UNITED STATES DEPARTMENT OF AGRICULTURE

+ + +

USE OF WHOLE GENOME SEQUENCE (WGS) ANALYSIS
TO IMPROVE FOOD SAFETY AND PUBLIC HEALTH

+ + +

October 26, 2017
8:00 a.m.

U.S. Department of Agriculture
South Building, Jefferson Auditorium
14th & Independence Avenue, S.W.
Washington, D.C.

USDA:

DR. DAVID GOLDMAN
Assistant Administrator
Office of Public Health Science
Food Safety and Inspection Service
U.S. Department of Agriculture

CARMEN ROTTENBERG
Acting Deputy Under Secretary for Food Safety
U.S. Department of Agriculture

DR. UDAY DESSAI
Senior Public Health Advisor
Office of Public Health Science
Food Safety and Inspection Service
U.S. Department of Agriculture

DR. PETER EVANS (Moderator)
Office of Policy and Program Development
Food Safety and Inspection Service
U.S. Department of Agriculture

Free State Reporting, Inc.
1378 Cape St. Claire Road
Annapolis, MD 21409
(410) 974-0947

SETTING THE STAGE

DR. MARTIN WIEDMANN
Gellert Family Professor in Food Safety
Cornell University

DR. DAVID GALLY
Professor of Microbial Genetics
University of Edinburgh

DR. NORVAL STRACHAN
Chair in Physics and Chief Scientific Advisor to
Food Standards Scotland
University of Aberdeen

FEDERAL/STATE COLLABORATION

DR. JOHN BESSER
Deputy Chief, Enteric Diseases Laboratory Branch
Centers for Disease Control and Prevention

DR. STEVEN MUSSER
Deputy Director for Scientific Operations
Center for Food Safety and Applied Nutrition
U.S. Food and Drug Administration

DR. DAVID GOLDMAN
Assistant Administrator
Office of Public Health Science
Food Safety and Inspection Service
U.S. Department of Agriculture

MR. DAVE BOXRUD
Molecular Epidemiology Supervisor
Minnesota Department of Health

DR. PATRICK McDERMOTT
Director, National Antimicrobial Resistance Monitoring System
Center for Veterinary Medicine
U.S. Food and Drug Administration

DR. WILLIAM KLIMKE
Senior Scientist
National Center for Biotechnology Information
National Institutes of Health

DR. GLENN TILLMAN
Chief, Microbiology Characterization Branch
Office of Public Health Science
Food Safety and Inspection Service
U.S. Department of Agriculture

DR. CHRIS BRADEN
Deputy Director, National Center for Emerging and Zoonotic Infectious Diseases
Centers for Disease Control and Prevention

ALSO PARTICIPATING

DR. MUNA NAHAR

DR. MELANIE ABLEY
Food Safety and Inspection Service
U.S. Department of Agriculture

DR. CHRISTINE ALVARADO
U.S. Department of Agriculture

DR. CATHERINE CARRILLO
Scientist
Ottawa Laboratory (Carling)
Canadian Food Inspection Agency

MR. STEVEN ROACH
Director, Food Safety Program
Food Animal Concerns Trust (FACT)

DR. MARC ALLARD
Center for Food Safety and Applied Nutrition
U.S. Food and Drug Administration

DR. ALEX BRANDT
FSNS

DR. BETSY BOOREN
OFW Law

DR. JORGEN SCHLUNDT
Director
Nanyang Technological University Food
Technology Centre, Singapore

MS. SHERRI McGARRY
Centers for Disease Control and Prevention

JOSEPH HEINZELMANN
Neogen

INDEX

INDEX (cont.)

M E E T I N G

(8:07 a.m.)

DR. GOLDMAN: Good morning. If I could ask everyone to please take their seats. We're about to begin. Hopefully you can find a seat.

Good morning. My name is David Goldman, and I'm one of the Assistant Administrators here at FSIS, the host of this meeting, along with our multiple partners, and I want to introduce to you our Acting Deputy Under Secretary for Food Safety here at USDA, Carmen Rottenberg, who wants to provide you with the official welcome. Thank you.

MS. ROTTENBERG: Thank you, David. I want to thank all of you for coming here today and also for those of you who are participating by the webinar. This has been a busy week in the life of the Agency. We have had the privilege of hosting two public meetings about really important information that furthers our public health goals and really shows the deep levels of collaboration that we have with our public health partners.

As we've heard the last couple of days,

whole genome sequencing technology has become a routine part of the NARMS surveillance screen for resistant genes in enteric bacteria. And, as you all know, whole genome sequencing is now regularly used in outbreak investigations, but always in the context of other available evidence.

Many of you that are in the industry have been asking us to have a meeting like this for the last year, year and a half, and we have a really robust agenda to go through with you with really talented scientists and really folks from all areas where whole genome sequencing touches.

We at FSIS nearly completed building our own whole genome sequencing capacity, and we do intend to have WGS fully implemented into the sampling programs this fiscal year. So the timing of this meeting is appropriate, and there's been an incredible amount of work on behalf of the agencies to put this together, and I just want to thank all of the folks who have worked on that.

It's our intention at FSIS to analyze the whole genome sequencing data using validated and

transparent methods, which is also why this 2-day public meeting is so important.

As we move forward, with utilizing whole genome sequencing, we continue to collaborate with our public health partners, and I think that you're going to be able to really see that showcased herein the next couple of days.

So I didn't want to take a lot of time this morning, but I did want to welcome you all and thank you again for coming, and again thanks to all of the speakers who are going to be sharing with us the next couple of days.

So thank you and welcome.

DR. DESSAI: Good morning again, and before we start, I'd like to kind of do a few housekeeping things.

All right. So a couple of things. No food or drinks are allowed here. Number 2, you have your restrooms on this side in the Fifth Wing, and the cafeteria is in the Third Wing. So you have the reception over there. If you have any questions, please feel free to ask for help.

Now, all the time those who are non-feds should have their badges when you are moving around in the building.

With that, I just want to say this is an exciting time and especially this meeting is very important, and like Carmen said, we are in a phase here where technology has come to a level where we can think of inserting some components of that, possibly into the regulatory process. And later on, you will see from various speakers where we are today, where we will be in 5 years, where we go in 10 years, and how this technology can take us or help us go to our 2030 goals potentially.

So we have a lineup of speakers who will basically take you through what is whole genome sequencing and what it can do and how it can take us where we want to be. What is whole genome sequencing and risk connection? And many topics which are of interest as well as where you need clarity in terms of moving forward.

So we have three speakers in the first module, and that is called Setting the Stage, and

these threes speakers will basically lay the ground for you.

The first speaker we have is Martin Wiedmann from Cornell. The second speaker we have is David Gally from Edinburgh Institute, UK, and Norval Strachan is our third speaker who will be talking about risks and WGS.

Now, Norval and David have traveled. So if they are sleepy at times, just kind of bear with them because they have 6 hours of jetlag, okay.

Having said that, a couple of things to keep in mind. There are two units here. One is for you guys here. The second is it's a webinar. So your speeches are streamed. The slides are being streamed, and keep in mind please speak in the speakers so people on the line can also hear you very clearly.

Now, we don't have a clicker yet. So what you have to do, the speakers is, just indicate to our wonderful folks here, and they will change the slides for you. Okay.

Having said that, over to Martin.

DR. WIEDMANN: Thank you. Good morning, everyone. So what I'm going to do is give sort of general overview of the whole genome sequencing.

Next slide, please.

What I want to emphasize, before we get started, in more into details, is that we really need to look at whole genome sequencing and omics tools in a grander context of what might be actually pretty disruptional technology changes in food safety, and I'm going to call the place that we're going to go to, precision food safety. It's not very innovative, but I think what I'm going to try to sort of impress on you is that we can't look at whole genome sequencing technology by itself.

We need to look at it in terms of the other changes in tools and technologies that we will have available to assure food safety and how these tools will help us to improve food safety from improved outbreak detection, to improved source tracking, improved identification of pathogens and will help us to move from a typically reactive approach to a much more proactive but ultimately predictive approach.

And some of the other tools that are important in this framework are really the ideas of machine learning, artificial intelligence, GIS or global information systems technologies, that will give us this greater precision in food safety, and it will work along with this whole genome sequencing.

Next slide.

So here's what I'm going to run through today. I'm going to provide a brief overview and then go into a few case studies or a few areas where whole genome sequencing already has a considerable impact or will continue to have a major impact.

Next slide.

So first I'm going to just go over some sequencing technology, data analytics at a sort of very high level to set us all up at the same page.

The reason we're here is that sequencing technologies have developed dramatically from first generation, Sanger type sequencing, that were fairly costly and did allow us to do whole genome sequencing but not anything what we can do now, to next-gen sequencing tools that some people divide into sort of

a second generation, with Illumina being the work horse of what we do today, to now moving to third generation sequencing tools, and these are the tools that allow us to do large-scale whole genome sequencing, but they also are used for applications in food safety that are not actually whole genome sequencing. People sometimes use WGS and NGS, next-gen sequencing interchangeable, but I'm going to try to separate that out a little bit for you.

Next slide.

The innovations we see, and we're going to continue to see are probably best illustrated with this picture. We're not just going to scale down costs, but also scale down size of the equipment. This is a picture of the MinION or MinION, however you want to pronounce it, which is really where I see and think a lot of others see the future of whole genome sequencing, that's miniaturized to a spot where we can get sequence data more quickly with this smaller equipment and, you know, continuing to reduce cost. So this is sort of the future and you can see the scale of this sequencing equipment here.

Next slide.

So as I said, whole genome sequencing has been performed as traditional sequencing but the workers are these next-gen sequencing tools, but the important thing is that next-gen sequencing can be used for other applications.

And some examples of, there are metagenomics, which I'll quickly touch on, where you can basically sequence all DNA found in a sample. All genetic material, take it all and sequence it.

And you can also use these same tools to highly parallel sequence many genes at the same time. We're used to PCR-ing one gene and sequencing it. You can do highly parallel sequencing, that can address some of the issues with whole genome sequencing that we're going to get hits, we're going to get information that we don't know what to do with. So we can use to target hundreds of thousands of genes and really only look for pre-defined targets.

And then we move into the area of RNA sequencing where you don't sequence all the DNA but

you actually sequence RNA. RNA is an unstable genetic material. It's not consistently unstable, but it might give us some better ideas of what these organisms do and may also help us in some cases differentiate live and dead organisms.

So these are all applications of next-gen sequencing, but they're not whole genome sequencing per se.

Next slide.

I'm going to move a little bit into the data analytics now, focusing really on, you know, application of whole genome sequencing to bacterial genomes. It can be used for parasites. It can be used for viruses, but the main application that I want to emphasize is really whole genome sequencing of bacteria.

So the average bacterial genomes that are relative to food safety range somewhere from 2 to 10 Mb per organism. So that's the range of what we deal with.

And if you look at bacterial genomes, we typically can different the genomes in terms of a

core genome. So if you take something like *Listeria*, these are the genes that are found in all *Listeria monocytogenes* and the accessory genome. These are hundreds of thousands of genes that are found in some *Listeria* or some organisms, but not others.

These can play very important roles because they might provide an antimicrobial resistance as we mentioned before, but at least in a number of the analysis we use, we may ignore and not use that information. So this is a set of data that can provide lots of very valuable information that may not always, and I emphasize, not always, be used.

Also the question always comes up, you know, so you scientists call this whole genome sequencing. Do you really sequence the whole genome? And, the reality is, we sequence the whole genome but we don't assemble the whole genome. So very often there are certain pieces of the genome that are missing for a variety of reasons including genes that you have repeated multiple times. So, yes, it is not the whole genome that we analyze even though we call it whole genome sequencing.

Next slide.

So as we look at sequencing, if you have this 2 through 10 Mb per genomes, as we sequence it, it's basically a giant puzzle. We take this genome, cut it into small pieces, this is what you see on the top, and it goes through some experimental parts to make these pieces ready for sequencing, put them on a sequencing platform and then we end up with millions of pieces of DNA that we now need to put back together, which is what we call assembly.

Broadly speaking, there are two different approaches to it. One of them would be reference guided assembly, where we say, this is a *Listeria monocytogenes* as we sequence. Let's pick a similar *Listeria monocytogenes* and use that to put these pieces back together to get a sequence that we can analyze.

Now, we can also analyze these SNPs by itself but in most cases, at some point at least, we're going to do some sort of assembly.

And you can do a *de novo* assembly, where you put the pieces together without that sort of

template reference guiding and then obviously relies on a more massive computational power.

Both of these tools are used. There's pros and cons. I'm not going to get into all the details of it, but often it's a combination of the two of the -- different approaches to assembly, but typically they will lead us to very, very similar answers.

Next slide.

The big discussion when we talk about whole genome sequencing, what most people look at really as an output is single nucleotide polymorphisms, SNPs, and I'm sort of simply going to explain those, and this picture, what we're really looking at here is a difference of one A, C, T, and G. As you see in the top, you might have one sequence that has a C, the other one that has a G, and if you have that over 3 million nucleotides of *Listeria*, you say these two *Listeria* differ by one SNP.

Now, there's a similar difference or another difference that you can find between two genomes that is also important as we look at the analysis used, which are insertion or deletion

indels. And we call them indels because you really don't know whether the C in the sequence on the top was deleted and therefore got us to the reference or whether the reference acquired it. So it could be an insertion. It could be a deletion. It's a chicken and egg question.

When we have these differences, they're often referred to as single nucleotide variance as opposed to SNPs because it's a variant but it's not a polymorphism. So if you hear people talk about SNPs, it typically does not include indels unless people use the nomenclature somewhat imprecisely.

Next slide.

The simplest presentation of how we analyze these data is here. We have four sort of hypothetical sequences. In red, you can see differences. So you can see Isolate 1 and 2 differ by a single T that is marked in red in the middle, and the outputs you will typically see from these are trees, shown on the left, where you can see that 1 and 3 are identical, 2 is similar to it, and 4 is more different, or I can do what we call SNP

matrices, where we come to pairwise comparisons, and I can look at 1 versus 3 and I see a 0 meaning they're the same, there's no SNP differences, or I can look at 1 versus 4 and I can see there's 3 differences.

And you look at this with a few sequences, it looks pretty straightforward. I can tell you same or difference, right.

Next slide.

But as we expand this and not just have four isolates, but have 100, 1,000, 5,000, 10,000, and we build these trees, they obviously become more difficult to interpret. Now, we have to curl them up, so we can put them on one slide. Otherwise, no one can see them, and even when we curl them up, no one can see them. So it makes interpretation and use of these trees difficult and challenging.

Obviously, you can zoom in on sub-trees of that and just look at a specific plate, a specific subgroup of interest, but even that can become challenging partially because some of these branches will change as we add new information and new

isolates to it.

So next slide.

So one approach that is used and is increasingly used and gets us around, is this idea called multilocus sequence typing. I call this whole genome multilocus sequence typing. There's really two subsets to this. One of them is whole genome MLST that uses all the genes. The other one is core genome MLST, and as I mentioned before, it's a core genome. Those are the genes that are common to all organism. In the case of core genome MLST, I will just use these core genes that are common to all organisms.

Those tools are being used, are being further developed and different groups look at different tools. Obviously, when we look at our core genome MLST, we will lose some information because we're not using some of those genes. They're only found in some isolates, but having core genome MLST makes some things in terms of analytics simpler.

So how does MLST work? We have a database. We define unique loci. So those will be genes but

also some other similar types of structures such as non-coding RNAs, and then any change in this gene, whether it's a SNP, whether it's a insertion or deletion, equals a new allele and we can name these alleles.

You can see this most easily on the bottom. So we have hypothetical locus 1. We have allele 1. Allele 2 differs by 2 SNPs. So we call it 2. It's different. The next one differs by 1 deletion from allele 1, also differs from allele 2. *Therefore,* we call it allele 3.

In this scheme, the numbers don't indicate similarity. You cannot go saying, well, this is 1 and 2, so they're similar, while allele 1 and 100 are very different. The numbers are simply added as we identify alleles. So 1 and 2 can be very different, and 1 and 100 might only differ by 1 indel.

Next slide.

So what we can then do on a larger scale and what is done at a larger scale is we put all these data together. I've got your hypothetical 3 isolates, A, B, C. I've got your number of loci. So

locus 1, all three isolates are the same. Locus 2, isolate C is different. So it's named number 2. Same for locus 2,005 and we then look at all of them together. So if two isolates are the same in all of these loci, they're given one single name A or a longer numerical designation.

So it takes this mass of information or relatively mass of information for genome sequencing and really drills it down to one number or a set of numbers whereas we will hear later as type of zip code. Okay. And so it really takes this information at least as a superficial level and makes it much easier to deal with it and compare between labs, to tell me same or different, and we can even try to come up with some of these naming schemes to make them so that the similarity of numbers at least gives us some idea how closely related those isolates are, just like with zip codes.

So these are the two high quality SNP and MLST are the two sort of core tools that are currently predominantly used to really analyze this whole genome sequence data.

Next slide.

Now, regardless of which method we do, one of the core questions that it comes down to is, so how quickly do two isolates become different, and to me that's one thing that's cause many of the questions. And these are a couple of datasets that were derived on reasonably large datasets to get us at this question.

So the first one is a large study on *Listeria monocytogenes,* which estimates about 2.5 x 10^{-7} substitutions per site, so per nucleotide, per year. What does that mean? You're going to get one SNP difference in the core genome every 2.5 years. That's a reasonably long time for some people and for evolutionary biologists, it's probably a pretty short time.

If you look at core genome MLST types, when does a type become different as organisms multiply? In *Listeria*, it takes about 0.2 alleles per year. So that means in 5 years, an organism will change.

That obviously has huge and important implications for use of these tools, right. If we

have the proverbial goose that flies from Florida to Canada, along the flyway, and leaves little drops of poop all the way along with *Salmonella*, that *Salmonella* is probably the same all the way along the way, and went it comes back, it's still the same.

If you look at a breeder flock in poultry, and we have *Salmonella* in that breeder flock and that *Salmonella* is transmitted across the chain, down to a slaughter house and maybe a retail establishment, that *Salmonella* probably could stay the same and we could find that identical *Salmonella* in different locations. I think that's an important thing that we really need to consider.

Another estimate on *Salmonella* Cerro, a specific *Salmonella* Cerro type, you see it's in the same range of substitution rates that were estimated. So that gives you *Salmonella* genome slightly larger, gives you some of the same ideas.

So the challenge here is that evolutionary biologists will talk about most recent common ancestor. That's when two isolates have this most recent common ancestor defined by sequence. Two

isolates with one SNP difference, that most recent common ancestor might be 4 or 5 years ago or longer, and that's why we need these tools along with epidemiology and other evidence to interpret it.

Next slide.

So here's the summary on the data analysis. Obviously, it does involve many steps, many of which I skipped over. We have done considerable advancements in standardization and transparency of those tools which will probably be covered later.

We have a number of different approaches for analysis, but in most cases, they will lead us to the very same answer.

The other important thing to remember, once you have a whole genome sequence, I can analyze it with a high quality SNP, I can use the same raw data and run a core genome MLST, whole gene MLST, whatever I want to.

And the important, you know, caveat there is obviously that these analyses establish recent common ancestors but do not establish relevant epidemiological links. They point us in the

direction, they can help us support some conclusions but recent common ancestor, at least in my book, does not equal cause and effect.

Next slide.

So now I'm going to move through some of the sort of areas where whole genome sequence has significant impact, and the first and foremost is obviously outbreak detection.

Next slide.

And we've had PulseNet where we use basically barcode type fingerprints to track foodborne diseases to do surveillance for a long time, had tremendous impact in improving our ability to detected foodborne disease outbreaks, which you'll hear more about later.

Next slide.

But there's been some challenges with this, and this is a slide from some work we did that probably illustrates the best. This is a number of *Salmonella* Montevideo isolates that we collect by PFGE. When you look at them, they all look the same. Only problem was that some isolates came from

pistachios in California and some came from an outbreak, and it was linked to sausage produced in a facility in Rhode Island and raw isolate from the pepper, but they're all the same. So very, very hard to use these data to sort of give us some insight.

When we did whole genome sequencing on this, shown in next slide, all the isolates that came from the pepper, the sausage and associate human cases, clustered in this tree on the top right. We put them all together. We can clearly show that they are different even though they have all of these isolates shown in this tree have the same PFGE pattern.

More importantly, as is shown with the red circle, we find a set of five isolates that are very closely related by SNPs that are short time frame in the same state that likely represents another outbreak within this large cluster of similar PFGE types they would have never recognized with PFGE alone. But with this whole genome sequencing, we can separate them out and see that.

So this illustrates, in one very specific

example, why whole genome sequencing is so powerful for outbreak protection.

Next slide.

The long-term impact of this is probably best illustrated by *Listeria* where we have a lot of history using whole genome sequencing. You can see before '97, before PFGE, one outbreak every 3 years. Once we started using PFGE, about two to three outbreaks a year. Once we started whole genome sequencing, somewhere in the range of 7 to 10 outbreaks a year. One outbreak every year, let's say 10 outbreaks a year, 30 times more outbreaks detected per year. It's not that we had more *Listeria* problems. It's simply it detected outbreaks better.

The important part is that there are the size of the average outbreak went from about 7 year outbreak to somewhere for 3 to 4 per outbreak with whole genome sequencing today. So we detect more outbreaks, but we detect smaller outbreaks. That means we detect earlier on. We detect outbreaks that we previously would not have.

Next slide.

And so obviously this came along as routine implementation of whole genome sequencing in September 2013.

Next slide.

One other organism that this is going to have a major impact on is going to be *Salmonella* Enteritidis. *Salmonella* Enteritidis, about 50% of *Salmonella* Enteritidis have the same PFGE type. PFGE, the routine surveillance method. Molecular surveillance is not very good at differentiating Enteritidis. Once you apply whole genome sequencing to this organism, shown in the next slide, we're going to differentiate these isolates with whole genome sequencing to the point now where we can detect larger number of clusters and ultimate outbreaks very similar to what I've shown you in Montevideo.

So at the beginning of this paradigm shift, we have seen it in *Listeria,* but I think we will see some tremendous impact on *Salmonella,* and I predict *Salmonella* Enteritidis is one of the organisms that we will particularly see these impacts but also other

Salmonella serotypes.

In addition to better detection of the outbreaks, whole genome sequencing also provides tremendous opportunity for better trace-back, and food processing plants that can be used by both industry but also is used by regulatory agencies.

One early example of this that actually goes along well with this sort of data on SNP differences was some work we did in collaboration with CDC published in 2008 where we actually sequenced *Listeria monocytogenes* from a listeriosis case and a listeriosis outbreak that happened 12 years apart but were linked to the same facility. We sequenced the *Listeria monocytogenes* from those two years, and we found that at least some of them 1988 and 2000 isolates, 12 years apart, differed by a single SNP. Very stable, probably because this was a ready-to-eat food processing facility survived in this facility over time and differed and changed very little over that time.

Now, if you apply this to processing facilities, and this is an example of a non-whole

genome sequence based data, but it illustrates the point, you can see with subtyping and here we colored different subtypes, provide different subtypes in different colors, you can in this picture see which is a 2-year surveillance of a processing facility that we have a very specific *Listeria* shown in green, more or less survive in this facility over 2 years. That's obviously a challenge, and if you do proper root cause analysis in this case, we could traceback the persistence of this *Listeria monocytogenes* to a very specific location in the facility, which ended up being floor mats which obviously provided us with the opportunity to just remove these floor mats and see if our hypothesis was right, and lo and behold, once we removed these floor mats, that type of *Listeria* was not found in the facility over a 6 month follow up.

So it shows the power of using whole genome sequencing and subtyping data to traceback and identify sources that's in a facility or throughout a supply chain.

Now, that is obviously one thing that, you

know, causes some people concern and heartaches and headaches, right. And this is sort of the hypothetical case study I want to speak through, talk through, if you take this to the next step.

