ILSI 2021 Annual Symposium Session 4: Advances in Enhancing the Microbiology Safety of Foods

Transcript of the presentation, Applications and Future Potential of Next Generation Sequencing in Food Safety, Marc Allard, PhD, Food and Drug Administration (FDA), United States

Thank you very much. And so, everything I'm going to be talking to you today about is using genomics to provide safe, wholesome, and sanitary foods. And you heard about the Kelly talked about the CDC and PulseNet, and essentially what I want to talk about is the landscape that were FDA and USDA also share all of our pathogens, all of our genomic data. And essentially, and then we also uploaded to the SRA at NCBI. And so essentially one of the first lessons we learned is the need to share your data in a timely way. And the sooner you share it, and the more metadata, the bigger impact you'll have on reducing public health problems. And essentially a one health approach genomics can be compared across, regardless of hosts, they can be directly shared. And so, it integrates with this one health method.

The... Where's my slide here. So once the data's at NCBI, it can also be used by many, many other software tools. And so, because NCBI can only do so much, they're sort of thinking about it in a global way. And they're also storing and sharing the data and essentially curating it for the long term. And these other groups can all add value to their particular stakeholders and their particular questions. And so, and this includes...

Particular questions. And this includes global databases. Some of these are wide open and others, you have to be a member to participate, but this way we get more value for the data that we upload. So, in particular, the FDA wants to make sure that all the data can be optimized for use across many platforms. This idea is talked about as the fair principles, which means findable, accessible, interoperable, and usable. And so, this is sort of a constant work in progress of keeping standards up and as new platforms and new data files come out in the research industry is to integrate those so that the data can be shared across multiple platforms. You mentioned Galaxy Trakr. What FDA has done is we've used an open platform galaxy as an open platform, a cloud-based computing, and the Galaxy Trakr is just software that FDA uses that are essentially tools, custom workflows. These are all the things that the Genome Trakr members do. They check the quality of the data. They can upload. They can do many of the basic pipelines.

This allows anyone in the world to fully see in a transparent way exactly what our data analysis pipelines are doing. And so, not only is all the genomic data publicly available at NCBI, but all the bioinformatic methods are transparently shared. And then lastly, we've been starting to use, this is a work by Ruth Timmy and Maria [Bulky 01:08:16], the protocols.io is a place where we can share all our literature and share all of the protocols, all of our standard operating procedures. Essentially, anyone with a sequencer, anywhere in the world, can essentially plug in and use these tools, get access to how we do it. And then, they can also either be a part of our network or they can do it independently. They may want to do work that they don't want to share. This covers all kinds of levels of protocols include how

you do the quality checks, how do you do the metadata, which is the description of the isolate that you're sequencing, and then detail on data submission and data curation. This way we have standards and so fewer people make mistakes and you receive data in a sort of organized way.

Essentially, the lessons learned is that when you do the standardized methods, you see fewer problems and, essentially, open tools, particularly open software tools, promote global adoption, because generally they're less expensive than buying piece of software, maintaining your own computational costs. For all of our partners, we just pay the computational fees for running the Galaxy Trakr. In this case, there's no cost to our collaborators by sequencing and uploading their data and participating in the Genome Trakr, we found that this is a good way to, essentially, provide more bioinformatic tools to our partners.

There are other cloud-based options. This also, you could give a copy of all the software to give to another group, another food safety agency, and they can either run it themselves and pay those computational costs, or they can be part of our network depending on what we work out. It also controls the data analysis for version control because a lot of this software has a lot of choices. And so, this is a way that we can define and verify a platform as doing what we think it can do and then share it so that different groups aren't using different versions of the software. And so, this is mostly genomic epidemiological analysis tools that we share, as well as quality.

The total number of sequences in the Genome Trakr database has continued to rise exponentially. I think we just broke 600,000 foodborne pathogens. There's another 200,000 of other pathogens, and this, I want to reaffirm, this isn't just FDA. This is all of their partners, the CDC, the FDA, the USDA, but it also includes all of our international and domestic partners, as well as anyone who essentially publishes a salmonella genome and puts it in GenBank, that goes into the database. We're upwards to 12,000 isolates a month or more. We continue to expand, essentially, the diversity of pathogens and the total numbers of pathogens. I will say we primarily focused on bacterial foodborne pathogens, but one of the future goals is to start also collecting more foodborne virus and parasite data. But we were just at the beginning stages of that.

