detection Technologies

BioWatch is the Nation’s only biodetection capability that provides early warning in the event of an aerosolized biological attack. It consists of a nationwide network of air sampling units operating continuously, from which samples are extracted and analyzed for select agents using detection assays. The Program owes its success to an extraordinary network of federal, state and local stakeholders who rely on the data generated by the Program to make decisions pertaining to preparedness and response. To ensure the accuracy and defensibility of its data, the Program adheres to rigorous, established standards in qualifying the performance metrics of its assays and reagents. BioWatch recently tested the Luminex Magpix Multiplex PCR platform against SPADA agent panels and associated Standard Method Performance Requirements (SMPRs) in an effort to achieve higher throughput and better efficiency in daily laboratory operations without compromising the sensitivity and specificity of its detection capability. Further, it is exploring the potential offered by amplicon and metagenomic sequencing technologies for operational use, and as tools for characterizing jurisdictional environments. The degree of granularity provided by Next Generation Sequencing (NGS) assays could help the Program detect intentional releases of threat agents and identify deliberate genetic manipulations to agents, accurately distinguish threat agents from near neighbors, identify antibiotic resistance genes, and overall, yield more robust genomic information to better inform response decisions. Hence, there is interest within the Program in collaborating with other federal, state and local entities to define robust standards for NGS of pathogens in environmental samples.

Alexa McIntyre *Cornell University*

Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers: Short and Long Reads

One of the main challenges in metagenomics is the identification of microorganisms in clinical and environmental samples. While an extensive and heterogeneous set of computational tools is available to classify microorganisms using whole genome shotgun sequencing data, comprehensive comparisons of these methods are limited. In this study, we use laboratory-generated and simulated controls covering 846 species to evaluate the performance of eleven metagenomics classifiers. We also assess the effects of filtering and combining tools to reduce the number of false positives. Tools were characterized on the basis of their ability to (1) identify taxa at the genus, species, and strain levels, (2) quantify relative abundance measures of taxa, and (3) classify individual reads to the species level. Strikingly, the number of species identified by the eleven tools can differ by over three orders of magnitude on the same datasets. However, various strategies can ameliorate taxonomic misclassification, including abundance filtering, ensemble approaches, and tool intersection. Nevertheless, these strategies were often insufficient to completely eliminate false positives from environmental samples, which are especially important where they concern medically relevant species. We show that experimental design and analysis parameters, including depth of sequencing, choice of classifier or classifiers, database size, and filtering, can reduce false positives, provide greater resolution of species in complex metagenomic samples, and improve the interpretation of results.

Jason Kralj *NIST*

The Mixed Microbial DNA Reference Material for Pathogen Detection

Jason Kralj, Sam Forry, Nathan Olson, and Scott Jackson

Next-generation sequencing-based metagenomics has enabled the characterization of complex microbial samples and mixtures. However, these analyses rely upon a combination of sample pre-processing steps, NGS sequencing, and computational tools that bias the results. We utilized the feedback from the 2015 NIST/FDA SPIN workshop to select a cohort of 25 bacteria whose DNA make up the basis for our RM.

The DNA for each isolate will have a known concentration, and with an assembled genome will allow us to infer the genomic concentration. This will enable end users to mix the DNA in defined ratios and determine the sources and magnitudes of process biases, and move metagenomics towards quantitative results with intercomparability. To date, we have acquired and performed preliminary characterization of half of the components making up the material. These include near-neighbor organisms, high and low G+C content, genomic repeats, and antimicrobial resistance genes. We present our characterization process and preliminary results. Additionally, we performed sequencing and analyses of a Latin square design of DNA mixtures for examining limits of detection, computational tool performance, and biases.

Nancy J. Lin *NIST*

Yeast Cells as a Candidate Reference Material to Support Training in On-Site Biological Agent Sampling and Detection

Nancy J. Lin, Sandra M. Da Silva, James J. Filliben

Routine training and proficiency testing in on-site collection and assessment of biological materials typically use near neighbor organisms or attenuated/inactivated biothreats to simulate biothreat agents. These materials can have real and perceived health and safety risks, result in false positives

during true events, require specialized training facilities, and have limited availability. We have developed a surrogate material based on modified Baker's yeast (*Saccharomyces cerevisiae* NE095) to challenge the entire biodetection process, including nucleic-acid detection technologies, while minimizing concerns. Vial-to-vial homogeneity and real-time and accelerated stability in terms of total and viable cells demonstrated the suitability of lyophilized yeast cells as a candidate reference material. Quantitative polymerase chain reaction (qPCR) confirmed the stability of the inserted nucleic acid sequence. Fitness for purpose was demonstrated via interlaboratory studies and a full-scale functional exercise. Overall, our results support lyophilized yeast as a promising material for safe and effective quantitative workflow evaluation and field training to increase confidence in the response to potential biological threat incidents.

Scott Tighe *ABRF Metagenomics Research Group*
Technical Strategies for Oxford Nanopore Sequencing and Rapid Pathogen Detection and Bio-surveillance

Since the advent and commercialization of nanopore sequencing, the ability to rapidly characterize full bacteria genomes has dramatically increased. However, as with every new disruptive technology, the need for innovative supporting reagents and protocols are needed. While the Oxford Nanopore is a wonderful new tool for all areas of biology, including pathogen detection and bio-surveillance, the need for rapid DNA extraction that generates high molecule weight DNA as well QC protocol is a must. Here we will present several new protocols (include metapolyzyme) to accelerate DNA extraction and enable rapid detection and identification of bacterial populations from mixed populations and sample types.

Scott Jackson *NIST*
Initial Results from The Metagenomic MVP Challenge: An Interlab Study Designed to Assess Bias Associated with Shotgun Library Preparation Methods and NGS Platforms
Scott Jackson[1], Nur Hasan[2], Kelly Moffat[2], Manoj Dadlani[2], Scott Tighe[3]
[1]NIST, [2]CosmosID, [3]University of Vermont and ABRF

Here we discuss the preliminary findings of an interlab study (Metagenomic MVP Challenge v1.0) that was recently carried-out by NIST, CosmosID and the ABRF-MGRG.