Let's say you have a facility that has a *Listeria monocytogenes* positive in a finished product, one day's production. Typically you end up with a recall of one lot that was produced that day.

Let's say 11 months later, you have another positive for the same facility, a different type of product, *Listeria monocytogenes* again. Now, you're in sort of a tougher spot because it's like it could be a repeat positive of the same subtype. It could be a separate issue. If you now have whole genome sequence data on the January and December isolate, match by whole genome sequencing, let's assume zero differences, that will lead you down to the conclusion that that *Listeria monocytogenes* probably perhaps persisted in that facility.

Obviously, you need some additional information. Obviously, could be a reintroduction in that facility, but a conclusion that some people

could draw and might draw is that any of the food produced within January and December was produced under unhygienic conditions. If I'm going to reintroduce the same *Listeria* time after time, that a food processing facility has under control, that's arguable.

The challenge then becomes when we take this sort of information and extrapolate to non-ready-to-eat foods, let's, for argument's sake, say raw poultry or raw meat, where we can now have reintroduction, *Salmonella* is endemic. It's found regularly in poultry farms potentially, in dairy farms where we get ground beef, and so it could be truly a reintroduction. So it could not be an issue with the facility that cook upstream.

So it illustrates that we can't just extrapolate from ready-to-eat facility to non-ready-to-eat facility. We need to consider the overall supply chain as we interpret these data and really do our epidemiological investigations, even when we're just talking about food contamination, not human disease cases.

The next one I want to move to is how to use whole genome sequencing to better understand and define pathogens.

Next slide.

So the case study on this one is something that some of you may be familiar, large recall of Fonterra because they found *Clostridium botulinum* in their powder. The story there was that after 4 to 6 weeks later, they suddenly discovered it wasn't *Clostridium botulinum*. It was actually *Clostridium sporogenes*. How did they identify that?

Ultimately probably with whole genome sequencing. The challenge here is that *Clostridium botulinum* and *Clostridium sporogenes* are very, very similar. Unless you do PCRs or mass experiments, you cannot differentiate them. If you use whole genome sequencing, you can differentiate these close-related organisms very quickly, very easily. This is not just a one time deal.

We published this paper recently where we had a similar incident. We found a *Clostridium* species that some people could have worried about

being *Clostridium botulinum*. With sequencing, we actually found it was a new species that fell somewhere in the proximity of *sporogenes* and *botulinum*.

So a great tool to rapidly differentiate organisms and give us better species classification or classification into pathogen or not. This just doesn't apply to *Clostridium*. Same issue with *Bacillus*, *Bacillus cereus*, *Bacillus anthracis*, *Bacillus thuringiensis*, a group of closely related organisms, difficult to differentiate, but with whole genome sequencing, you can differentiate them quickly and decide food safety hazard, yes or no, and sometimes obviously it's more of a gray zone, but in some of these cases, it's very easy to decide but only if you use these tools.

Now, where it's going to get more exciting than just taking existing pathogens and saying, is it one or is it not, if we now apply these tools to non-pathogen groups, say *Listeria monocytogenes* and say are all *Listeria monocytogenes* the same or can we differentiate different subgroups that are less

likely to cause disease.

The work on *Listeria monocytogenes* I'm showing you here was actually built on some initial work where we found certain subtypes of *Listeria* that were not defined by whole genome sequencing, that were very common in food, about 30% of food isolates, but very rare in human isolates, about 2% of human isolates.

The question was why do these isolates show up in food but not humans? When we looked at DNA sequence data, we could identify a mutation in one key gene in *Listeria*, *inlA* which allows it to attach to human cells. These *Listeria* which were found common in foods, rarely in human cases, had a different form of this protein that did not allow *Listeria* to attach to human cells. *Therefore,* they were much less likely to cause human disease. So now we can take a known pathogen and say they're not all the same hazard. There's considerable differences, and we beat this horse to death with a number of studies including animal experiments, etc., to show that this *Listeria* by a thousand-fold less likely to

cause human disease. So tremendous improvements there.

It doesn't stop at *Listeria*. Here's an example of *Salmonella* Cerro which is an organism which is very common in cattle. And as shown here, we never found it, rarely or almost never found it in human cases. Performed whole genome sequencing on *Salmonella* Cerro, identified a number of mutations in very specific genes that are important for this organism to cause human disease. Mutations there probably means it can cause human disease, consistent between epidemiology and whole genome sequencing, much less likely to cause human disease.

So we can use these tools and hazard characterizations and say, not all *Salmonella* are the same. Some of them we can define probably pretty well that they're a reduced human health hazard. So that's another great application that might be a little bit more in the future, but I think a very important one.

And what I want to end up with is metagenomics. So now we're not sequencing whole

genomes, but sequencing other DNA in an organism, and what I want to envision there, and this future already is there, I want to envision a new type of audit. Now, you audit your facility in a far away foreign country, you collect samples of an ingredient that you source, for example, pepper, you characterize them and then your incoming lots are characterized at regular intervals with that same type of analysis to just find out if what you're getting is similar to what was in the facility, was produced in the facility, when you did an audit.

And what you might end up there, and this is a hypothetical example is hypothetical example, these could be four samples collected during the audit. You get different bacterial species in them, you get a bacterial species profile. This is your first lot that comes in your plant, looks similar. This is your next lot that comes into your facility to test, but then you suddenly get this lot of pepper. Looks different, right. So obviously something is different. We don't know whether it's a food safety hazard or not, but we know it's a

significant deviation.

Is our audit for that facility still valid or do we need to re-audit that facility? We need to know what's going on.

We could look at these data, but we can also use advanced tools such as machine learning to define those deviations. That's where the future is.

So that's where future application of whole genome sequencing and next-gen sequencing will go.

So what are the challenges which is obviously one of the reasons we are here? I've outlined most of them, but I'm going to try to summarize them.

One key challenge is obvious. We can find bacteria with very few or no SNP differences in different locations, food and food associated environments. WGS rarely will give the final answer. It will point us to a certain point, but we need epidemiology, we need other evidence and the evidence needs to be combined.

How do we combine that evidence? Very often this might require new tools, right. Now, it's

preponderance of evidence. We find an unhygienic facility. We find a whole genome sequence match. We find this and we find that, but I think some of these new tools of artificial intelligence and machine learning potentially can help us to better combine this piece of evidence.

Another challenge is metagenomics which detects both live and dead cells. Presence of certain genes is a public health hazard. We find an antimicrobial resistance gene in our food, but unknown if it's a live organism or even a pathogen, do I need to worry about. You may need a new risk assessments for presence of genes.

And then obviously still considerable uncertainty around data interpretation, different data analyses approaches, and these affect industries' willingness and ability to use those tools when they probably should use them. Define two *Listeria monocytogenes*, 12 months apart in their facility, they should use the best tools to find out whether it's the same *Listeria* or not.

In the current climate, sometimes people

are afraid to use those tools.

One of my ideas is that it might be time for a moratorium where we simply allow industry to use these data, not as a fear of having these data required to turn over to agencies, but be able to use them themselves to figure out how they can best use it in their context and then after while, start to come back to discussions about constant sharing of these data. But that's just one of my opinions.

Conclusions: Precision food safety is here. Improved outbreak detection, improved surveillance, improved source tracking, and improved bacterial identification due to whole genome sequencing is the new reality. It's happening already, maybe not at the penetration some of us would like to see, but it's happening already.

The roadmap for other uses of whole genome sequencing and next-gen *Salmonella* is less clear. Will metagenomics and whole genome sequencing replace hygiene indicators? I've given you some of these ideas in terms if you find the same *Listeria* over time, if you find the same metagenome over time, what

does that tell us?

Will whole genome sequencing change the approach to defining hazards where we move away from bacterial species but move to bacterial, to clonal groups, subtypes, presence of genes, presence of mutations, to define hazards? Will that give us better information and better risk-based tools for management of food safety?

No matter where we're going to go with this, these are new tools that will require new people with new training. So it's going to be very, very important that we train not just food scientists, but everyone who works with the food industry and in public health, around food safety, to use these tools.

Thank you very much.

DR. GALLY: Okay. Right. I thank you for the opportunity to speak today. It's been 24 years since I've been in Washington. So it's a long time. I was last here when I was doing a postdoc in North Carolina and drove up here in an old VW Rabbit that kept breaking down, a VW Golf as they're known

everywhere else.

So I'm David Gally. I'm based at the Roslin Institute, just outside of Edinburgh. Famous for Dolly the sheep. We like to pride ourselves in the fact we have expertise in genetics and genomics in livestock but there's a whole host of strange microbiologists that hang out there as well, and they're very interested in the genetics and genomics of bacteria, and certainly that's what I'm really going to carry on the theme of Martin's introduction.

I think you're going to get some similar things that was asked to set the scene, predominantly in a One Health perspective and just really to make the obvious point that in terms of One Health, of course, food safety, we're very interested in transmission from animals to humans, but also where is the role of the environment in that, and certainly the transient role and the persistent role of the environment in transfer of bacteria.

So as you've heard, I'm trying to work out the best way to deliver this with what's been heard.

So the key is that diversity is

understandable from the sequencing. We're getting amazing insights into the diversity of the bacterial world from sequencing, and it's really challenging, the whole taxonomy of bacteria, in fact, but it's a beautiful insight into that diversity.

The bottom line as we've already heard is the precision, where you have an organism that has 5 million bases. We can really look down at differences of just a few, and that's way more precision than we've ever had before. So that's absolutely critical.

Within that before, as you've heard, we can identify very related organisms and certainly identify sub-clusters that are more of a threat to human health, and that becomes important for prediction capacity which is really the second half of my talk, and will really differ from what Martin's told you so far.

The tracking as well, very, very important. To bear in mind, it's not just a one-way street. We've got plenty of examples of where we're getting flowback from humans to animals. So this becomes

very, very interesting. It becomes testable and trackable with whole genome sequencing where we can actually identify flowback into livestock species.

We can therefore also with this precision identify the origins and vehicles of transmission, and really the key, the second half of the talk as I said, will be how can we use this information to improve prediction value of this information? To actually worry about the risk of all subsets, a particular same threat to us and the answer's often no, and we should be able to understand that more in more detail using these technologies.

And, of course, it's not just about tracking whole bacteria. As you've heard as well, we're particularly interested in, at the moment, it's very high on the agenda, in terms of antimicrobial resistance genes and being able to identify and attribute really sources for those and transfer of those. And again that's possible through these technologies.

So, traditionally, we would have had our microbiology, if we're lucky, and grow our organisms

that we're actually talking about which we obviously can't for many, but we have our sample, be it directly from the animal or food -- we can carry out some classic microbiology in the lab and identify our colonies. It's going to take us a little while to maybe determine through subsequent testing, often serotyping, PCRs, additional tests, exactly the subtypes that we're dealing with, and obviously with those bacteria, we can determine antibody resistance using sort of standard plating and techniques.

As we have heard, for whole genome sequencing, we really still need to focus on the first agar plate there. We still need to get hold of our individual isolate and then sequence it, so we know exactly the sequence related to the isolate that we have.

To make the obvious point, that the more information we can get on the sequences of specific isolates, the more we'll be able to type that information into a database and understand what's present when you analyze samples in a metagenomic way. So where you can go direct to the more complex

sample. That's going to be a challenge in terms of the databases and how we share that information which is going to be critical to how we progress in this science.

We then have the analysis side which Martin has described in detail. And what the intent of the potential, the idea is here that from the sequence of the organism, we can get obviously what bacteria it is, what subtype, potentially what virulence genes it has and what resistance genes, and really start to fit it into the epidemiology of previous exposure to that organism. So that's fairly obvious. We've had that covered.

So a lot of the work that we do in Scotland is based around enterohemorrhagic *E. coli* O157, and I'm going to use that and *Salmonella* as my two key examples to explain some of the basic, of some of the more futuristic ways of going about this.

So just to show on the left here, the kind of orangey thing is a cell. We've got *E. coli* 0157 colonizing that cell. You've got the concept that it produces Shiga toxin, and that's the main

pathological determinant in terms of human infection.

It colonizes arsenic. It colonizes cattle and other ruminants using a type 3 secretion system which injects proteins into that cell to help it colonize.

We have about 1,000 cases in the UK each year. It's harder to get the estimates of the whole of USA but well over 10,000, anything up to 50,000, depending on the literature that you read.

Originally known as the "burger bug," but actually in the UK now we have a lot more cases associated with direct contact with animals and cases where produce particularly from vegetables, etc., that have been contaminated potentially with irrigation water and that's the source of human infection.

So with the sequencing, we can get information that feeds back to the very, I suppose, quite straightforward and additional sequencing methods. So shown here on the top left is the inner and outer membrane of the bacteria or the polysaccharide. So the actual O type can be

determined successfully from the sequence. Okay. So that's good. So you can say it's an O157, O26, one of the gang of six, etc. So that can be determined on the basis of the sequence.

Other aspects as well in terms of flagella type, etc., can all be determined that way.

You can also go in and look if it carries type 3 secretion system, affect the proteins, different types of Shiga toxin that are involved in different pathologies, in humans and certain subtypes related to more serious disease in humans.

To some extent, we can relate that to also previous typing methods. Phage typing, we can do to some extent. PFGE becomes more difficult which I'll kind of relate to later.

Okay. So we've heard about single nucleotide polymorphism, SNPs. One way to try and explain the level of granularity we have now, and we've heard from Martin as well, there are different ways of using the information where the genome differs. This is really the core genomic information. There are different ways that we can

use that, but it's about that precision.

And one way that's used by public health England at the moment and others, is to use a SNP address which is shown at the bottom here. We're starting from the right-hand side. You really assign the particular isolate into groups based upon the level of relatedness on the differences of number of SNPs that they have.

Without a simplistic level and Martin has outlined, you have to really know the potential for variation over time with your genome in terms of error rates, but at the moment, simplistically if you're within 5 SNPs, you can consider that you have very related bacteria that may be associated in an outbreak.

And one way of doing that precision, as Martin mentioned, is sort of a zip code way of thinking about it. If you here look at the diversity here say of households, the distribution of households in the USA, this could be the distribution of bacterial variation we have within our *E. coli* O157, and using the SNP address, we can focus that

down to a particular set of states and then smaller regions on that, get it down to a town level, get down to a block level, and eventually to a final kind of household address.

So level of precision is right there in terms of what we can do with that sequence data, and it is way more than we were able to do with these previous techniques, okay, and it's absolutely critical to take that point home.

So how do we apply that with something like *E. coli* O157. What we have here are 2,527 sequenced genomes, arranged in a ring, but it's hard to fit it in, and otherwise, I kind of like that one. The colors represent traditional lineages that have been designated for O157, 1, 1, 2 and 2.

And the branching here is shown at the level of 25 SNP relatedness. It stops there. It doesn't break it down any further than that. As you can see, the relatedness of 250 SNPs, you've got 79 clusters; at 100, 240; at 10, 1,423; and so you can identify an isolate at a specific, precise level within that tree.

So how do we actually apply that? This is an example here from an outbreak in England and Wales, and the key point here is that in the end, there were 49 cases that could be confidently linked to the packed leafy salads when initially there was nothing -- I mean the epidemiology is absolutely critical, but initially you weren't necessarily getting all the cases from the same product source, okay. So you are able to then associate these clusters based on a very related SNP address.

One thing also to say about this is that a year later, there are a number of further cases, this time associated with lamb products and those are the ones shown in pink higher up, and as Martin alluded to, what we can do there is say these are kind of related common ancestors to this outbreak. It does not mean therefore that the contamination may have occurred from lamb or from sheep as a source of contamination and of that produce.

Furthermore, what we're trying to do in the UK is sample across the country. This is from beef farms and the locations of the farms that we're

sampling and then we can get our prevalence studies from this, but it also allows then to begin to associate regions with particular subtypes of O157. So we're actually getting a sort of locality to particular types.

Where that's useful is that we can actually plot on our scheme of all our organisms, those causing human infection, those that are related with cattle. We can understand why they can converge, where we're really getting isolates that are coming from our local cattle into human populations, but we can also spot imported infections as well, where we don't have those particular organisms in the local cattle population or the local ruminant population.

So it becomes very useful for understanding imported threats versus those we're generating locally.

But also we want to be able to predict from this which isolates are more of a threat to human health.

As was mentioned, we ideally want to use as much of the information as possible, so not just the

core level. We want to make sure that we're using core plus accessory genome both, for that prediction analysis.

So this is an example of looking at antibody resistance characteristics, now general *E. coli*. This is a project we're involved with around Lusaka in Zambia, and it's looking at *E. coli* isolates from cattle, small holders, and from a human population. And you can just about make it out, but the cattle isolates are in red, the human isolates in blue in this tree of *E. coli*. And the bars all around the outside are the number of antibody resistance genes. Okay. Shown from *n* equals 1 -- from 0 to 17.

And just from this type of analysis, we can get to see when we have our blocks of human isolates, we have many more significantly higher levels of antibody resistance than we have in these *E. coli* that are coming from cattle. We have again data on what these animals have or haven't been treated with. So all this comes from the WGS data in terms of actually very easy to use databases such as ResFinder

that you can plug either your de novo assemblies or your actual basic reads into.

Taking sort of a step forward, to see what we'd like to do is have more complete DNA information. I appreciate this is a horrible slide. Each of these lines represents a completely assembled O157 genome. Okay. So we have 14 *E. coli* O157 sequences, and this is really what we do with Jim Bono in USDA Nebraska, and this is all based on -- sequencing. So you're getting long read sequencing where you can actually fully assemble the chromosome of the organism.

And the key point I want to show here is these blocks of color are prophages or X bacteriophage regions that at times have moved into the O157 genome. You can see some of them are very similar in very similar places, but a few, especially the green and red ones, sorry if you're color blind, but the ones that lay on the right side, very much more. We worry about that because these particular prophages carry the toxins that cause serious damage to human health.

Now, these are very difficult to actually position and identify using short read sequences because there are very strong similarities between them.

So it really helps to have this type of long read sequencing information to understand the isolates that we're dealing with. I mean examples also of outbreaks that we've had where the number of SNP differences in the core genome might be say 5 or 6, but actually the organism over a year, and this is one example I think that was in a restaurant that had two outbreaks, separated by a year, where the organism had acquired over the difference in time, a plasmid, and it rearranged prophages in its genome. So while at the core level, it was very, very similar, it actually had something like an additional 250,000 base pairs of information. So you have to bear that in mind when you're just using core SNP, SNP-based information.

At another level, we can look at where insertion sequence elements are within the genome, and again Martin gave some really nice examples of

how is the threat the same, dependent on mutations that have occurred. Well, here we have an insertion sequence element that has jumped into the Shiga toxin 2a gene. This inactivates this and it means this strain is less of a threat to human health. So this is impossible to spot with short read sequencing really, and you have to have the long read sequencing to identify it.

Okay. So when I finish off the talk, we're looking at the potential use of machine learning and prediction of both pathogenesis, zoonotic potential and host attribution.

So machine learning has been originally described as the capacity of the computer to learn from experience, i.e., to modify its processing based on newly acquired information. And the first algorithm, first work, we were doing it back in about in the 1930s, pre-computers.

We should be aware that a lot of what you do now, your activity is monitored, right. You are watched. When you type in your searches into a web address, all that's fed back to Google, etc., and

they're crunching that and they're using machine learning approaches to really nail exactly you need that and you -- etc., in your world.

It's used to exam changes in patterns for bank fraud. It's used, pattern recognitions are used, for examining images for identifying tumors, etc. AlphaGo was recently in the news in terms of DeepMind computer teaches itself to become world's best Go player as well. So watch out.

So one of the ways we've used this recently is a supervised machine learning method which I'm going to very quickly take you through with these really stolen from the web tutorial. So very simplistically, two sets of data. Height and weight and we have data for men and women in this case, okay. So very, very binary system, and we have this training data. So in this methodology, you need training data. So this is our training data, and then you can assign a rule that will best separate that training data based on the information you have. Okay. And it's all about getting the rule in the right place in terms of separating, giving you the

optimum distance in terms of separating your data.

You then come along with -- sorry. This is very, very fickle.

You then come along with your test data. So it's new data that you haven't seen before and you plot that and then you apply your rule, and the idea here is you then can assign whether you are dealing based just on height or size in this case, and weight, whether you're dealing with a man or a woman.

Okay. It's obviously flawed, and we get very skinny short guys clearly, but this is only two bits of information. We now start to think about applying that to hundreds or thousands of genes and the prevalence of those genes or the predicted proteins of those genes across isolates. Then we get into the proper world of multidimensional support back to machine analysis which is way out of my league, but we use it.

And the idea here is you're still able to draw a separating line in that data. Okay. So you're still able to assign A to B, even though you've got very complex patterns of data, and that's

the beauty of this supervised machine learning approach.

So recently we've applied that to looking at O157 strains across all the lineages. We have a particular problem in the UK on the left there with our lineage 1 isolates, and that's really horrible and really hard to see, but basically the human and the bovine isolates are really mixed up in the lineages. So it really becomes difficult to say, if you have an isolate that fits in there, is it more or less likely to be a threat to human health.

So we applied the support vector machine process to this, train on the subset, test on the remainder, repeat the process and obtain statistics or prediction scores. So the left axis here, you're looking at a probability. This is actually what you want to come out with in terms of prediction capacity. Probability based on isolates, this is all Illumina sequencing, based on an isolate whole genome sequence of whether it contains more bovine or more human information in terms of its comparison to the other isolates in those groups. Okay.

And you can see here in the green box on the right-hand side, the bovine isolates are on the right, the majority score well for being bovine, but there are subsets that score well into the human zone. And the proposition is that those are the isolates that are more of a concern. They are more of a threat to human health. Okay.

How can we test this? I mean you get what you look for, right. This is really about the concept of it, not necessarily whether it's right or wrong at the moment, but what we can do, if we have large amounts of data, and actually have the metadata associated with that sequence data.

So, for example, I mean one way we've tested this is to take two outbreaks, one was a milk outbreak on the left there, the sort of light blue, and then the pink is a food outbreak. And what you're trying to do there is take the sequences of those isolates and to score them. Now, those that are coming from food or milk or animal, where are they fitting on our 0 to 1 probability score? And you can see that all of those isolates that come from

hamburger, cattle, or milk are all scoring high for human, even though they have that animal source. So, indicating again, this is just a small sample size, the possibility of predicting subsets that are more of a threat to human health.

The way we're trying to take this further, and again complex slide, but this is again using now all *E. coli* that we can get our hands on. We don't have enough yet. We're nowhere near it. We're just dealing with a few hundred. This is standard *E. coli* that come from cattle and *E. coli* that come from humans. We have the whole genome sequences of those, and then used that same information to try and train the machine to learn -- train the computer to come along. If I come along with a new *E. coli*, I say where does this come from? Does it come from a cow? Does it come from a human? And it will assign, it will assign red for cattle there, and blue for human, in terms of the scores. And we can get that about 90% right at the moment based on the small sample set we have.

Again, looking on the right-hand side, the

black dots are where O157 fits in that scheme. So it is very interesting, that something we know that has enough potential, that has the capacity with its gene content to move from cattle to humans, actually generally goes above the intermediate line. It contains genetic content that as far as machine learning goes, can be ascribed to both cattle and human, but puts it up towards the human category.

What's interesting is to understand some of the other *E. coli* in the group and whether they represent a zoonotic threat. On the left there, we can plot the scores for the different hosts as a bar graph. So the bovine score is in red and the human score is in blue.

Obviously, this is the very beginning of things. We need, you know, a phenomenon amount of data to make this more accurate, but it is a start for us.

We can do this as well -- the last few slides are on *Salmonella enterica* serovar Typhimurium. We know for *enterica* we have different serovars that are fairly host restricted or very host

restricted in different cases, but the Typhimurium, there are known subtypes that are much more associated with human, but it's generally considered an organism is able to traffic well between hosts and if we identify Typhimurium, we generally think it has the potential to cause human disease.