At NCBI, essentially, the first application is that the phylogeny identifies novel linkages for outbreak detection and infectious disease control. And so, if you upload genomes to NCBI, they produce a daily cluster of the very tips of the tree. This is only the very tips of the tree within 50 snips or less, but this is the application that FDA and their public health partners are most interested in. Here's an example where an FDA inspector swabbed and found two positives at a mung bean sprout facility, and we uploaded that in the cluster shows that it's clustered with some clinical illnesses from PulseNet. All these pNUSA are the PulseNet USA clinical samples. And so, there's really three things we're looking for, is the first is the new inspection data positives.

Did they link to anything? Any other food or any clinical? That may be the first tier. The second would be, has it then made anybody sick in the past, or... I mean, once again, this is a hypothesis that the contaminated food may be getting and making people sick, but even an old isolate suggests that there's a risk that that pathogen has caused disease in the past and may in the present. And then, the third sort of tier of concern is whether there are current recent clinical cases linked. And so then, that means that we need to really dig in and look carefully at this cluster. At NCBI, there's over 36,000 clusters examined daily, about 4,000, and maybe almost 5,000 of the clusters has an FDA isolate sitting in the cluster. The clusters, the first piece, and this is what genomics provides, fewer clinical cases are needed to have confidence that they are interlinked.

This early unambiguous determination of the scope of the cluster. The clusters that don't group in that didn't share the same common exposure to a food and any unrelated clinical isolates' evidence that there's an independent co-occurring contamination. It's either completely independent or there's poly-clonality, meaning multiple lineages of pathogens that come from the same contamination event. As spoken by—a lot of this will overlap with Kelly Hise' talk earlier. It essentially allows previously indistinguishable serotypes and PFGE types. It allows it now to be tracked and genetically linked, and essentially, this higher confidence incentivizes early actions. I mean, early actions, both by the food producing firm, as well as the public health agencies. Essentially, the genomic evidence, what I would call the microbiological laboratory piece of evidence, the clustering, is just a part of the evidence that's typically used to identify an outbreak vehicle.

There's also all of the epidemiological evidence, which is once they define the scope of the outbreak, they can determine the association between the illness and the food exposure. And then, the third piece of evidence is the actual, the additional trace back and investigators. This is food inspectors going back to facilities, and essentially, trying to determine the common source, and more importantly, determining the root cause of the illnesses, how the pathogen is getting into the food supply. This may lead them depending on where the first positives occurred, or they may go back to a processing plant, and that may lead them back to a farm or even specific ingredients. The FDA inspectors are using the genomic data and their own inspection to look for the pathogen. This is what we call trace forward and trace backward because the FDA wants to know, essentially, if there's a contaminated ingredient. It may have been shipped to multiple food production facilities.

And so, you'd want to check which of the facilities... Some of those facilities may have received contaminated ingredient, but they controlled it. Whereas other food production sites or places, somehow, it got past their preventative controls. And so, we want to know, not only where did all the potentially contaminated food gets shipped to, but also going back as well. And so, this is a detailed part, and essentially, for any outbreak, they would want to go do this kind of study in a post-outbreak response to understand what happened and how the pathogen entered into the food supply. Essentially, the source tracking and root cause event investigation does a very important thing. It identifies whether this is a resident or a transient pathogen. We know that the fields are not a sterile place. There are animals going through here, there's water and soil, there's pathogens on the farm. You can't keep that from happening really. It's difficult.

But what can happen is that there could be a farming practice that essentially causes a manmade occurrence of creating more pathogens or that you could control. This is often directly related to animal waste, fertilizers, soil amendments, and the treatments of water. Within the facility itself, if you get one of those pathogens from the field get lodged as a resident pathogen in the facility, then you have a serious problem because if it's in equipment then it can regularly sort of spill out and get onto the food product, and then the company's actually producing more contaminated food because of this activity. Determining the resident and transient stature of the pathogen is important and then figuring out a preventative control to interrupt it either on the farm or in the food production.

