Meeting Summary

Using WGS to Protect Public Health and Enhance Food Safety

Purpose

In April 2016, the Foundation for Meat and Poultry Research and Education, National Cattlemen's Beef Association, the National Pork Board, and the U.S. Egg and Poultry Association convened a joint two-day meeting with researchers and regulatory representatives to discuss "Using WGS to Protect Public Health and Enhance Food Safety." The United States Department of Agriculture Food Safety Inspection Service (USDA-FSIS), the Food and Drug Administration Center for Food Safety and Nutrition (FDA-CFSAN), the Centers for Disease Control and Prevention (CDC), and the National Center for Biotechnology Information (NCBI) were present. The purposes of the meeting were to (i) update the meat and poultry industry on the status and trajectory of whole genome sequencing (WGS) as a regulatory tool, (ii) observe alignment among the federal agencies using or supporting WGS, (iii) facilitate a Q&A session among industry, regulators, and researchers, and (iv) begin discussions on future research priorities that address data gaps.

Meeting Format

On day one of the meeting, representatives from CDC, USDA-FSIS, FDA-CFSAN, and NCBI summarized the current status and trajectory for WGS in their agency. This was the first event where all agencies that represent the Interagency Collaboration on Genomics and Food Safety (Gen-FS) were present in a stakeholder meeting. Each individual presentation concluded with a Q&A session followed by a panel Q&A session. At the end of day one, the meeting conveners challenged industry stakeholders and researchers to (i) reflect on the information provided by the agencies, (ii) ask what stakeholders need in order to do business in a WGS era, and (iii) suggest perceived research needs and/or data gaps. Day 2 was limited to researchers and industry stakeholders; it began with summaries, critical analyses and a Q&A session with researchers with expertise in bioinformatics, epidemiology and data management, and WGS application in surveillance studies. Outcomes from the meetings were summarized into residual and new questions on future policies, bioinformatics needs of stakeholders, potential research opportunities, and remaining concerns regarding the impact and ramification of WGS.

Using WGS to Protect Public Health and Enhance Food Safety – Agency Perspectives

CDC Perspective

Dr. Robert Tauxe, Deputy Director, Division of Foodborne, Waterborne and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, CDC, began the regulatory perspective delineating the roles of CDC, FDA, and FSIS in foodborne disease outbreaks. Dr. Tauxe emphasized that foodborne disease outbreak investigations require epidemiological, microbiological, trace-back & environmental assessment data; no one source is sufficient to attribute cause. He underscored that WGS provides a **n**ew window into outbreaks and sporadic cases (analytics to predict likely sources), and has the potential to replace traditional microbiological characterization (e.g., speciation, serotyping, virulence and resistance profiling). He stated that the 2013 *Listeria monocytogenes* WGS pilot program demonstrated the power of WGS when paired with epidemiological and trace-back data as it (i) resulted in greater confidence in matches between food and clinical isolates, (ii) decreased the size of the average listeriosis outbreak, and (iii) identified new foods associated with listeriosis (e.g., caramel apples, ice cream, bagged lettuce). Current priorities include developing a WGS database for pathogenic *E. coli*, using WGS to rigorously investigate antibiotic-resistant and multidrug-resistant (>3 antibiotics)

Salmonella, and onboarding all 50 states to WGS. At the time of the meeting, 30 states had validated pipelines to contribute genomes into GenomeTrakr, the distributed network of labs sequencing foodborne pathogen genomes; all 50 states are anticipated to be online by 2019.

