Draft to be submitted to Microbe. 06/05/2012

Meeting Report: Breakouts on Outbreaks.

Marc W. Allard[1]*, Steven M. Musser[1], and Eric W. Brown[1]

* Correspondence: marc.allard@fda.hhs.gov

[1]Division of Microbiology (HFS-710), Center for Food Safety & Applied Nutrition, U.S. Food & Drug Administration, 5100 Paint Branch Parkway, College Park, MD USA

On March 1st and 2nd, one-hundred thirty scientific experts in bacterial genomics and molecular epidemiology, representing 57 institutions around the world, gathered for a workshop in Arlington, VA, entitled, *Disease Outbreak Detection in the Genomics Era: A Global Road Map Forward*. The meeting was hosted by researchers from the U.S. Food and Drug Administration's Center for Food Safety and Applied Nutrition (CFSAN) and jointly coordinated by public health scientists on both sides of the Atlantic including FDA and The Danish Food Institute at the Technical University of Denmark. The majority of those in attendance were experts in the areas of global public health, animal health, infectious disease genomics, bioinformatics, and computational sciences and represented 24 different government agencies across 12 countries including The United States, Canada, Denmark, Germany, The United Kingdom, Sweden, New Zealand, France, Portugal, Japan, Mexico, and China. The primary purpose of this meeting was to define and develop solutions to questions and challenges surrounding the deployment of next-generation DNA sequencing tools for public health and disease outbreak detection on a global scale.

Whole-genome sequencing (WGS) technology is contributing long anticipated solutions to what were once viewed as insurmountable challenges in the genetic analysis of bacterial pathogens. Complete genome sequences from multiple bacterial strains can now be collected and analyzed in just a few days, underscoring the potential of this technology as a molecular epidemiological tool to assist in disease outbreak investigations. Recent examples in the literature illustrate the ability of WGS to discern high-resolution genetic relatedness of otherwise indistinguishable isolates based on the genetic changes that accrue within individual bacterial strains. Proof-of-principle studies have been undertaken successfully using the technology at the U.S. FDA, the CDC, Northern Arizona University (Dr. Paul Keim's laboratory), Public Health Canada, Harvard and Cornell Universities, The Sanger-Wellcome Trust in the United Kingdom, The Danish Technical University and Danish Food Institute, The University of Muenster in Germany, and various industry colleagues engaged in WGS technology development.

The first meeting responsible for organizing a coordinated and global direction for the use of WGS in the public health arena was hosted by the Danish Food Institute/Danish Technical University and occurred in September 2011 in Brussels where a select group of 30 experts from around the world assembled for two days to adopt and endorse the concept of a single global pathogen identification and tracking system based on WGS technology. In full, eleven articles were adopted by the committee that *in toto* provided a conceptual framework for a movement forward using WGS as the basis for a global pathogen identification network. (http://www.genomeweb.com/sequencing/next-gen-sequencing-shows-promise-public-health-faces-technical-political-social). The meeting was framed in the following summation: *"The*

rapid advancement of genome technologies holds great promise for improving the quality and speed of clinical and public health laboratory investigations, and for decreasing their cost. The latest generation of genome DNA sequencers is now capable of providing highly detailed and robust information on disease-causing microbes, and in the near future these technologies will be suitable for routine use in national, regional and global public health laboratories. With additional improvements in instrumentation, these next- or third-generation sequencers are likely to replace conventional culturing and typing methods to provide point-of-care clinical diagnosis, providing essential information for quicker and better treatment of patients. Provided there is free-sharing of information by all clinical and public health laboratories, the comprehensive understanding these genomic tools provide on infectious disease agents could spawn a global database or a system of linked databases of pathogen genomes that would ensure more efficient detection, prevention, and control of endemic, emerging and other outbreak occurrences world-wide."

The follow-up meeting in Arlington, earlier this month, shared the Brussels accords with the larger global public and clinical health and health policy community and provided a more detailed plan forward for countries capable of deploying this technology. Specific objectives that were addressed included an expanded follow-up to the 2011 Brussels meeting and a detailed debate concerning the short and long term obstacles and solutions for a global system for identification of microbial pathogens based on genomic information. The meeting also provided an overview of ongoing initiatives in this area and discussed how worldwide collaboration can be achieved to establish a globally distributed WGS system.

