NIST Special Publication 1183

Standards for Pathogen Identification via Next-Generation Sequencing (SPIN) Workshop

Summary Report

October 20-21, 2014
Gaithersburg, MD, USA

Nathan D. Olson
Scott A. Jackson
Nancy. J. Lin

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.SP.1183

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

NIST Special Publication SP1183

Standards for Pathogen Identification via Next-Generation Sequencing (SPIN) Workshop

Summary Report

October 20-21, 2014
Gaithersburg, MD, USA

Nathan D. Olson
Scott A. Jackson
Nancy. J. Lin
Biosystems and Biomaterials Division
Material Measurement Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.SP.1183

June 2015

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. Opinions expressed in this report are the authors' and do not necessarily reflect the policies and views of the DHS, NIST, or affiliated venues.

National Institute of Standards and Technology Special Publication SP1183
Natl. Inst. Stand. Technol. Spec. Publ. SP1183, 30 pages (June 2015)
CODEN: NSPUE2

This publication is available free of charge from: http://dx.doi.org/10.6028/NIST.SP.1183

Abstract

Advances in DNA sequencing over the last decade have lowered the barrier to whole genome sequencing (WGS). However, additional challenges must be overcome before widespread adoption of WGS for pathogen identification in applied settings such as biothreat detection, outbreak monitoring, and clinical diagnostics is realized. In an effort to identify priority areas for standards development and facilitate adoption of new sequencing technologies, the National Institute of Standards and Technology (NIST) convened a two-day workshop composed of representatives from Federal Government agencies, academia, and industry, all of whom are utilizing or developing methods for pathogen identification using next-generation sequencing (NGS). The workshop objectives were to identify current and anticipated measurement challenges hindering the implementation of NGS in pathogen identification, and to propose avenues to address these challenges including recommendations for standards development. First, leaders in the field presented their efforts and thoughts on WGS challenges related to specific areas including metrology, sample preparation, outbreak and antimicrobial resistance surveillance, large-scale genomic sequencing projects, genome sequence databases, bioinformatics methods, and biomarker development. Then, attendees met in breakout groups and identified a number of measurement challenges and potential standards-based solutions for NGS-based pathogen identification. The challenges covered all steps involved in NGS, from sample to answer. Proposed solutions focused on reference materials, reference data, documentary standards/guidance, and interlaboratory studies for method validation and proficiency testing; data analysis and interpretation; database quality and curation; and sample processing. Next steps are to further prioritize the challenges identified at the workshop and convene groups of experts in collaboration with NIST to work toward addressing these issues.

Key words

metrology; next-generation sequencing; pathogen identification; reference materials; standards; whole genome sequencing

Sponsorship

This workshop was sponsored by NIST and by the Department of Homeland Security (DHS) Science and Technology Directorate under Interagency Agreement HSHQPM-13-X-00190 with NIST.

Workshop Organizers

Scott A. Jackson
Nancy J. Lin
Nathan D. Olson

Biosystems and Biomaterials Division
Material Measurement Laboratory
National Institute of Standards and Technology

1 Introduction

Recent and continuing advancements in DNA sequencing technologies have transformed the field of microbial genomics and the use of whole genome sequencing (WGS). Sequencing whole microbial genomes once took large sequencing centers months to complete prior to 2003 and is now regularly performed by small laboratories in a few days. The ease in which WGS is achieved by next-generation sequencing (NGS) platforms has resulted in a desire to move NGS technology from the basic research setting to the applied setting. Most notable are the areas of molecular epidemiology, clinical diagnostics, and microbial forensics.

A real-world example of how WGS is currently being used for pathogen identification is in the case of foodborne outbreak investigations. WGS is considered state-of-the-art in molecular epidemiology. Typically epidemiological outbreak investigations involve determining whether two strains are the same or different, or whether a group of strains form a cluster. Indeed, an epidemiologist might ask whether a food isolate is the same strain as a clinical isolate associated with an ongoing foodborne outbreak. The ability to answer this question rapidly and confidently has significant implications for public health as, in this example, contaminated foods may be quickly identified and recalled, thereby further preventing the spread of disease. Furthermore, identifying the source of the pathogen has legal implications in the forensic setting and enforcement of safe food manufacturing practices. Arguably, strain typing via WGS is the ultimate in discriminatory power as every genome position is identified. The challenge to using WGS for outbreak investigations is that replicate colonies sequenced using NGS procedures will produce non-identical genome sequences. Differences may be as subtle as a few base changes or as blatant as large-scale structural differences. Whether the differences are true biological mutations introduced during culturing or measurement errors associated with sequencing or data analysis is unknown. The application of metrological tools including reference materials and guidance documents will help define the source of these differences and better inform outbreak investigations.

A further impediment that must be recognized when considering standards for pathogen identification via NGS is the fact that NGS technologies are rapidly improving. Updates to current sequencing platforms occur as frequently as every six months. As a result, laboratory workflows, including multiplexing procedures, can and do change regularly. Technology and chemistry updates also often include longer sequencing reads. This change in the nature of raw data necessitates changes in bioinformatics analysis and interpretation. Standards developed for WGS technologies must be technology and platform agnostic to maximize their applicability as the field evolves.

The current procedure for WGS includes a number of steps that differ based on the particular NGS technology being utilized. In general, DNA is extracted from the sample, the extracted DNA is manipulated to generate a sequencing library, the library is sequenced, and the resulting data are analyzed using a bioinformatics pipeline. Each step of this measurement process has associated challenges and inherent biases that impede the adoption of WGS to the applied setting. Furthermore, there are implementation challenges including establishing comparability with current methods, evaluating and certifying end-user proficiency, developing appropriate data storage/analysis/interpretation procedures, and applying quality control measures. The development and use of a measurement infrastructure, including guidance documents, reference materials, reference data, and interlaboratory studies, will help address these challenges by increasing confidence in WGS results and improving decision-making related to pathogen detection, identification, and attribution. Standards will also serve to gauge the stability of rapidly evolving sequencing technologies, including instrumentation, sequencing chemistries, library preparation, analysis routines (bioinformatics pipelines), as well as data interpretation.