So can we apply the same approach to have a look at Typhimurium? The moment we got hundreds of isolates across avian, bovine, human, and swine, obviously there will be other reservoirs where we can train the support vector machine on this and actually then take samples and tests where the host attribution may lie. So what is the likely host for the organism?

And you get these very funky kind of looking graphs here. So top left is the score for looking at avian isolates. What's interesting here is that you can the majority of avian isolates score very well for being avian and actually don't have a lot of other increased color for other hosts.

We're very interested here in which hosts are maybe housing Typhimurium isolates to have more

capacity to move to other animals or colonize us, okay.

Again, I wouldn't want to say this is reality, but this is what's possible in terms of thinking about the way this technology can be applied. If you think about taking a water sample and sequence *E. coli* or something or other that comes from it, and actually say where do these organisms come from, we can attribute the likely contamination in terms of whether that's human or whether that's a local farm up the road. This allows us this degree of prediction, obviously combined with what we already know about the phylogenomics of many of these organisms.

So at the moment, we can use the machine learning to predict the host source and zoonotic potential. We have to prove it. It's still conceptual but it's interesting at the moment that the majority of these Typhimurium isolates show pretty host restricted signals. So it is only again a small subset that really make it into the human domain in terms of the way the genomic information is

being analyzed. And so therefore potentially only specific subsets are a risk to human health.

With larger datasets this type of approach can inform us of where human infections originate from to inform risk assessments.

What we now need to do is go back and ask exactly which genes, which combinations are being used to make these decisions so we can understand the biology behind this type of prediction.

Okay. So, in summary, as you heard from Martin as well, whole genome sequencing is very powerful. It offers -- it's that where we're all kind of using at the moment, transformative in terms of its capacity to track infections and trace sources of bacteria.

Considerable information can be extracted from whole genome sequencing including basic taxonomy and identification of virulence genes and AMR genes, and this helps determine the threat represented by isolates.

But again as the technology improves, the longer read sequencing will give us even more

potential to be specific about this, but at the moment, the costs are very high still for that. So that's really where the short read sequencing can be bolstered by long read methods to more accurately assemble the genome and predict these phenotypes.

The issues are really around getting hold of data related to the sequences. I completely understand that many industries, it really should be released information. How do we use it? Do we use it ourselves? The more information we can have, the more sequencing information, the more related to human disease, the more related to which animal it comes from, time of isolation, place of isolation, can really transform our capacity to be precise about this understanding.

And there are amazing bacterial collections at the moment that are now being sequenced and as we go forward, that information will be available. Again, we're only scratching the surface of diversity that's out there.

And then the obvious point, the more that we have in those databases of individual isolate

sequences, the better the metagenomic approaches can be and that's certainly to sort of reiterate what Martin said in terms of the capacity of the metagenomics as the costs come down, and it's clearly from a diagnostic level, the monitoring level is going to be a clear way forward in terms of picking out the threats that exist within deeper sequencing potentially of air samples, water samples, etc., within facilities.

Okay. Just quickly to thank the fact that the machine learning is really a serious Ph.D. student's work, and this is just written up as Nadejda Lupolova, the University of Edinburgh, and I've been lent some slides from Tim Dallman at Public Health England in terms of the zip code and SNP mapping.

Thank you.

DR. STRACHAN: Okay. All right. Thanks very much for inviting me to speak today. My name is Norval Strachan from the University of Aberdeen. My training is as a physicist, but I worked the last 20 years on risk assessment and molecular epidemiology

of gastrointestinal pathogens, which I guess is the reason I've been asked here today. I'm also the Chief Scientific Advisor for Food Standards Scotland. The views that I'm giving today will be my own.

Okay. So in terms of talk, what I want to do is I want to outline what risk assessment is and the way I understand it and the way it's drafted out roughly by Codex. I want to provide some examples of whole genome sequencing which are applied to the different steps of risk assessment. And then the final part, which I want to speak about and say a little bit about, source attribution, how this relates to risk assessment and basically how we can use whole genome sequencing to help us with that.

So here's a diagram of what I want to say, a little bit about risk assessment. First of all, this isn't food. This is a hazard in front of us. We've got a rock on the top of a cliff. Okay. So because that rock can cause some damage if it falls from that height, the rock's a hazard. A hazard is something that causes a negative impact particularly in our case for considering health.

But also it's well for one to think about what risk is. Okay. So risk, there's really two dimensions of risk.

One is a probability. It's a probability that that rock's going to fall down and cause us harm. So that's one dimension of risk.

The other dimension of risk is the severity of risk. So that's a pretty large rock, and if it fell on my head, I don't think I would get up from that. So it would probably kill me. If it's a much smaller rock, maybe I might survive that. So we need to think about severity as well.

And so what risk assessment is, it's looking at the hazards in a technical, scientific way, looking at knowledge associated with that, to determine what the risk is.

Okay. So we'll move onto something which is a big more in our topic area. So there's a picture there of a beef burger, and it's a rare beef burger. Okay. So you can think about what the hazard is that might be associated with that. So, for example, *E. coli* O157 could be a hazard

associated with that rare beef burger. The vehicle is, of course, the beef burger itself. So a hazard is a biological agent in the food for a potential to cause an adverse health effect. Right.

So in terms of thinking about risk associated with that, we try and think if we eat that beef burger, what's the probability of us falling ill from consuming the beef burger which may or may not have that hazard in that, and then the second aspect is severity as well. So if we fall ill, how ill will we fall? If it's O157, we could have hemorrhagic uremic syndrome, or perhaps we could have mild diarrhea perhaps. So that's the two aspects.

Then this risk assessment itself is using the scientific knowledge that we've got available to us and putting it in a form to make up an opinion on terms of the risk associated with eating that rare beef burger meal.

Okay. In terms of Codex, there is four steps in risk assessment, but before I actually go into those four steps, probably the most important thing actually is a statement of the purpose of the

risk assessment, and this is usually defined by the persons who want to manage the risk. So basically it's looking at -- well, what we're interested in perhaps is what's the risk of consuming the beef burger, but we'd be interested across a population, you know, across the U.S. How many of these beef burgers are eaten? How many fail ill? What's the severity of illness? And so we're hoping that the risk assessment will answer these questions so that the risk managers can then look at this and then decide, yeah, okay, there's not really much there. I don't need to do anything about it or else maybe we do need to do something about it, and then put some mitigation strategies in place to try and reduce the risk.

So the risk assessment itself, the four steps associated with that. There's hazard identification which has already been mentioned. So identifying the hazard, but in this case, it's the *E. coli* O157 in the burger.

There's then also the exposure assessment. What exposure assessment is doing is it's looking at

whether there is a organism in the beef burger that we actually ingest, and that we might fall ill from.

The third part is the hazard characterization which looks at -- there's two aspects in hazard characterization. One is how many bugs does it take for us to fall ill? So we need to think about things like dose response. And the other aspect of hazard characterization is well as the severity associated with those organisms, so the severity of disease that we're likely to get.

In all these first three steps, we can use whole genome sequencing to help us among other things, of course, but in the fourth step, risk characterization, is putting this all together, so looking at the risk across the whole population, for example, that probably doesn't involve whole genome sequencing per se because we will use the database from that.

Okay. So what I want to do now is I just want to go through one or two examples for the first three steps of risk assessment.

So hazard identification itself is

generally qualitative process to identify the hazards. So basically we know some particular types of food, those particular types of agents that we know of previously have causes disease. So it would be the hazard associated with those particular types of food.

However, as has already been mentioned today is that in terms of risk assessment, how we've routinely done it before, we ignore heterogeneity between organisms. We think that all the *E. coli* O157 are the same, all the *Campylobacter* are the same, all the *Listeria monocytogenes* are the same, and we tend to treat them all in the same way, and that's the traditional way that it's been done.

But the great opportunity with the whole genome sequencing is that we are now able to characterize these different organisms. So we're able to find out the variance between them. It makes it more complex but using this knowledge hopefully can help us.

So we're going to look at two examples. The first hazard identification example is one here

which is published by a Dutch Group in the Netherlands, and it's an *E. coli* O157 example and basically what this did, they studied 38 strains of *E. coli* O157 and how they attached to human epithelial cells.

So one of the points we're seeing in this analysis was it caused disease. The organisms need to attach to human epithelial cells. They had basically a model system. So basically they grew the organisms up, they put them through gastric fluids to simulate the stomach, and then through intestinal fluid to simulate transfer through the gut, and then they looked at the attachment of these cells, the microorganisms into the epithelial cell line.

And these are some of the results that they got from the attachment assay. So the graph at the top right there, it shows along the bottom is fractionate adhesion to the Caco cells, the human epithelial cells. On the vertical axis is the frequency. The top graph on the right, this is actually the ones from humans although it's actually misnamed in the paper. We see the part highlighted,

the red circle, it's where there was very good attachment to the epithelial cells.

The example at the bottom for the animal examples that they had, that had a poor adhesion to the Caco epithelial cells.

Okay. So we've got this phenotypic data and all throughout the two talks previously, both David and Martin said the importance of the wet biology being done, but also how we can link this to the whole genome sequencing data. So it sequenced all the data and for *E. coli* O157 which this is for, there's 5.5 Mb. What you need to do is reduce that down to something more manageable. So they reduced it down to SNPs for the core genome and they got about 28,000 SNPs. So that's made it simpler but obviously there's still 28,000 SNPs there. So it's not quite simply enough to deal with.

So what they did next was quite smart which is doing like a genome wide association study and the graph that we've got here, along the bottom, we've got -- basically it's all the SNPs listed along the bottom and then on the vertical axis, if the bar's

higher, it basically means that there's a better attachment of the organism associated with that particular SNP. And there's a little red line along the middle there, I'm not sure what the statistical significance is.

So all those black bars that go above that red horizontal line are the ones that's of interest. Those are the SNPs of interest.

So they initially found 17 SNPs, but then after correction for a sample structure, only 1 SNP stood out as a potential biomarker. So we started 5.5 million basis and gone to 8,000 SNPs and now they're down to 1 SNP, and what they're seeing is this particular one, this particular SNP is a marker for strong attachment in the *E. coli* 0157 cells. The SNP itself was found actually to be an enterogenic area.

So this shows you potentially what whole genome sequencing can do. You can identify SNPs which can give you an indication on whether the *E. coli* can attach to human cells well or not. Further work needs to be done because this was only done with

38 strains. We need to be doing further strains to make this validated.

The second example I want to present on hazard identification is I just actually want to mention the one that David gave already because I think it's quite important. One of the things he actually said with the machine learning work that he spoke about. And what he said was we demonstrate only a small set of bovine strains is likely to cause human disease even within previously defined pathogenic lineages. And if I remember rightly, within the paper it's about 1 in 9 of the strains that he tested were likely to be pathogenic. So this is telling us, you know, although O157 potentially are not pathogenic to humans, maybe 1 in 9 are, and that's important in terms of taking into risk assessment calculations.

So thanks for that, David. You did all the hard work to explain that example for me.

I want to go now and say a little bit more about exposure assessment and give an example associated with exposure assessment. And we're going

to go back to *Listeria* for this example. So what you can think about, looking at production of dairy products. Imagine that *Listeria* can get into dairy products and contaminate the milk within the dairy itself and then we've got a number of steps. So there's a heat processing step in terms of pasteurization. Then it gets packaged maybe into various different products, sent to the supermarket, and then it's eaten and consumed by humans.

So what we can look at, is consider the different strains of *Listeria* that you can find perhaps within the original population of *Listeria*. So think about the wild type first of all, and that's that blue line there.

So the first step in transport, it might grow within that step. Then the pasteurization step, it's been heat treated. There's a big die-off, a big kill associated with that. Then during storage, it may grow a little bit again perhaps and then stomach passage, because we've got low pH there, then it may die-off. So this would be a typical *Listeria* strain which goes through this process, and throughout the

process there's considerable die-off. However, that's the wild type.

But we can think about is the resistant type, and what you'll see in the graph here, the resistant type, it starts at a much lower concentration on the left side, shown in red, maybe grows a little bit during transport. In terms of pasteurization, the die-off may not be so much because it's resistant to heat. In storage, it could increase a bit perhaps and stomach passage may also be resistant to pH. So what you end up with is actually quite a lot more of the resistant type being present at the end of the process, compared to the wild type of *Listeria*.

So looking at this in a little bit more detail, an example that was given in the literature, what we can do now is just look at the graph here, by Metselaar, and this is just the pH example of that. So they've got a culture of *Listeria* which have grown up to a higher level. We then put it into liquid medium which is pH 3.5, so acid for a considerable period of time, up to 200 minutes.

So start off with, there's a big die-off. Okay. That big die-off is from those wild type strains which are sensitive to pH, but then there's this long tail, and this long tail, what you find in the long tail is the resistant strains and those are the resistant strains that are surviving, and actually recultured organisms from the resistant, from the tail, and the 23% then were stable resistant. So this is all very interesting, but what we need to think about now is how whole genome sequencing can help us with that.

So what they did, they sequenced the resistant strains and they sequenced the wild type, the ones that were non-resistant. And so we have this graph here. So at the very top, we've got the wild type, and below we've got all the resistant strains, and this here is a sequence which is upstream from *rspU* gene which is associated with stress tolerance within other organisms.

And what they found in the ones that were resistant by a number of different mutations in this upstream region, and the hypothesis is that these

mutations are causing the strains to be able to obtain resistance.

So what would be interesting is to look at this group of changes, to look across the population to see if this is a common thing in terms of the human population and also if you find those strains within a particular product, to look to see if they are resistant, if they have this mutation as a biomarker to look for that.

Okay. What I want to say very briefly is a little bit of an example on exposure assessment and metagenomics, and again Martin's mentioned very nicely a little bit about metagenomics. The thing about most microbiology studies that have been done previously is that they're based on culture. We can only culture some of the organisms, not them all, and this leads to the biases and, of course, in the real world as well, there are always a community of organisms stand to be present. So metagenomics allows culture independent analysis of microbiological populations.

So there are strengths and weaknesses with

metagenomics. Martin outlined them, but I don't want to go into that here. I just want to go into the one particular example, and this example is an example which is going to be from cheese. So what we have is our sample, and it could be human or whatever, but we're going to speak about cheese.

We're going to extract the DNA or rRNA and in this particular example I'm going to think about is extracting the 16S, extracting rRNA and from that doing 16S rRNA sequencing. And from that what you can do is you can, on the bottom left-hand side there, that small graph, which physically shows -- identifies the species and relative sequences of the microorganisms that are within your sample.

Okay. So we do that, and so this is done for an example in terms of Italian cheese, and what's shown here on the left-hand side, we've got a list of all the organisms or families of organisms that are found. And these are repeated along the diagonal on the right-hand side.

And one of the interesting things with this is what you can do is you can work out which

organisms occur together, which organisms do not occur together. So if they occur together, they're a red [sic] dot. If they don't occur together, they're a blue [sic] dot, and if it's somewhere in between, the colors in between yellow, blue, light blue, white and so on.

And what I've highlighted here is for *Listeria* here, and here we're looking at *Listeria* species along the horizontal and what you can see, lots of blue dots. So it's organisms occur together with *Listeria*, and other vertical down is *Lactobacillus brevis*. So it's very common in this type of cheese to have *Listeria* with *Lactobacillus brevis*. They tend occur together in this type of cheese.

Ideally, what we're wanting is actually red dots there where there's exclusion because if there are red dots, what that means is that maybe the organism *Listeria* may be competing with is producing some compounds that's inhibiting it. So that's what we would be looking for.

If you're able to find those sorts of

organisms, what you might be able to do in your startup culture, for example, is include those organisms so that they are then in your cheese. So you then design a product which is less likely to inhibit any *Listeria* that may be present.

So this is an example using metagenomics where there is a potential for it to help in the design of food products.

Okay. I'll say a little bit about hazard characterization now. So as I mentioned previously, hazard characterization, there are two aspects to it. There's the dose response and also the severity in terms of the human response.

Okay. So those responses both have to do with ingestion of the pathogen and colonization associated with that. We have a graph on the right-hand side there. What I want to show here is just the large variation in doses. This happens to be for O157 but you get fairly similar things for other organisms as well. We've got dose along the bottom and we've got probability of illness on the vertical axis.

There are a number of circular dots in that, and they represent outbreaks. So, for example, the dot at the top right-hand side there, the dose that the people had in this outbreak of between 10 to 4 and 10 to 5 organisms, and 80% of the people fell ill.

You can see, there's a lot of variation. This is from a number of outbreaks from different parts all associated with *E. coli* O157. The best dose response model fits through the 0.50 one in the middle there. You can see a huge range. There's a huge variation that's involved due to differences in the *E. coli* itself, but also there will be differences in the human cases as well because we know humans are of different susceptibilities if you're young or immunocompromised and so on.

The example I'm going to go on and speak about now in particular to whole genome sequencing is about the severity of disease. I'm sticking with O157 for this and for O157 or for Shiga toxin producing *E. coli*.

What we do know is that there are two main

types of Shiga toxin, Shiga toxin 1 and Shiga toxin 2, and these have subtypes 1a to 1f and 2a to 2g. So there's all these different subtypes involved.

And David mentioned it briefly, that there's a difference in potency in these and from work done previously in terms of a mouse bioassay, there was stx2d, stx2a were the most pathogenic, and then stx1 toxins were less pathogenic to the mouse.

What's great about whole genome sequencing now is that we can do the Shiga toxin typing, just directly from the next-generation sequencing reads, and Phil Ashton and colleagues from Public Health England published this a couple of years ago now. So we're able to get this information just directly from the reads from the genome, not from the sample genome itself because the sample genome, it has problems assembling the toxin genes because they're paralogous.

Okay. So leading on from that, I'll go onto this example or information on being able to sequence these toxin genes, is able to help us understand about disease in humans. So this

phylogeny here is for O157, was from 105 Scottish clinical samples which was carried out by Anne Holmes and colleagues at the Scottish *E. coli* reference laboratory. And what they were able to do was they were able to look at the Shiga toxin types, and they were also able to look at the severe disease.

Now, let me just point, run quickly across, because I can't point. So this last column here, the red bars we're interested in is the HUS cases ,and here is the difference in colors on the left-hand side, the Shiga toxin types.

Okay. And what they found was so basically look at the right-hand column, and it's the bottom half which we have all these HUS cases, all the really nasty cases, and what they found was that 8 out of 10 of those involved the stx2a gene. So basically the stx2a is a good indicator of severe disease. So we can use whole genome sequencing to help us with that. This is for O157, and it may be helpful to use this for other organisms as well.

Okay. So that was about hazard characterization.

So what I want to spend the last few minutes on is speaking about the sources of human infection and source attribution. So I'm going to speak about *Campylobacter* and *Campylobacter* can be found in lots of different sources, chickens, sheep, wild birds, pigs, cattle, etc. And there's also cases of *Campylobacter* in Scotland, America, and many places across the world, but one of the questions is where are the cases coming from? Where are most cases coming from?

So risk assessment along the top, which we have been speaking in the first three-quarters of the talk, which basically follows the organism say from the cattle through the food chain to the infected person and all the way along there like that.

But source attribution, using microbial subtyping is sort of a cheat, but it's very powerful. What it does is it basically -- it just looks at the types of organisms in the animal sources and looks at the type of organisms in humans, and it does a comparison, and it compared to see which is more similar to each other. It's a bit like what David

was speaking about, machine learning approach, but the approached used here are based on population genetics for this. That's what this is based on.

So here I've basically shown at the bottom left, chicken, yellow and red types, in human, yellow and red types. If you look at the top for cows, you get yellow and red types with black types as well. So there's some overlapping crossover, but there are some differences as well, and the idea is to link the two using these populations genetic methods.

The method I'm going to speak about is MLST and Martin again really nicely explained MLST. So you isolate the DNA, you sequence it, and I'm going to speak about seven locus MLST to start with, and for the seven loci, you get the numbers and then from that, you combine them together to get a particular sequence type, sequence type 257 there at the bottom for this particular sequence type of *Campylobacter*.

So an example to quickly look at is in Manawatu region in North Island of New Zealand, and had really big problem with *Campylobacter*. They did the source attribution. So what they were able to do

with the source attribution was to predict the source, and it's colored there. So over a period of years, 2005 to 2007, poultry is in yellow, bovine is red, ovine is blue and environment is green. And as you can see, poultry was the main source according to the source attribution.

Okay. So they felt that they needed to do something about that. And so this is where risk management comes in. So if we go back to our rock example, yeah, maybe we need to do something about that rock. So we can think about our risk mitigation strategies. We put up a sign, a warning sign so people maybe don't go past that area or they take care going past, and that can be their risk mitigation.

But in New Zealand, in terms of the poultry interventions, what they did was a number of different things. They improved procedures for catching birds and cleaning crates. They improved their immersion chilling. They produced mandatory targets for *Campylobacter* on poultry after primary processing. So they put a whole set of interventions

in, and then they continued the source attribution to see what would happen.

And this graph here, this is what happened in effect. So I've now included the year 2008 and on into 2009, and what you can see is that the number of cases reduced, that was a smaller percentage of poultry cases. In fact, there was a 74% reduction in *Campylobacter* in poultry cases.

So this is a way of monitoring how successful or otherwise the risk management strategies were. It doesn't actually throw a whole raft of things out there to try and solve the problem. It doesn't actually tell you which one actually was most important.

In Scotland, we also follow *Campylobacter* using this type of methodology as well. Our colors aren't quite the same as the New Zealand colors, but the yellowy orange one is the chicken, and so chicken we find is the most important source of *Campylobacter* in Scotland. It varies in the model it used. It varied between 55 and 70% for that.

What I also wanted to say just very briefly

is a little bit of work that we've done on source attribution of *Listeria monocytogenes* in Europe, and this was an EFSA project that was led by Eva Moller Nielsen, and our group was involved in doing the source attribution work. It was a fairly small dataset for doing this sort of work, but it can potentially show what the potential of it is. So basically we had isolates of *Listeria monocytogenes* that we sequenced from fish, swine, ovine, bovine, and poultry sources, and then we compared that to what's in the human population and we attributed it to the human population.

So as I say, the database was fairly small. There was about 700 isolates in total that were here that were sequenced, but what you can see from that is basically we used three different models and that's the different colors there. That tends, from what we got in this dataset, it tends that bovine sources are maybe a bit more important than others.

Obviously, this was a start. I think this is first work that's been published in source attribution for *Listeria* using subtyping data. Here

we had 1748 genes from the core genome MLST for that.

So I think this is something as well that I'm sure in the U.S. you will be able to carry out these sorts of studies as well. Indeed, I think you already are starting to do so.

Okay. So just some take home messages. I've said about the steps in risk assessment, and I've spoken a little about some examples where whole genome sequencing can be used in risk assessment. I feel that we're only at the tip of the iceberg in this yet, and there's a lot more that can be done in this.

I think as well, as already been said, David mentioned about the epidemiology data being important to link with the next-generation sequencing data but also as well as that, the biology data, the phenotypic data is really, really important to combine that.

I think potentially we're getting lots and lots of whole genome sequencing data and trying to tie these datasets together is really, really important.

And the last bit I want to say about is source attribution. I think source attribution is really quite helpful for understanding the sources of disease and tracking that over time, and also I think in terms of when risk management strategies are put in either by companies at regional or national scale, that they can be evaluated to an extent using this type of methodology as well.

So that summarizes what I wanted to say. So thank you for your attention.

DR. DESSAI: Okay. While the speakers settle down and the projector gets turned off, I just want to state that setting the stage was the session where we were going to talk about the hazard, how the hazard is characterized, then different tools that are used to characterize this hazard in the context of WGS, and then we heard areas where we can do some predictions using newer approaches.

And then we went to the risk part of it which is very important, how to transition from defining the hazard, characterizing it further to turning it into some part of potential risk, and I

think that is a challenge here.