During the meeting, a draft road map forward for establishing a global disease outbreak detection system using shared genomic information for bacterial, viral, and parasitic microorganisms was developed. Workshop participants provided insight and perspective to a number of obstacles facing the successful development of such a system. Several important issues addressed at the meeting included: (1) where and how WGS data will be stored and curated for the world health science community; (2) the appropriate metadata to be attached to genome sequence submissions for disease detection and identification; (3) the specific computer resources required to implement a global genome-based disease detection network; (4) recommendations for data analysis pipeline design and determination of data types and categories to be included in such a database; (5) pinpointing potential political and legal restrictions for the sharing of genomic data on a global level; and (6) identifying essential steps for formatting of data necessary for point-of-care clinical utility and rapid detection for all aspects of public health.

To facilitate discussion on these important issues, an international cadre of subject matter experts addressed several topics related to the event including applications of WGS to solving foodborne and clinical disease outbreaks and other aspects of genomic epidemiology; available web-based tools and data management systems for desktop analysis of WGS data; health policy perspectives and the importance of global data sharing; and US and world regulatory perspectives on WGS technology.

Other highlights disclosed at the meeting included (1) a leading role for The National Center for Biotechnology Information (NCBI) at the National Institutes of Health to develop a portal to upload sequence read archives (SRA) of draft genomes and provide a rapid pipeline for identification and clustering to other draft genomes in the database.

This also will include rapid annotation and a brief report of important features relating to multidrug resistance and virulence in these pathogens. NCBI would work with global partners (EBI and DDBJ) to create a global network to upload data from local international laboratories. Agilent reported a collaborative effort with UC Davis, BGI, FDA and NCBI to sequence 100,000 genomes of human pathogens at the new genomic facilities at the UC Davis veterinary school. This genomic sequencing collaboration requested isolates from the attendees for draft sequences to help produce a large reference database to be housed at NCBI SRA with critical and informative metadata attached to produce a valuable publicly available genomic resource. Below we list the majority answers for the breakout questions that were asked to the participants.

Question. How would the development and availability of a global WGS pathogen database be used by the following stakeholders: (I) clinical point-of-care users; (II) molecular epidemiology public health users; (III) public health policy users. Clinical point-of-care users would want a database that is applicable to the clinic and would want information connected to the genomes that could someday be replaced by genomic data such as antibiotic resistance of the species and strain, virulence characteristics and linkage to other phenotypic and biochemical testing typical to the clinical work-up. Where available genomics should provide pathogen typing (VNTR/MLST) for existing databases. Clinicians want the

databases to hold actionable items for example whether the genes stx1 versus stx2 are present in the bacterial genome, Virulence profiles, AMR gene content with connected treatment guidance and known validated SNPs. Most importantly global clinical point-of-care users need point and click technology in data formats that easily integrate with their LIMS systems. The ultimate goal is to add the genomic actionable information to a report that goes to the clinicians. When sharing a genomic database numerous concerns were voiced about having high quality for using validated large scale clinical databases. If a suitable database can be built and maintained MD's would want to know how many other MDs have seen the same collection of markers. This database may be especially useful for the identification of new and emerging pathogens. Examples of successful databases share by clinical professionals include clinvar and the HIV database. To be especially useful this database needs to be populated with the necessary data to make treatment decisions, to predict clinical outcomes, and to assist in the management of patients. Hospitals are seeing value in understanding routes of transmission to implement infection control especially if these are endemic pathogens found in their facilities.

Molecular epidemiology public health users would use a large public genomic database to assist with investigations and in the confirmation of local outbreaks by determining the scope of the outbreak with more fine scale clustering. The goal is to define the transmission and determine what can be done to prevent and or minimize an outbreak in real time. Public health professionals expressed the need for a 2 tiered system and there was vigorous discussion about what to make public versus private. It was expressed that If each entry included information on who submitted the data and a case number was provided than private information could be discussed offline through the usual channels. The first questions a database should address are

have we seen this pathogen before and if so what metadata is associated with the sample to assist in locating the source of the contamination. The primary goal for this group is source attribution and trace-back of pathogens to monitor the spread of the outbreak and to reduce its impact.