Essentially, the increased identification of harborage and root cause of contamination is directly supported by genomics. The FDA notifies the responsible food facility, essentially, that they think there's a pathogen and that they're the cause of the problem. This increases awareness, and awareness incentivizes the food manufacturers to invest in preventative controls and essentially keep investing until they have preventative controls at work. And so, the new root cause discoveries also improve

knowledge of reservoirs and risks. Not just to that food industry, but the FDA will publish this in there they're good farming practices and good manufacturing, food manufacturing practices. It doesn't just affect the one particular firm that had a problem, but it notifies the whole food class. All the almond growers or all the leafy green folks will, hopefully, learn from one incident. And then, they have the Office of Compliance.

They're watching this. Just like Kelly talked about, the rep strains. We're watching known strains from past events and, essentially, we're watching that. And if that preventative control somehow fails again, or it's not really working 100%, then we'll see those pathogens show up in the clinical community or in food and we can rapidly trace back to its original source. Essentially, the combination of zero tolerance for many of these pathogens and the power of whole genome prediction is reducing foodborne illness in the U.S. so these tools are working. Some of the time, you can't figure out what the root causes and from just an initial inspection. And so, some of the time, the FDA gets involved with research, which are essentially longitudinal studies. And this would be, in this case, this is a study that went from 2010 to 2014 in the Delmarva Peninsula, where they collected many samples, essentially on the farm, looking for where the pathogens reside.

Where can you find pathogens? What are the risks? And in this particular study, all the green mean they found no salmonella, but the red were places where we regularly found salmonella. This was in the creeks and sediments and most standing fresh water. Some of the animals that are regularly in the water, like the waterfowl, you could find the pathogens. And then, this got us a very careful study of, well, if the risk is in the standing fresh water, then how are the farms using water? And what are the preventative controls in place to prevent foodborne pathogens in the water from getting into the crops? This helps with guidance, but it's typically a long-term surveillance and research. We also work with food communities, like the Leafy Green Task Force, and they're academic collaborators.

All of our data publicly facing all of our data's at NCBI, and we talked about it. You don't need a log in or account. You can go and see any of this information. This is our publicly available data. I wanted to point to the right side of the slide. There's lots of information here about anti-microbial resistance monitoring. They have whole databases of the 5,000 genes known to cause anti-microbial resistance, and you can actually get a daily check of what new genomes have come out with new anti-microbial resistance monitoring genes. All the clusters are here as well, but this is a set of tools that I encourage you to go and look at. One of the things we've been working on carefully over the last few years, but in particular, is the metadata. One of the important things to do is, essentially, fully describe your metadata in a standard way.

And so, we have various isolate attributes, but it includes the collection dates, the species, essentially who collected it, some geographic area. We typically list that's down to state. And then, the source is a critical piece. We break everything, either as a human clinical or a food or environment. This is a way where we can then have automatic software that says, "When do we see a match between a clinical and a food or an environmental sample?" And then, more detailed information about the labs and the, exactly, what kind of environment it was taken from. These are all deposited through the NCBI system of bio projects, and this is what's called a bio sample. The more we can organize the metadata and standardize it, the more powerful it can be used for statistics. And so, this is GenomeGraphR is a risk prediction software that was created with collaboration between France and the U.S. FDA and some academics.

This uses this ontology, where you say, "Is it food, or is it environment?" If it's a food, "Is it a plant or an animal?" It goes all the way down to a particular kind of fish or bivalve that you can... And there are different kinds of ontology, but essentially, this makes all of that data more valuable for risk assessment and any kind of statistical analysis. Essentially, recently we went back, we didn't do this originally, but we've been back to all 90,000 FDA isolates in the NCBI, and we added the Food Ontology and the IFSAC Ontology.

But there are other ontologies as well, the GenEpiO, it's exactly where in the environment do you collect and how you call, how you name the food. This allows the FDA to say and make predictions. Is there more risk from tree nuts or is there more risk from leafy greens?