So last 2 days, we had the NARMS meeting, and we talked a whole lot about AMR, but I think we were a little shy of making that transition from what a hazard is to what the risk can be because it's a challenging area.

So what we're going to do right now is we have about half an hour of question and answer session. You have microphones which are right there. Those of you who want handheld microphones, let us know. We can provide those.

After half an hour, you'll get a break, and we will be back on time for the next session which is going to be partners talking about whole genome sequencing. The most important thing about this meeting is although we are hosting it, it is a meeting of all the partners involved in whole genome sequencing. Let me just make that pretty clear here.

So we would like you to be back on time, and when you're going out, please don't forget your badges as well as if you need any escorts, let us know.

All right, so we open the floor to question and answers. What we realized yesterday is that those who are online sometimes cannot hear the questions and the answers very well. So please speak into the microphone very clearly. Thank you.

And those who are asking questions, please state your name and affiliation clear as well. Thank you.

DR. EVANS: We can start off with a question from a question on the webinar. There was a question about whether there were a national or international databases for microbiome data that could be used by scientists to study risk in the ways that you were talking about.

DR. WIEDMANN: So the question was is there an international or national database of the microbiome that allows people to study risk.

DR. EVANS: And metagenomic as well.

DR. WIEDMANN: And metagenomic. Well, I'll take a first stab at that. And to the best of my knowledge, there's no database on the microbiome but there are a number of databases on whole genome

sequences of pathogens and bacteria organisms that one could then use to look at microbiome dataset, which organisms are microbiome, and then use that to potentially assess risk.

Now, you have to step back on microbiomes. There are two ways of studying the microbiome. Number 1 is based on a 16S gene. That approach, if you want to assess the risk associated with a given organism, there's no better way of saying it stinks. Okay. 16S sequencing does not differentiate basic *Clostridium botulinum* and *sporogenes,* between anthrax and *Bacillus weihenstephanensis*, between *Listeria monocytogenes* and *Listeria* species.

So if you do 16S based microbiome sequencing, in terms of specific risk due to the presence of a microbial hazard, it's not going to work.

The second option is what people call shotgun metagenomics, where you sequence all the DNA in a sample. That's the one which you potentially can use to map it against these species databases and then potentially have enough information to look at

the risk associated with that food.

Now, that's fraught with a whole big area of problems. Number 1 is you don't know whether the organisms is alive or dead for starters, okay. So that's problem number one.

Problem number 2 is unless I have one gene equals risk, so I need a combination of multiple genes, O157:H7, enterohemorrhagic *E. coli*, textbook example, whereas the typical short read sequences, I don't know whether my stx gene and my eae intimin gene that allows *E. coli* to attach are in the same organisms or in two different organisms. So without that information, it's very hard to assess the risk.

So the best thing we can do with metagenomics data right now, in my mind, is use the species databases and try to infer risk from it but it is extremely challenging, and I would in the mass majority of cases be very, very cautious.

DR. STRACHAN: Yeah, I would agree with the metagenomic data on this 16S. You have to be very cautious about the resolution that's there.

There's also, I think as well as the

resolution that we actually want as well and hope maybe the food industry could make informed decisions based on, for example, if you're able to find a *Listeria* in your food product, that would be something that you're interested in acting on or thinking about even though it's *Listeria* and not *Listeria mono*, for example, but I'd be interested in food industry views on that.

DR. CARRILLO: Hi. Is this working? Yeah. Cathy Carrillo from Canadian Food Inspection Agency. I have a question I think for most of you. You brought up the idea that some *Salmonella* or *E. coli* are less of a problem than others. Without an animal model to test these assumptions, as regulatory agency, how do you think we can get to the point where we can say this *Listeria* is okay. This *Salmonella* is okay. What sort of evidence can we provide? How do we know something new didn't come into the genome, you know, that might be a problem? Where do you see us going with this?

DR. WIEDMANN: I can take a stab at the *Listeria* example first, but the question really is,

you know, if you have SNPs, SNP data and you have data among human disease cases, saying we have these SNPs and they're underrepresented among human cases, so based on the distribution of isolates for certain SNPs among human and food, it looks like this one is less likely to cause human disease. So that's an association, not a cause and effect.

The question then becomes how do we go from association to cause and effect without having a clear animal model where we can take that *Salmonella*, *Listeria*, stx and put it into animals and say it really has a reduced likelihood of causing human disease.

So there's obviously a couple of ways around it. Number 1, we have a range of animal models. No animal model is perfect, but they will help us. With *Listeria*, we have a guinea pig model that assesses at least a number of factors that are important very well. We can supplement that with doing experiments in human tissue culture. We can grow human cells. We can increasingly grow human organoids. So not just one type of cells, but sort

of something that resembles an organ and use that to assess the effect of some of these mutations.

If those data on, you know, exposure human disease cases, tissue culture and imperfect animal model all converge, that's about as good a scientific evidence as we will get and in many cases, people are going to probably look at it and say that's sufficient.

So what we've done is a case study with *Listeria* with these single nucleotide polymorphisms internalin A, we found that isolates with these SNPs are about 100 times less likely to show up in human cases as compared to the ones that don't have those SNPs. Risk assessments look at large sets of isolates where we had even exposure data.

When we infect human tissue culture cells, those isolates are about hundred to thousand-fold less able to infect human cells in tissue culture. When we put these strains into guinea pigs, they're about a hundred to thousand-fold less likely to cause disease. So we have convergence of different lines of evidence. So with *Listeria* that works pretty

well.

With *Salmonella*, it's going to get a little bit more challenging. We can do tissue culture studies, not perfect. Animal models that really mimic human disease, it's a lot more tricky, but we have some animal models that will get us there.

Obviously, if we then move to *E. coli* stx, it gets a lot more tricky. If we then move to other organisms like *Bacillus cereus*, for example, you know, we don't have a good *Bacillus cereus* animal model at all, how do we assess which genes in *Bacillus cereus* is really responsible for these? How do we differentiate the *Bacillus cereus* from *Bacillus thuringiensis* which is supposed to be non-pathogenic by species definition, not always is, it gets even more challenging. So those are the ones that are going to require better animal models. It's going to require better tissue models. It's going to require that space in between, where we grow organs and assess, you know, characterize some of these hazards and characterize some of these organisms in these models.

DR. GALLY: I mean just from the O157 example, I mean certainly we are now aware that there are particular regions and in collaboration with groups in Sweden, where a particular area in the country has the kind of biomarkers as such for the more serious strains, right. So I think you can then take that information and then try to intervene specifically in those regions. Of course, you have to have the methods to intervene.

So this is always the trouble with this. I think if you're detecting these particular organisms in food, you can't, at the moment, you're fear being able to say let's leave it alone. We're not going to bother with that one. Obviously, that's far too dangerous at the moment, but I think there are cases where, for example, we continue to work on vaccines for this work, and I think we can target particular herds in particular regions where the more highly pathogenic bacteria exists.

So I think it really has to go hand-in-hand with other ways that we can intervene with this knowledge.

DR. STRACHAN: Yeah, I would just comment very briefly. I think it's really important for scientists when they're doing this sort of work to explain what their lack of knowledge or uncertainty is, when we speak about any strains to the degree that they're sure that they're pathogenic or otherwise, because risk managers or people in food factories, have to make decisions based on that, and if they can get an understanding of what that uncertainty is, then that will inform them in making their decisions.

DR. WIEDMANN: I think -- the important things are also what decisions are you going to try to make with these data. If your decision is simply you bringing in raw material from five different farms and you have some information that poultry from farm X has a certain SNP profile that might indicate higher risk and you want to process that at the end of your processing run rather than the beginning, I probably don't need animal data. I can probably do that without that. I don't need that high level of evidence.

On the other hand, if I'm going to try to make some other decisions, regulatory decisions, for example, you know, the amount of evidence I need and supporting data I have is very, very different.

MR. ROACH: Hello. I'm Steve Roach from Food Animal Concerns, and my question is actually sort of related to the first one. And what I'm concerned about is when you take genomic data from one environment and then try to apply it to another one, probably where the question came in my mind is when you talked about looking at the resistant genes in West Africa. And are we sure they're going to be the same as the ones that we've collected in Europe or in the U.S.? Definitely there's overlap, but there may be some questions when we start kind of using genomic data from one environment and then trying to use it in another one.

And another paper that I looked at by Margaret Davis, several years ago, they looked at, compared resistant genes on dairy farms versus just resistant genes on feedlots or calf farms. And in one environment, there was a lot more resistant

selection pressure, and what you found is that genes were concerned more on where you used more antibiotics, and that you actually had the genes kind of drifting where you had less selection pressure.

And I'm just concerned about, you know, how do you actually address that, particularly when we talk about using genes from this environment or maybe looking at resistant genes in India that may be very different than in the U.S.

DR. GALLY: It's a huge area, as most of you will be aware. For resistance genes, there are particular alleles that, you know, the majority actually, it's global, in terms of all the different subtypes. However, there are specific examples where we can track and identify particular types and you can then begin to associate those with clades and surpluses of bacteria that have associations with particular environments or particular animals or humans.

And I think there it can be quite powerful in terms of saying that, for example, *Staph aureus* has moved back from humans and is now in chickens and

has come with this resistance. I mean that sort of assignment is possible now.

But I think that for the majority, yeah, we're chasing our tails with this. It's really, really difficult to know the direction of travel of resistant genes but again the more of this information we have, the more we can address these questions. I mean at the moment, all these studies are conducted independently, and then trying to compare them is more or less impossible.

I think the power of looking at, with metagenomics is, for example, we're doing a study on pig farms at the moment, is the deep sequence environmental samples and get a real readout relative to which antibiotics are used in those facilities as to what resistance genes are present and how that fluctuates with time.

It's not about blame. It's about understanding what's there, what's out there, and then what can come through in the product.

But I mean beginning to trace AMR genes in foods and things, that's a whole other massive issue,

isn't it, but it's one we might obviously be starting to do, and there's a push now for obviously supplying poultry and supplying various livestock that have got minimal interventions of antibiotics. So clearly to track that, we're probably going to be asked to do that, right. So it's a big area.

DR. WIEDMANN: I mean the databases we use to go from genome to phenotype or predicting phenotype is obviously a huge challenge. It's not just limited to antimicrobial resistance.

Part of the issue again goes, what are you trying to do with these data? If I'm going to try to pull out a genome sequence for antimicrobial resistance and predict resistance of an organism, if I have the wrong database, that can cause challenges and I might incorrectly predict an isolate as sensitive for resistance when it isn't, and we have all, I think, seen this. The databases get better, but every time they get better, there are continuous issues with some of that.

You need to have some subject matter expertise, and then particular, if you move into

different environments, you're going to run into problems with that.

So the databases need to continue to be built. We need to very, very careful about extrapolating. I think your example is a really great one if you're starting to analyze sequence, come from environments but has exposure to different antibiotics, different countries, different treatment of animals, for example, that we don't have to rebuild the database and we're going to make some wrong calls absolutely until the databases get better.

DR. BRADEN: Chris Braden, Centers for Disease Control. I want to segue into the previous question and discussion about the effort that a number of partners have made to build some of the databases for whole genome sequencing and the metadata that goes along with it. Certainly we use, you know, to make the public repository at the NIH NCBI.

And my question is, you know, have you really used that database or others like it? Is it

sufficient in the quality of the sequence data and the comprehensiveness of the metadata that go with it in order to conduct some of the studies that you're trying to conduct in these predictive models?

DR. WIEDMANN: So the NCBI database, I think is useful for some applications but I think in general, the metadata that are there are probably not there yet to allow some of these investigations. You know, it depends on what you want to do with it. If you want to do source attribution, for example, I'd be very challenged to say we can use the NCBI data where they are right now to really do source attribution. I think we can get there but I have not seen much people that have actually validated that the predictions of the source attributions are always correct. So I think we need better metadata, if we want to do source attribution.

And the other question is do we want to use these data to predict, you know, do hazard characterization or hazard identification? Typically, the metadata there are limited. I have not seen, for example, in human cases, a lot of

metadata on disease severity and disease sometimes at a level of resolution that will help us to then link SNPs or other genomic characteristics that are likely to cause disease, and for obvious reason. I mean I understand there's huge issues with regard to, can we track back to who that person is if we give enough geospatial plus symptoms plus age plus predisposing factors which we need in *Listeria* really because we have this interaction of, you know, human susceptibility plus food plus organism.

So I think the data are useful in some ways but with some of the questions we're really trying to ask right now, and more often than not, I don't think they get us where we want to go.

DR. STRACHAN: Yeah, I would agree with that, and also the metadata, I think for the reasons, if people aren't willing to put these up on databases and cases traceback to individual human cases, it becomes a problem or companies perhaps for that matter.

And plus I know with some other databases is that there are metadata that are hidden across the

front public end where particular users can't use that under particular agreements. So I think it is quite a big challenge.

And you know, for myself, you know, for research purposes, you know, if I find a sequence on like my NCBI database that will relate far back to the original paper it came from, I go back to the original paper and I dig the metadata out from the original paper which isn't a very efficient way of doing things in some cases but, yeah, that's just the way it is.

DR. GALLY: I was at IAFP in Florida in the summer, and probably the reason coming to that is you see what is available and some great conversations with folk around I mean a whole bunch of *E. coli* sequences just released actually, and we received the database of source attribution studies and linking into other groups in the U.S. to try and work on that. So from an *E. coli* point of view, that's been incredibly helpful.

But I think a key problem here is exactly what Norval's stated which is about, especially with

human data, it's really powerful if you know the degree of severity of disease, if you can have geographical insight into that as well. This is really difficult information to get out in the UK. I'm not talking about USA, but it is really powerful data if you can have it because it really allows you to link to the information you're getting from animal sources as well.

I think we're a really long way from that information becoming easily available. I think some of the barriers could be broken down quite simply by deciding what type of information and the level of granularity of that, that can safely be put on automatic databases. They do it in Scandinavia better, and I think there's some really nice models to follow out there.

So I think people are beginning to get the message that we have to do this, but there's still a huge number of hurdles, and it's going to be very organism specific as well. Fantastic *Salmonella* examples and other organisms we're really light on data.

DR. ALLARD: Thank you. This is Marc Allard, FDA, Center for Food Safety and Applied Nutrition. I just want to comment on the last set of comments, and then I have a question.

Essentially the Government's data at the FDA is all FOIA-able. So it's just a matter of how much is released to the public and how soon, but I think they're fully open to recommendations of what additional information should go into the metadata, and as long as it's legally allowable, I think the Government's willing to release that information.

And so my question is essentially going directly at risk. We have a model for phenotypic, genotypic prediction within NCBI where they built the bio project of the 4,100 resistant genes and then every new genome is blasted against it to call the presence and absence. I believe they don't do allelic differences yet, but they have presence and absence.

So my question for the risk group is, this would be easy to replicate, build genomic databases, bio projects, for specific genes. The question is,

it's not clear where to start. Which genes should we start? Which ones are clearly risk connected? We know there's been some publications with the STECs and the NACMPF documents as well as the French recently released a big study on *Listeria*.

So my question is do we think we should be building these databases? Do we have lists of recommendations? How do we move forward with rapid prediction? I want to build tools for the public.

DR. STRACHAN: I think that's a very good point. I don't know, because I've done a lot of *Campylobacter* source attribution work, but I think you speak about the French project as well with *Listeria*. I think whole genome MLST is a good place to start because you have basically all the core genomes and core genes and also possibly a number of accessory ones as well that can be mentioned, and I think that would probably be a good way to start.

DR. WIEDMANN: I think there's some, maybe a few examples where we might be ready, internalin genes of *Listeria*, pretty clearly linked to risk, a lot of data. I think FDA has some data. So, for

Listeria, we can do some of it.

I think as we move to *E. coli*, I rely on the two of you to tell me whether you're ready. I think the problem goes back to that very often it's not one gene determines, you know, risk, but it's interactions between different genes. How do you get at that, and that's going to be much more challenging.

So I would really want to get that before I'm going to put out a simplistic tool where we pretend one gene can ultimately predict risk and not looking at interactions of different genes.

The European or Germany *E. coli* outbreak was unusual. *E. coli* was an unusual attachment gene is a great example of where, you know, could we have used that? How could we have predicted that risk because all of these databases are a priority and maybe, maybe not. I don't want to make a judgment on it.

And I think the antimicrobial resistance databases provide some examples of the risks of it. I mean there are certain genes in there because you

don't have the resolution. You might predict that organism is antimicrobial resistant and it is not. We have seen it be published and others have published it, streptomycin resistant, very difficult to predict because there some allelic variance.

So I think we want to be very careful and I think put some uncertainty around these predictions, too, when we do it. I think there are some examples, but honestly I think there are only a few right now where I feel like we are ready to put something like that together and put it to use other than research use. Maybe we can put it to use for some sort of risk rankings, again back to my sample, you know, which poultry flock, which cull cows, are you going to process first, second, third? So I think we might be able to use some of that information.

So all of these things we need to think about what we're going to use these data for. We can use it for so many decisions, and the trick always is to use it, if you have the right database for decision A, but we start using it for decision C, we're going to run into trouble.

DR. STRACHAN: I guess another thing in passing, I was just thinking about STEC, there's a virulence finder which is developed by Flemming Scheutz from SSI in Denmark. So you can basically upload your genome, whether that be an assembled or the reads, and it will feedback a number of virulence genes, whether they're present or not.

DR. BRANDT: Alex Brandt from FSNS. I have a question with regard to, I know we're kind of talking about relatedness of presence/absence of genes and different alleles and even some of the phenotypic traits don't always match up. So is it enough to just look at presence/absence of genes or different alleles, or should we be looking deeper at like transcriptomics and really going to that level? I guess that's my question simply.

DR. DESSAI: Before you guys answer the questions, we are running a little over our time, and we'll extend the time by about 5 minutes if that's okay with the crowd here. We also have two questions online to address. Is that right? Okay. Go ahead.

DR. WIEDMANN: I think in some questions

you can have a gene present. There's examples. We have a gene present and it's not transcribed. It's not turned on. Therefore, you don't get the phenotype. I think looking at RNA, looking at transcriptomics can help you with this, and ultimately you should be able in many cases to then predict polymorphisms and SNPs and promoters in regions that drive transcription so that you're not going to have to do routine transcriptomics for it, where then it becomes obviously completely different issue is, if you're starting to assess risk in different foods.

Something every food microbiologist knows, *Clostridium botulinum* is only an issue if it's found in a food that's anaerobic conditions because that's when the genes are turned on. So if I want to assess that risk, I need to have transcriptomics and some other data because gene presence/absence does not equate risk at all. So we have this range, too, in the thing. So in these cases, whether we need transcriptomics or something else is a different question but gene expression if very, very important

and we can't forget that different foods based on anaerobic condition represent completely different risks.

DR. DESSAI: Okay. We have online questions.

DR. NAHAR: One of the online participants asked what about focusing hazard characterization on sampling physicians, veterinarians, nurses, farm workers, food preparers, etc.? Those populations tend to be at highest risk for the spread of pathogens as well as developing AMR.

DR. DESSAI: Will you repeat the question?

DR. NAHAR: So it's a sampling question. What about focusing on sampling this particular population, farm workers, physicians, veterinarians?

DR. WIEDMANN: I think the question is what about risk assessment, I think risk characterization might not have been right, and isolate collection from people at high risk, particularly on the primary production side, the farm workers and anyone who is working in these areas of a wide range.

So I think is -- I'll try first stab.

That's a tough one. I think it will give us different isolate sets that could be useful. I don't think that we not focus on it that much is that big an issue. I think they're going to be caught indirectly through the public surveillance system, too. We get into a whole slew of social issues, you know, illegal immigrant farm workers, are they going to seek? You know, what's the reporting among those? Are we missing cases because of whole set of other issues which I think is a very, very important one. Are we capturing all high risk individuals without surveillance? And I think that's something worth thinking about it, and it could be very good sentinels, particularly for some emerging zoonotic diseases.

DR. STRACHAN: Yeah, I think the point, I look at it sort of in terms of like source attribution, thinking about *Campylobacter*, workers working within a poultry factory might have particular exposures, but as you think about rural children living in Northeast Scotland, and also in the USA as well, you're more likely to have contact

with farm animals or private watch -- are going to get different spectrum of types of *Campylobacter* for example. So I think looking at the different groups will tell you something about what -- confirm different exposures they have in the pathways they're getting the disease from. So I think from a molecular epidemiological perspective, that could be quite helpful.

DR. GALLY: I mean there are obviously -- I mean I'm aware of quite a number of funded studies where that particular close association between product, livestock rearing and people that are working with livestock is being looked at. Obviously, studies throughout the last 50 years where the different typing tools have been used to do that. There's current studies doing that, using whole genome sequencing now, and it's really looking at transfer of particular organisms and AMR and virulence on those close quarters, but I suppose it's -- yeah, I mean it doesn't really necessarily help address the bigger issue we've then got of I suppose the distribution that can occur through product which

is really a much bigger scale.

I suppose you could develop -- I suppose there are the chances of those -- when we think about viral evolutions, there are the chances of maybe that being the epicenter of something kicking off, that obviously ends up being a lot more serious, do we give that -- I think the question is, do we give that extra attention? Do we do extra monitoring of all those folks that are in those direct environments, handling birds, etc., because there's more likely to be the start of a particular -- I think we're probably a very long way from that, I would suggest.

DR. DESSAI: Okay. We'll take our last question, and then we will close this session.

DR. NAHAR: Sure. Last question, can you speak to the pros and cons of hazard ID based on WGS data alone showing AMR gene prevalence versus how we've traditionally conducted such work using phenotypic AMR?

DR. GALLY: I can only speak to what I'm aware of with a couple of different pathogens but while there are obviously issues in getting phenotype

from genotype and it's particularly difficult for whether it can be single nucleotide changes that comes out, certainly the horizontal nucleotide resistant genes, it's actually pretty good. I mean certainly for the *E. coli*. There's a high degree of accuracy to presence of a particular allele and then giving phenotypic resistance. So it certainly can work very well in that space. But, you know, you can't have 100% accuracy with that, but it's a very good genotype to phenotype mix in terms of whether the horizontal nucleotide resistant genes.

DR. WIEDMANN: I may be a little bit less bullish on this, and the example I'm going to give, if you look at *E. coli* and *Shigella*, if you look at resistance measured with traditional approaches, so MIC, etc., the way you interpret them in terms of susceptible intermediate resistant or resistance, actually differs between *Shigella* and *E. coli* despite the fact that *Shigella* is actually an *E. coli*. So that translation from presence of genes to then treatment decisions, you know, does a certain treatment work in at a different level is challenging

in some very specific cases. So I think we really need to look at it from organism to organism, and again it depends on what you mean about what sort of risk. Is it the risk of a resistant infection if this gene is found in a certain organism? I think for some organisms we might be there. For a lot of them, we really need more data to make these linkages.

DR. GALLY: I suppose I'm thinking particularly the arena of real-time diagnostics and particularly human and animal infections where at least using that information from initial sequencing of a direct sample which is where we're trying to move with some of this, it's still much better than being in the dark. So it's at least having some information you can make an informed decision on. But, yeah, we still don't understand all the complexities to other things that will influence resistance, but certainly I think it's a step in the right direction.

DR. DESSAI: Okay. Just an announcement for those online: Please send your questions which

are short and focused so those will be easily transcribed and understood here.

Number 2, if you have any other suggestions, send those to us.

So I think we've set the stage for this conference pretty neat and covered a lot of areas. The speakers will be available for you to have more discussion or if you have any questions, and I think I would like to thank them profusely for the work that they have done this morning. Please join me with a big round of applause.

So we'll be back at 11:00.

(Off the record at 10:28 a.m.)

(On the record at 11:01 a.m.)