When using metagenomic methods to assess the content of a complex microbial sample, there are many steps/factors in the measurement process that may introduce bias. These include, but are not limited to i) sample collection and storage methods, ii) DNA/RNA extraction technique, iii) library construction and/or PCR amplification strategy, iv) NGS instrument/chemistry, vii) depth of sequencing and read length used, viii) raw data filtering and QC methods and ix) data analysis/interpretation. We intend to systematically investigate each of these steps/factors individually to understand how they each contribute towards the overall bias of the entire measurement process.

The Metagenomic MVP Challenge v1.0 is intended to be the first of many challenges that will each address a different step of the measurement process. The goal of Challenge v1.0 is to specifically identify and understand measurement bias associated with 1) shotgun library preparation techniques and 2) next-generation sequencing platforms. We recognize that this only addresses a small aspect of the larger measurement process. This restriction in scope is intentional and it is our intent to hold future challenges to address other steps of the measurement process. Moreover, the goal of this challenge was NOT to grade individuals or labs on their "competency", but to identify bias in the measurement process.

A metagenomic reference material was developed that consisted of a mixture of genomic DNA from 10 different microbes with varying GC content. The mixture was designed such that each genome is

equally represented in the mixture. This material was made available to laboratories that volunteered to participate in this interlab challenge. We asked each participating laboratory to 1) build a shotgun library 2) sequence the material on their NGS instrument, 3) return their raw sequence data (fastq) to us and 4) return metadata describing their entire sequencing technique. The only constraint we imposed was that the library had to be whole genome shotgun, not 16S amplicon, in order to make the data comparable. Otherwise, we encouraged participants to sequence the material using any and all available library prep techniques and sequencing platforms. It was our hope that this crowd-sourcing effort would return a large array of sequence data generated from many diverse library-prep techniques and from all current NGS instruments. An aggregated analysis of this data allows us to identify how the different library preparation techniques and sequencing platforms influence the sensitivity and specificity of the overall measurement.

We received 98 raw fastq data files from 49 samples that were generated across 16 independent laboratories. In total, over 400 gigabases of data was generated from our metagenomic DNA reference material. These data include 9 different library preparation methods and 5 different sequencing platforms. A preliminary overview of our aggregated findings will be presented at the conference.

Charles Chiu *UCSF School of Medicine*
Control Materials for Metagenomic Next-Generation Sequencing Assays

An unbiased metagenomic next-generation approach (mNGS) been shown to be useful in the broad identification of pathogens in clinical samples for infectious disease diagnosis, including viruses, bacteria, fungi, and parasites. Clinical adoption has been hampered, however, by the highly complex workflows inherent to metagenomics that can lead to analytical errors. Standardized reference control materials and databases are critically needed for quality control in the development of clinical metagenomic protocols and pipelines, although these are currently lacking. This talk will discuss CLIA laboratory validation of a clinical mNGS assay for identification of pathogens in cerebrospinal fluid (CSF), with a focus on the control materials that were chosen and used for the assay. These materials have since been adapted for validation and use with real-time metagenomic analysis pipelines, such as on the MinION™ nanopore sequencer (Oxford Nanopore Technologies). We will also discuss a 1-year nationwide, multi-site clinical study ("Precision Diagnosis of Acute Infectious Diseases", June 2016 - 2017) to evaluate the clinical utility and cost-effectiveness of mNGS versus conventional microbiological testing for diagnosis of infectious causes of meningitis and encephalitis from CSF, and ongoing efforts to expand into fever / sepsis (plasma) and pneumonia (bronchoalveolar lavage fluid).

Heike Sichtig *US Food and Drug Administration*
FDA's Role in Building the ID NGS Diagnostic Toolkit

This presentation will specifically focus on FDA's role in building microbial reference materials and genomes to support infectious disease NGS diagnostic development. FDA and collaborators established a publicly available dAtabase for Reference Grade micrObial Sequences called FDA-ARGOS. With funding support from FDA's Office of Counterterrorism and Emerging Threats (OCET) and DoD, the FDA-ARGOS team are initially collecting and sequencing 2000 microbes that include biothreat microorganisms, common clinical pathogens and closely related species. Manufacturers who develop sequence-based test to identify infectious agents and/or to detect resistance or virulence markers can use FDA-ARGOS to advance their development programs and to support the regulatory science review of such test. For more info, visit the FDA-ARGOS reference genome database project website:

https://www.fda.gov/MedicalDevices/ScienceandResearch/DatabaseforReferenceGradeMicrobialSequences/default.htm.

Shanmuga Sozhamannan *Technical Coordinator, JPEO-JPM-Guardian-DBPAO*
Next Gen Sequencing Based Biodefense Assays: The Need for Standardized Alternate Reference Materials

Nucleic acid based assays, such as real time polymerase chain reaction, are the mainstay of clinical diagnostics and bio surveillance. The dawn of Next Gen sequencing technologies, combined with decreasing sequencing costs, has heightened our expectation for NGS-based diagnostic/detection assays. However, even after a decade of significant improvements to the technology and decreased costs overall, translating NGS technology into a diagnostic capability has not been realized. I will discuss some of the idiosyncrasies associated with assay development in general and how they specifically play out when applied to NGS technology. Additionally, I will discuss some new challenges to biodefense assay development regarding access to reference materials that has created a heightened need for robust, *in silico* methods and synthetic biology approaches. Taken together, these steps may hasten the path towards development of NGS diagnostics.

Appendix D: Abstracts from Contributed Lightning Talks

These contributed abstracts were accepted for both a poster presentation and a five-minute podium presentation.

Nadim Ajami *Baylor College of Medicine*
VirMAP, a Viral Metagenomics Analysis Platform for Virus Taxonomic Classification

Nadim J Ajami[1,2,*], Matthew C. Wong[1,2,*], Matthew C. Ross[1,2], Richard E. Lloyd[2], Joseph F. Petrosino[1,2.]
[1]Alkek Center for Metagenomics and Microbiome Research, and [2]Molecular Virology and Microbiology Department, Baylor College of Medicine, Houston, Texas.
*Authors contributed equally to this work.