Those results may help us decide what to do. I just want to say that it's not just anti-microbial resistance in the AMRFinderPlus [inaudible 01:27:21], there's also lots of genes related to other genes relating to stress and virulence, such as the persistence of a biofilm, or the resistance to a particular kind of disinfectant, quat resistance, chlorine resistance. These genes can also be called for every genome uploaded into the database. This is another set of tools that we hope to help the food industry to understand their resident pathogens and to then also look carefully at preventative controls to, essentially, control this natural evolution, or well, some of it's selected, it's living in these environments that some of the bacteria become resistant. Another thing we can do, and this is sort of future vision is, we know how to do genomics. We know how to do source tracking and root cause investigation.

We can, essentially, educate a broader community on the power of these genomic tools. And so, we have a big program the next three years in building food safety and capacity through the Asia Pacific economic cooperation. This is a longstanding global network that we can leverage to essentially adopt genomics and to get them to all be integrated in this great food safety. Here's a 21 economic... Some of these are countries and others are collaborators, but across the Pacific rim, and we'll be doing a lot of meetings like this to provide awareness, but then it'll get down to surveillance and actual laboratory training and bioinformatic training. The last thing I really want to talk about is, the question is why should you pay for this expensive genomic technology? And so, we did an economic evaluation of whole genome sequencing programs in the U.S. and, essentially, we're comparing two data sets. We have all of the data at NC...

Speaker 2:
Mark. We've lost your voice.

Marc Allard:
... has been watching. It's two data sets, NCBI and the clusters and the metadata. And then, the other data set is the NCBI's tracking of all the outbreaks and the food associated and the number of people. What this shows is that as we increase sequencing, essentially, we see fewer illnesses. We see a reduction in the burden of illness, and this occurred a little slowly in the beginning. It takes one or two years for your investment to pay off. But by 2019, we saw a reduction and the burden of illness of almost half a billion dollars. And I suspect today, it's probably a much closer to a billion. The pathogens, you can see the Listeria is showing the greatest amount of reduction in the burden of illness. Essentially, what this shows is that whole genome sequencing is definitely an important link to food safety and the most heavily sequenced pathogens, those illnesses are decreasing faster that related to other pathogens.

Essentially, the heavily sequenced pathogens are getting smaller. The outbreaks are getting smaller, though we're seeing more of them. We think this is related to faster and more precise outbreak investigation with all of our collaborators within the U.S., and even considering the additional costs the program likely paid for itself. In addition to the direct effects on public health to program increased accountability, it increases effectiveness and efficiency of our compliance and enforcement, and it facilitates root cause analysis, essentially risk assessment and risk management. We believe that this would work in any other country, even though their own food safety culture and programs may be a little different. Like I said, we're not covering all of the foodborne pathogens. We're only sequencing sort of the top ones, but there's even more impact that we could potentially do.

And then lastly, we're showing this evidence, we're saying, if the size of the circle represents the amount of funding, if you want more maximum benefit, then you need to do more funding in the food and environmental sampling and surveillance. Essentially, as all the speakers have shown, WGS is reproducible, accurate, it's increases in surveillance, should uncover more reservoirs, and essentially, we're incrementally improving our food safety. This will compound to fewer contaminations that reoccur every root cause we solve and instituted preventative control. And then, there's all these other tools that Hank was talking about, emerging technologies like laundry portable, [minaian 01:33:25] technology, metagenomics. These are all mobile devices. These are all possible things. We know how to transfer this information. In fact, WGS, can should be expanded to other human, animal, and plant pathogens.

Essentially, it can also be translated and transferable to other infectious disease control applications like hospitals, nursing homes, medical manufacturing, waste management, composting, agricultural use and reuse, and many of the genome tracker labs pivoted and used their equipment and data sharing and uploading directly to address pandemic response. In fact, the FDA is going to also be sequencing for COVID-19 from sewage waters connected to, and to protect food safety workers around the U.S.

Anyway, thank you. This is a huge endeavor with a lot of people. This is just an acknowledgement for the water surveillance. And then, there are lots of whole agencies where we're directly involved and collaborating. Thank you.