Public health policy users wanted to use the database to prospectively identify the next pathogen or clone that is out there and rapidly recommend treatments. They expressed a need for a sentinel global system to assist in identifying what is circulating out there and what pathogen may be coming next as an outbreak or pandemic. The idea is to use the database to be ready with treatments such as antibiotics or vaccines for the new or emerging pathogens. Like the clinical professionals they are concerned most with patient confidentiality and a clinically validated database that includes virulence profiles, AMR gene content, relevant SNPs, and source attributions. Public health officials want to know if the pathogens are endemic to their nation or coming from outside, and whether they are newly modified or older endemic strains. Broader goals voiced include using the database for global outbreak investigation and response in identify and implement national interventions to aid in regulatory decision making. A carefully vetted genomics database could assist in global cooperation on infectious disease events. Cautions were voiced however concerning what metadata should be publicly versus privately made available for the WGS pathogen database as this powerful information could lead to the shutdown of a critical company for a nation, or broader trade barrier decisions, or even evoke international travel issues to name just a few policy decisions. Powerful information should be carefully vetted and released so it is a difficult problem about what metadata to

share even though the benefits are equally valuable. Policy makers are more broadly trying to make risk assessment decisions and source attribution is a large part of these determinations for resource allocation. Some broader goals may also include: using the database to assist in adherence to international health regulations; to help develop education programs; and to develop better international communication strategies. A database with historical samples also would help public health researchers globally in understanding such things as how antibiotic resistant genes have spread from animals to humans, or its environmental effects. The bottom line is that building such a database could potentially help governments to put better public health policies together.

Question. In the case of whole genome sequencing, what kind of data should we filter, query and analyze? Should we focus on changes in core genes only or all variable sites, or should we focus on kmers and snps? It was thought by some that there will need to be different filters for different questions. For example, for outbreak detection and trace-back SNPs will be critical, but for species and serotype gene presence/absence will be useful. Speed of analysis is also critical especially for clinical reporting or for outbreak investigations. Timely delivery of useful data is more important when people are getting sick than understanding the detailed evolution of traits. Several participants recommended mining genomic database to populate existing databases such as re-sequencing arrays, MLST and known SNPs. Some attributes for good database design is the need to be self-organized, having a tiered system, with known define useful targets that are well validated and having a consistent annotation. Until clear standards are in place, the community recommends maintaining the raw read data like the NCBI SRA. A

tiered approach might include a k-mer approach for microbial identification and a more fine scaled SNP approach for sub-lineage identification and clustering. Good consistent gene annotation will be critical for identification of gene function. The regular and consistent curation of the database is essential and thus many participants warmly accepted the support to be provided by NCBI and other international database partners who are well trained and staffed to do this work.

Question. What is the minimal standards of quality that we should expect from genome sequencing data to be shared in a pathogen surveillance database? How can we ensure data integrity? Rather than defining a minimal standard the opinion is that any data can be submitted as long as the quality is attached to the SRA data file as currently required at NCBI. This also goes for the quality of the metadata provided by people uploading draft genomes. Raw data reads will be required to be uploaded so that annotation, assembly and confidence scores can be calculated in a consistent way. This will also be technology agnostic and allow for future advances in genomic technology. This strategy puts the power in the hands of the end-user to sort out what level of coverage fits best for their particular application. While the system will allow for lower quality contributions, the community still emphasizes the need for continued development of high-quality reference genome sequences. As the field of genomics testing matures quality control parameters for sequence quality should be established and used in proficiency testing and accreditation for centers of excellence. This quality level is largely based on the reproducibility of the data and generally for clinical draft genomes is around 30-40x coverage. For foodborne pathogens the qualities scores are similar with >4 passes, >95%

SNPS detected and 15-20x coverage. Several participants are actively involved in the capture of raw data and metadata from multiple sources to facilitate the most common possible analyses, present or future. Global standards of data quality control and strain nomenclature need to be coordinated with already established consensus standards organizations. National inter-agency cooperation on data management and interpretation is recommended as is a transparent set of standards for curation that are public availability.