FDA-CFSAN Perspective

Dr. Steven Musser, Deputy Director of Scientific Operations, FDA-CFSAN, provided an overview of the advantages, challenges, and difference between WGS and pulsed-field gel electrophoresis (PFGE). He noted that PFGE, while the workhorse of the PULSeNet system for 20 years, does not provide sufficient resolution to differentiate some *Salmonella* serovars such as Newport. One of the major differences between the platforms is that WGS is in a public database whereas PFGE data are not, which segued into the challenge that comes with managing the identifiers in the data. GenomeTrakr (available at http://www.ncbi.nlm.nih.gov/biosample/), is the NCBI database that houses public WGS and its respective metadata that are "scrubbed" of many identifiers (e.g. company name). While Dr. Musser maintained that the metadata would not be traceable to a company, several examples exist of instances where the company is highly predictable due to available data on sample type, location, and organism if an outbreak occurred in that timeframe. The FDA is working to build an environmental isolate database; their position is that human clinical isolates can be potentially matched to putative sources more quickly, thus averting illnesses. Dr. Musser used NSpriled Natural Foods salmonellosis outbreak as an example where WGS data from environmental isolates from that company in the database matched new human clinical cases, thus flagging that company as a potential source of the clinical isolates. However, he maintained that the pathogen would have to be found in the plant for regulatory action. Further benefits from WGS include concurrent determination of antimicrobial resistance, virulence potential, and serotype, which is a significant laboratory cost savings. Yet the challenge remains of defining how similar is similar enough for clinical and food isolates to be considered the same strain because there is no single threshold that is applicable to all species. Dr. Musser concluded by suggesting that industry use WGS (i) as a supplier management strategy, (ii) as a method to identify resident pathogens in their facilities (he termed this biomapping), and (iii) for spoilage organisms to become familiar with data management and interpretation.

NCBI Perspective

Dr. David Lipman, Director, NCBI, overviewed the history of the NCBI, a federal service organization, and its interests which are to (i) create automated systems for knowledge about molecular biology, biochemistry, and genetics, (ii) perform research into advanced methods of analyzing and interpreting molecular biology data, and (iii) enable biotechnology researchers and medical care personnel to use the systems and methods developed. NCBI houses the WGS database. Historically, the NCBI database has been a repository for a large number of species; recently, it has been leveraged for foodborne pathogens where the interest is in small differences (e.g. single nucleotide polymorphisms (SNPs) in a very limited number of species. A SNP is a single DNA molecule difference. The cumulative number of SNPs throughout the genome is a measure of relatedness between an isolate of interest and the reference strain; fewer SNPs translate to higher relatedness. Phylogenetic trees (visual representations of the calculated relatedness of isolates in the database) are built everyday with new data uploaded to the NCBI database to see if there is a trend among isolates; similar clinical isolates or isolates that match environmental isolates in the database signal potential relatedness. Two major methods are used to define differences and similarities among isolates --SNP analyses (FDA and FSIS) and WgMLST (whole genome multilocus sequence typing; CDC). WgMLST investigates differences between 1500-2500 loci between a reference strain and isolate of interest. Both methods have advantages and limitations, and utility is species-dependent. Regardless of comparison approach, NCBI will soon have the ability to look for antimicrobial-resistance pattern trends within the database.

USDA-FSIS Perspective

Dr. David Goldman, Assistant Administrator, Office of Public Health Science, USDA-FSIS, provided an overview of WGS in the FSIS strategic plan, which

includes complete phase-in of WGS as the primary subtyping method by 2023 (began in 2012). In 2017, FSIS anticipates sequencing and uploading 5000 isolates representing multiple pathogens isolated from food, food contact surfaces, and non-food contact surfaces; they are currently on target to achieve this goal. Furthermore, FSIS is developing capacity to sequence the isolates and process the data in-house prior to uploading to NCBI. Dr. Goldman emphasized that FSIS recognizes the need for epidemiological and trace-back data to associate cause and effect. As reported by FDA and NCBI, FSIS will use WGS to supplement serotyping and antimicrobial-resistance characterization, and these data will be reported to the company for pathogens isolated from their foods. FSIS is continuing to partner with the National Antimicrobial Resistance Monitoring System (NARMS) to investigate extended spectrum beta-lactamase (ESBL) resistant *Salmonella,* which were recently detected in the US.

All speakers convened for a Q&A panel session immediately following the agency presentations.

Researcher Perspectives and Q&A.