To aid in identifying priority areas for standards development, NIST convened a two-day workshop composed of stakeholders representing federal and state agencies, academia, and industry, all of whom are utilizing or developing methods for NGS-based pathogen identification and characterization.

The workshop had three primary objectives:

1. Identify current measurement challenges hindering the implementation of WGS in pathogen identification.
2. Identify future measurement challenges anticipated to arise as WGS technologies evolve and culture-independent diagnostics become more feasible.
3. Identify avenues to address these challenges, including recommendations for standards development.

Workshop attendees included representatives from federal and state government agencies, industry, as well as academia (Fig. 1). The complete list of attendees and affiliations is provided in the appendix.

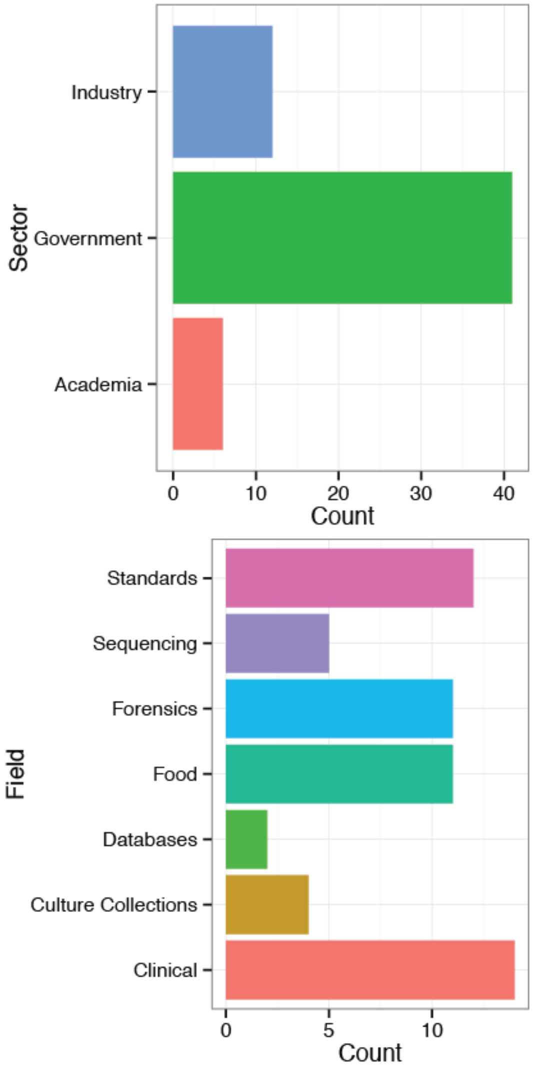

Figure 1. Breakdown of attendees by sector (top) and field of research (bottom).

2 Workshop Overview

Opening remarks were provided by Dr. Laurie Locascio, Director of the Material Measurement Laboratory (MML) at NIST, and an introduction to the workshop was provided by Mr. Scott Jackson (MML, NIST).

The workshop proceeded with a series of invited presentations by representatives from various government agencies and universities covering areas including metrology, bioinformatics, data access and sharing, sample preparation, phylogeny, reference materials, reference data, and current research efforts at NIST. Despite the diversity of topics, several recurring themes were identified including the need for standard methods and reference materials for sample processing and data analysis, performance metrics for validating data analysis methods, well-curated and diverse databases, approaches to define limits of detection, and guidance on results interpretation.

The majority of the second day of the workshop focused on small and large group discussions to identify current and future key measurement challenges associated with implementation of NGS for pathogen identification and potential solutions for these challenges including standards development activities. Attendees were divided into breakout groups for small group discussions. After the breakout sessions, the full group reconvened and individual group findings were summarized and a general discussion followed.

It should be recognized that two different and unique applications of NGS for pathogen identification were discussed at the workshop. The majority of the discussion focused on strain-level identification and discrimination of pure isolates via NGS-based WGS. The second application covered was culture-independent diagnostics using shotgun metagenomic sequencing to identify pathogen-specific signatures directly from complex samples (e.g., stool, blood, suspicious powder, etc.).

The workshop agenda is provided in the appendix.

3 Presentation Summaries

The following section provides a brief summary of each talk in the order they were presented. The slides for a number of the presentations are available at http://www.slideshare.net/, tagged "nist spin workshop."

3.1 Metrology for Identity and Other Nominal Properties[1]

David L. Duewer, PhD, NIST

Dr. David L. Duewer is a Research Chemist in the Chemical Sciences Division at NIST where his research addresses numerous aspects of chemical metrology. Dr. Duewer's presentation set the stage for the workshop by introducing the concept of metrology as related to identification in biological systems. The focus of metrology is establishing measurement infrastructures to define confidence for a given measurement, thus enabling data informed decision-making. This process is supported by the "calculus of confidence," as proposed by Dr. Duewer and his colleague Dr. Marc Salit, and is based upon the primary tools of metrology: traceability, validation, and uncertainty. Dr. Duewer presented how these three components, typically applied in traditional quantitative metrology, would translate to the metrology of identification. Methods for quantitative metrology are well established and outlined in the "Guide to the expression of uncertainty in measurement" (GUM)[2], NIST Special Publication 811[3], and the "International vocabulary of metrology - basic and general concepts and associated terms" (VIM)[4]. For quantitative metrology, traceability enables comparisons over time and place from the International System of Units (SI) to a result. Validation ensures that sources of error and biases associated with a measurement process are well understood, establishing method uncertainty and scope. Uncertainty enables meaningful comparisons of results and determines whether a measurement method is fit for purpose based on estimates of precision and bias. Unlike quantitative metrology, the metrology of identification has not been established. Dr. Duewer presented analogies for traceability, validation and uncertainty for identification measurements. For traceability applied to identification, pure substances can be used to certify unambiguous "barcodes" or unique identifiers that define a substance's identity. The unambiguous barcodes would be catalogued in an authoritative database, and unknowns would be compared to the database, similar to the comparison of quantitative measurement values to SI units. In identification, metrological validation would determine if the method is fit for purpose through the use of defined inclusivity and exclusivity panels[5]. Confidence is used in place of uncertainty

[1] http://www.slideshare.net/nist-spin/spin-duewer-20141017

[2] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML. Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization, Geneva. ISBN 92-67-10188-9, First Edition 1993, corrected and reprinted 1995. (BSI Equivalent: BSI PD 6461: 1995, Vocabulary of Metrology, Part 3. Guide to the Expression of Uncertainty in Measurement. British Standards Institution, London.)