Question. What is the desired metadata important to be attached to genomic data submissions from the clinical health point-of-care perspective, the public health perspective and the health policy perspective? A long list of desired metadata was discussed by participants including: passage information, sample type, timeline of potential outbreak, time and geography of isolates, sample source (human, animal, food, blood, fecal), environmental conditions (temperature and salinity), GPS/lat/long+/-, clinical presentation (actual clinical values), co-morbidities, co-isolates, structured searchable editable format (ontologies), treatments, outcomes, phenotypes, desirable patient information such as age, gender, environment information, specimen type, phenotypic data (AST), infection site, sample source, sample description, part of outbreak, outbreak data, storage location, contact information, and publication. Also, to integrate with current metadata requirements for existing databases it was recommended to closely integrate with standards like GSC, Genbank and ClinVar.

Question. What is the level of IT interface that should be expected to utilize a global pathogen WGS network? Many participants wanted web-based, drop-down menu driven, point click

systems for end users. Database should allow for both input and export to facilitate extraction of data by specific investigators who want to do additional analyses. For input, five fields of minimal metadata are recommended for all entries but optional fields will be essential. A great challenge is verifying and standardizing the optional metadata fields, which is work currently ongoing by many participants involved. It was recommended to encourage very high submission quality but not to require it initially. The community would carefully monitor the technology and recommend enhance submission requirements as capability increases. To be a global database there is a wide spectrum of needs requiring multiple interface options. For example, some advanced bioinformatics laboratories and clinics want command line driven formats for rapid uploading of draft genomes out of their existing LIMS systems while the developing world health encourages slow connection tolerance for bioinformaticly challenged communities. There is also the challenge between public health verses researcher needs, where the former wants automatic graphic interface reports on actionable known genetic information and the latter wants full access to the data for more detailed analysis. Some clinics are moving toward an Interface that gives users choices for different answers whether they only want the results for a specific test or answers for all known genetic variants. Usability, stability, portability, free, open access, and the ability to mine multiple databases are all important attributes for a global pathogen genome database.

Question. If we are to develop a public health surveillance toolbox for molecular epidemiology and personal medicine point-of-care toolbox for endpoint users, what analytical output tools would be considered helpful in such a tool-box? Numerous tools were requested by meeting

participants including: clustering features; queriable default report of summary information; mapping of metadata onto the phylogenetic tree for temporal analysis; transmission network for likely path of dissemination or evolution; translate genomes into useful or signature biomarkers; listing of virulence markers; “geo-genomic” mapping of geospatial information with incidence and frequency of sequence type and including custom drawings of local site information; emerging pathogen modeling-predictive tools and requiring global connections; lists of phenotypes-core biochemical signatures and antibiotic resistance markers across broad instruments in use, clinical outcome for submissions; that the database provide ways to drill down into complex data strata; to have some tools that measure the confidence of the results; and to provide intuitive graphical presentation to indicate what is it (e.g. speciation markers); whether it is dangerous (e.g. pathogenicity index); a summary of resistance gene information; whether we have we seen it before (i.e. strain relatedness and phylogenomics); and naming feature (geogenomic IDs). The goal is to put a system in place that compress the time line between the initial outbreak occurring and its global notification so it is critical that there is an easy mechanism to alert public health officials. It was clear to many that this will need to be developed as more users weight in on what is needed and what is working. Hope was expressed that this process will offer more guide-lines for surveillance. It should help connect the dots of individual cases. For many real time data should be available for query.

Question. What support do we expect from the world’s major sequence curation centers such as NCBI, EBI and DDBJ. Participants requested standardized submission and curation of WGS data supported by clinical grade databases, data archiving, data quality control/validation,

development of new genomic analytical tools, ensure open access and "searchability", provide programmatic interface to mine data quickly, provide tools to facilitate submission, maintain metadata associated with sequence data; annotate genome differences. It was also recommended that there be a two pathway system built; a separate conservative-fast path for outbreak response, and a detailed-slow path for more in-depth analysis. It was also recommended that more pathogens should be included in the reference database, as well as more non-coding regions should be annotated as some are thought to be very important (e.g. CRISPRs). An important issue for an international data base is that the research conclusion should be constantly updated.