Recent advances in sequencing technologies have enabled deep interrogation of metagenomic samples. Much attention is driven towards characterizing bacteriomes and mycobiomes and more recently viromes. Current approaches have significant drawbacks when attempting to classify wild-type viruses whose genome sequence, but not genome information, have greatly diverged from known and classified database sequences. This problem is exacerbated by the fact that viruses can also share high-level protein homology across various taxa. We present an approach capable of accurately assigning serotype-level viral taxonomies based on coherently sorting and merging nucleotide and protein information, creating pseudo-scaffolds via a tiered mapping assembly, and taxonomically classifying based on aggregate information. We validated our viral metagenomic analysis platform, VirMAP, using proprietary and public datasets and compared it to recently described pipelines. VirMAP generates accurate viral taxonomic classification without the need for targeted sequencing, high coverage, read overlap, or pair-end sequence data.

Marc Allard *FDA/ORS*
GenomeTrakr Database: WGS Network for Foodborne Pathogen Traceback

Timme Ruth FDA/ORS, Sanchez Maria FDA/ORS, Allard Marc FDA/ORS, Stevens Eric FDA/ORS, Hoffman Maria FDA/ORS, Kastanis George FDA/ORS, Lindley Sabina FDA/ORS, Muruvanda Tim FDA/ORS, Strain Errol FDA/OAO, Payne Justin FDA/OAO, Pightling Arthur FDA/OAO, Rand Hugh FDA/OAO, Pettengill James FDA/OAO, Luo Yan FDA/OAO, Gonzalez-Escalona Narjol FDA/ORS, Melka David FDA/ORS, and Brown Eric FDA/ORS.

Introduction: In 2012 a pilot project was set up using whole genome sequence data to track foodborne outbreaks. In this network, public health agencies collect and publically share WGS data in real time. This high-resolution, rapidly growing database is actively being used in outbreak investigations at state, national and international level.
Purpose: The GenomeTrakr network demonstrates how desktop WGS data can be used in concert with traditional epidemiology for source tracking of foodborne pathogens. Along with the paradigm shift in technology this new "open data" model allows greater transparency between federal/state agencies, our industry partners, academia, and international partners.
Methods: Ten new labs were added to the network in 2016 in an effort to grow and diversify the foodborne pathogen database. Two new surveillance efforts were added to collect food and environmental isolates of *Escherichia coli* and *Campylobacter*. And multiple data analysis pipelines were tested on benchmark datasets in an effort to validate our analysis methods.
Results: Our partner, NCBI, is currently producing daily cluster results for ten pathogen surveillance efforts: *Salmonella enterica, Listeria monocytogenes, E. coli*, and *Campylobacter,* Acinetobacter,

Klebsiella, Serratia, Elizabethkingia, Providencia and Morganella all of which are publically available. The hardware and software implemented in GenomeTrakr allowed us to compare and cluster genomes of 10s of thousands of taxa at a time. The high-resolution WGS data in concert with solid epidemiological evidence has drastically enhanced our ability to identify the food source of current outbreaks for *Listeria monocytogenes*, for which the CDC is also contributing clinical isolates in real time. Details will be provided for one of these outbreaks where WGS provided the lead in a 2015 Virginia sprout outbreak.
Significance: These results demonstrate two major contributions of GenomeTrakr: WGS as a high-resolution sub-typing tool and the global benefits of having an open data model. Understanding the root causes of foodborne contamination will assist our academic, public health and industry partners to develop preventative controls to make food safer globally.

James Pettengill *US FDA*
Segal's Law, 16S rRNA Gene Sequencing, and the Perils of Foodborne Pathogen Detection within the American Gut Project

James B Pettengill, Hugh Rand
Biostatistics and Bioinformatics Staff, Office of Analytics and Outreach, US Food and Drug Administration, College Park, Maryland, United States

Obtaining human population level estimates of the prevalence of foodborne pathogens is critical for understanding outbreaks and ameliorating such threats to public health. Estimates are difficult to obtain due to logistic and financial constraints, but citizen science initiatives like that of the American Gut Project (AGP) represent a potential source of information concerning enteric pathogens. With an emphasis on genera Listeria and Salmonella, we sought to document the prevalence of those two taxa within the AGP samples. The results provided by AGP suggest a surprising 14% and 2% of samples contained Salmonella and Listeria, respectively. However, a reanalysis of those AGP sequences described here indicated that results depend greatly on the algorithm for assigning taxonomy and differences persisted across both a range of parameter settings and different reference databases (i.e., Greengenes and HITdb). These results are perhaps to be expected given that AGP sequenced the V4 region of 16S rRNA gene, which may not provide good resolution at the lower taxonomic levels (e.g., species), but it was surprising how often methods differ in classifying reads – even at higher taxonomic ranks (e.g., family). This highlights the misleading conclusions that can be reached when relying on a single method that is not a gold standard; this is the essence of Segal’s Law: an individual with one watch knows what time it is but an individual with two is never sure. Our results point to the need for an appropriate molecular marker for the taxonomic resolution of interest, and calls for the development of more conservative classification methods that are fit for purpose. Thus, with 16S rRNA gene datasets, one must be cautious regarding the detection of taxonomic groups of public health interest (e.g., culture independent identification of foodborne pathogens or taxa associated with a given phenotype).