[3] http://www.nist.gov/pml/pubs/sp811/

[4] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML. International Vocabulary of Basic and General Terms in Metrology Second Edition 1993. International Organization for Standardization, Geneva.

[5] LaBudde, R. A., & Harnly, J. M. (2012). Probability of identification: a statistical model for the validation of qualitative botanical identification methods. *Journal of AOAC International*, *95*(1), 273–85.

in identification measurements when comparing measurement results and is based on the amount and quality of evidence supporting the data[6]. Overall, the presentation described how the metrological tools of traceability, uncertainty, and validation could be applied to support the use of identification-based measurements in decision-making processes. These concepts provide a framework for developing a metrologically valid measurement infrastructure for WGS-based pathogen identification.

3.2 Forensics: Human Identity Testing in the Applied Genetics Group[7]

Peter M. Vallone, PhD, NIST

Dr. Peter M. Vallone is the Leader of the Applied Genetics Group in the Biomolecular Measurement Division at NIST. The Applied Genetics Group is at the forefront in the development of standards to support the forensic community's use of DNA typing for human identification. Dr. Vallone's presentation described NIST's support of DNA typing-based human identification through standard reference materials (SRMs), interlaboratory studies, education and training, as well as basic research. Short tandem repeat (STR) markers are used by the typing community for human identification. However, there are a number of challenges regarding the use of STRs for human identification including method reproducibility and obtaining identification information from degraded samples. The Federal Bureau of Investigation (FBI) DNA Advisory Board Quality Assurance Standards (QAS) for Forensic DNA Testing Laboratories define requirements that laboratories must meet for accreditation[8]. To support the QAS accreditation program, the Applied Genetics Group developed polymerase chain reaction (PCR)-based DNA profiling standards, including DNA extracts and cells certified for STR alleles based on polymorphism lengths. The Group also developed standards for human DNA quantitation, mitochondrial DNA sequencing, and heteroplasmic mitochondrial DNA mutation detection. Through coordination of numerous interlaboratory studies to compare performance and demonstrate method reproducibility, the Group was able to help laboratories evaluate their performance relative to peers and establish confidence in measurements made by accredited laboratories. The Group also supports the human identification laboratory community through developing training materials[9], hosting workshops, and publishing textbooks[10,11,12,13,14]. The Group's basic research has also helped advance DNA-based human identification methods. For instance, a set of miniSTRs with smaller PCR product sizes relative to typical STRs were developed at NIST to facilitate typing with degraded DNA. The miniSTR technique was used to help identify 20 % of World Trade Center victims that were unidentifiable using traditional

[6] Climate Change 2007, Synthesis Report http://www.ipcc.ch/publications_and_data/ar4/syr/en/contents.html

[7] http://www.slideshare.net/nist-spin/vallonespinoct2014-141209161936conversiongate02

[8] http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas_testlabs

[9] http://www.cstl.nist.gov/strbase

[10] Butler, John Marshall. *Forensic DNA typing: biology & technology behind STR markers*. Academic Press, 2001.

[11] Butler, John M. *Forensic DNA typing: biology, technology, and genetics of STR markers*. Academic Press, 2005.

[12] Butler, John M. *Fundamentals of forensic DNA typing*. Academic Press, 2009.

[13] Butler, John M. *Advanced Topics in Forensic DNA Typing: Methodology: Methodology*. Academic Press, 2011.

[14] Butler, John M. *Advanced Topics in Forensic DNA Typing: Interpretation*. Academic Press, 2014.

techniques. The group is also working on rapid PCR assays to reduce sample-processing time, new STR markers for increased confidence in identification, and NGS assays for human identification. In summary, the Group's efforts over the years have resulted in a forensic DNA typing capability built on a metrological foundation based on science, standards, and reference materials to ensure high quality measurements and establish confidence in results.

3.3 Separation of Bacteria from Complex Samples

Javier Atencia, PhD, University of Maryland

Dr. Javier Atencia is Research Faculty at the University of Maryland College Park Bioengineering Program and a Guest Researcher in the Biosystems and Biomaterials Division at NIST who specializes in the development and application of microfluidic devices to address critical measurement needs. Dr. Atencia's presentation provided an example of cutting edge research using an engineering approach to facilitate rapid recovery of pathogens from complex matrices. Currently sample processing, specifically isolation of a pathogen from a sample, is the most time intensive step in pathogen detection and identification. Dr. Atencia presented his work on pathogen enrichment methods to reduce sample processing time. His approach leverages the pathogen's natural motility and utilizes microfluidics coupled with chemoattractants and chemorepellents to separate the organisms. Feasibility studies have successfully isolated pathogens from complex matrices such as food products. Overall, microfluidics offers a potential method for rapid, simple enrichment based on pathogen motility and chemotaxis for applications including food safety and microbial forensics.