DR. EVANS: Welcome back from the break. My name is Peter Evans. I'm with the FSIS Office of Policy and Program Development, and I want to welcome you to our second session which is on Federal/State Collaboration. We're going to have eight speakers today, and this is basically going to take us from now to the lunch and then after lunch to the end of the day. We're going to start off with three

speakers, have a very short question and answer, 5 minutes, go to lunch. And then we come back from lunch, we'll have three more speakers, a break, and then two speakers.

So we're going to hear presentations about how whole genome sequencing is being used in the United States both at the federal and state level, and so how WGS capability is being increased and also examples of how organizations are using the data. And then we're also going to learn about how the agencies are working together, depositing data in a common database at NCBI and then also developing common procedures and standards through the NFS Consortium.

So first I'm going to welcome Dr. John Besser from CDC, Dr. Steven Musser from FDA CFSAN, and Dr. David Goldman from FSIS, to speak about their experiences within their organizations. Thank you.

DR. BESSER: Well, thank you. I'm very pleased to have the opportunity to speak to you today. The organizers gave me a long list of suggested topics. So I'm going to move along pretty

quickly here. I really hope I never need to use that picture in a dating app.

So I've been around long enough to remember the beginning of the PFGE era, and what was really interesting was a lot of the concerns I hear now from laboratorians, epidemiologists, regulators, industry, are very similar to what we heard 20 some years ago, and so I'm going to talk about what hasn't changed, what's really an extension of what we've been doing for a very long time.

I'm going to talk about the WGS infrastructure that we're building at CDC, but I'm going to talk about then what's different, what can we expected different about whole genome sequencing and then I'll touch very quickly on where we see this all going.

PulseNet, as you've heard, has been around for 21 years now. We just celebrated the 20th anniversary last year, and over those years, we've seen investigations with our colleagues in FSIS, FDA, and the states and with industry, impact virtually every commodity of food. It's been tremendously

impactful.

And the way it works, of course, is that we have combined in a One Health model streams of data from food monitoring programs, animal monitoring programs and human disease monitoring programs, all into this one system. And the way it works is we connect cases that may be geographically distant from each other by this common DNA fingerprint.

We released a study just last year that looked at the cost and benefit of PulseNet. This was during the PFGE era, and we found that at a minimum, it saves about 270,000 cases of disease per year and about $500,000,000 in cost to society. And that's really the portion of the benefits that we could measure which I suspect is really a very tiny proportion of the total benefits.

We investigate about between 30 and 60 national clusters per week and at the state level, there's about 1,500 to 2,000 investigations per year. So this is a constant, very big activity.

And I'll describe a little bit about our network which has driven how we've built the

infrastructure. It's based on local testing, local control of the analysis using standard computers, centralized quality assurance, so that we're all able to communicate in a standardized way, and in the WGS era, we're going to have centralized bioinformatics and centralized high performance computing.

And the rationale for having this distributed testing network is that it provides functionality for both local investigations, national and international investigations, and it helps with turnaround time which in our world is important, and I'll show you later how important it is. It gives us an enormous body of resources. With 86 laboratories in the network, we can expand or contract the work as needed because there's so many independent parties that can do the work. And it gives us local control of patient identifying information and commercial confidential information. That's all held, and you'll see more later.

So what's the same as it's been? Well, the subtyping methods, be it performing or whole genome sequencing, primarily work by grouping together cases

that are most likely to share a common exposure, such as food. They work the same way.

Matches between cases and food and environmental samples in both circumstances provide a hypothesis as you've heard many times already today, not a proof, but they prove a hypothesis in both of these circumstances for both methods.

And in both methods, the historical database is routinely examined for matches to current clusters. So, for years, we've been identifying food or environmental or human cases that match current clusters. So that's actually not new.

And how we interpret this data also has many similarities. A match with both methods, what we call a match, which is another subject, means that an association is more likely than if there is no match. You notice I said more likely. So it's not absolute.

A mismatch means that an association is less likely but not impossible, and I'll show you why that's the case.

And as you've heard over and over again,

other types of data are needed to support any of these conclusions be it whole genome sequencing or PFGE, such as epidemiological data and traceback data.

I'm going to focus on this for a moment. Why does the mismatch mean -- does not absolutely mean there there's no association? Well, it turns out that outbreaks are all different from each other. They can be very clonal as we sometimes see when an individual food handler contaminates food and a single strain gets into a population, but there's a whole spectrum of different outbreak ecologies, if you will, that span everything from that to mass contamination events. I used irrigation water here, but there's negative pressure events where sewage gets into drinking water, and there's no reason that in that circumstance you'd expect there to be a single strain and everything in between. So outbreaks themselves are not necessarily purely clonal.

And we've seen that in the PFGE era. This was an major outbreak of listeriosis associated with

cantaloupe a few years ago, and we saw multiple different serotypes, multiple different PFGE types because the product was contaminated with an environmental source that had many different germs in it.

And we've seen it in the WGS era. This is a phylogenetic tree of cases that were all associated with having consumed chicken from a particular processing plant and we can see that there's a very variation between these cases in terms of SNPs, high quality SNPs, all associated with the same sources. This is an example of a mass contamination event.

So what kind of infrastructure are we building? Well, this is our master diagram, and it really can be divided into two sections, one that's closed to the public, where information that is potentially identifying to individual patient level is kept or commercial confidential information, information which can't be released to the public, and everything below that line is information that will be available to the public.

Isolates are sequenced, controlled by the

local user. They go through a calculation engine that pulls out different types of information which I'll show you in a moment, and then also looking at allele databases, looking at different alleles and reference identification, etc.

And we're building the system to coordinate not only with the other federal agencies, but with global partners and with the states and local governments and all these other institutions listed on the left. So a lot of different parties that we need to coordinate with.

And we're building the core genome MLST databases in collaboration with global partners, because we're looking at global systems for food which is inherently a global commodity, and we're also paying close attention to the developments at NCBI.

And recently, a meeting was held with PulseNet International to hash out some of the global issues. How do we collaborate around the world and they came up with this vision paper which you should read when you can.

So these databases automatically pull up all this different kind of information, the genus and species, serotype, pathotype, the virulence information, the resistance information, all types of different subtyping information from core genome and whole genome MLST and plasmid profiles, and also a SNP analysis is available if desired. It will be shortly. All this information automatically populates the local databases.

This is an example of the type of report that we can expect from individual isolates. We can have all the virulence information, serotype information, resistance information, etc., so that we can start properly learning about risk based on different, for instance, non-O157 subtypes.

This is where we're at with building the infrastructure. As of yesterday, 49 states now have sequencers, and 37 states have been certified by PulseNet as able to perform these methods.

This is a list of all the different databases and where we're at with their development and release. *Listeria* is the only one that is fully

released, but the rest are soon to follow.

The number of genomes that have been sequencing has been going up very dramatically but hold onto your hats. It's going to be a lot more very soon. Within, by the end of FY18, we hope that most *Salmonella*, all STEC and all *Listeria* in the United States from clinical cases will be sequenced.

So what's different? Well, the PFGE was very specific for some organisms and low for others as you saw for *Listeria,* Enteritidis, whereas sequencing is very high for the most part in all organisms. We have limited ability to evaluate the closeness of strains with PFGE where we can do that with whole genome sequencing. The data is categorical pretty much with PFGE. It matches or it doesn't, where it's continuous, there's different shades of closeness with whole genome sequence data.

And here's how we actually used it in the PFGE era compared to how we used the data in the whole genome sequence era. These are all cases associated with the consumption of *Listeria* associated with a particular type of ice cream. In

the past, we would have focused on a single PFGE type, and we would at the onset, ignore the other cases that we couldn't draw together, and use that information to make the association as part of the investigation.

Now, in the whole genome sequencing era, we can look at all the different types that are related to each other and in this particular case, there was 11 PFGE types and only 2 sequencing types that were used, making the investigation much simpler.

But the impact of all of this is that we can detect outbreaks when they're smaller, which means that we catch them earlier, they're smaller, and we can solve more outbreaks than we could in the past.

Also, one thing that hasn't been discussed is that this information is invaluable for ruling out likely outbreaks. I say that probably happens more than not, where we can say, well, this product probably isn't involved in this situation.

And when we interpret the cases, matching of cases with products or environment, in the past,

when we had a very new or a very rare PFGE subtype, it formed a strong hypothesis, but when it's a common subtype, we really couldn't say it was a weak hypothesis. With whole genome sequencing, a product match is uniformly a strong hypothesis.

So you've seen this slide already, but this is the different phases of our *Listeria* program, a collaboration between the agencies. In the early days, there was few outbreaks represented by the blue bars. When we turned on PulseNet, a lot more outbreaks, and the outbreaks got smaller, a median of 69 to 11. When we did an epidemiological project called the *Listeria* Initiative, more outbreaks still and the median size went down to 5. Now, in the whole genome sequencing era, more outbreaks yet, and it went down to 4.

So what we've seen, summarizing all this data, is a dramatic increase in the number of solved outbreaks and a lowering in the size of outbreaks. So this means to industry that these things are found with more surgical precision. They're solved quickly when there's less cases, less potential impact.

And so we get the question a lot about what kind of cutoff values? Well, we have these rule of thumb cutoffs but really what I'd like to tell you, if you remember this diagram, there's really no such thing as an absolute cutoff. So absolute cutoffs are not possible, and really we need to combine epidemiology, traceback data and the other data as a part of the total solution for defining what's important and what's not.

And these give strong hypotheses, WGS does, but again it doesn't absolutely mean that a match between a cases and a product means causation as Dr. Wiedmann said earlier. The food chain is extremely complex as this slide from FDA shows you, and if we look at transmission mechanisms, the chain of transmission can be even more complex and WGS doesn't tell you anything about the chain of transmission.

Okay. Just a few moments in my last couple of minutes. What's ahead? Well, we're working on these advanced analytics, machine learning, disjunctive anomaly detection, etc., and we're also

exploring the use of metagenomics for food safety as many other institutions are. And as you may know, there's a big collaboration between Mars Company and IBM to look at these very issues.

There's a wide range of potential applications for metagenomics in our field. We're focusing on two, pathogen discovery and in situ or direct from specimen pathogen characterization. And we're looking at several methods. I've listed here, amplicon sequencing and shotgun metagenomics, but much work remains.

But there's good reasons we're doing all of this. Most pathogens that cause disease, in the United States and worldwide, are thought to be unknown. In fact, our PulseNet pathogens are only 4% of the total causes of disease. So there's a tremendous amount to be learned by doing pathogen discovery.

And these direct from specimen methods have the opportunity for us, majority shortening the timeline between when a patient consumes a food and when there's an actionable result, which is currently

quite long. In addition, then we have to find a cluster. We have to identify a food. These direct from specimen efforts can cut weeks off of the whole process potentially leading to more outbreaks solved more quickly and more illnesses prevented.

And finally, there's a problem in the U.S. from what's called culture-independent diagnostics. We're potentially losing isolates because of changes in diagnostic testing in both PFGE and whole genome sequencing dependent isolates. So we're putting a lot of effort into developing tests that don't depend on isolates.

And that's it. Thank you very much.

DR. MUSSER: I'm happy to be here. My name is Steve Musser. I'm the Deputy Director for Scientific Operations at the Center for Food Safety and Applied Nutrition of FDA.

We've been doing sequencing and applying it to food safety applications for a number of years, and what I'd like to try and do today is walk you through FDA's approach, and I would also like to caveat a lot of that with, while FDA and FSIS and CDC

have very similar approaches, FSIS' and FDA's are legal and regulatory requirements are also different which means that we respond and act in different ways in cases of finding either outbreaks or cases of contamination in facilities.

So with that said, if I could have the next slide. So I'd like to talk a little bit about GenomeTrakr and the GenomeTrakr laboratories and the network and the database. All of these technologies, and it would mean whether PFGE or whole genome sequencing or whatever, are useless unless you have some way of comparing them, and some way of putting them into a place that you can look at them and then you also need people to be supplying information to those databases.

So GenomeTrakr was essentially begun by FDA, and it's essentially a larger dataset of information than anyone of the other single datasets. So if you were to look at, you know, what was just collected by somebody's personal academic group or PulseNet, for example, so GenomeTrakr contains everything in PulseNet as well as lots of other

things outside of PulseNet.

And so it would have public health, government, private, academic sources of information and just to give you some idea, I know John mentioned there's about 40 to 50,000 sequences in the PulseNet database. There's over three times that many in the GenomeTrakr database. So it's just a larger collection, and there's advantages and disadvantages to each, and I think based on the mission of the different agencies and groups, there's good reason for that.

So we established this network in 2011 with a fairly small investment in state laboratories and our FDA field laboratories, FSIS and CDC joined as well as dozens of other collaborating organizations following that, and I would also like to highlight the National Center for Biotechnology Information, NCBI, which we would not able to do any of this without because they not only serve up the data but they provide analytical tools and ways of looking at the information which you wouldn't be able to do.

And when we began this network, we were

also trying to answer two questions. The first one was, we need to provide this information in a way that is public. We were under a Presidential Executive Order to do so, not just ours, but all Government supplied information. We also didn't have the money to actually supply all this or maintain all this. So we needed someone to help us. And NCBI was more than willing to do that, and they've been with us from the start, really helping and making this a useable database.

So unlike the rest of NCBI, which is, you know, you could just upload and you wouldn't be part of GenomeTrakr, these actually, people that are part of the GenomeTrakr Consortium actually do fill out a form and there is some information there. If I could have the next slide.

So one of the issues that we had when we first began looking at what would be used instead of PulseNet or in supplementing PulseNet was we have this great body of clinical information that CDC has gained by using the PulseNet system, and while there were some food isolates in it, the majority of

isolates were from clinical sources.

So what we found out when we were trying to do these investigations is you know that a whole bunch of people are sick, but you don't really have any idea of where to look. And the reason for that is there's no way to look for that information because there aren't any sequence data.

So we've concentrated a lot on obtaining and getting food and environments isolates in addition to clinical isolates because we recognize that the maximum benefit, particularly in outbreak situations is kind of knowing where to look or helping understand where to look.

And if you look at PFGE in the past, it was primarily driven when we found a group of people that were sick, and then we do a lot of epidemiology and we really solved a lot of outbreaks and intervened in very positive public health ways.

What we're seeing now with whole genome sequences is that sometimes our information in sequencing helps drive the epidemiology or helps refine the epidemiology. So it's not a one-way

street. It's actually information that can provide data in both ways, and really help inform our investigations. Can I have the next slide?

So currently the GenomeTrakr Network includes all of the FDA labs, all the CDC labs as well as the labs that CDC would fund and contribute to PulseNet. FSIS's laboratories, we fund 19 state agriculture, health and university labs in the U.S., 1 hospital lab and 17 labs located outside the U.S. so we have 4 continents and 10 countries contributing information.

There's approximately 150,000 sequences in the database now. It's growing at a rate of between 5- and 8,000 sequences a month. So that number, it's a static number now. By the end of the month, I'll be quite different.

And then we partnered with CDC and FSIS in 2012 to do all clinical *Listeria monocytogenes* and environmental. So if FDA or FSIS found environmental samples, we uploaded them and sequenced them and likewise, if CDC and its network found clinical samples, they sequenced and uploaded them. It's

probably the best *Listeria monocytogenes* database in the world in terms of completeness over the last couple of years.

You can find out more about the GenomeTrakr website at the link at the bottom or you can just go to fda.gov and search GenomeTrakr or Google.

Next slide.

So this is just a cartoon way of visioning what happens. At the top, anyone that's part of the network or the consortium performs a sequence, they upload it into the NCBI database, or the European version of it, EMBL, or the Japanese version, DDBJ. These databases are copied within each other every night. So there's a version of all of these sequences available throughout the world, and so if NCBI were taken down through a power outage or other natural disaster, you could get at the data through EMBL or DDBJ. So it's part of the redundant system that's been built at NCBI.

This is a system that's open. It's available to industry. It's available to academics. It's available to public health agencies. There is

no restriction on who can look at this information.

Next slide.

We always get asked, so the sequence information is only valuable with the metadata. I think you heard a little bit of that discussion this morning. This is the metadata that's provided in the case of a food/environmental or a clinical on the right.

Basically the information that's circled is really the information I wanted to highlight today, and that's a specific FDA number. So if we did an inspection of your facility, you would get a report at some point that would have this number associated with the samples that were collected. You could, and I'm hoping that the NCBI -- Bill Klimke's in the back. I see him there. So he will tell you about how to do this, but you can type this number into the search engine that NCBI has built, and you can see where your isolate lies in the tree. And you can then download those sequences and you'll have information specific to your isolate which we think is great because you can then verify what we're

doing.

In the case of other information, the geographic information, we list it only by country. We also list the type of food or if it's an environmental facility, it would just say environmental, that it was collected by FDA. If there's PFGE information, that's also provided, and then who it was provided by. So if it was provided by FDA, there's contact information there. So if you are doing research and you don't have access to the full list of metadata, obviously we know what state and where the place was that we collected the information. That we could provide to folks.

Then on the clinical side, very similarly it lists if someone got sick in the U.S. is about what you have. We don't have any other information.

We spend a lot of time working with various people in industry, academia, and public health and other folks throughout the world trying to get this very limited meta dataset and although as it was correctly pointed out, it is a very limited amount of information. Some very clever people have figured

out how to take advantage of this information, and while not having completeness, be able to develop a lot of very cool models on predicting disease and predicting risk and by simply just knowing that someone got sick and it was from this organism, and maybe from this particular food or this particular area. And then they can ask a lot of questions like, well, I only need to know this little piece or that little piece.

Next slide, please.

So this didn't come out quite right, but the point is that you have one data record, you can get lots of things from it. The most interesting thing to me is that we're kind of myopic in our approach. We're looking for very specific things, but there's lot of other information there that we haven't even begun to mine. So we really are trying to encourage people to put this information up there and be able to use it in ways that we as regulatory agencies may never have thought of.

Next slide, please.

So again these are questions we get all the

time. Yes, it is more discriminatory that PFGE. Like all living organisms, there is stress and pressure when they adapt to their environment. This has been known for a long time. They mutate more rapidly and so we get a clue to the geographic origin based on the stress and the pressure and the environmental response that the organism sees. And that's what's really so important about this particular technology and the use of it.

In clinical applications, it's slow. I mean you're really looking at days to get the analysis done, and so it's probably not acceptable for that, but for us, in the regulatory community and the food community, it's really very, very good.

The point that I'd like to make on this slide, in particular, is that when we have some environmental information, we know where to look as opposed to not looking. We really spend a lot of resources, very ineffectively, by not even knowing where to look, not even, for example, knowing, you know, should we be looking at imports from Southeast Asia or should we be looking at domestic samples. So

all of this really helps and aides in a much faster response.

Next slide, please.

What data analysis tools should be used and why? Another question we routinely get. Well, that depends. It depends on who you are and what you're trying to accomplish.

So if you're -- I don't want to speak for CDC, but I'll just, having listened to their presentations, they're interested in a very fast, kind of easy way of pushbutton looking relationships, which gets you to a certain point, and you're looking at sort of a limited dataset.

Because we're a regulatory agency, and we know that there are consequences to our results, legal and other, we use the gold standard SNP method. Both methods for the most part give very similar results and again it depends on what you're approach is.

As John pointed out, there's no single threshold. Generally if it's less than 20 SNPs away, we'd be taking a closer look at it. That doesn't

mean that we do anything with it. It just means we take a closer look.

All of our tools are validated. They're all on the website. You can look at the validation documents. They're all validated according to software standards, well accepted standards, and so you can go and you can pull down the software, you can look at it, you can look at all the validation statistics and all of those things with it.

What software should I use? Well, that depends on your level of comfort. There are commercial versions like BioNumerics, which is the standard that CDC and PulseNet uses. We've heard from numerous people that that's too expensive, that they don't use it that much, that it's too difficult, on and on and on. So we have been working for more than a year now, and I think we are going to have a Galaxy version, a free publicly available Galaxy version that uses our SNP pipeline that will work in an Amazon cloud service. Stay tuned for that. I thought it would be more advanced than it is, but it's a little behind because of security concerns but

we are getting very close to having something that you can plug your own information into and get the same answer that we get we hope.

And then how do I access the data and do analysis on the NCBI site? There's a really simple website. It's basically the NCBI site/pathogens, and that will take you right to the viewer and you can do whatever you like with it. Very simple.

Next slide.

What happens with a match? Reanalysis of the data happens first. So if we thought we saw a match, we would take a much closer look at the data. We would pull down the information. We don't take any regulatory action based purely on the match. We would resample, we would reinspect, and then depending on those reinspection results, a number of things can happen because there's two possibilities.

We can take and we do take regulatory actions routinely on samples where there's no epidemiological evidence. If you have pathogens in your facility, they're in the food or they're in zone 1, zone 2, you're in violation of the law of having

unsanitary conditions and so you're going to hear from us and trust us, that is really the best place for you to be because the next step is not where you want to be, where you've made people sick, and we do rely on epidemiologic evidence because when you get into civil lawsuits, you get into legal issues that you really don't want to be involved in.

If you're in that first half, unsanitary conditions, we can at least work with you and we would like to work with you on solving that problem and helping you understand the information.

Remember, we've taken maybe dozens, hundreds of swabs of your facility, and we've done the sequencing free, you're not paying for it, and so we can provide all that information as well as where they came from and what to do about it.

Next.

We do this because we know that there's a supply chain and that if we found a sample, perhaps in the processing facility, it may have actually come from an ingredient or from the manufacturing facility. There's upstream and downstream

information, and so we really do not want to take regulatory action unnecessarily against someone who's not involved.

There's also a movement of processing equipment. One company goes out of business. They sell their equipment to another company. The material is contaminated and so the contamination moves from one facility to the next.

Next slide.

A note about clinical matches and epidemiology. When we go into facilities, what we see is this. We see an incoming stream that's relatively free of contamination and an outgoing stream that has loads of contamination. And if you think about it, it makes senses. If you had, you know, a million bags of flour or a million bags of lettuce, and you're producing that every day and only one of those bags were contaminated, we wouldn't detect any illness.

When we start detecting illness, you've got multipliers of 40 or 10 or 5 and you actually have to hit those susceptible populations. So when we go in,

we usually find very significant contamination.

Next.

This is a slide that our industry has been begging us for and asking us for, you know, what are the implications of whole genome sequencing? What do you see? How does this work? And I would just like to caveat a lot of this because we don't keep our information this way necessarily and exactly how all of these things would have worked, but the total number of facilities inspected where we would be looking at high risk facilities that we were either in before or were indicated as being involved in some kind of outbreak, or we've had matches in the past, so we've done an inspection, we've done another inspection and we see some relationship there.

There were 167 requests for additional information done by our bioanalytics groups to look at the information, and then of those 600 inspections and 167, it was actually a very small number of regulatory actions with some of them being more significant than others. The regulatory meeting were generally the most common occurrence. There were

three injunctions and one mandatory recall, and one suspension with is very severe. But, for the most part, there was a small number of regulatory actions, not a huge number based on this technology, and I think in many cases we would have arrived at some of these situations without sequencing at all.

Next slide.

Should we start using sequencing? How much time do I have? One or two minutes.

Contribute sequences to GenomeTrakr database, if you're industry I tell you to please contact your lawyers and then get a second opinion. There are consequences to doing this which you may choose not to take. I should say there's risks to doing this, not necessarily consequences.

If you're doing routine environmental and high volume product sampling, you don't want this technology. It's too costly.

If you're doing supply chain management and say PCR doesn't work for you any more, you can't figure out what the source of the problem is, you probably want to look at this because it is an

invaluable tool in tracing where things came from and how they there.

If we've been in your facility and you've got a positive, you may as well kind of throw your hands in the air and just engage in this because when we look at civil litigation and other litigation that the Justice Department may take, it's when you knew about it and what you did about it. So if 6 months ago we were in your facility, we inspected, you have this and you haven't done anything and you have no documentation, I'm sorry, I'm going a little over, you know, you can't hide it any more. We already know you have *Listeria mono* or *Salmonella* there, and it's really what you did about it during that time frame. And the next slide which I hope is the last one.