Question. Who might sequence for those unable to acquire sequencers? Would core sequencing centers be a useful approach for those who cannot? What about the future of twinning between larger centers and smaller labs with access to critical isolates. In general it was recommended that large sequencing centers use their excess capacity to involve smaller partners by following the GDD sites setup already in developing countries though this likely will require additional funding. Collaborations between scientists are still an important way and bringing developing countries into the network for all applications as well (i.e. ReAct Network). Twinning seems like a very good idea though more planning is needed for defining how sequencing countries would find and join up with non-sequencing countries. The area laboratories concept will be important where several regions are served by one larger sequencing center. Pilot international projects should be set up with the help of international activities agencies such as between US FDA and some other international partner (i.e. Mexico).

All genomes would require the isolate to be deposited and thus it is important to include both collaborating centers as well as reference labs.

Development and deployment challenges were discussed in our third breakout session. Question 8. In Brussels, several barriers were identified to participation and data sharing Including : sequestration prior to publishing; regulatory fears across international boundaries; patient privacy; and Industry concerns (i.e. food production facilities) for self –implication. Please provide mitigation strategies for each of these barriers. In general, how can we foster active international participation in contributing to this database? It was judged that one pressure for delay is a natural expectation for holding data until a paper is accepted for publication, so one solution is to attach strings to funding from government and other scientific funding agencies to encourage early release of data. HHS already has sharing policy already in place for providing access to data into GenBank in the human genome project and a similar model could be adopted here. For industry adoption one solution would be to allow people to prove compliance by acquittal from high resolution source typing technology like WGS. Other solutions are well known such as continuous education and benefit to augmenting economic burdens from outbreaks and importation bans; reportable foods registry is already ongoing for industry so we could add the submission of the isolate for the WGS database. The community also can write documents that clearly explain what governments and industries and growers would get out of such a system such as fiscal benefits or work burden for proving compliance, and resource allocation support. Sequestration prior to publishing is primarily an issue for academics and less so for public health laboratories. To insure patient privacy it was

recommended that a genome database could be model after pulsenet. To allay industry concerns, participants encourage collaboration with industry to support compliance and consumer trust. Sequence data can absolve as well as implicate which is important as the current blunt tools used to link industry to outbreaks is not rapidly providing detailed source attribution like genomics has shown to produce. Agencies could also collaborate with product liability companies and insurers to incorporate genomics into preventive controls and guidance.

Some of the delays in global surveillance are under our control such as speeding up the CDC clearance process, access to important strains from field labs and the NCBI data accession process times for uploading data. As these are stream-lined a more rapid response will ensue. For more sensitive data it is recommended that standard security be put in place such as IHR, controlled access to databases; export control for DITRA select agents, standard RHISC security can be applied. Readily available bioinformatics tools will improve the speed of analysis. We could also make agreements for international brokers for collection and distribution of sequence data (data escrow). It was deemed that personal relationships in the scientific community, working with the regulatory community will improve the process. We also must be proactive about building these relationships across international boundaries. Standard methods are to remove PID or restrict access at NCBI, with the simpler of these is to keep PID separate and link back to the data using a isolate identifier.

Much of the support requested from an international database is already in place. We recommend following the NIH guidelines for sequence communication. Once sequence passes QC give

a limited delay (i.e. 45 days) before public released of the data. Currently, there are different time expectations for providing metadata associated with genome data but this should also improve with minimal metadata initially associated with data upload.