3.4 Microbial Genomics at NIST[15]

Nathan D. Olson, NIST

Mr. Nathan Olson is a Research Biologist in the Biosystems and Biomaterials Division at NIST where he and his colleagues are developing methods and reference materials to support the use of NGS in microbial identification. Mr. Olson's presentation provided an overview of the genomics and microbial work at NIST to demonstrate how NIST meets the measurement needs of research communities. Examples of NIST working collaboratively to address a community's measurement needs included the External RNA Control Consortium[16], Genome in a Bottle Consortium[17], and standards development for biothreat detection. Each of the projects uses reference materials together with basic research, documentary standards, and related tools to increase measurement assurance. Mr. Olson's talk also included an overview of the microbial genomics work at NIST including the use of NGS to evaluate purity and identity of microbial samples and reference materials. To assess purity of a microbial material, taxonomic sequence classification algorithms were used to identify low levels of contaminants without assumptions regarding contaminant identity. To determine material identity, requirements for evaluating algorithms used to identify single nucleotide polymorphisms (SNPs) were defined. Finally, NIST

[15] http://www.slideshare.net/nist-spin/spin-workshop-microbial-genomics-nist-46427604

[16] https://sites.stanford.edu/abms/events/ercc2

[17] https://sites.stanford.edu/abms/giab

in collaboration with the Food and Drug Administration (FDA) is developing microbial genomic reference materials for use in evaluating and improving sequencing platforms and bioinformatics pipeline performance. Mr. Olson's talk highlighted how reference materials and documentary standards can be used to address measurement needs related to microbial genomics and that current efforts at NIST will help establish the measurement infrastructure needed for NGS-based pathogen identification.

3.5 Sequencing and Informatics for Microbial Forensics

Adam M. Phillippy, PhD, NBACC

Dr. Adam M. Phillippy is a Senior Principal Investigator at the National Biodefense Analysis and Countermeasures Center (NBACC) in Frederick, MD where his team develops bioinformatics hardware and software for microbial forensics. In microbial forensics, the primary goal is attribution where results and supporting measurement science must stand up in the court of law. Dr. Phillippy's talk focused on the workflow and quality management system his team developed to ensure that results generated in his laboratory are suitable for attribution purposes. One way to ensure results are suitable for attribution is by following consensus standards; however new technologies are rarely standardized. To take advantage of new technologies not yet standardized, they follow the Federal Rules of Evidence and the Daubert standard as guidelines for the use of scientific evidence in court. To comply with the Daubert standard, the team produces peer-reviewed publications on software they develop and makes the software publicly available. Dr. Phillippy also described that all samples are metagenomes comprised of a population of cells with varying degrees of diversity, novel pathogen discovery is complicated by gaps in the current databases, and microbial identification should be viewed in the context of phylogeny focusing on degrees of relatedness and not taxonomic matches. Measurement challenges highlighted in the talk included contamination during sample processing, integration of multiple quality control measures, methods for phylogenetic analysis, correlation of phenotype and genotype, and the quality of public data.

3.6 Bacterial Pathogen Genomics at NCBI[18]

William Klimke, PhD, NCBI

Dr. William Klimke is a Staff Scientist at the National Center for Biotechnology Information (NCBI), part of the National Library of Medicine, National Institutes of Health (NIH), where his group is collaborating with federal and state public health agencies on pathogen informatics. Two major initiatives are 1) the FDA GenomeTrakr project which includes federal, state, industrial, and international partners to sequence food and environmental isolates, starting with *Salmonella*, but expanding to other foodborne pathogens, and 2) the real time *Listeria* project that is a collaboration with FDA, Centers for Disease Control and Prevention (CDC), US Department of Agriculture (USDA), and state public health agencies where all of the *Listeria* collected in the US are sequenced. Both of these projects require cultured pure isolates and are

[18] http://www.slideshare.net/nist-spin/bacterial-pathogen-genomics-at-ncbi

intended to aid surveillance projects related to rapid detection and source tracking of foodborne illness outbreaks. Sequence data generated by the participating laboratories are sent directly to the sequence read archive (SRA, an international public sequence database), and NCBI has developed automated assembly, k-mer, and SNP-based phylogenetic pipelines for analyzing the sequence data. During these analyses a number of errors have been discovered including taxonomic errors or misidentifications, contaminated samples, and anomalous sequence data and assemblies in both the reference as well as the surveillance real-time datasets, and NCBI is working actively with data contributors to fix the data. NCBI also has ways to flag anomalous data as a warning including data with unverified taxonomic identifications and also to mark reference data. Dr. Klimke also highlighted the need for additional proficiency testing of bioinformatics pipelines to help identify sources of error in data processing, which would help increase measurement assurance for pathogen identification. With respect to SNP pipelines, Dr. Klimke suggested a set of SNPs that have been independently verified with Sanger sequencing would be an essential dataset for making comparisons.

3.7 Next-Generation Sequencing for Identification and Subtyping of Foodborne Pathogens[19]

Rebecca Lindsey, PhD, CDC

Dr. Rebecca Lindsey is a Microbiologist in the Enteric Diseases Laboratory Branch (EDLB) of the CDC where she works on the development and use of tools for monitoring Shiga toxin-producing *Escherichia coli* (STEC) outbreaks using NGS. Dr. Lindsey's talk described her group's efforts to facilitate the adoption of WGS for pathogen detection as part of the Advanced Molecular Detection (AMD) initiative. The vision for this project is to bring WGS to public health laboratories to characterize foodborne pathogens, replacing multiple workflows with one single efficient workflow. For adoption of WGS in public health, tools must be simple, comprehensive, and operational in a network of laboratories, allowing for comparisons to central and local databases. Whole genome multilocus sequence typing (wgMLST) allows for definitive subtyping with easily communicated results, but implementation requires standardization of areas including methods, analysis, and nomenclature. To this end, EDLB is working with national and international partners toward recommended protocols, sequencing quality metrics, a customizable database and analytical software package to develop a standardized analysis procedure, and allele databases to support standard naming for wgMLST patterns. Although there is no consensus at the CDC yet, these efforts are moving EDLB toward equipping public health labs to use wgMLST for pathogen identification. Overall, Dr. Lindsey's talk highlighted the need for a single standardized workflow with simple analysis tools and easily communicated results but acknowledged that traditional "shoe-leather" epidemiological data ultimately play the key role in identifying outbreaks.