The only thing I'd like to say about this is that there's more to this technology than simply doing outbreak detection, particularly if you're interested in the effectiveness of sanitizers in your facility, if you're interested in supply chain management, it really is a technology that can get

you to places that you couldn't get by any other technology.

And with that, I think I'm done. Yes. Thank you.

DR. GOLDMAN: Well, good morning. I'm David Goldman again for those who weren't here at the very beginning of this meeting, and I'm going to represent the FSIS exploration of this technology in our regulatory programs. And what I'd like to do is to tell you a little bit of a story, show you a little bit of the data, and I would say at the outset, that the way to put this is we've identified many more questions than answers, and I think you've heard that theme already, and I think you'll see it in the slides here as well.

So I want to acknowledge my partner, Uday Dessai, that you've heard from -- he's moderated some of the sessions already -- in a team of people here at FSIS who have really in less than 5 years put together quite an enterprise in terms of whole genome sequencing and its application.

Here's the outline, and I won't spend any

time here. We're going to cover these things in fairly short order and hopefully leave a few minutes for questions of us from a federal perspective.

So just briefly about our Agency for those who are new and don't know about FSIS. We are an inspector intensive operation. We have inspectors in every plant that operates in the U.S., and you can see the map that sort of depicts the establishments. We have about 6,000 establishments that we regulate and inspect each day, about 7,000 plus inspection personnel. And one of the consequences of that work and the way we operate is to do lots of sampling of food products as part of our verification activities, and you can see here, we generate almost 10,000 bacterial isolates per year, and so that's sort of the foundation of the work that we're able to do with whole genome sequencing.

You can see on the right our authorities and just to be clear here, we regulate meat, poultry, and processed egg products, and FDA regulates all the rest.

So we made a very purposeful move towards

whole genome sequencing, and I'm not going to go through all the details here, but we started several years ago to begin to plan for the use of this new technology in our regulatory programs. It's become a prominent part of our Agency's strategic plan which was issued last year, as well as our annual plans, all of which are available to you if you're interested. And again, this is something that has taken some special attention and time to do. It was a high priority in terms in both the budget and the resources that needed to be realigned within the Agency.

Just a brief mention here about our workflow: You've heard about others, and their workflow and this slide is maybe a little hard to read but it compared the workflow using pulsed field analysis against whole genome sequencing, and the bottom line is to just show you that whole genome sequencing really takes only a couple of extra days for us to do it to get that sequence uploaded to NCBI. So the turnaround time is pretty good. It's a one day extra for *E. coli* or STECs, but really in

less than a week, we can have our data uploaded.

Here's a snapshot. We use this snapshot at various venues to kind of mark our progress in terms of bringing up the technology and adding sequences to the NCBI database, and you can see the bar graph on the upper left, shows you the number of isolates that are sequenced per quarter and our steady state, our goal was going to be about 2,500 isolates per quarter. So you can see that toward the end of the last fiscal year, we approached that, we exceeded 2,000 isolates per quarter.

On the right you can see sort of our history starting in July, and I should point out here that we really depended a lot on FDA in the initial phases to help us get our program up and running.

On the bottom, the table just shows you the various sources of the isolates and by pathogen and you can see that in the history of our efforts, we've now uploaded more than 10,000 isolates.

I did want to provide a little perspective. You can see the *Salmonella*, just over 6200 isolates have been sequenced and uploaded, and to give you

some context, there are about 110,000 *Salmonella* sequences in the NCBI database. So that will give you a little bit of background there.

Okay. Just another fact sheet really about our capacity building. We have three regulatory labs. We now have sequencers in all of those three labs and are producing sequences there.

This past year, which just ended in September for us, we had sequenced and uploaded over 7,000 isolates, remember out goal being about 10,000 a year. We do intend now in this current fiscal year, starting in October to sequence every single pathogen isolate and some proportion of the indicators which are isolated during our NARMS work, and I'll come back to that in just a minute.

And on the right you can see the metadata that companies each upload to NCBI, and it's somewhat similar to what you've seen from the FDA, but I do want to point out that the product and source, of course, is really important but we just upload the year that the sample was collected and the state in which it was collected. There's no more specific

information than that that's uploaded as part of the metadata.

So this describes sort of the rest of the talk which is kind of where we're going with our analysis, and just outlines some of the different types of applications that we've begun to explore within FSIS, and I'll cover each of these in turn.

So, first, I want to show you two slides that are related. We used whole genome sequencing in an outbreak investigation in retrospect. So I wanted to point that out, and let me see if I can orient you to the important pieces here.

So this is an outbreak. This is actually two outbreaks that occurred in the same facility, separated by just about a calendar year, and what you can see here is most of these isolates represent clinical and food or environmental isolates from the second of the two outbreaks, but there is one clinical isolate from the first year's outbreak and you can see how tightly related those isolates are. So this is the first of two slides related to the same outbreak.

Now, here's another analysis which looks at the simply commodity isolates. There are no clinical isolates in this depiction here, and what you can see also, at the point here, is that there are some isolates here that represent at least in one case a different PFGE pattern and 2 different years' worth and see again very tightly clustered.

The next area that we've started to apply this technology is in exploration of harborage in ready-to-eat facilities. We had been using pulsed field analysis to help us understand the extent to which *Listeria* might be harbored in a plant from one year to the next or from one sampling event to the next, and we've now overlaid that work with the use of whole genome sequencing to help us understand the extent to which there could be harborage in plants. We know, of course, the potential is there.

I also want to point out, as was mentioned earlier, there are about 1,000 ready-to-eat facilities that are dually inspected by both FDA and FSIS. We call them dual jurisdiction establishments. So this information is obviously of interest to both

FSIS and to FDA.

Now, in this slide, there are two points I want to make. One is that this is from one plant and I want you to focus on this cluster. I know it's hard to read, but the red triangle here represents a sampling event in 2012, and all the rest of these isolates were from the environment or food contact surfaces in 2015. So you can see that spread over a 3 year period, here are isolates that are very closely related. I think it says 0 to 3 SNPs there.

The bottom part of the slide just has reassured us that in most instances, there is widespread agreement in our analysis between what pulsed field analysis and whole genome sequencing is telling us, whether using SNPs or whole genome MLST. There are some differences however. There are some instances that we've encountered in which different PFGE patterns can be aggregated through a common genome sequence.

Okay. Now, this slide, we're turning our attention now to *Salmonella* from chickens. So it's specifically about those, the upper part anyway, is

about that in particular. And we looked at the serotypes that we commonly find in chicken products, in our chicken sampling, and you can see the serotypes which are familiar to all of us and the numbers of isolates from those various serotypes. And then we looked at this number and saw how closely they aggregated through a SNP analysis, and then we compared them to clinical isolates.

And so if you looked at, going out to 20 SNP difference, you can see the agreement or the sameness of those clinical isolates to the product isolates, and you'll note a couple of things that are of interest, and then we did the same analysis including just 10 SNPs difference.

If you look at Kentucky, which is commonly found in chicken, there's 0% of both whether you go to 10 or 20 SNPs of relatedness to clinical isolates. And you can see some of the same thing with Typhimurium, and the story with Typhimurium is different because we find Typhimurium in every single one of the products that we regulate. So that's the explanation there, but this sort of analysis can help

us as we move forward.

On the bottom right, you can see this has to do with resistance. We did an analysis of 1700 plus *Salmonella* isolates and about half of them were pan-susceptible and had no resistance genes and then you can see a list of those commonly isolated or detected genes in those samples.

There are several slides here I'm just going to kind of run through quickly on geographic distribution. We have undertaken some analyses to see whether we can determine, and these are all related to *Salmonella* in cattle. So we looked at Dublin, Montevideo, and a couple of other serotypes you'll see in subsequent slides, and I'm just going to kind of scroll through these.

Here's some more on Montevideo -- let me just go back one second. The story here for both Dublin and Montevideo is that these are highly diverse serotypes which we sort of understood already, and there doesn't seem to be a lot of geographical clustering.

This slide's just a snapshot of the extent

of the diversity. So you've got the same PFGE pattern depicted in this box here and yet there's significant diversity within *Salmonella* Montevideo.

Here's a look at *Salmonella* Newport which is another cattle adapted serotype, and again lots of diversity and very little geographical clustering. Now, of course, for us, we take the samples at the slaughter plant, and cattle movement is a big factor. So that's one of the confounders in trying to do this sort of analysis. The cattle may have come from lots of different regions and that may be why we are not finding the geographical relationships that we expected.

I'm going to end up by talking just briefly about some of our work with NARMS. Now, I think you heard the last 2 days, prior to this meeting, we had a NARMS meeting. It was a public meeting, and there was a lot of discussion about the use of whole genome sequencing in the resistance context. You've already heard a little bit about this at this meeting today, and this is our analysis looking at cecal isolates of *Salmonella* for one calendar year, and the point of

this slide is that this is sort of reaffirmed what has been shown now for several years, in that there is a high concordance between the genetic elements found using ResFinder or other tools to determine the resistance genes and the AST that's done more traditionally.

We've also been able, within the context of NARMS, to do some work to identify specific resistance genes and you can see the list of the genes that have been detected as a result of the collaborations across NARMS. Certainly the CTX-M-65 is of great interest because it confers ESBL as well as resistance to some other antimicrobials as well.

What we've tried to do in our Agency is to let the producer community, the slaughterhouse operators know when we find these genes so that they can be aware that we have been doing this work as well as to let them know that they may want to examine their food safety systems.

Two other quick applications: We have begun looking at *Salmonella* pathogenicity islands. We are very interested in the Agency in virulence.

So we will continue to be interested in trying to isolate virulence determinants through the application of this technology. And you can see on the left, I'm not going to say any more about it, but the *Salmonella* pathogenicity islands have been well described, and we can use this approach to helping us understand the isolates that we obtain in our regulatory programs.

On the right side, there has previously been described heat resistance genes in *Salmonella*, and then when we use that knowledge and looked at our beef derived isolates of *Salmonella*, we determined that there were no heat resistance islands in those samples, but again that's an application of a specific approach we can take looking at the isolates that we generate.

I'm just going to end up here with a couple of slides. Here's sort of a graphical depiction of where we are and where we hope to go with the application of this new technology. So on the left, you know, we've spent a lot of time learning about the power of whole genome sequencing. We've had

relatively modest goals, both in terms of developing our capacity, whether it's the machines themselves or our human capacity, our resources, in terms of personnel, and then beginning to apply the technology as I've just described it.

You know, as we move toward 2030, and we're always focused on the healthy people goals, we hope to do more of the applications, some of which are described here and have greater goals, all of which we hope will help us to decrease foodborne illness related to the products that we regulate.

This is just a reminder that we're part of a big collaboration. I think that almost all of these organizations are represented in the room here, and we've attended all of these meetings that you see outlined here. I think it's important to say that a few times during that meeting, to reassure everyone in the room that we are in this together, we're learning together, and we need to have a shared understanding about the meaning of the work that we do.

And there's just one summary which I've

covered pretty much. We have the capacity. We now have the resources in terms of the staff we have assembled. We'll continue to hire out a little bit more in terms of folks who can do both characterization work in the lab as well as bioinformatics, but we're approaching that capacity.

And then, you know, the applications within NARMS are important, and there's a lot of good work being done and will continue to be done there as we are contributors to NARMS.

And then finally, just to reiterate, as has been said many times here in this meeting already, this is a tool. It needs to be placed in the context of our other findings and certainly for FSIS, traceback is a real key finding. We haven't talked much about traceback, but the ability to traceback to the plant that has produced product that may be contaminated and cause illness is a real key factor in this, and so we want to use whole genome sequencing in that context. And with that, I will end, and we'll have a few minutes for questions.

Thank you.

DR. EVANS: Thank you, Dr. Besser and Musser and Goldman, and I will just ask David, if you could come back to the stage, and we're going to have -- this is not on our agenda, but we're going to have 5 to 10 minutes where we give opportunity for the folks in the room and folks on the webinar to ask questions from our first three presenters. So did we have any questions on the web?

DR. ABLEY: Okay. We do have a question from one of our online. To what extent will random sampling of retail product be pulled, tested and positive isolates ran on WGS? Will a risk-based approach be taken in recalls if nothing relates to illness, ready-to-eat versus non-ready-to-eat?

DR. MUSSER: So as a matter of routine, we don't like collecting retail samples. We sometimes do surveys to look at things other than microbial contamination but we don't generally pull retail samples because it's extremely difficult to figure out where the contamination may have come from based on retail sample analysis. It doesn't mean it's not done. It's just not a good way of finding more

problems occur. And what was the second part? Read the second part again.

DR. ABLEY: Will risk-based approach be taken in recalls if nothing relates to illness, ready-to-eat versus non-ready-to-eat?

DR. MUSSER: I'll take a stab at it, and then I'll let David do it. I think, yes, typically in non-ready-to-eat, there's less likelihood of illness and we tend to focus our efforts on high risk commodities. So we don't typically test lower risk commodities. We still do, but not as much, and so our risk analysis and our risk-based inspections are just that. So it would be on the highest risk which would be ready-to-eat and the non-ready-to-eat would be less tested.

DR. GOLDMAN: I would just add that at FSIS, we have both risk-based sampling as well as randomized sampling, and sometimes they interact within the same sampling program, looking at the same type of product. So our sampling is principally for verification of a good safety system. So that's the biggest driver here, but within that, again and

Listeria would be one example, ready-to-eat products, we do do risk-based sampling within that based on some algorithms we've develops.

And so in terms of the recalls, you know, once there are illnesses, it doesn't matter whether it's from a ready-to-eat product or a raw product. We would still proceed accordingly.

DR. EVANS: Any question in the room?

MR. McDERMOTT: Hi, Pat McDermott from FDA, Center for Veterinary Medicine. I have a question for Dr. Besser about the confidential information. I understand some states don't allow isolate level information to be put into a public domain. Will that apply to whole genome sequence data from PulseNet as well? In other words, will some of the whole genome sequence data remain behind the CDC firewall or can it be de-identified enough that all the states will allow that isolate level information?

DR. BESSER: Good question. Thanks, Pat. It's my understanding that all states have signed onto the MOU with whole genome sequencing, that the data that we request through the PulseNet database is

uploaded into the PulseNet data which is kept behind the state firewall. Some of that goes to CDC level, and then the minimal dataset that the agencies have all agreed upon gets uploaded to NCBI.

So I'm not sure that any of the states have declined to put in the information. I could be wrong, but I think they've all agreed to it.

DR. EVANS: Are there any other questions in the room for our first set of speakers?

(No response.)

DR. EVANS: So we're going to break for lunch, and just to remind you, the lunchroom, as you leave this room and take a right, go down to the third wing and take a right and the lunchroom is halfway down that corridor. We'll be back in this room at 1:30 p.m. with three more presentations on Federal/State Collaboration. Thank you.

(Whereupon, at 12:06 p.m., a luncheon recess was taken.)

A F T E R N O O N S E S S I O N

(1:31 p.m.)

DR. EVANS: Okay. We're going to start again. Welcome back from lunch. I hope everybody had a great lunch and a great opportunity to talk with colleagues and maybe some new people, meet some new people in the meeting. That's one of the purposes of these public meetings as well.

So we're going to continue on in this session on Federal/State Collaboration, and we're going to dig a little deeper hearing from the Minnesota Department of Health on how they're using whole genome sequencing in their public health practice, and then also how the FDA and NARMS is using whole genome sequencing to look at the public health problem of antimicrobial resistance in bacteria. And lastly, we're going to hear -- I guess this whole session was intended to kind of illustrate how federal and state agencies are working together and it's really going to crystallize in the last presentation. We're going to talk about Gen-FS, which is an opportunity for agencies to get together

and work on these problems, standardization, common protocols, all to improve the quality of the data that's going into the common databases at NCBI.

So first we're going to hear from Dr. Dave Boxrud from Minnesota, and Dr. Patrick McDermott from FDA's Center for Veterinary Medicine, and lastly from Dr. Bill Klimke at the National Center for Biotechnology Information.

And again, we're going to ask for folks to hold their questions. We are going to have a question and answer period at the end of the day after our break.

So Dave.

MR. BOXRUD: Thank you.

DR. EVANS: Thank you.

MR. BOXRUD: Hi. I'm going to talk about whole genome sequencing at the Minnesota Department of Health. I was asked to talk about how sequencing is being done at the state public health laboratories which is a pretty difficult task because there's 50 states and then many local health departments. So I'm going to give a little bit of background about

how public health laboratories nationally are incorporating whole genome sequencing but then really kind of focus on how we're doing it in Minnesota as just one example.

So this talk is going to give us a status update on whole genome sequencing at public health laboratories, talk about the role of whole genome sequencing at public health labs, how we communicate whole genome sequencing inform at public health labs, a little bit on how we've evaluated whole genome sequencing and so we can understand and interpret the data at public health labs and then lastly, I'll close with an example of the utility of sequencing.

So John Besser showed this slide a little bit earlier, but I wanted to point out that in 2014, Association of Public Health Laboratories did a survey and at that time, 21 state public health labs had a sequencer in house. Now, we have 43 labs and 37 states are certified which means that they not only have the sequencer and have the training but they've been able to show proficiency with their sequencing.

All states in the country have been financed for a sequencer, but by the end of 2017 or early 2018, all states will have a sequencer in house.

But there are still a number of issues that state public health labs are dealing with. Right now we're in this incredible transition time, and we're continuing to do pulsed-field gel electrophoresis which is a conventional or traditional subtyping method while we're transitioning into this new subtyping method. It puts a tremendous resource crunch on the laboratories, probably our main challenge right now in public health labs.

Many public health labs have IT issues, either with storage or their IT departments don't allow them to use certain types of software. Some of that has been largely resolved, but there are still some issues in some states.

Training on both the west lab side and on the bioinformatics side, using the data, is still a little bit of an issue, but it's been resolved in a lot of areas.

Bioinformatics resources is a little bit of a challenge. John talked about how PulseNet is doing their analysis and that's going to be very helpful for a lot of public health laboratories, but right now we're still in a little bit of a waiting mode.

And lastly, ordering reagents is a bit of an issue for public health labs. Reagents are very expensive and sometimes in certain states, it's really hard to put five or six figure orders in and that causes a lot of problems with our ordering issues.

So I just want to start with the very basis of foodborne disease testing. This is traditionally how we've done our testing, and I really think that this is the linchpin of how we identify outbreaks.

On the laboratory side, we get human samples in. We analyze the samples and subtype them, subtype the bacteria in them, the *Salmonella* or the *E. coli* and then report our results to epidemiology, really focusing on clusters so that they can investigate them.

On our epi side, we interview cases and

then investigate clusters and lead investigations.

And I think what is really important is this process, this work on both sides, but also the relationship between the two, there should be a sense of urgency with all of this work and with how this data is being communicated back and forth.

But, of course, we have a lot of other partners that we work with regularly with other state and local health departments, USDA, FDA, CDC, and PulseNet is certainly been a huge partner, and also with our environmental health team in the state for investigations.

Just to talk a little bit more about epidemiology information, these interviews that they do, I'm not going to go into a lot of details, but these are really, really important to try to get as good of exposure data as possible and to get a really good interview takes a lot of time. You're really asking a lot of in depth questions of what people ate, where they ate, specific brand names, when they ate certain foods, restaurants they went to. Each interview takes anywhere between 30 minutes to 60

minutes.

And one of the ways in Minnesota that we have for routinely sharing subtyping information is we created an automatic daily report through our laboratory investigation management system, or LIMS system, and it's broken up into a few different pieces but the idea is that we're able to have a standardized methodical way of getting information from our laboratory to our epidemiologists.

And I'm just going to talk a little bit about a couple of these parts. The first part is just essentially giving a report of what was seen the previous day, very simple information. Some background demographic information on the cases, but also the subtype information. In this case, it's the PFGE pattern. So every PFGE pattern gets a name so that we can communicate that with our epidemiologists.

The next part that is really key is that we take a look back at the subtypes for those serotypes and say, have we seen that for the last 30 days, and that information is also sent to the epidemiologists.

So in this case, we had seven isolates that had the same PFGE pattern that were seen in the last 30 days. Obviously, this is a cluster that our epidemiologists would be very concerned about.

So now that we're going to a sequencing based method, a lot of things stay the same, and there are a few differences. On the public health lab side, we're still looking at human samples, but instead of doing pulsed field, we would be sequencing those samples, but we still need to report that information to our epidemiologists and from the epi side, they're still doing their normal stuff. They're interviewing cases and investigating clusters and looking for outbreaks.

We do have some more partners, actually the same partners initially, but we have another partner which is NCBI, which is where we have a sequence data repository. I would also be remiss to say that FDA has been a much bigger partner with us, with sequencing. We are a GenomeTrakr lab, and we are sharing data back and forth with FDA and they've been a tremendous partner for us.

So now that we're working with whole genome sequencing data, the communication is a little bit more challenging. Some of the previous speakers talked about nomenclature and how they're naming their patterns. So far to the public health labs, we don't have that available. So we created a number of spreadsheets to give this information back and forth and essentially the most important thing is that we have a cluster ID. We tell our epidemiologists what is a cluster and then they investigate those cases, and we also send them an email that lays out what a cluster is and gives a little bit of background information. In this case, it's two cases, the same PFGE pattern, gives some information that's very easy for them to understand and to create an action on this.

The negative side of this is that this is not going to be sustainable for a large number of organisms. It works for the short term, but it's not possible to work in the long term.

So now we want to focus about how we determine what is a cluster and how we have used that

information that we learned in the past going forward. So I'm going to talk a little bit about *Salmonella* Enteritidis, a study that we did with *Salmonella* Enteritidis. It's the second most *Salmonella* serotype in Minnesota but it's very clonal.

Here's the major pulsed field patterns. There's about five patterns that make up about 80% of the isolates in Minnesota. So pulsed field really does not do a good enough job of providing diversity on *Salmonella* Enteritidis.

So we decided to do a retrospective study on SE, and we wanted to look at stability, how stable are the sequences over time, typability, are we always able to get sequence data from every isolate, discriminatory power and also epidemiologic concordance. Epi concordance is essentially defined as if it's an outbreak, they should look very similar. If they're not related, they should look very different.

So our laboratory and epidemiologists met and they came up with a study set, and we used some

very well characterized isolates. We look at isolates from 7 outbreaks but also looked at 22 sporadic isolates. We also looked at some in vivo isolates. So multiple isolates from one person over time. And we worked with the New York Department of Health for the analysis for this.

And what we saw was that within an outbreak, there were very few SNPs. There was a maximum of three SNPs within an outbreak, and when there were sporadic SNPs, they looked very different from the outbreak isolates, even when they were the most common key of key patterns. So that provided us some information on how to interpret this data going forward.

Our conclusions were that the sequences are stable within a person over time. All isolates were able to be typed. There was a lot of diversity, and there was really good epidemiologic concordance. Outbreak isolates looked the same. Isolates that were not epidemiologically related looked very different.

So we used this data to do a prospective

study in which case we looked at whole genome sequencing and pulsed-field gel electrophoresis clusters in real time and tried to identify the source of them. And what we saw was that with sporadic isolates, by whole genome sequencing, they looked very different. There's an average of 93 SNPs different between the most common PFGE pattern that were sporadic, and going back to what we saw from the outbreak isolates, there was 0 to 3 SNPs difference.

So from a sequencing perspective, it was quite easy to see what is like to be related and where we should focus our investigation.

And also the number of clusters that we saw with whole genome sequence compared to pulsed field went up dramatically but the number of isolates within each cluster went down dramatically.

So I think this provides us with a great opportunity to identify outbreaks earlier with fewer cases.

And, lastly, I just want to talk about how we use this data going forward, one great example.