Peter Thielen *JHU Applied Physics Laboratory*
Optimization of High Throughput Total RNA Sequencing for Acute Encephalitis Diagnostics

Thomas Mehoke, JHU Applied Physics Laboratory, Thomas.mehoke@jhuapl.edu
Craig Howser, JHU Applied Physics Laboratory, craig.howser@jhuapl.edu
Briana Vecchio-Pagan, JHU Applied Physics Laboratory, briana.vecchio-pagan@jhuapl.edu
Jared Evans, JHU Applied Physics Laboratory, jared.evans@jhuapl.edu
Arun Venkatesan, Johns Hopkins Medical Institute, avenkat2@jhmi.edu

Peter Thielen, JHU Applied Physics Laboratory, peter.thielen@jhuapl.edu

Background: Metagenomic sequencing has significant potential to replace traditional clinical diagnostic assays. Two major challenges for this area of research include 1) establishing baseline microbiome composition for a specific clinical sample type across many individuals and 2) establishing detection sensitivity for pathogens in comparison to standard diagnostics methods. This is particularly relevant for clinical conditions such as meningitis and encephalitis, in which fewer than half of patients are positively diagnosed, making treatment challenging and compromising patient recovery. **Objective:** We sought to establish unbiased RNA sequencing in the Johns Hopkins Hospital system for pathogen detection in the cerebrospinal fluid (CSF) of acute encephalitis patients. Our goal was to generate methods that could be executed in under 24 hours for maximum clinical utility, to include sample processing, data acquisition, and data analysis. **Methods:** We developed robust sample preparation and data analysis workflows for sequencing of total RNA from small volumes of cerebrospinal fluid (CSF). Using these methods, we have generated datasets that include a panel of non-infected patient samples, infected patient samples of known and unknown etiology, and a dilution series of an RNA virus in normal patient CSF. **Preliminary Results:** This workflow can be executed in under 24 hours from sample receipt with limited user intervention. In this presentation we will discuss the observed benefits and limitations of total DNA or RNA sequencing in clinical settings, specific implementation considerations when a pathogen is detected, and the high sensitivity of metatranscriptomic sequencing in comparison to quantitative PCR detection methods. Additionally, we discuss the power in large scale analysis of this data using ultra-rapid sequencing read classifiers, as well as statistical testing to determine positive predictive value in metatranscriptomic sequencing datasets from new patients. **Preliminary Conclusions:** As sequencing and data analysis methods advance, the feasibility of expanding these methods into clinical environments and non-traditional treatment settings is promising. We will discuss the challenges associated with detection of unexpected human pathogens in clinical samples, in which traditional diagnostic tests are not available to make a clinically actionable diagnosis due to IRB limitations.

Ruth Timme *FDA/ORS*

Bacterial Benchmark Datasets for Comparison and Validation of Phylogenomic Pipelines

Ruth E. Timme, Hugh Rand, Martin Shumway, Eija K. Trees, Mustafa Simmons, Richa Agarwala, Steve Davis, Glenn Tillman, Stephanie Defibaugh-Chávez, Heather A. Carleton, William A. Klimke, Lee S. Katz

As next generation sequence technology has advanced, there have been parallel advances in genome-scale analysis programs for determining evolutionary relationships as proxies for epidemiological relationship in public health. Most new programs skip traditional steps of ortholog-determination and multi-gene alignment, instead identifying variants across a set of genomes, then summarizing results in a matrix of single nucleotide polymorphisms or alleles for standard phylogenetic analysis. However, public health authorities need to document the performance of these methods with appropriate and comprehensive datasets so they can be validated for specific purposes, e.g., outbreak surveillance. Developing such standards is the task of the Genomics and Food Safety group (Gen-FS), a collaboration among the FDA, NCBI, FSIS, and CDC. As members of the Gen-FS WGS Standards working group we present a set of benchmark datasets to be used for comparison and validation of phylogenomic pipelines.
We identified four well-documented foodborne pathogen events in which the epidemiology was concordant with standard WGS phylogenetic analysis. These are ideal benchmark datasets, as the trees, WGS data, and epidemiological data for each are all in agreement. We have placed the sequence files, sample metadata, and “known” phylogenetic trees in publicly-accessible databases and

developed a standard descriptive spreadsheet format describing each dataset. Our "outbreak" benchmark datasets represent the four major foodborne bacterial pathogens (*Listeria monocytogenes*, *Salmonella enterica*, *Escherichia coli*, and *Campylobacter jejuni*) and one simulated dataset where the "known tree" can be accurately called the "true tree". The "Gen-FS Gopher" downloading script, and associated table files are available on GitHub: https://github.com/WGS-standards-and-analysis/datasets.
These five benchmark datasets and validated SNP set will help standardize comparison of current and future phylogenomic pipelines, and facilitate important cross-institutional collaborations. We welcome additional benchmark datasets in our recommended format, and will publish these on our GitHub site. Together, these datasets, dataset format, and the underlying GitHub infrastructure present a recommended path for worldwide standardization of phylogenomic pipelines.

William Klimke *NCBI*
NCBI Pathogen Detection: Facilitating Traceback and Outbreak Investigation of Pathogen Genome Sequences in Real-Time Using an Automated SNP Clustering Analysis Pipeline

William Klimke, Mike Feldgarden, Arjun Prasad, Martin Shumway, Richa Agarwala, Mike DiCuccio, Lewis Geer, Avi Kimchi, Tatiana Tatusova, Jim Ostell, David Lipman

The NCBI Pathogen Detection pipeline was created in 2013 to facilitate the analysis of genome sequences for foodborne bacterial pathogens. Federal agencies responsible for food safety initiated two projects to utilize next generation sequencing: the FDA GenomeTrakr and the FDA/CDC real-time Listeria project. With the USDA joining shortly thereafter all federal agencies became involved. Starting in 2013, all Listeria collected from clinical patients and food and environmental sources in the US were sequenced in real-time through a network of state public health, federal field labs, and the national agencies and the data uploaded to the sequence read archive at NCBI. The three other foodborne pathogens, Campylobacter, Salmonella, and Shiga toxin-expressing E. coli (STECs) have started to be sequenced during that same time period with the eventual goal of having all isolates in the US sequenced by the end of 2018 (approx. 90 000 isolates per year). The NCBI Pathogen Detection pipeline assemblies the incoming raw sequence data in SRA and clusters the assembled genomes together with those in GenBank using both k-mers as well as SNPs. Single linkage clustering of the SNP distances is used to generate tight clusters in order to facilitate outbreak and trace-back investigations. The specificity of using whole genome sequencing along with epidemiological data was shown by CDC to result in an increase in the number of clinical cases associated with food sources, a decrease in the case number per cluster, and with more outbreaks were resolved for Listeria during the first pilot year of the project. The analysis results from the NCBI pipeline are made publicly available both in a web interface as well as on FTP which provides open access to the organisms impacting food safety in real time and has aided public health and food safety labs and organizations in doing real time comparisons of pathogen sequences to speed up outbreak and traceback investigations and improve public safety.