[19] http://www.slideshare.net/nist-spin/next-generation-sequencing-for-identification-and-subtyping-of-foodborne-pathogens

3.8 Validation, Standardization, and Application of the FDA WGS Pipeline for Foodborne Contamination Traceability

Eric W. Brown, PhD, FDA

Dr. Eric W. Brown is the Director of the Division of Microbiology in the Office of Regulatory Science at the U.S. FDA's Center for Food Safety & Applied Nutrition (CFSAN). His group develops and uses microbiological and molecular genetic strategies to detect, identify, and differentiate bacterial foodborne pathogens. Dr. Brown's talk on applying WGS to foodborne pathogen investigation included recent examples demonstrating the use of WGS to support real-time outbreak investigations as well as a description of their role in the development of the GenomeTrakr project. In collaboration with the CDC and NIH, his group was able to use WGS and SNP-based phylogenetic analysis to provide actionable results from real-time outbreak investigations that took place during the spring and summer of 2014. GenomeTrakr is a network of laboratories with WGS capabilities working in close collaboration to further expand foodborne illness outbreak response. Efforts are underway to validate and standardize sequencing and data analysis methods used by the network through documentary standards activities, publically available data analysis pipelines[20], and intralaboratory and interlaboratory comparison studies. Dr. Brown's talk highlighted efforts to develop the infrastructure for real-time outbreak investigation using a network of laboratories and noted that further standardization of both sequencing and data analysis methods would help advance the attribution capability of the network.

3.9 The Prospects for Nextgen Surveillance of Pathogens: A View from a Public Health Lab[21]

William Wolfgang, PhD, New York State Department of Health

Dr. William Wolfgang is a Research Scientist at the Division of Infectious Diseases at Wadsworth Center, New York State Department of Health where he leads a group implementing WGS for pathogen surveillance. Dr. Wolfgang's talk covered how and why his group is beginning the transition from using pulsed-field gel electrophoresis (PFGE) to WGS for outbreak surveillance and investigation. PFGE offers low discriminatory power compared to WGS, as his group demonstrated in a proof of principle study on a *Salmonella* serotype Enteritidis outbreak. In that study, WGS was able to identify outbreak isolates where the discriminatory power of PFGE was unable to differentiate outbreak isolates from non-outbreak isolates. Currently WGS is more expensive than PFGE and most labs are still using PFGE, so maintaining backwards compatibility is critical. Dr. Wolfgang also highlighted the need for standards related to quality and reproducibility for both sequencing and data analysis to accelerate the development of new sequencing technologies allowing for more accurate and meaningful comparisons.

[20] http://snp-pipeline.readthedocs.org/en/latest/

[21] http://www.slideshare.net/nist-spin/the-prospects-for-nextgen-surveillance-of-pathogens-a-view-from-a-public-health-lab-46427338

3.10 Whole Genome Sequencing and Antibiotic Resistance Surveillance

Patrick McDermott, PhD, FDA

Dr. Patrick McDermott is Director of the National Antimicrobial Resistance Monitoring System (NARMS) at the FDA Center for Veterinary Medicine (CVM). NARMS is an interagency collaboration with the USDA and CDC to track antibiotic resistance in bacteria from food-producing animals, retail meats, and clinical disease cases. Dr. McDermott's talk covered the role of NGS in surveillance for infection control. Surveillance is fundamental to infection control and requires the monitoring of baseline resistance levels and the spread of resistance genes in order to develop attribution hypotheses regarding sources of resistant bacteria and to support education and policy-making. The challenges to ongoing surveillance include sampling design and prioritization, data collection and analysis, and reporting. Use of WGS simplifies the process as information about serotype, resistance patterns, genetic relationships, and resistance mechanisms can be obtained from a single method. A proof of principle phenotype-genotype correlation study indicated that WGS is able to reliably predict resistance in *Salmonella* and *Campylobacter*, with the degree of agreement varying slightly with the drug class. This application of WGS requires new standards for reporting and interpretation. It also relies on a curated database of resistance genes with known mechanisms. The FDA is developing such an antimicrobial resistance database based on information collected from public databases and literature.

3.11 Development of FDA MicroDB: A Regulatory-Grade Microbial Reference Database[22]

Heike Sichtig, PhD, FDA and Luke Tallon, University of Maryland

Dr. Heike Sichtig is a Principal Investigator and Regulatory Scientist at the FDA in the Division of Microbiology Devices, Office of In Vitro Diagnostics and Radiological Health where she leads efforts to develop regulatory grade reference databases for clinical use. Mr. Luke Tallon is the Scientific Director of the Genomics Resource Center (GRC), Institute of Genome Sciences at the University of Maryland School of Medicine where he is responsible for the scientific oversight of the GRC and also provides guidance to researchers using sequencing technologies. Dr. Sichtig began the presentation describing the formation of an interagency group to generate information to evaluate sequence data quality in the public domain, with an emphasis on defining a genome sequence quality threshold and required metadata for inclusion in a high quality database. As sequence quality requirements vary with application, multiple levels of reference databases will likely be needed, for example databases with only high quality genomes for validation and clinical use, and databases with high quality and other available genomes for testing and development. Additionally, in collaboration with her co-presenter Mr. Tallon, she is developing the Microbial Reference Database (MicroDB), a high quality, public microbial sequence database for clinical use. Dr. Sichtig and Mr. Tallon presented on their efforts toward developing the MicroDB and provided example data submission cases. To ensure high quality,

[22] http://www.slideshare.net/nist-spin/development-of-fda-microdb-a-regulatorygrade-microbial-reference-database

MicroDB includes requirements for extracted genomic DNA, biosample metadata, sequencing data, sequencing metadata, as well as suggested phenotype metadata. Mr. Tallon presented on his group's work developing methods for generating high quality genome assemblies for the MicroDB. The sequence data, assemblies, and annotations are deposited in the GenBank archive with an FDA interface to access the data. Submissions include isolate metadata based on the NCBI minimal pathogen template as well as sequencing metadata such as library, platform, submitter, coverage, and assembly and annotation pipelines. Dr. Sichtig and Mr. Tallon's talk highlighted the need for genome assembly standards, methods to document assembly quality, and metadata standards that include both analysis methods and sample characteristics.