Dr. Henk den Bakker, Assistant Professor, Texas Tech University; Dr. Zaid Abdo, Associate Professor, Colorado State University; Dr. Matt Stasiewicz, Assistant Professor, University of Illinois; Dr. Haley Oliver, Associate Professor, Purdue University; Dr. Steven Ricke, Professor, University of Arkansas; Dr. Jim Bono, Research Scientist, USDA-MARC; Dr. Tommy Wheeler, Supervisory Research Food Technologist, USDA-MARC.

Bioinformatics

Although in agreement with the usefulness of WGS in general, the bioinformatics experts felt that the above presentations were not fully transparent in describing the bioinformatics methodology used in their data processing and whole genome assembly and comparisons to identify origin and history of contamination. In addition, the presentations failed to address limitations of these bioinformatics methods and the sequencing technology used.

Transparency requires that all parts of the methodology be available and well documented for either the industry or academic third parties to be able to scrutinize, benchmark and assess against possible alternatives. Transparency also requires highlighting the approach used to assess confidence in the called SNPs used in constructing the phylogenetic trees describing relatedness and the approaches used to evaluate false positives to assess such relatedness, which requires an understanding of the limitations of this proposed methodology. Limitations of the short-read Illumina sequence methodology include, but are not limited to, the use of a core genome, common genes between isolates, to address relatedness. Leaving out mobile genetic elements or components that are not considered part of that core can give false indication of close relatedness that might not be justifiable. Conclusions based on the resulting methodology is species-specific; the number of SNPs that might distinguish different strains or serotypes will differ between bacteria with a conserved genome such as *Listeria* and a divergent genome such as *Salmonella*. No clear approach was identified to demonstrate how to deal with this. Depth of sequencing and the use of short-read sequencing technology as opposed to long-read sequencing might provide different results in the ability to assemble and close the genomes of the obtained samples. The quality of the genome assemblies and SNP calling depends on the methods used for assembly, *de novo* vs. alignment to a reference. Accuracy of alignment to a reference also depends on the quality of the reference genome used and closeness of that genome to the aligned isolates. Sequencing machine contamination might give a false indication of the residence of a pathogen. Evolution can limit the ability to identify relatedness of pathogens over long periods of times between sampling. This needs to be taken into consideration when comparing samples to evaluate persistence or residence of a pathogen. Based on the above, consensus was reached that WGS data will have to be combined with, and not replace, other data, especially epidemiological, in determining and tracking contamination.

The expert panel confirmed a need for outreach in educating the industry about the different aspects of the methodology and technology used in WGS and for future research to benchmark and identify alternative methods for WGS in the area of food

safety. Also, transparency is paramount, especially as these methods are used in tracking the source of pathogens, to allow the industry and other parties to challenge and benchmark accuracy of these methods.

Epidemiology, data management, and surveillance studies

A number of strengths, weaknesses, opportunities and threats were identified pertaining to the use of WGS and to the initiation of the discussion and involvement of industry and academia at this point in time. The timing of this meeting was identified as a strength that could allow the industry to be involved in the early stages of implementing this new technology and influence its future use. The panel believed this to be possible in two ways: (i) by emphasizing to government policy makers and policy enforcers that WGS alone is not sufficient to predict virulence and that the epidemiological and environmental context should still be considered in tracking pathogens to the origin of contamination, and (ii) given that industry does and can collect more samples than the regulators, by investing in research to set a parallel WGS pipeline that utilizes best practices in bioinformatics including, but not limited to, those used by government labs. This last point can open the door for industry to invest in research to study pathogen-specific nuances including scope of diversity, niche and virulence potential. The lack of historical data available to both regulators, especially FSIS, and some indication of disconnect between the current use by regulators and available environmental studies of *Listeria moncytogenes*, for example, were highlighted as an alarming weakness of the current state of regulator use of WGS. This highlighted a threat of an increased uncertainty for the industry in trying to anticipate the use of this technology. This is in addition to the fact that the accumulating database based on the samples obtained by regulators is biased towards food pathogens of regulated products, which might impact the chance of false positive identification of bacterial isolates not closely related to those in the database.