Tim Mercer *Garvan Institute of Medical Research*
Synthetic Microbial Communities Provide Internal Reference Standards for Metagenome Sequencing and Analysis

Simon Hardwick[1], Bindu Kanakamedala[1], Ted Wong[1], Chris Barker[1], Tim R. Mercer[1,2]
[1]*Garvan Institute of Medical Research, Sydney, Australia*
[2]*Altius Institute of Medical Research*

Metagenomics can reveal the size and diversity of environmental microbial communities and identify novel, previously uncultured microbes. However, the complexity and novelty of microbial communities, combined with technical biases in next-generation sequencing can confound accurate metagenomic analysis. We have designed a synthetic set of DNA standards, termed *sequins*, that represent the diversity of natural microbial communities. Each DNA standard 'mirrors' a representative microbe genome, thereby retaining the same nucleotide content, sequence architecture and analytical performance of the original microbe genome. Sequins are then titrated to form a ladder by which to measure quantitative features of metagenome analysis. Here, we validate the use of sequins as internal reference controls to assess library preparation, sequencing and metagenome assembly and analysis by comparison to known reference and natural samples. We also demonstrate the use of sequins to assess the diagnostic performance of individual sequenced libraries, and for the absolute and relative normalization and comparison of multiple samples. Together we provide metagenome sequins, as well as accompanying resources and software toolkit, as a reference standard to aid in metagenome studies by the research community.

Appendix E: Abstracts from Contributed Posters

Large Scale Genomic DNA Isolation from Human Pathogenic Bacteria

Lucy W. Burns, Luke Burnham and Bruce C. Hemming, Microbe Inotech Laboratories, Inc., 11754 Westline Industrial Dr., St. Louis, MO 63146-3402

There is a need and an increasing demand for DNA standards to validate microbiological research particularly in studies comparing microbiome analyses. DNA standards are obtained from homogeneous sources processed with large scale DNA extraction protocols. These protocols have been developed to isolate large amounts of genomic DNA (10 mg) from different species and strains (>20) of human pathogenic bacteria. In addition, the method development and production of standards under contract from NIST has focused on a number of parameters to ensure DNA quality and maintain fragment length integrity. Data gathered during the production of such genomic DNA standards is presented.

AXSIM: A Benchmarking Software for Metagenomic Data Analysis

Corey Chen[1], Crystal Icenhour[2], Yuan Chen[2*]
[1]Department of Computer Science, University of Virginia, Charlottesville, VA, [2]Aperiomics, Inc, Ashburn, VA.

Background: In metagenome study, taxonomic profiling is the key component for biological data interpretation. Recently there have been an increasing number of ultrafast and accurate programs that can map taxonomic labels to metagenomic DNA sequences. How to estimate of performance of different software and how to choose the right one based on the users' requirements becomes a critical problem. Thus, it is necessary to develop a benchmarking tool that users can use to easily generate customized in-silicon reads and to estimate the sensitivity and specificity of the final result.
Results: We introduce AXSIM; a new approach to developing a standard for metagenomic reads classifying software and a tool that simulates its own reads. AXSIM exhibits an ergonomic GUI interface and can be installed through a cloud web service and local machine. The software provides all necessary functions to generate and compare reads including: input validation, software exception/error handling, direct downloading of genome sequences from NCBI, phylogeny tree extraction of all sample genomes, metadata analysis, and summarization and depictions of results. Also, users can now take output from different software and determine the specificity and sensitivity of the metagenomic reads. Currently, this software is able to benchmark the following reads simulation software: CLARK, Kraken, Pathoscope, and itself. We have extensively tested and validated the consistency of AXSIM using different groups of identical parameters across these simulation platforms.
Conclusions: AXSIM is a versatile, efficient, and user-friendly open source software that allows users to benchmark and decide which commercial simulation software is the most reliable in the industry.

Exploiting the Cytoelectric Properties of Microbes by Dielectrophoresis to Isolate, Detect, and Characterize Microorganisms in Complex Specimens

Lorenzo D'Amico, PhD[1,2], Peter RC Gascoyne, PhD[2], Joseph F. Petrosino, PhD[1]
[1]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA
[2]Advanced Electrofluidic Systems, LLC, Houston, TX, USA

The prevailing paradigm in microbiology and microbiome science attempts to monitor microbial community dynamics by analyzing massive quantities of biomolecular data *in silico*. This requires that the entire ensemble of cells within a sample be disrupted to release biomolecules that can be analyzed without explicitly observing any viable organisms. These methods require highly trained technicians to work for days in sophisticated laboratories, and delaying intervention (antibiotic treatment or product recalls) and imposing significant social and economic costs. Furthermore, disrupting cells to extract nucleic acids and metabolites confounds efforts to unravel *in vivo* microbe-microbe and microbe-host interactions by preventing the study of viable consortia *ex vivo* and by mixing molecular information with that of the host. These shortcomings represent significant barriers to advancing our understanding of the interactions among pathogens and indigenous microbiota within various niche habitats. The absence of improved experimental techniques has hindered our ability to extract from large descriptive datasets a reliable biomathematical framework that predict the impact pathogen exposure and microbiome dysbiosis have on human and environmental health.

As part of a broader solution to improve and augment existing microbiological analytical tools, we are creating technologies that harness microscale phenomena to 1) isolate microbial communities from clinical and environmental samples; 2) characterize and sort natural and synthetic communities based on biophysical properties; 3) sustain and expand microbiomes *ex vivo*; and 4) isolate etiologic agents of infectious disease from clinical samples for rapid diagnosis. These miniaturized technologies are both preparative and analytical, capable of directly measuring biologically-relevant properties and exploiting subtle differences in these properties to isolate and sort viable subpopulations of microorganisms prior to molecular analysis.