3.12 From Genome to Biomarker: The Path Forward

David A. Rasko, PhD, University of Maryland School of Medicine

Dr. David Rasko is an Associate Professor in the Department of Microbiology and Immunology and a member of the Institute for Genome Sciences at University of Maryland where his lab uses comparative genomics to examine pathogen evolution. Dr. Rasko presented his group's work using WGS to support the use of comparative genomics and phylogenetics to identify new biomarkers and improve the accuracy of diagnostic *E. coli* isolate characterization. This genomic epidemiology approach has identified new associations between genomic features and virulence factors, and new markers for identification. Dr. Rasko also described the development of MIGEN-Dx, a cloud computing, microbial diagnostic bioinformatics resource that is sequencing platform agnostic. MIGEN-Dx takes raw sequence reads and input and compares the sequence reads to numerous databases, including *E. coli*, Salmonella, and resistance marker databases, and generates a clinically relevant summary report. The summary report is formatted to facilitate clinical interpretation and includes quality control for sequence format, coverage, human contamination, species present, virulence profile, potential resistance mechanisms, and presence/absence of individual markers. Dr. Rasko's talk highlighted the need for expanded sequence databases with extensive metadata for biomarker discovery, the challenges associated with defining the limit of detection for complex (e.g., metagenomic) samples, and the importance of providing summary reports amenable to clinical interpretation to enable WGS use in clinical diagnostics.

3.13 High Throughput Sequencing Pipeline for Diverse Organisms

Bart Weimer, PhD, University of California, Davis

Dr. Bart Weimer is a Professor in the Department of Population Health and Reproduction in the University of California Davis School of Veterinary Medicine and Director of the 100K Genome Project. His laboratory focuses on microbial physiology and function for food, animal and environmental applications using systems biology approaches. Dr. Weimer's talk described the 100K Genome Project and its goal of creating a reference database of 100,000 pathogens associated with food, animals and humans. Isolates used in the project are first authenticated and archived. Dr. Weimer's laboratory has observed that ~15% - 20% of the cultures are multispecies when received, indicating that culture contamination is a widespread problem. Samples are then sequenced using Pacific Biosciences and Illumina with optical mapping used for additional quality control when necessary. Workflows developed for sample processing are published as application notes by Agilent[23,24,25,26,27], providing a valuable reference for sample methods. Currently, the project has almost 10K genomes and is extending internationally to increase database diversity. Dr. Weimer's talk highlighted the need for robust biomarkers to enable fast and actionable results, the need for standardization of DNA modification and methylation reporting, and the value of RNA sequencing for metagenomics samples to provide insight into functional activity.

[23] Kong, N., et al. (2013). Production and Analysis of High Molecular Weight Genomic DNA for NGS Pipelines Using Agilent DNA Extraction Kit (p/n 200600). Agilent Technical Note 5991-36722EN.

[24] Jeannotte, R., et al. (2014). High-Throughput Analysis of Foodborne Bacterial Genomic DNA Using Agilent 2200 TapeStation and Genome DNA ScreenTape System. Agilent Technical Note 5991-4003EN.

[25] Kong, N., et al. (2014). Automation of PacBio SMRTbell 10 kb Template Preparation on an Agilent NGS Workstation. Agilent Technical Note 5991-4482EN.

[26] Jeannotte, R., et al. (2014). Optimization of Covaris Settings for Shearing Bacterial Genomic DNA by Focused Ultrasonication and Analysis Using Agilent 2200 TapeStation. Agilent Technical Note 5991-5075EN.

[27] Kong, N., et al. (2014). Quality Control of High-Throughput Library Construction Pipeline for KAPA HTP Library Using an Agilent 2200 TapeStation. Agilent Technical Note 5991-5141EN.

4 Measurement Challenges and Potential Solutions

The entire NGS process, from sample to answer (Table 1) is challenging and requires a measurement infrastructure to increase confidence in the final answer. Each step in the process has measurement challenges, biases, and uncertainties requiring measurement solutions and standards including physical materials, data, written guidelines/methods, and interlaboratory studies. Table 2 describes these primary classes of measurement-based solutions that can be applied to address measurement challenges.

Table 1: Steps in a typical sample-to-answer workflow for sequence-based pathogen identification.

Step	**Description**
Sample collection	Collection and transportation of raw material (sample) to the laboratory.
Sample processing	Initial laboratory processing of the raw material, including sample enrichment and/or culturing, followed by DNA extraction. For culture independent methods the DNA is extracted directly from the raw material.
Sequencing	Preparation of sequencing libraries and generation of sequencing data.
Bioinformatics analysis	Computational analysis of sequencing data, including genome assembly and comparative genomic analyses.
Results reporting and interpretation	Interpretation and presentation of results for data-informed decision-making.

Table 2: Classes of metrology-based solutions to support a measurement infrastructure.

Class of solution	Description
Materials (M)	Physical materials used to evaluate or validate components such as new or existing laboratory methods, training protocols, and capabilities. These materials can include: • primary reference materials[28] for calibration or method/protocol validation • secondary reference materials, characterized in relation to a primary reference material for traceability • quality control materials for routine assessment of run to run performance
Data (D)	Data generated for use in evaluation and validation of bioinformatics pipelines and algorithms. Datasets could be generated from reference materials, existing reference data[29], or simulated data with defined properties.
Guidance documents (G)	Community accepted guidance documents such as standard operating procedures, standard guidance, or standard methods. These documents could be formal voluntary consensus standards[30] or community accepted best practices. Though not yet defined, standardized bioinformatics pipelines or data analysis methods would fall into this category.
Interlaboratory studies (I)	In interlaboratory comparisons, round-robins, external validations, and proficiency testing, a common sample and protocol are distributed to participants for analysis. The study results are used to validate the measurement process and/or the participant's ability to perform the measurement.