Concerns also were raised about the cost of catching up while addressing individual industry challenges and acknowledging industry's efforts and simultaneously maintaining anonymity of sample sources and privacy of industry participants. To circumvent these concerns, a VolunteerNet-like structure was proposed, where industry can come together and create a central resource that maintains privacy while allowing for WGS research of importance to the industry. The concern of privacy was also raised in relation to government public data sharing of their WGS samples, thereby highlighting the need to maintain rigorous assessment of the metadata accompanying these samples to guarantee privacy of the sampled industries.

Other discussed issues echoed those presented in the Bioinformatics' methodology critique, includinga disappointment in the level of discussion of transparency and limitations of the proposed methods. The panel proposed engagement of a third party, such as the National Academies of Science, to help assess the direction and usefulness of the WGS approach proposed by regulators.

Summary of industry needs, residual questions and concerns, and potential research opportunities to help navigate WGS as the next generation of DNA fingerprinting:

Research Needs and Opportunities

- What level of confidence needs to exist in the differences among strains?
- How different do isolates need to be to be considered different strains? Should isolates be compared based solely on shared parts of the genome (the core genome)? How much impact will including the non-shared parts of the genome have on confidence to assess similarity of isolates? How should species-specific differences and evolution be taken into consideration in these comparisons?
- Can WGS be used to determine virulence?
- What role should WGS play to complement the epidemiology to assess an outbreak?
- What impact can completeness (or lack thereof) of the available databases have on correct identification and matching of sequenced isolates?
- Can FSIS data be mined for similarities among organisms by location?

- Should an industry database be developed outside of regulatory scope?
- What are the limitations of the bioinformatics tools and pipelines used to assess similarities and differences between isolates?
- Should an industry bioinformatics pipeline be also developed outside of the regulatory scope? Who should develop and maintain this pipeline?
- Industry needs user-friendly bioinformatics basics (e.g., what are SNPs, differences among SNP- and wgMLST-typing schemes)
- How common/prevalent is a given strain in the environment? Can the same strain be found in multiple environments that are unrelated? Alternatively, how confident could we be that finding it on a farm would exclude other farms for example?
- What are the limitations and biases of the short-read sequencing technologies currently used in WGS? Would new long-read sequencing technologies overcome these limitations? Should there be a mix of the two technologies to complement the strength of one another?

Policy Issues and Questions

- How can industry do research without regulatory ramifications?
- FDA and USDA have seemingly different value for epidemiology in outbreak investigations, which needs to be resolved.
- Is the definition of an outbreak changing given that attribution may be determined in single cases, which used to be considered sporadic?
- Where is FDA funding for sequencing and database construction coming from?
- The National Academy of Science should be involved to review WGS systems before policy is made.
- A public meeting on WGS systems among the agencies would be appropriate.
- Engage trade associations (e.g., Food Marketing Institute, Grocery Manufacturers Association, National Restaurant Association) that are downstream users of meat and poultry industry products.
- Can a company search the NCBI database for their isolates or isolates from their suppliers?

Concerns

- What are the legal ramifications of testing and the data?
- How will consumer groups use WGS data?
- What are the implications of predicted virulence potential?
- Industry stakeholders and/or their trade associations need to schedule time to meet with FSIS and FDA.
- How is WGS changing outbreak investigation strategies? Will epidemiology play a more significant role in outbreak investigations?
- FDA is "retrospectively" sequencing isolates; FSIS is only sequencing newly collected isolates.

Writers:
Haley Oliver, hfoliver@purdue.edu
Zaid Abdo, Zaid.Abdo@colostate.edu
Steve Ricke, sricke@uark.edu
Reviewers:
Henk den Bakker, Henk.C.den-bakker@ttu.edu
Matt Stasiewicz, mstasie@illinois.edu
Tommy Wheeler, tommy.wheeler@ars.usda.gov

Funded in part by the Beef Checkoff.