At the core of these technologies is a programmable microfluidic platform that uses dielectrophoresis, gravity, and hydrodynamic forces to manipulate microbial cells within a disposal cartridge. Dielectrophoresis (DEP) is the motion of a particle (or cell) suspended in a fluid and exposed to a time-varying and spatially nonuniform electric field. By observing DEP behavior one gains a direct measure of the intrinsic cytoelectric properties of biological cells, which emerge from biologically-relevant features such as cell size and morphology, the composition and structural form of biomembranes, and the composition of the cell interior. Microfluidic technologies using DEP and other physical forces have been successfully applied in our lab to physically separate bacterial spores from vegetative cells, and to isolate bacteria from blood, urine and stool slurries. In addition, different bacterial taxa could be separated from one another using this prototype platform. Indeed, the cytoelectric properties of microbes add to the list of existing biomarkers that can be exploited for analytical or preparative purposes. Like other chromatographic techniques, different microbial cells elute from the microfluidic cartridge at different time points, and could be analyzed downstream using cultivation, automated microscopic observation (for real-time detection) and DNA sequencing techniques.

It is anticipated that these fundamentally new capabilities in sample preparation and sample processing will augment existing laboratory techniques by allowing high-throughput and sensitive bioanalytical tools be applied more elegantly to dissect complex biological systems like the microbiome. Furthermore, it may be possible to bring to market advanced analytical systems that can be deployed outside of centralized laboratories to provide actionable results earlier than current methods, thereby reducing the economic burden caused by batch contaminants or infections in humans and livestock.

Cell-Based Reference Material for qPCR: Stability Study

Sandra M. Da Silva, Nancy J. Lin, James J. Filliben
Material Measurement Laboratory, NIST

Microbial quantification and detection face a series of practical and technological challenges including those related to the nature of the samples involved in the analysis (e.g., matrix from clinical

and environmental samples). Despite efforts to improve methods, confidence in the measurements is lacking, especially for measurements made at the point of need or point of care where results are used to inform critical decision-making (e.g. biothreat detection). To address the need for measurement confidence, we are developing a reference material based on whole cells as a low-risk surrogate to challenge nucleic acid-based detection technology workflows, including sampling, DNA extraction, and detection. We stably inserted DNA sequence External RNA Control Consortium-00095 (ERCC-00095 from NIST SRM 2374) into a Saccharomyces cerevisiae strain to convey specificity. Feasibility as a reference material for quantitative polymerase chain reaction (qPCR) was demonstrated previously via interlaboratory study. Currently, we report the on-going stability study of a dry-format of the material by measuring cell number, cell viability, and DNA integrity as a function of time (up to 4 months) and temperature in Celsius (-20, 4, 20, 50). These conditions represent deviations that might occur during shipping or storage and help establish shelf-life. Preliminary results suggest acceptable cell number stability, with no statistically significant change in cell number at any temperature over time. In contrast, ~90 % loss in cell viability was observed at 50 °C after just 30 days, while other temperatures had no viability change, indicating 50 °C should be avoided to maintain viability. DNA integrity is currently being assessed by qPCR and pulsed field gel electrophoresis. Overall, this engineered yeast holds promise to support measurement assurance for the analytical process of nucleic acid-based detection technologies, encompassing the method, equipment, and operator, to increase confidence in microbial detection results.

Development of NGS Sequencing Standards to Aerosol Filter Samples

Shannon L Johnson, Grace Vuyisich, Emily Alipio-Lyon, Attelia Hollander, Cheryl D Gleasner, Kimberly McMurry, Norman Doggett, Alina Deshpande
LANL

Development and Evaluation of Whole Cell- and Genomic DNA-Based Microbiome Reference Standards

Juan Lopera, Ph.D., Monique Hunter M.S., Megan Amselle, M.T., Brian Chase, M.S., Stephen King, M.S., Maria Mayda, Ph.D, Kevin Zinn, B.S. and Dev Mittar, Ph.D.
ATCC, Manassas, VA, USA

Advancement and accessibility of next-generation sequencing technologies have influenced microbiome analyses in tremendous ways, opening up applications in the areas of clinical, diagnostic, therapeutic, industrial, and environmental research. However, due to the complexity of 16S rRNA and metagenomic sequencing analysis, significant challenges can be posed by biases introduced during sample preparation, DNA extraction, PCR amplification, library preparation, sequencing, or data interpretation. One of the primary challenges in assay standardization is the limited availability of reference materials. To address these biases and provide a measure of standardization within microbiome research and applications, ATCC has developed a set of mock microbial communities comprising fully sequenced, characterized strains selected on the basis of phenotypic and genotypic attributes, such as cell wall type (Gram stain classification), GC content, genome size, unique cell wall characteristics, and spore formation. These mock communities mimic mixed metagenomics samples and offer a universal control for microbiome analyses and assay development. Moreover, these standards have been developed with different levels of mock community complexity (10 or 20 strains per community) with even or staggered relative abundance, including diverse strains that are relevant to a broad range of applications. In addition, to minimize the bias associated with data interpretation, we have developed a data analysis module in collaboration with One Codex. This module provides a user-friendly output in the form of true-positive, relative abundance, and false-negative scores for 16S rRNA community profiling and shotgun metagenomic sequencing.