So this is the most common PFG pattern in

Minnesota in August and September of 2014, with each box representing one case of the most common PFG pattern. So there are 19 total cases. So by PFG, it's really hard to understand what is going on. We just know that there's a lot of cases.

When we did whole genome sequencing, we found that there were eight of the isolates that were 0 SNPs from each other. So we investigated those to try to understand what the relationship between those cases were, and so our epidemiologists did a great job and they would interview almost all of the cases and found that six of the eight, all ate a frozen chicken Kiev product. One of those eight was a secondary case and another one of the eight may have eaten a frozen chicken Kiev product. They ate a lot of frozen stuff. So they didn't remember it specifically.

Of the 11 that were not part of this cluster by whole genome sequencing, none of them ate this product. So we were able to use a whole genome sequencing to identify a product that would have been very difficult to identify by pulsed-field gel

electrophoresis, and it certainly would have been delayed.

So the future of whole genome sequencing at public health labs is going to continue to increase. WGS will provide more information methodically compared to current methods, and it's very close to replacing some of our current methods such as pulsed-field gel electrophoresis and serotyping.

When we have a standardized WGS nomenclature that CDC is producing, it will greatly improve our communication and make the data sharing much easier.

So, in conclusion, WGS is a tool that can help us identify clusters and outbreaks better than our traditional methods, but there are current challenges and there will be additional challenges as we implement this technology, but speed and communication will be vital to continue these aspects for outbreak investigation.

Thank you.

DR. McDERMOTT: Good afternoon, everyone. Thank you to USDA for the invitation to join this

conversation and show you some of the ways in which we're exploiting this very power next-gen sequencing data in our antibiotic resistance surveillance work at FDA.

So a little context: I think everyone in this room is well aware of the global threat posed by antimicrobial resistance and a great deal has been done even in the last few years at the highest political levels to encourage and lead countries in developing and I should say dedicating resources to addressing it on a global/international level, and WHO came out with their global action plan for antimicrobial resistance back in 2015.

And some of the summary points in that report that helped set the context here, that they point out is that the development of resistance is linked to how often antibiotics are used. I'd say that's true in nearly every case, not every single case. Because many antibiotics belong to the same class of medicines, resistance to one specific antibiotic can lead to cross resistance to others, and I think if you hear these simple facts, you will

start to see quickly how whole genome sequencing can resolve some of these issues in more detail.

Resistance that develops in one organism or location can also spread rapidly and unpredictably and can affect antibiotic treatment on a wide range of infections and diseases, including those that spread between animals and humans. So it's important to keep in mind that the same classes of antimicrobials are used in food animal production, in treating our pets that's used in human medicine.

Now, there's some restrictions on that in pet and animal production. There are no restrictions on that in the antibiotics our veterinarians might use to treat infections in our companion animals.

And so we know resistant bacteria can be found in food animals and food products destined for human consumption including those same genes and strain types as we know from illness in humans.

And certainly my mentor in my postdoc years I thought just put it perfectly when he said antibiotics are societal drugs. It's the only class of essential medicines or any medicines which used in

one environment can compromise or affect how effective they are used later on in a different environment.

And so that brings with it a sort of new set of responsibilities, different ways of thinking in One Health, which I'm sure you all are also familiar with, has been the framework in which WHO has tried to build capacity in countries to do surveillance for resistance and also to address it on the policy level.

In the past, what we do in our NARMS program has been defined as integrated surveillance, and WHO defined that as the coordinated sampling and testing of bacteria from food animals, foods, and clinically ill humans and subsequent evaluation of resistance trends throughout the food production and supply chain using harmonized methods.

And I'll show you how that's what we do in our NARMS program, but one thing that we've been contemplating that came out of both a recent review of the NARMS program and our meeting in this same place, yesterday and the day before, is what are the

prospects and value in moving towards more of a One Health paradigm, and that One Health paradigm means sampling beyond that integrated food chain and looking into the environment and looking at animals and not just zoonotic bacteria from animals but animal pathogens and elsewhere. And again, whole genome sequencing technology, especially metagenomics, is going to make this more possible than ever I think.

So the purpose of NARMS, what we've said publicly for many years, is to monitor trends in resistance, get this information to people who can act on it, conduct research to fill in the gaps. It's very difficult to sample yourself to every root cause or to every solution and research has been an important part of filling in some of those gaps.

More recently, it's become adopted by CDC to prioritize outbreaks based on, now more than ever, whole genome sequencing data, and it's an important part of FDA's regulatory processes whereby a new animal antibiotic being proposed for review by a sponsor goes through qualitative risk assessment

steps that include mechanisms of antibiotic resistance, current resistance to other compounds, and so on and so froth.

So another way to look at this process overall and how surveillance fits into it, obviously surveillance is a key component to any public health action if you're going to get ahold of the magnitude and nature of whatever that hazard is, in this case, resistance, and understand the benefit of any interventions that were targeted based on the baseline and trend data.

So I describe NARMS as beginning with establishing baselines for resistance in different pathogens from different sources; how that resistance spreads; what it looks like over time. Can we get to where it might be coming from so that our decision making, whether it's regulatory or not is more of a scalpel than a hammer, as some described that? Understanding the contribution or the relationship of antimicrobial use in resistance is a tough one, that we're still struggling with but ultimately it hinges on human health impact and the burden of resistant

infections in humans.

And when that situation becomes untenable, then the interventions that are put in place again give NARMS, if we're doing our job right, an opportunity to measure that impact.

So this is the basic structure of the program. It is an interagency program between the FDA, the CDC, and the USDA. The Centers for Disease Control is getting isolates from the field, if you will, from clinical labs and physician labs around the country. Every 20th *Salmonella* is subjected to standard in vitro antimicrobial susceptibility testing but as we've heard, eventually all, and soon, all of the *Salmonella* reported will also be sequenced, but we'll continue to have phenotypic data on 5% of them.

USDA-FSIS has a random sampling of national production at slaughter of the four major food animal species, and then at FDA, where they retail meat testing is done in 2018, we'll have 22 states sampling 80 products every month of beef and pork and chicken and turkey, and then isolating *Salmonella* and

then characterizing them phenotypically and sequencing right now is being done at the FDA labs.

So that the integrated portion and structure right now. We combine these data into integrated reports where we try go harmonize our reporting as well as our methodology and our latest integrated report came out this Monday.

I think the CDC human report which looks beyond *Salmonella* and *Campylobacter* to other enteric pathogens, should be coming out very soon.

So the issue about what does this mean for resistance monitoring and beyond that, what does this mean for susceptibility testing and maybe some day clinical care?

You all know that currently what we do and what we've done for many decades to assess susceptibility in an organism is we do a biological response within serial dilutions of antibiotics and look for the lowest concentration that inhibits visible growth.

There's a lot of talk about how good is genomics going to be abridging to these old methods,

and I think I would start out by saying the old methods aren't as good as you might think they are. It's a biological response in an artificial situation and it's similar to what Dr. Besser said about PFGE. It's a predictor of the likelihood of success, what it's correlated with, with other parameters. The method itself, the laboratory method itself allows the three drug dilution range for QC for most organisms. Well, that could be 1 or 4 µg/ml. That's a pretty broad range, and in some cases, CLSI permits a 4 drug dilution range to be in QC. So it's not, you know, the chemists would always make fun of us when we tell them about this because it's very imprecise in a lot of ways, and it also doesn't always reflect all the things that go into resolving -- well, it doesn't reflect all that goes into resolving an infection.

So I just want to begin by saying that while the gold standard is MIC testing for susceptibility, it's got its own perils and so I think that's a good place to start.

And so when we start looking at how well

genomics can predict resistance or be useful in surveillance or in guiding clinical therapy, it's good to keep in mind that that's not a perfect method either. It doesn't always correlate well with clinical outcome as well.

And in a lot of cases, the studies haven't been done at all, and so breakpoints will be borrowed from one pathogen or animal into another without the clinical outcome data.

So you have this MIC. So what do you do with it? Well, you know, the title of this is the art and science of drawing a line somewhere, and it is an art and a science, and there's two ways right now which the data is considered.

One is how well does it predict clinical resistance? One MIC set with using a lot of data, using wild type MIC distributions, using clinical outcome information which is really the key one, using PK/PD data that addresses the concentration of the drug, at the site of infection, and then common standard of practice in medicine goes along with that. And FDA is responsible for setting those

breakpoints.

Another approach is what's called epidemiologically cutoff values. Let's identify everything that's no longer wild type and use that marker as a way to monitor resistance, and we could have a long conversation about fighting over what that word means because to some, it only means the former. It only means clinical resistance and clinical breakpoints but in some cases, to other people it means non-wild type or decreased susceptibility, depending on the questions you're asking.

And so in one case here, the breakpoint for epidemiological cutoff value in this mockup would be 0.25, or in this case, 0.5. So here's a wild type distribution. You might have intermediate MICs in here that used to be indeterminate, recognizing the limitations of the method and then a clinical breakpoint. And so where you set that point can affect the data that you are describing over time.

Well, CLSI has tried to capture all these different definitions into one and describe resistant

strains, as those not inhibited by the usually achievable systemic concentrations of the agent with normal dosage schedules, and/or fall in the range where specific resistant mechanisms are likely, and here's our genomics, or clinical efficacy has not been reliable in treatment schemes. So you can see they're trying to really make everyone happy and sometimes the definition is as loose as the data.

So you've all seen this slide in some iteration probably before, but when genomics became routine and affordable, the first question in our surveillance system is, how do we incorporate it? How well can we bridge from the old to the new? How well does it predict resistance in pathogens? And we started off with a very basic question, what is the correlation between the presence of known resistance genes and isolates with a MIC above just the clinical breakpoint?

And then I'll say a little bit about how we're exploring metagenomics as well. And we've done three studies, a main study looking at this, and as Dr. Goldman pointed out, the correlations for the

bacteria under surveillance in NARMS is very high, 95 to 99% correlation between MICs above the clinical breakpoint and the presence of known resistance genes, and we've done it for *Salmonella*, *Campylobacter* and *E. coli* and Greg Tyson who has done a lot of this work and is working on *Enterococcus*.

When NCBI and we started conversations about building these analytics into their pipelines, we expanded it to over 6,000 *Salmonella* now. The first study was 600, and the data holds up. It's very highly correlated with this one MIC, the resistant MIC in the presence of known genes.

We did a collaboration that's ongoing with Argonne National Labs where they took a machine learning approach and just said, can we predict MIC from the genome, blinded to the presence of any known resistance genes, and I couldn't explain the computer part of it. That's a little beyond me, but basically it was a k-mer based approach looking at correlation with onefold serial dilution above or below the MIC, and again it holds up incredibly well. This actually surprised me to see that you could actually predict

all the MICs for which we had ranges with the same high degree of confidence.

Now, I should emphasize, this work's ongoing and we're still trying to get some more dilution ranges to improve the data but it's incredible how well you can use the genomic data alone.

So what CDC's doing in sequencing all the *Salmonella* is a resistance monitoring system in my mind because these correlations are so good. So the dataset has just taken off with every *Salmonella* that's isolated. We can consider that in our risk assessment and regulatory processes at FDA because the correlation is solid. I'm going to skip one slide.

On the metagenomics side, just a few quick points. We have started to apply this technology both looking at animal cecal samples in NARMS and looking at the retail meat isolates to get an idea of really to sort of stress test the technology and explore its limitations for doing routine surveillance. And Daniel Tadesse in our group has

been working on this and we have kept every cecal sample collected in NARMS for the last 3 years and have some 20,000 of them, nationally representative randomized samples of U.S. food animal production.

So we didn't want to let these samples go in part because new policy changes were coming, and so we saved them all and have started to look at them metagenomically and started as this slide shows, through associate specific resistance alleles with different animal sources, and also in the retail meat samples as well. So stay tuned for updates on that.

In some of our earliest studies, back in, oh, my goodness, it must have been in '05, '06, when we did our first sequencing and it was very expensive, we got some information that I think is still a very good illustration of what the data mean to us and different ways in which it can be presented, and this plasmid just happens to show one of our first outputs which was a multidrug resistant *Salmonella* Newport on a backbone that was essentially identical to a strain from a child in Madagascar who had plague. So this plasmid, when you start thinking

about One Health, you see the importance of global One Health. What does it mean? Well, it certainly means these things can develop quickly and spread rapidly around the world and what is seen halfway round the world can become fairly common in the U.S. food supply, and this just shows different arrangements of genes but another part of it that's interesting and important is we see associations with resistance to decontaminating chemicals used in processing plants.

So new drivers of resistance emerge in these data, new associations and arrangements of genes with some indication of where they might be around the world.

And the next event is here, the next slide just shows a spike in gentamicin resistance that we saw in NARMS data, and it just illustrates the fact that we had never gotten to the bottom of the genes behind this without genomics because the PCR primer is available for aminoglycoside-resistant genes at that time didn't have all the alleles we found. So the new discovery is an obviously part of what we get

from whole genome sequence data routinely now. So what we called research and ad hoc PCR studies is now routine surveillance.

Another important event that showed the value of this technology was plasmid-mediated colistin resistance that emerged in China in the fall of 2015, Errol Strain was really helpful in this in going into NCBI's database and saying have we ever seen this in a domestic isolate of any bacteria in the United States, and at that time, screened 155,000 genomes at NCBI and said it wasn't there. Incredibly powerful to do retrospective surveillance like this without opening the freezer.

We later did studies of selective enrichment and found *mcr-1* in 2 out of 500 swine samples, and the metagenomics tools we were developing at the time picked it up as well. So that bodes well for the future of a metagenomic component to the surveillance.

So that's just a few illustrations of how we're building it into NARMS. I would invite you to go look at our last report that came out Monday where

we've put genomics now into our dataset. We went back and sequenced every retail meat *Salmonella* back to 2002, and as David Goldman pointed out, they're sequencing all the animal isolates. Next year we'll have all the human isolates. So knowing how good the predictions were of genotype to phenotype, we put quite a bit of effort into interactive data displays where you can look at resistance over time and see the genes that go with it from the different sources. And it's really powerful, and it gives people access to the data in new ways and allows them to ask and answer their own questions.

One of the good ones is we've gotten beyond this really crude metric of MDR, multidrug resistance, being resistance to three or more drugs. Well, what drugs, right? So now you can go in and see what drugs and refine the analysis and have more confidence in what you're talking about in the trend.

Another thing we've developed is, to take the next obvious step, we harvested all the NCBI data, screened it for all its resistance genes and took advantage of whatever metadata that were there

with its strengths and limitations, and you can do the same thing now with this tool we called Resistome Tracker, where you can look in any source according to the metadata categories, at any resistance gene in this case, an aminoglycoside resistance gene, *aadA1*, in chicken. You can see where it is at NCBI by biosample, and then what our intention was, just to set it on top of NCBI. You can click on any of those isolates within any of the genes you're interested in and go right to the SNP tree.

So this just seemed like an obvious thing to do based on NARMS data and taking advantage of the sequencing that was coming out.

And then the last tool we built into Resistome Tracker is, well, what's new? We want to know the wallpaper in the room, if you will. What's the backdrop for analyzing our domestic data in NARMS? So any new gene that comes up in any submission to NCBI, we've put an alert system on there so we can track it and see the data in which it was reported.

And Heather Tate who is here has developed

this Resistome Tracker tool, and we're going to try to launch it. It's not out yet. We got busy this week, but maybe in the next week or two, you'll be able to play around with this and see if you like it. It includes a mapping tool where you can look at over time when these different resistance genes appeared around the world, and you can apply this, of course, to any pathogen.

So we've put a lot of effort into making sense of the complex data that we are all generating through genomics.

So, in closing, One Health - One Method. I know it's a big of an exaggeration, but it sure feels like it. We can predict so much from the genome. We can predict resistance in our target pathogens so well from the sequence data. It overcomes so many past limitations with the metagenomics, you know, our reliance on cultivatable organisms which is expensive and obviously very narrow, and what we can say about the Resistome.

The best part to me, the second bullet, permits us to look farther with fewer resources and

lower costs. So now we can get out and look at the costal waterways and look at surface water systems and start incorporating companion and add on other testing because we can glean what we need from these samples, and that's where we're trying to go next.

I've noted how it reveals new associations including determinants perhaps of animal origin of other drivers of resistance and new alleles, greater confidence in decision making. That's important. We would like to note just where a pathogen comes from, but where a resistance comes from.

And with that, I can't thank everyone, but it's been a tremendous opportunity to collaborate broadly, and it's been a real satisfying experience working with NCBI, who we will hear from next, and CFSAN and CDC and USDA on this. Thank you.

DR. KLIMKE: So this is great. Everyone's talked about NCBI. So I don't have much to talk about. Go to the next slide please.

So for those of you not familiar with what NCBI is, we are the premier biomedical informatics institute on the planet. We are the host for all the

PubMed, all the medical literature over 23 million publications, all the clinical data, and all the sequence data that you've heard about from today is being submitted to our databases, and then we have specialized databases and tools, and I'll talk about one in particular today called our pathogen detection system.

Next slide.

And again to reiterate what someone has already said, we share all the sequence data with the European Bioinformatics Institute and the DNA database in Japan within 24 hours of the data being released.

Next slide. Go to the next one.

This is a rough schematic of how our system for the pathogen detection pipeline works. We have submitters on the left including, you know, FDA GenomeTrakr and CDC PulseNet and USDA, submitting data to us, into the public databases. And then we have a pipeline which I won't go into a lot of detail today. There will be one other slide on that. We basically do an assembly, a clustering, and we

produce phylogenetic trees, and we make the reports publicly available to all of you.

Next slide, please.

I want to talk briefly about the metadata. This is the template that we built, and you seen that from a few other people already this morning. This came up from discussions with FDA about the sort of metadata they wanted to see submitted to the databases that would enable them to do their job. It's basically four categories of information. What is the sample, including the organism, a unique identifier such as a strain or isolate number, it's categorization, whether it comes from a clinical/host-associated sample or from a food or environmental sample, and then a few other categories such as when, where, and who.

The ones I've colored in yellow are the absolutely minimal sample fields that need to be filled out in order to submit using this template, and that's what CDC is using right now, and 6 months after the sequence gets deposited, they update with some additional metadata fields.

The USDA and FDA, and you've already heard from some of them this morning, they are willing to submit things like the geographic location, the year, and month the data was submitted as well.

This has been an incredibly successful template. So we have many, many other templates for bacteria, for example, at NCBI. This one has been used by over 256,000 submissions to date, and that's not just samples from PulseNet and GenomeTrakr. That includes other academic laboratories and other state actors that are willing to submit data to it, and you can see the breakdown of the clinical versus the food and environmental. I'll talk about some of these metadata fields later on.

Next slide.

So basically what we're trying to do in this system is we're taking very large volumes of data and reducing it to relevant data. So we might have hundreds of millions of base pairs in a short read sequence for one particular isolate. We do an assembly, and then we're turning that around into a phylogenetic tree along with associated antimicrobial

resistance genes, and we want to do that within 24 hours of the sequence being deposited, you know, so turning big data into useful data.

Next slide, please.

And we have to do this because we know that the sequences that are coming in are going to increase in the future. So this is just a snapshot of the last 6 months of the data that's been submitted to our pathogen detection system, and you can see it's dominated by *Salmonella* in that orange slice of the pie, and then the other three foodborne pathogens are basically making up the other 95% of that. And so we have a few clinical things that are coming in, but basically it's the foodborne isolates that are coming into our system.

And we know that the *Salmonella* that are being collected in the U.S. should all be sequenced by the end of 2018, and so actually we should see the rate of that double within the next 18 months or so. So we should expect, I believe it's 90,000 isolates per year for all the foodborne pathogens in the U.S.

And you can see at the bottom is the graph

showing the sort of number of submissions per day and you can see, when we first turned this on in 2013, that's when we first started the Real Time *Listeria* Project, where all *Listeria* in the U.S. were being sequenced and submitted to us, and that's ramping up for the other foodborne pathogens. You can see both the spiking nature, sometimes we get large volumes of submissions from, you know, Public Health England is an international partner that sometimes submits in batch mode. They might submit 1,000 or 10,000 in a day sometimes.

But you can see that it's growing, and we would expect that rate to increase as we move forward into the future.

Next slide, please.

This was a snapshot of data that I presented at NIST as part of their standards for pathogen detection. This was basically looking at the total number of submissions and whether they're clustered or not for the four foodborne pathogens, and again you can see *Salmonella* is predominant on this slide.

The interesting thing about the clustering, so I didn't mention this, I'll probably mention it in the next slide or so, we make clusters of 50 SNPs or less. So very closely related isolates, and you can see most of the *Listeria* and most of the *Salmonella* clustered, *E. coli* and *Shigella* not so much. So that's an important byproduct of the biology of those organisms, and I'll come back to this in some of the analysis I do.

Next slide, please.

So this is the last pipeline slide you'll see. So those of you who were at my other talks can decide if you don't need to see any more of those, but basically the data comes into us, you know, and it's several days before it gets sequenced and submitted to NCBI, but we do an assembly and we are starting to do wgMLST allele calling and producing a table of nearest neighbors. We're doing that now for *Salmonella* and *Listeria,* and that's been running for 2½ months, and we're aiming to get that data back to FDA within 1 hour of the sequence being deposited. So that's an incredibly fast turnaround time, where

the sequence comes off the machine and comes to NCBI, and if Errol does it right, he says, you know, he'll submit it in the morning, have his coffee and then he can look at the results, you know, by lunchtime. And he can make a decision about the inclusion/exclusion within that 1 hour.

We're also doing SNP clustering. This slide, you know, we're trying to replace the initial clustering we do using whole genome MLST and we aim to get the clusters of these nearest neighbors into phylogenetic trees and on the web-based interface, and I'll show you some examples, within 24 hours.

So the rapid reports just gives them an extra day to say, you know, these isolates are related or not related and make decisions based on that.

Next slide, please.

So just for the rapid reports, like I said, we're doing *Salmonella* and *Listeria*. We're reporting the five nearest neighbors in all neighbors with less than six allele differences, and that cutoff may change. You know, we're right in the pilot phase of

this project, and we've already gotten some preliminary feedback that we may want to change that threshold, reporting the number of difference, number in command and the SNP accession if it exists. So if they can actually see within the list of nearest neighbors that some of those are already participants in an existing SNP cluster, they can actually take additional steps.

And we put this into one file per run. Most of those reports are already being made available on the FTP site. So I won't show you an example of that, but you can go take a look.

Next slide, please.

And all of this data is being made available into our publicly available pathogen detection website. So you can go to this page right now and see all the data being submitted from GenomeTrakr and PulseNet, and you can interrogate the data, and I'll show you some example of that. There's a couple of ways into it, but we already have over 174,000 pathogens and at least 142,000 with either an acquired or chromosomal antibiotic

resistance gene.

Next slide.

So I just wanted to do a quick analysis, just basically for my own benefit. Using the metadata that's being submitted, I wanted to look at the isolation source. The isolation source, the definition of that is it describes the physical, environmental and/or local geographical source of the biological sample from which it was derived. So for a clinical isolate, that might be blood, stool, urine. For a food or an environmental isolate, it might be a river bed. It could be a food contact surface. It would be an environmental swab. It could be cheese.

So next slide.

So I simply wanted to ask the question, if I look at all the submissions coming into our system from CDC, FDA, and USDA, including the state labs that are submitting under GenomeTrakr and PulseNet submissions, and if I want to look at all pairs within 10 SNPs, what am I going to see? And I want to categories each one of those pairs, either they're

clinical versus clinical, clinical versus food or environmental and that could be potentially the smoking gun for causing clinical illness, and then food and environmental versus food and environmental. So go to the next slide.

This is a summary of those pairs. We're just looking at counts. We have the four foodborne pathogens at the bottom, you know, so *Campylobacter*, *E. coli/Shigella*, *Listeria,* and *Salmonella*. And then the pairs are those three categories I just told you about. Clinic/clinical are red. Clinical and environmental are green. And environmental and environmental are blue.