Specific examples of challenges and solutions frequently identified in small and large group discussions are provided in Table 3 and are organized according to the workflow steps. This list is not an exhaustive list of ideas discussed at the workshop nor is it intended to reflect a comprehensive examination of gaps in knowledge; rather it serves to highlight some of the recurring discussion items. In general, the solutions point toward the need for a measurement infrastructure to increase confidence in the entire process and ultimately the final results. To further support development of this infrastructure and the microbial genomics field overall, increased national and international sharing of data, protocols, and analysis methods are needed. While the challenges and solutions in Table 3 are application-independent, application-specific needs were also discussed, such as standards to support attribution in microbial forensics, standards to define the assay limit of detection for culture-independent diagnostics, and data

[28] VIM definition: Material, sufficiently homogeneous and stable regarding one or more properties, used in calibration, in assignment of a value to another material, or in quality assurance.

[29] VIM definition: Data that is critically evaluated and verified, obtained from an identified source, and related to a property of a phenomenon, body, or substance, or a system of components of known composition or structure.

[30] ISO/IEC Guide 2 (1996) Standardization and Related Activities -General Vocabulary. Geneva. *Voluntary consensus standards:* General agreement, characterized by the absence of sustained opposition to substantial issues by any important part of the concerned interests and by a process that involves seeking to take into account the views of all parties concerned and to reconcile any conflicting arguments.

reporting tools to enable clinicians to easily understand and interpret sequencing results. Further, there is a need to incorporate functional biomarkers, such as antibiotic resistance or pathogenicity, into curated databases to develop connections between genotype and phenotype.

Table 3: Examples of challenges and possible solutions identified in the workshop.

Step	**Challenges**	**Possible Solutions***
Sample collection	Lack of robust sampling design	(G) Standard documents or established best practices for application-specific sampling procedures
Sample processing	Biases in DNA extraction efficiency, including matrix effects; DNA quality	(M) Cell mixtures at known ratios (M) Control matrices spiked with target material for assay validation (G) Guidelines for in-house development of cell mixtures as secondary RMs (I) Interlaboratory study to evaluate method repeatability and reproducibility
Sequencing	Lack of consistency in protocols used among laboratories; variations in platforms	(M) DNA or RNA mixtures at known ratios (G) SOPs and community accepted best practices (G) Metadata standards (I) Proficiency testing to evaluate laboratory performance
Bioinformatics analysis	Lack of agreement between methods; method validation is difficult; new tools constantly emerging	(D) Reference data for evaluating and validating pipelines (G) Performance/quality metrics to compare pipelines (G) Metadata standards for pipeline reporting (I) Interlaboratory study to compare bioinformatics methods
Results reporting and interpretation	Results presentation is not always designed for successful interpretation by non-bioinformaticists	(G) Standardized vocabulary for results communication (G) Guidelines for data interpretation with application-specific levels of confidence (G) Metadata and metrics to include with results
Databases	Lack of consistency in data quality; lack of comprehensive databases; inconsistent metadata reporting	(D) Collection and identification of reference-quality data within databases (G) Metadata standards for submission and data curation

* Solution classes include M (Materials), D (Data), G (Guidance documents), and I (Interlaboratory studies)

During the workshop, several specific examples of standards-based solutions were discussed. Some of these solutions were identified as potential starting points for standards development based upon their anticipated high impact, expected usage, and relative ease of preparation as compared to other activities. This list is not exhaustive but serves to identify examples related to the four categories of measurement-based solutions: data, materials, guidance documents, and interlaboratory studies.

Reference data: Workshop attendees recognized the need for “ground truth” to evaluate bioinformatics analyses of complex samples and discussed that development of *in silico* reference data might be a simpler starting point than physical reference materials (such as cells or DNA). Datasets could be created by combining known genome sequences and used to benchmark and compare bioinformatics pipelines as well as results reporting and interpretation. These datasets could include relevant pathogens as well as environmental contaminants, such as host DNA, other microorganisms, and viral genomes.

Reference materials: Matrix reference materials (e.g., blood, soil, milk, etc.) well-characterized in terms of their existing genomic content, microbial and non-microbial, are needed to enable users to challenge their measurement workflows with relevant matrix contaminants. In addition, target organisms could be added to the matrix material in order to validate parameters such as limit of detection. It is noted that this type of reference material is application specific.

Reference materials: Microbial reference materials with known mixtures of multiple cell types are needed to provide a known input for metagenomic studies. It was noted during the workshop that all sequencing measurements should be considered metagenomic analyses because even pure samples are composed of a population of cells and can have contamination. Microbial RMs could be intact cells or genomic DNA and could be designed to include a number of bacterial species along with relevant viral and host genomes.

Guidance documents: Guidance documents describing general methods to develop, characterize, and use reference materials are needed to enable industry and other users to prepare quality control materials in house with organisms relevant to their application space. There is a need for reference materials specific for each application area, yet development of such a vast set of RMs is too large for any one organization. Further, certified RMs can be expensive and might not be feasible to use on a daily basis.

Interlaboratory studies: An interlaboratory study to evaluate the contents of a mixture of well-characterized microbes at known ratios is needed to support metagenomic analyses. Results from the study would be used to characterize sources of bias and uncertainty associated with metagenomic sequencing, including sample processing, sequencing, and data analysis.

5 Conclusions

The SPIN workshop brought together the stakeholder community to identify challenges to the adoption of WGS in applied settings and metrology-based solutions to those challenges. WGS offers the advantage of a single laboratory method for pathogen identification. However, a measurement infrastructure is critically needed to validate sequencing and analysis protocols and instill confidence in results used to support decision-making. Further, the associated data analysis is often organism and application specific and presents unique challenges requiring novel validation methods. Overall, the solutions identified during the course of this workshop will help guide the development of a standards-supported measurement infrastructure for pathogen identification using WGS. This infrastructure will help enable WGS to achieve its full potential in revolutionizing pathogen identification.