A Novel Way to Control PCR Chimera in the Library Preparation of 16S Targeted Sequencing

Mikayla Mager, Shuiquan Tang and Larry Jia
Zymo Research Corporation, Irvine, California, US

16S rRNA gene targeted sequencing is a popular technique for microbial composition profiling in microbiomics because of its simplicity and robustness. However, this technique suffers from the formation of PCR chimeric sequences, which stem from the recombination of different PCR templates. In a PCR amplicon with 10 ng of microbial DNA, running the PCR reaction for 30 cycles can cause the PCR chimeric sequences to occupy ~35% of total PCR products. The most effective way to control chimera is to limit PCR cycles. Controlling PCR cycles for a large number of samples with varied quantities of templates is difficult. This is why many popular protocols choose to use a high number of PCR cycles for all samples. E. g. Human Microbiome Project used 30 cycles and Earth Microbiome Project recommended 35 cycles. We developed a different strategy to solve this problem: we perform PCR reactions on real time PCR systems rather than regular PCR machine, which allows us to monitor the progress of PCR reactions all the time. The strategy consists of three steps: (1) controlling initial input of PCR template, (2) run all reactions for 18 cycles and withdraw samples that pass a fluorescence threshold, (3) for remaining samples, continue PCR reactions for 5 cycles and withdraw samples that pass the fluorescence threshold, so on and so forth. Using this strategy, we are able to keep the percentage of PCR chimeric sequences below 2% for all types of samples. Real time PCR also allows direct library quantification in the end, facilitating the subsequent step of library normalization.

Challenging a Bioinformatic Tool's Ability to Detect Microbial Contaminants Using *in silico* Whole Genome Sequencing Data

Nathan D. Olson, Justin M. Zook, Jayne B. Morrow, Nancy J. Lin
Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD

Microbial materials (genomic DNA and cultures) free of contaminants are needed to validate pathogen detection assay. For example, when validating a pathogen detection assay using near-neighbor negative control strains contaminated by pathogens can result in false positives. Current methods contaminant detection in microbial materials are not sensitive enough for metagenomic pathogen detection assays. Whole genome sequencing (WGS) is a promising approach for microbial contaminant detection due to its sensitivity. Further, there is no need for *a priori* assumptions about the contaminant. Before WGS can be used we must first understand the method's limitations for detecting contaminants and potential for false positives. We demonstrate and characterize a WGS-based approach for detecting organismal contaminants using an existing metagenomic taxonomic classification algorithm. Simulated WGS datasets from ten genera as individuals and binary mixtures of eight organisms at varying ratios were analyzed to evaluate the role of contaminant concentration and taxonomy on detection. For the individual genomes, the false positive contaminants reported depended on the genus with *Staphylococcus*, *Escherichia*, and *Shigella* having the highest proportion of false positives. For nearly all binary mixtures the contaminant was detected in the *in silico* datasets at the equivalent of 1 in 1,000 cells. Though *F. tularensis* was not detected in any of the simulated contaminant mixtures and *Y. pestis* was only detected at the equivalent of 1 in 10 cells. Once a WGS method for detecting contaminants is characterized, it can be applied to evaluate microbial material purity, to ensure that contaminants in microbial materials used to validate pathogen detection assays generate genome assemblies for database submission, and benchmark sequencing methods.

Olson ND, Zook JM, Morrow JB, Lin NJ. (2017) Challenging a bioinformatic tool's ability to detect microbial contaminants using *in silico* whole genome sequencing data. *PeerJ*5:e3729 https://doi.org/10.7717/peerj.3729

EDGE Bioinformatics: Updates and Extensions for Democratizing HTS Analysis

Casandra Philipson[1,2], Logan Voegtly[1,3], Chien-Chi Lo[4], Po-E Li[4], Regina Cer[1,3], Kimberly A. Bishop-Lilly[1], Pavel Senin[4], Yan Xu[4], Karen Davenport[4], Theron Hamilton[1], Patrick Chain[4]
[1]Genomics and Bioinformatics Department, Biological Defense Research Directorate, Naval Medical Research Center-Frederick; [2]Defense Threat Reduction Agency, Fort Belvoir, VA; [3]Leidos, Reston VA; [4]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM

Background: Identifying infectious agents with rapid certainty is indispensable for effective public and force health protection. High-throughput sequencing (HTS) offers unparalleled resolution for blind pathogen detection and characterization. The majority of HTS workflows require users to be proficient in programming and command-line interfaces, limiting the ease of field-forward sequencing. To address this, the Naval Medical Research Center-Frederick (NMRC) and Los Alamos National Laboratory (LANL) developed EDGE bioinformatics (Empowering the Development of Genomics Expertise). EDGE v1.0, released January 2017, is an intuitive web-based bioinformatics platform designed for scientists with little to no bioinformatics experience to democratize the analysis of microbial and metagenomic HTS data.
Methods/Results: The EDGE bioinformatics suite combines vetted publicly available tools, and tracks settings to ensure reliable and reproducible analysis workflows. Since the initial release, we developed a step-by-step video tutorial series that walks users through each module, accessible at http://tutorial.getedge.org. The EDGE workflow begins with raw sequencing reads. Users upload in-house data, or run analyses on samples deposited in SRA. As default, EDGE performs quality control, assembly, annotation, and taxonomy classification. Additional modules are available to execute host removal, reference-based analysis, phylogenetic analysis, and PCR primer analysis. Default settings offer a robust first-glance and are often sufficient for novice users. All results are compiled and available for download in a PDF-formatted report. We caution that results still require in-depth scientific understanding for confidence in interpretation, however report visuals are often informative to even to novice users. The active development version, EDGE v1.5, incorporates three new features: specialty genes profiling module, 16S/18S/fungal ITS analysis using QIIME, and a comparative batch analysis feature for side-by-side view of samples. The specialty genes module implements the ShortBRED pipeline to search ARDB and Resfams for antibiotic resistance genes, and VFDB for virulence genes. Specialty genes results are visualized in tabular form as well as Krona plots.
Conclusions: NMRC and LANL co-developed EDGE bioinformatics to offer scientists with little to no bioinformatics expertise a point-and-click platform for analyzing HTS data in a rapid and reproducible manner. Ongoing development has expanded EDGE (v1.5) to include new modules that enable detection of antimicrobial resistance and virulence, batch comparisons, and the QIIME pipeline. Future versions may include analysis of amplicon data, identification of pro-phages and plasmids, as well as differential gene expression for RNAseq datasets. An outward facing demo version is available for testing at http://hobo-nickel.getedge.org.