What metadata is shared is also very important and may vary from country to country, so it was requested that a global database make it easy for the users to contribute some data with the upload. Also, users request the ability to easily update data files with new or modified metadata as some data may be too sensitive to release during an event but later will have historical value. Some participants also recommend that some data be embargo of up to a year after deposited in protected area for the most sensitive data. Most journal requirements are to have datasets made available upon publication, and during review. It was widely recognized that regulatory and international considerations require more sensitivity. In general more limited metadata will be shared due the proprietary nature of the data, particularly with regulatory cases or CDC and outbreak cases. In many of these events a strain ID number could be attached to any new draft genomes and all data will be associated with a particular government agency so that if matches are found by others during comparisons then those investigators can contact regulatory groups to solve these events in the existing ways that they are handled currently. Public database will be an image of certified version but may have different information available. Patient privacy has a separate yet similar set of privacy issues such as HIPAA to make anonymous versus de-identified versus limited versus clinical and identified datasets. There are also healthcare institution and system concerns as well as potential changes per the ANPRM were projects require consent up front for any use of clinical sample and data. Industry can in a similar way participate and not self-implicate through anonymize or de-identify data.

Question 9. What additional political or economic challenges do you envision aside from the above? Should those who seek access to such a network also be expected to actively contribute

in populating its database? The general opinion to the above question was no as this depends on funding and a strong case will need to be made under many of the budgetary constraints that are here or predicted for most countries and agencies. Some countries or industries may pool resources for sequencing and populating a database. Either way, we should keep the database open and accessible for all. It was widely viewed that we wanted participation from developing countries by recommending collaborations between developed and developing groups. This includes recommending support through philanthropy such as the Gates Foundation. Interactions between developed and developing regions can be a win-win situation and collaborations should support commercialized data analysis and encourage tech transfer. Sequencers have to share credit with those who generate and provide materials. Federal agencies also need to develop an outbreak MTA or IRB agreements with partners. Added Values and carrots for physicians should include covering costs, as well as economic ROI with benefits defined.

Question. As part of a road map forward for the ultimate establishment of a genomic pathogen sequencing and tracking system, what would you envision for a real-world pilot proof-of-concept deployment of whole-genome pathogen surveillance? How large? What stakeholders? Who should be included? What are potential funding sources? What are the key pathogens or sectors (i.e. foodborne outbreaks and clinical transmission)? One recommendation is to copy existing roll outs of new technology into regulatory laboratories. For example, some suggest following the example of the Luminex deployment assay starting with training and deployments of technology. Regionalization for an early network will be important so that one sequencer will

serve several states or agencies. How do we get core sequencing laboratories on board in big medical centers and in government core centers that currently can generate many pilot sequences quickly? Working within the established countries and working at the national level may facilitate budding networks. We will need pilot studies to convince the political systems that further expansion is of value. We need to establish the objectives of what pilot projects should deliver. This can be modeled from some pilot projects that are already in place, such as existing international collaborations and consortium. Several phases of preparation and roll out are envisioned. Simulations should be run for emergency preparedness because this enables Data Acquisition Coordination Centers (DACC), bioinformatics, clinical microbiologists and infection control experts to speak to each other in sharing the data. Proficiency testing of the same isolates sent to multiple centers for sequencing and analysis is also standard methods validation. It may be easiest to start with certified government (i.e. US FDA, USDA, CDC) and other larger well-funded agencies in countries ready to begin now, and later expand to state, provincial and local academic and industry partners. We can start by defining pilot tests to evaluate the system's performance and scalability. We need to define clearly what data is to be communicated under different use cases for consistency across sites and comparability of genomic data and metadata. Funding for these projects should be similar to RFAs to ClinVar, NIAID and other funding mechanisms. Others that might support different aspects of this work might include: AHRQ, DoD, DoE, DARPA, DITRA, European agencies, WHO, Gates Foundation, NSF, NIH, CREES USDA. Primary users of this database might include outbreak isolates that also involved in veterinary and human disease, and present in environmental reservoirs (*e.g. Salmonella, E. coli, Campylobacters, Listeria*), military health care system and emerging infection surveillance partnered with host countries. By putting sequencers into field laboratories, we can jump start a global genome network. To make these

systems efficient we think that they could be dovetailed the network with existing PFGE PulseNet laboratories and databases that have been using other technologies for similar purposes. The future looks very bright for these new genomic technologies though there is much work and validation needed to implement a global network.