6 Appendix

6.1 Workshop Attendees

- Marc Allard, FDA/CFSAN
- Adam Allred, Thermo Fisher Scientific
- Kevin Anderson, DHS
- Kimberly Armstrong, FDA/Center for Devices and Radiological Health (CDRH)
- Javier Atencia, NIST
- Don Atha, NIST
- Paula Baker, NIST
- Nicholas Bergman, NBACC
- Eric Brown, FDA/CFSAN
- Heather Carleton, CDC
- Greg Cooksey, NIST
- Keith Crandall, The George Washington University
- Sandra Da Silva, NIST
- Toni Diegoli, Flinders University
- David Duewer, NIST
- John Elliott, NIST
- Mark Eshoo, Abbott
- Jay Eversole, Naval Research Laboratory
- Tamara Feldblyum, FDA/CDRH
- Sam Forry, NIST
- Katherine Gettings, NIST
- Yan Guo, Pacific Biosciences
- David Hodge, DHS
- Maria Hoffman, FDA/CFSAN
- Kelly Hoon, Illumina
- Robert (Chris) Hopkins, Booz Allen Hamilton
- Scott Jackson, NIST
- W. Evan Johnson, Boston University
- Gil Kaufman, American Dental Association Foundation, Volpe Research Center
- Elizabeth Kerrigan, American Type Culture Collection (ATCC)
- William Klimke, NIH
- Melissa Laird, Pacific Biosciences
- Pascal Lapierre, New York State Department of Health, Wadsworth Center
- Jodie Lee, ATCC
- Nancy Lin, NIST
- Rebecca Lindsey, CDC
- Laurie Locascio, NIST
- Teresa Lustig, DHS
- Patrick McDermott, FDA/CVM/NARMS
- Timothy Minogue, U.S. Army Medical Research Institute of Infectious Diseases
- Nathan Olson, NIST
- Adam Phillippy, NBACC
- Arjun Prasad, NIH/NCBI
- David Rasko, University of Maryland
- Scott Sammons, CDC
- Mary Satterfield, NIST
- Heike Sichtig, FDA/CDRH
- Frank Simione, ATCC
- Kevin Snyder, FDA/CDRH
- Shanmuga Sozhamannan, Department of Defense, Critical Reagents Program
- Arlin Stoltzfus, NIST
- Robert Stones, Food & Environmental Research Agency, UK
- Luke Tallon, University of Maryland
- Živana Težak, FDA/CDRH
- Eishita Tyagi, Booz Allen Hamilton
- Peter Vallone, NIST
- Bart Weimer, University of California, Davis
- William Wolfgang, New York State Department of Health, Wadsworth Center

6.2 Workshop Agenda

Standards for Pathogen Identification via Next-Generation Sequencing (SPIN) Workshop

NIST, Administration Building 101, Portrait Room

October 20-21, 2014

Day 1: Monday, October 20, 2014		
Time	**Speaker**	**Topic**
8:00 AM - 8:30 AM	Arrival	
8:30 AM - 9:00 AM	Opening Remarks and Workshop Overview	
8:30 AM - 8:40 AM	Laurie Locascio, MML Director, NIST	Welcome
8:40 AM - 9:00 AM	Scott Jackson, NIST	Workshop Overview
9:00 AM - 10:20 AM	Session 1	
9:00 AM - 9:30 AM	David Duewer, NIST	Metrology
9:30 AM - 10:00 AM	Peter Vallone, NIST	Forensics
10:00 AM - 10:20 AM	Javier Atencia, NIST	Engineering
10:20 AM - 10:35 AM	BREAK*	
10:35 AM - 12:05 PM	Session 2	
10:35 AM - 11:05 AM	Nathan Olson, NIST	Standards
11:05 AM - 11:35 AM	Adam Phillippy, NBACC	Bioinformatics
11:35 AM - 12:05 PM	William Klimke, NIH	Databases
12:05 PM - 1:15 PM	LUNCH*	
1:15 PM - 2:45 PM	Session 3	
1:15 PM - 1:45 PM	Rebecca Lindsey, CDC	Clinical
1:45 PM - 2:15 PM	Eric Brown, FDA	Food Safety
2:15 PM - 2:45 PM	William Wolfgang, NY State Dept. of Health	Public Health Labs
2:45 PM - 3:00 PM	BREAK*	
3:00 PM - 4:30 PM	Session 4	
3:00 PM - 3:30 PM	Patrick McDermott, FDA	Antimicrobial Resistance
3:30 PM - 4:00 PM	Heike Sichtig, FDA & Luke Tallon, U. of MD	Clinical
4:00 PM - 4:30 PM	Michael Smith, CRP	Reference Materials
4:30 PM - 5:00 PM	Wrap-up	
4:30 PM - 4:50 PM	All	Open Discussion
4:50 PM - 5:00 PM	Scott Jackson, NIST	Debrief and Intro to Day 2
5:00 PM	Adjourn	

*Coffee, snacks, and lunch are available for purchase at the NIST Cafeteria

Day 2: Tuesday, October 21, 2014		
Time	**Speaker**	**Topic**
8:00 AM - 8:30 AM	Arrival	
8:30 AM - 8:40 AM	Welcome and Day 2 Overview	
8:30 AM - 8:40 AM	Nancy Lin, NIST	Day 2 Overview
8:40 AM - 9:50 AM	Session 5	
8:40 AM - 9:10 AM	David Rasko, University of Maryland	Clinical
9:10 AM - 9:40 AM	Bart Weimer, UC Davis	100k Genomes
9:40 AM - 9:50 AM	Nate Olson, NIST	Breakout Instructions
9:50 AM - 10:05 AM	BREAK* - Walk to breakout rooms	
10:05 AM - 12:00 PM	Breakout Sessions	
10:05 AM - 10:55 AM	Breakout Session 1	Measurement Challenges
10:55 AM - 11:10 AM	BREAK*	
11:10 AM - 12:00 PM	Breakout Session 2	Potential Solutions
12:00 PM - 1:00 PM	LUNCH*	
1:00 PM - 3:00 PM	Breakout Session Results	
1:00 PM - 2:45 PM	All	Discussion
2:45 PM - 3:00 PM	Scott Jackson, NIST	Concluding Remarks
3:00 PM	Adjourn	

*Coffee, snacks, and lunch are available for purchase at the NIST Cafeteria