NGS-Based Phylogenomic Analysis of *Mycobacterium* Type Strains Supports Reclassification of Many Species and Subspecies

M. A. Riojas[1], K. McGough[1,2], M. H. Hazbón[1]
[1]BEI Resources/ATCC, Manassas, VA, USA [2] George Mason University, Fairfax, VA, USA

Background: The genus *Mycobacterium* currently consists of 177 defined specific and subspecific taxa, many of which were defined decades ago using morphological observations and biochemical assays that are often misleading. However, the more recent advent of next-generation sequencing (NGS) and powerful bioinformatics techniques allow the reexamination of the taxonomy and phylogeny of these taxa using modern methods. These techniques have the potential to elucidate the true structure of the genus *Mycobacterium*.
Methods: The genome sequences of the species type strains were obtained either by Illumina-based NGS for previously unsequenced strains or from GenBank for previously sequenced strains. The genomes were compared by digital DNA-DNA hybridization (dDDH) using the Genome-to-Genome Distance Calculator (GGDC).
Results: The pairwise genome-to-genome distances (GGDs) calculated indicate that many of the currently defined taxa are not supported by whole-genome analysis. Selected examples are listed below.
Numerous subspecies should be consolidated:

- *M. avium* subsp. *avium*, *M. avium* subsp. *paratuberculosis*, and *M. avium* subsp. *silvaticum* represent a single taxonomic entity (GGDs: 88.0-99.3%): *M. avium*.
- *M. fortuitum* subsp. *fortuitum* and *M. fortuitum* subsp. *acetamidolyticum* represent a single taxonomic entity (GGD: 88.4%): *M. fortuitum*.

Numerous species should fall entirely within the circumscription of other species (which carry nomenclatural priority):

- All the species of the *M. tuberculosis* Complex (*M. tuberculosis*, *M. africanum*, *M. bovis*, *M. caprae*, *M. microti*, and *M. pinnipedii*) are very closely related and in fact represent a single species (GGDs: 95.9-97.9%): *M. tuberculosis*.
- *M. conceptionense* falls within the circumscription of *M. farcinogenes* (GGD: 84.3%).

Numerous species should be demoted to subspecies of other species which carry nomenclatural priority:

- *M. yongonense* should be considered a subspecies of *M. intercellulare* (GGD: 78.1%).
- *M. vanbaalenii* should be considered a subspecies of *M. austroafricanum* (GGD: 79.8%).

Conclusion: The application of NGS and dDDH allows species to be comparatively analyzed using the entirety of their genomes rather than a few misleading biochemical characteristics or even a few genetic loci, e.g. 16S, *hsp65*, *rpoB*, etc. Using these techniques to examine the taxonomy and phylogeny of the genus *Mycobacterium* clearly shows that numerous existing taxa should be reclassified.

Novel Panel of Multi-Drug-Resistant Gram-Negative Clinical Isolates for Use as Standards for Antimicrobial Research and Measurement
Joyce Sutcliffe, John Pace, Raul Cano, Juan Lopera, Cynthia Long.
ATCC

Addressing Bias with Reference Materials in Microbiomics and Metagenomics Measurements
Shuiquan Tang, Ryan Kemp and Larry Jia
Zymo Research Corporation, Irvine, California, US

The field of microbiomics has developed rapidly in the past several years. However, this field has been criticized for poor data reproducibility across labs. A striking example of this problem was revealed in 2014 [1] by a study that evaluated the substantial inconsistency between the gut microbiome profiles of two populations: 1) a US population measured by the Human Microbiome Project and 2) a European population measured by the Metagenomics of the Human Intestinal Tract consortium (MetaHIT). The studies concluded that the large discrepancy in community structure was

methodologically influenced by the DNA extraction process which is a major source of bias in microbiomic workflows. To objectively assess the performance of different microbiomics workflows, it is essential to have an accessible, well-defined, and accurately characterized mock microbial community standards to serve as reference materials for optimization, validation, and controls for microbiomic workflows. Acknowledging this deficit, Zymo Research created the first commercial reference material for microbiome measurements, named the ZymoBIOMICS® Microbial Community Standard. Using the ZymoBIOMICS® Microbial Community Standard, we assessed the performance of several of the most cited DNA extraction protocols used in the Microbiomics field and the effect of various library preparation techniques for 16S and shotgun sequencing. We found that the three most commonly used protocols in this field for DNA extraction, including the HMP fecal DNA extraction protocol, are significantly biased. They over-represent easy-to-lyse organisms, such as Gram-negative bacteria, which explains the inconsistencies between the gut microbiome profiles derived from HMP and MetaHIT projects. Using the DNA standard, we were also able to accurately characterize some bias in the library preparation steps, such as GC bias in shotgun sequencing and PCR chimera in 16S sequencing. Zymo Research has since been standardizing and validating our workflow from collection to conclusion using the ZymoBIOMICS® Microbial Standards.

1. Wesolowska-Andersen A, et al. *Microbiome* 2014, 2:19.

The Detection, Classification, and Collection of Airborne Pathogens Using Electro-optics

Philip J. Wyatt
Wyatt Technology Corporation & Wyatt Aerosol Systems, 6330 Hollister Avenue, Goleta, CA 93117

The on-going development of a real-time aerosol detection, classification, and collection system, is intended to address a variety of applications. These include the accurate measurement and certification of NIST Standard Reference Materials, detection of dangerous environmental contaminants such as asbestos and carbon/soot particulates, and even the presence of a potentially dangerous hospital-sourced contamination involving a variety of possible bacterial species. The latter capability would provide also for the immediate warning of a bioterrorist attack in progress. Bacterial threats, be they accidentally produced within a hospital environment or the consequence of a terrorist attack, involve the dissemination of sufficient airborne agents to produce infection following a “reasonable” period of exposure. Accordingly, the electro-optical detection system must be capable of detecting the presence of multiple members of each such threat agent. Details of the extant system are described together with the need for real-time sample collection as well as interacting replicate devices. Specific examples are presented.