So you can see obviously *Salmonella* clinical cases completely dominates this slide. I'm not telling you anything new. You already knew that *Salmonella* was one of the biggest problems in the U.S.

What you will see though is that the number of cases of those pairs is not uniformly distributed. So *Campylobacter* is pretty uniform between the three categories. *E. coli/Shigella* and *Listeria* are not.

So I think this again informs you about something about the biology of those pairs and the distances that I use. I use a 10 SNP threshold. That was an arbitrary pick. A lot of people that come to these meetings, you know, say what is your threshold? You cannot, again I'll reiterate, you cannot simply just pick one threshold and get the answer. There's additional contextual information. Go to the next slide.

By looking at the number of samples within those thresholds, you can see most of the samples in *Salmonella* submitted are within that 10 SNP threshold to something else. That's not the same for *Listeria*. About 1/3 of the clinical cases are within that SNP threshold, and so it varies again between all the organisms.

The last column is also interesting because I said we made these SNP clusters of 50 SNPs or less. There's a number of cases that are just not clustered with anything or not clustered with a food or environmental isolate. So there's actually over 2,000 clinical *Salmonella* not linked to any food or

environmental isolate at all in our database and, you know, 1500 *E. coli*, and I'll come back to that just near the end of my talk.

Next slide.

So these next four slides are just looking at the four foodborne pathogens and what that isolation source breakdown is, and this is free text. So I'm just basically doing a naïve sort of smushing together of some of the terms into these sort of categories so we don't have a 100 rows. We only have 20 rows or so.

And again, you can see *Salmonella*, predominantly chicken, you know, we have some beef, pork, water. You have papaya, and it's high in yield, and I'll come back to that. And then things like research strains. So somebody in Micro 101 is just not doing, you know, the proper technique that they're supposed to be learning.

Next slide.

Listeria, again swabs, environmental swabs, cheese and various food products.

Next slide.

And then *E. coli*/*Shigella* dominated by soy nut butter and then beef, milk, goat, etc.

Next slide.

And then *Campylobacter*, of course, milk is a big problem, unpasteurized milk and beef again. So go to the next slide.

And this, of course, completely matches what you see on the multi-distributed outbreaks for CDC. You know, you see soy nut butter for *E. coli*. You see papaya. There's a huge papaya problem from Mexico, and you see cheese as a problem for *Listeria*.

So I'm not trying to present this as like a detailed scientific analysis of the metadata. I'm just simply saying that we make all this data available. We can do these types of analysis. This only took me a couple of hours in an afternoon. You could do similar sorts of things, and I'll dive into some more detailed analysis that you can do. So next slide please.

So, for example, a lot of people talked about isolates browser. Here I'm just doing a search for papaya into our isolates browser. You probably

can't see it in the back, but there's 225 isolates that come back with a search term of papaya in our existing database when I did this. That big table in the back is called the isolates browser, and basically every row in that table is an assembled gene that's come through our system or come through GenBank. It's got the associated metadata and you can actually, you know, add columns to that based on what you want to do.

We do a separate thing where we intersect the number of searches with the number of clusters and that's the little table at the bottom which is actually on the default page on the upper left. If you can't read that it says, you know, at the top row, 23 of the total 32 isolates in this cluster have the search on papaya, and if you go to the next slide.

So this is our new tree viewer which is not available to you. This is being alpha tested by FDA right now, and so if you went to our page and clicked on the SNP trees and you saw some examples of that earlier, you wouldn't see this exact page. But this

is something that we're testing at FDA because we want to make tools available to them that will help them make a decision based on the genetic distance.

So this is one example. This is a cluster of 710 isolates. Eight of them have the search on papaya. But only one of them is from 2017. If you see on the left, there's a breakdown by date, and if you look at that part of the tree where the food or environmental isolate hits, it's basically on a completely separate branch of the tree. It's not hear any clinical isolates. It's at least over 30 SNPs away from any clinical case, and so that particular papaya probably is not a positive of any clinical illness that's been recorded. If you go to the next slide though, we see the opposite.

This is another *Salmonella* cluster. It's got over 1,100 isolates again from papaya, but 3 of them are from 2017. And if we look at the SNP tree, you can see one environmental isolate is incredibly close to a large number of clinical isolates. Go to the next slide.

And if you look at the entire subtree,

you'll see there's 112 isolates in total, and they're incredibly closely related. The max SNP distance is 17. The average SNP distance if 3, and if you'll go to the next slide.

If we just look at the food or environmental isolates in that total list of 112, 3 of them exist. One is that papaya from Mexico, and two are papayas from the U.S.

And so what I showed you is I think that you can quickly see whether something is inclusive or exclusive, make a decision very quickly using just genetic distances with some limited metadata and then you can go onto the rest of whatever your job is with your public health or regulatory agency.

And again, so we're building tools used to facilitate that. We're not the agency that makes those decisions, but we want to make tools so you're 95% of the way there without any extra effort.

Next slide.

So just going back to unlinked isolates. Again clinical isolates are not within 50 SNPs of any food or environmental isolate. They're not just

singletons. So I thought when I first did this last week, that these would all be completely unlinked to anything but it's not true. Some of them were actually in some very large clusters.

So my question, and I'm going to ask these questions, I don't know the answers to them: Do we need a new sampling strategy? Are there new food vehicles that are waiting to be discovered? What does the epidemiology tell us? And I can't answer those questions because I don't have that information. You people have that information, and so you should be asking these questions and be able to answer them, again pointing out the number of isolates. And if we go to the next slide.

If we just look at one of those, it's an extremely large *S. e.* cluster of 506 clinical cases. There no food or environmental isolates in these cluster at all. They're all collected from 2013 to 2017. They come from the U.S. They come from Public Health England. So there's isolates in the UK that also cluster with these isolates. They come from state labs. So what is the cause?

I've heard something. I've heard a rumor, but I'm not going to tell you because I don't think I can share that, but obviously these are the sort of things we would be looking at, you know, based on the questions that I asked earlier. Using our interface, you can do that but also the places that produce this data can be asking these questions and answering them.

Next slide.

Just to basically end on the antibiotic resistance. You've already seen some of the talks using some of our databases and some of the analyses that we're doing. This came about because the President at the time decided that we needed to combat antibiotic resistance. So we put into motion this thing called CARB Report, you know, from the President's scientific committee Combating Antibiotic Resistant Bacteria. And NIH has mentioned twice in there, these two critical elements. One is to produce a reference database, a well-curated reference database and maintain a national sequence database of resistant pathogens. So go to the next

slide.

So we've done that in a number of ways. First, we've actually built a template to capture the susceptibility that people are collecting. It's called antibiogram. In addition to the sample metadata, the minimal template that we're collecting, you can actually submit AST data as well. We have almost 5,000 submissions using that template right now.

And we have put together a reference database of antibiotic resistance genes. Now, this is not something that we just did on our own. There's a lot of people out there that say, well, which database should I use. Well, we have a collaboration going with CARD. We may have a collaboration going with ResFinder depending on some upcoming discussions we have. We've taken over the Lahey database. So they were the place where you would submit novel beta-lactamase alleles to, and they would assign a new novel beta-lactamase allele, let's say SHV or TEM. They're retiring, and so they've asked us to take over that responsibility,

and so we're the place where you would actually submit those for, you know, when you want to make a publication into *JAC* or *AAC*, you would actually submit to us, get the novel allele number, and then this goes into your publication. Once it gets released, it gets fed into our reference database and used.

And so we're implementing tools for identifying those genes using that reference database, and if we could go to the next slide.

We're actually integrating that back into the isolates browser that I just showed you. So you can actually see the list of genes per isolate integrated directly into the list. So this is an example of another large *Salmonella* cluster. We have 50 isolates encode in *mcr,* so mobilized colistin resistance gene, and you can see just one example, probably you can't see it, but there's a cluster of four isolates from Thailand that all have a *mcr-1* gene.

And so this system will allow you to interrogate for not only metadata but also the

presence or absence of different resistance genes.

Next slide.

So just to summarize, we're enhancing our existing analytical pipelines to improve the turnaround time to answer, you know, basically these two fundamental questions. Are these isolates colonally related? Is there a point source for clinical illnesses?

And we're improving these interfaces, enhancing the information that's layered on top of them including antibiotic resistance genes, making the system much more easy for you to use to make determination of inclusion/exclusion, and in the future, we'll be adding things like virulence genes, heavy metal resistance, point mutations, mobile elements.

Next slide.

So that's it. If you have any questions, you can email that email address or come see me here. These are all the people that I work with on the pipeline at various points. So you can basically see it's an army of people that NCBI that are involved

with this, although the day-to-day operations, the number of people actually involved are just a few people.

I'd like to thank all of our colleagues at FDA, CDC, and USDA for making the data available because we wouldn't be able to build these tools if they didn't make the data available, and thanks to David Lipman for helping to push this system out even though he's now gone off to work on food products actually. I'll be happy to take any questions at any point.

DR. EVANS: Thanks, Bill. So we're going to for a quick coffee break, and we'll be back here at 3:00 with two additional presentations and then an opportunity for questions and answers.

(Off the record at 2:30 p.m.)

(On the record at 3:00 p.m.)

DR. EVANS: If everybody can take their seats, we're going to start up again with a Demonstration of Tools for WGS Analysis.

Okay. I want to welcome Glenn Tillman with the Molecular Characterization Branch of OPHS at USDA

FSIS, and he's going to talk about tools for a WGS analysis.

DR. TILLMAN: Hello, and good afternoon. Thank you for the invitation to come today. It's a very good pleasure after hearing all these great talks this morning. You really set the stage well. The first three talks went into great detail about where we're going with whole genome sequencing. So did the rest of the talks. So I hope to have something good for you all to add. So without further adieu, that's me. I don't like the picture very much.

Okay. And this is what I'm kind of planning on talking about, short and to the point. Give you a little background, but number 2 is the bullet point, what are we doing with this data? And that's what most people are here to hear about. So we're go straight into that.

So some of these things Dr. Goldman already spoke of on behalf of our Agency. Why are we using WGS? To support foodborne illness and investigations, to support our mission goals as part

of FSIS and mainly and more importantly the whole point of this is alignment with public health partners. You see a lot of familiar names who have been up here speaking over the last day, 2 days considering the NARMS that was last week, I mean earlier this week, excuse me.

So where are we currently right now? We currently have 12 MiSeq sequencers in our laboratory system. Eight of those are in our Eastern Lab in Athens, Georgia, where I'm located. Two are in our Midwestern Lab, and two are in our Western Lab.

We, in Fiscal Year '17, really ramped up and sequenced around 7,200 isolates. So it's quite a jump from our capacity building perspective.

We do collaborate really well with our public health partners, and we do consider epidemiological information in all that we do with these new emerging tools as many of the speakers have alluded to this morning. It's one big package of information that you get.

Finally, we do work really strongly with our NARMS partners. Dr. McDermott gave a great talk

earlier on how this tool can enhance what we're going and we're trending with antimicrobial resistance genes.

So I want to point this out to you as a reference guide. Dr. Klimke gave a nice talk on NCBI. Well, here are some of the top bioprojects that we and FSIS contribute to. Most of these are GenomeTrakr-specific projects. A few of these are set up specifically for us with our efforts with the NARMS program at FDA CVM. So many of these are from the cecal environment of our four major animal commodities. So please take note of those and peruse those bioprojects as needed.

And this last just shows kind of an upwards trend of how many isolates we've gone from kind of a low level in fiscal year 2014 to where we're at now with over 7200 in fiscal year 2017. We plan on going even further this year.

We talked a lot about metadata here today. Here's a nice example of one of our biosamples. We do release who collected it, of course, the collection date and the state in which it's

collected, and isolation source. We try to be very descriptive, as Bill was mentioning in his last talk, of how we could collapse some of those commodities into a certain product type, non-meat product swab.

Okay. Now, onto more of the analytical tools that we may use. I was actually asked to do a demonstration, but I couldn't bring myself to do that in command line in front of you all and streaming over the web. So we're just going to stick with some screenshots.

We'll start with how we do quality control assessment, antibiotic resistance gene detection, *Salmonella* and STEC serotype determination and STEC virulence gene characterization, and finally phylogenetic comparisons. So it will be a lot of stuff coming at you really quickly.

Currently, our group is responsible for characterizing upwards of 12,000 isolates per year, and that includes antimicrobial susceptibility testing, serotyping, PFGE, and whole genome sequencing. So we're doing all these tests in parallel. So it's a pretty busy group and a lot of

results coming out of there.

The goal over the next couple of years is to continually develop one single workflow because as we know, we can get all these types of information directly from whole genome sequencing, and we want to continue to try and pursue that as much as possible.

So an overview, we start with the sequencer. I think this might be the first picture you've seen of the MiSeq today. Again, we have 12 of these. If you can see up here, you'll see kind of what a FASTQ format looks like if you're not familiar with them. It's a different type of format than most of us are used to working with over the past 5 years. It's highly compressed, but even at compression rates, it's still around 300 megabytes per 2 files together for a single isolate. So you're talking large datasets.

That data goes one of three ways actually. You input the FASTQ into an assembler. In this case, I show a picture of a CLC Genomics Workbench. We also use Spade to do assemblies, de novo assemblies. We do a quality control pipeline where we assess

coverage from the raw files, the average quality and nucleotide balance. All these are worked out in our Gen-FS partnerships with NCBI and FDA and CDC.

FASTQ files can go into the whole genome MLST, in the BioNumerics 7.6, which was talked about by Dr. Besser earlier. Lyve-SET and SNP pipeline are two types of high quality SNP analyses that are actually publicly available, and I'll have some screenshots of that later on. And then finally, the NCBI Pathogen Isolate Browser. That one's nice because all the heavy crunch is done on the NCBI side, and we don't have to do that.

Okay. So once we've got our assembly, we do another set of QC, quality control. We look at file size. That's important. One byte equals a base essentially. We want to have some level of correlation that our assembly is very correlated with the actual bacterial genome size, 3 megabytes for a *Listeria monocytogenes* assembly. That equates nicely with 3 million bases, essentially what the take home on that is. We look at certain other metrics, N50, number of contigs. The lower the number of contigs,

the better the assembly, the better the sequencing quality that went in. We want low number of contigs.

Finally, correct organism, that's very important to us. We have multiple things in house to assess that, and on the NIH side, NCBI, Bill Klimke has talked at numerous conferences about there's still some level of sample mismatches and mix ups. We try and avoid that all throughout our pipeline here and with our other assays by double checking at multiple points.

Now, here's the FASTA form as you can see. It's a little bit different than the FASTQ file. Again, you've gone from 300 megabytes to around 3 megabytes. So you've highly compressed really complex data into a much smaller workable format.

Then you're going to input that FASTA into our downstream tools, determination of MLST, multi-locus sequence type, antibiotic resistance genes, virulence profile and serotype determination and potentially you can even do different types of file genetic comparisons using MASH Tree.

So to start off, I'll talk a little bit

about the antimicrobial resistance we use, the whole genome sequencing data, that's been talked about a lot again over this week.

We want to work to identify new genes of concern. That's been a big part of last couple of years, and FSIS efforts have been part of that with the CTX-M-65 and other genes. As I mentioned, the CTX-M-65. We have colistin we've worked to identify. Do we have any of these sequences in any of our sequences that we're putting upon NCBI? Quinolone resistance and the spread of plasmid-mediated quinolone resistance, *qrnB19*. We work with our partners at NARMS with that. Linezolid resistance as well.

Okay. So the overall kind of workflow, you saw what the FASTQ file looks like. So it goes into the QC in the assembly pipeline and then from there, the assembly goes into our blast database, and the output is a resistance gene profile. So that's a really quick run over of what we're doing.

But where do we get the antibiotic resistance database from? We use ResFinder

currently. We get that from the Center for Genomic Epidemiology, and much like our partners, we update that on a very, very frequent basis.

Here's the output. Here's the black box in which we work. Most of this is done in command line, and there's a reason for that. All these tools are publicly available and you can do them in most browser type formats. You can do that for 1 to 2 isolates, but when you do it for 7,000, you've got to have a much, much better and more efficient way of doing it. That's where in our hands, command line comes in.

This is kind of a typical output you'll see and then in the end, we do a lot of formatting with our bioinformaticist. A lot of their work is in formatting how we can get it into a file that's usable for other people.

So in this particular isolate, we identified five different genes confirming multiple levels of resistance including *bla* CTX-M-65, tet genes and sulfonamide and aminoglycoside genes.

All this can be found again publicly, for

the download portion at bitbucket.org, which I actually think can be linked to from the genomic epidemiology site at this point.

Okay. So one of the things I wanted to show was that phenotype and genotype comparison has been talked about. So we did contribute to the NCBI and FDA initiative recently to identify what that correlation was, but we did our own work in house with the 2016 cecal NARMS data which is essentially 1190 isolates. Dr. Goldman showed this earlier, and you can see and hopefully appreciate around 97 to 99% correlation. Like everyone else has mentioned, there's some gentamicin issues but overall this is a pretty tight correlation.

Okay. *Salmonella* serotype determination, very similar to what you saw before, assemblies key, putting it through our custom made BLAST database based off of SeqSero developed by the University of Georgia and the CDC.

That one is actually available on GitHub. A lot of these different tools are available on GitHub. You can see really faint screenshots up here

now against this backdrop, but you will see they are available and you can use those.

So to give you a little bit more background on that, several years ago when we first started this initiative, SeqSero was just coming into its own. So what we did was an exact matching algorithm using some python-scripts that we developed. So now we currently still use those python-scripts in conjunction with SeqSero, remove all the redundant factors and use that against the dictionary to identify the serotype.

So in our hands, we looked at over 4,200 isolates, and we found about a 96% rate where WGS matched that of the reported serotype result.

We had about 3.8% where WGS did not match the serology result, and that typically was just incomplete genetic factors that we did not get allowing us to call the serotype. In those cases, those are sent to NVSL to do traditional serology.

Okay. This is getting a little familiar to you at this point. One thing I want to highlight is how important it is to do the assembly from the FASTQ

files, and also to utilize already existing databases to do your work with.

The virulence typing and MLST typing that we do for STEC we find is very important. We use that from the Center for Genomic Epidemiology as well. We also use a virulence finder and serotype finder which were both mentioned earlier in previous talks.

Another output for that, and then here's an O26 strain. One of the nice things I like to point out, too, is previously in FSIS, we didn't get the H type. It's not part of what we used to do. Well, now we can get the H type. A lot of them are H11, H7, whatever they may be.

Sequence type is very important to us as well. Actually I'll give you another example later on how it helps us kind of predict what the O type might be.

The stx type, that was talked about earlier. You can actually look down to the stx1a in this case and then the eae allele, beta in this case as well.

This is an analysis that we just recently did right before we came. This is around I think 3 or 400 isolates. We looked at our top six non-O157 and O157, and what did we get out of the characterizations. We looked, what are the top MLST types? So looking at sequence types, sequencing allows us to go back any time retrospectively and do this. It's another very important point about sequencing. You can always do retrospective analyses.

So you start to see that we do have within each serogroup, we do have a predominant sequence type within each one of those. One of the ones I like to bring attention to is ST11 on the *E. coli* O157:H7. In our hands, our commodities, our isolates, those tend to be very familiarly known as ST11.

Various Shiga toxin types, in combination or alone; eae types, we tend to see gamma as the top in all of our *E. coli* O157 isolates as well. And our top serotype as you might expect, O157:H7, not unexpected.

But then we do have within our other serogroups, we start to see 100% or H11, H8, and so forth.

One of the nice things whole genome sequencing allows us to do, that previously we couldn't do, was we can start to identify H types on serogroup O157 that we normally would have said was negative for O157. We previously have had some H11s and H29s. In both of those cases, neither one was eae positive nor stx positive.

Okay. And finally for phylogenetic comparisons, all the pipelines have been developed by our public partners. Many are here. Lee Katz's Lyve-SET, which is on GitHub. The pathogen pipeline that Bill talked about was a great tool. FDA SNP pipeline also is on GitHub. And then wgMLST which is a very nice tool, BioNumerics 7.6 in the CDC PulseNet, and there's a screenshot.

Here's some output from the pathogen isolate browser. You saw this slide earlier from Dr. Goldman and about where do some of our type serotypes in our commodities, where to do they line

up as far as within 10 SNPs of a clinical or 20 SNPs of a clinical or are they in a SNP cluster? And you can see these numbers and you can see some of the things like Kentucky, 0% within clinical. Again, this is all hinging on all these clinicals are being sequenced in real time. We're kind of ahead in that we're sequencing everything in real time, and as soon as we're caught up in the next year with every clinical going on NCBI, maybe we'll see different numbers at that point, but that's just something to consider.

This was also brought up by Dave Boxrud. This is some work we did with pattern 4 in Enteritidis, very predominant. Fifty to sixty percent of our Enteritidis isolates are pattern 4. Well, you can start to see a delineation just looking at SNPs alone between those 4 pattern clusters which on your screen have more of a reddish orange tinge to them. You can start to see that they kind of break apart. With just PFGE pattern alone they've been clustered together.

Okay. And the last example, I'm almost

done here, we did some work with harbors of *Listeria* strains and some processing environments. That's one example of one establishment where we had about 14 or 15 isolates collected over a 4-year period. With this, we used three different pipelines to show the concordance between those three pipelines and to look, what did we see here?

So we used a SNP pipeline, Lyve-SET and wgMLST, which is hidden behind the placard. We did see a very strong concordance between all three of those pipelines. They did cluster all the same isolates together. I defined in this one a cluster of 20 or less.

One of the interesting things that you can see is there are several clades. Those are starting to break apart, these from here, which it makes sense because these actually have nothing to do with any of these events and this establishment right here. So that's very, very important.

The nice part is, with PFGE, all these had the same primary and secondary PFGE patterns, and we were actually able to start breaking those apart,

teasing those apart using some of the newer tools that were developed for whole genome sequencing.

Okay. And finally, I have one more slide. How else are we trying to use genotypic data? Again, this is only looking at genomics, not using transcriptomics which is always very vital for this kind of work, but this is locus of heat resistance. We did a query using certain genes for locus of heat resistance of all our assemblies that we've had over the past year or 2 years, and we ultimately found around 11 *Salmonella* isolates of varying serotypes that did have this actual locus of heat resistance.

So this is a pathway that we're interested in going into. What else can we find from this data, heat resistance, acid resistance, those type of intrinsic components.

So, in conclusion, we are focused. We've definitely invested a lot as an agency in moving forward with this type of knowledge and working with you all as our partners.

We've built sufficient capacity. We feel we're well beyond the just capacity building point at

this time. We're moving forward. What else can we do and add?

And we want to continue engaging our national and international partners. We have a lot of presence in both national and international meetings, including GMI.

And finally, we want to use WGS as we always have with PFGE, use it in conjunction with all the epidemiological information that we have and bring in that totality of evidence to any type of investigation.

With that, I'd like to thank all the people in the room and thank you for all the people on the slide. A lot of the collaborations I know for us, just standing up the program since 2014, took a lot of collaboration with everyone here and some of the people that aren't here that were here earlier in the week for the NARMS meeting. So I'd like to thank everyone.

DR. BRADEN: Well, good afternoon, everybody. My name is Chris Braden, and I'm from the Centers for Disease Control. I'm going to be talking

a bit about how the different agencies that are implementing this new technology are actually coordinating, and I think this is an important component of what we do.

So an approach we've taken in order to establish this interagency coordination is really to build upon a history of collaborative programs that we've had, and some of you here may know of or been involved with the Interagency Food Safety Analytics Collaboration that was established before this particular collaboration and really builds on that type of structure.

Of course, we want to apply advances in technologies and the one that we've been concentrating on is next-generation sequencing, but we want to scan the horizon to see what opportunities continue to evolve in the technology fields.

Of course, even within next-generation sequencing, as Martin had said before, there's going to be the next next-generation sequencing and how that's going to change what we do.

We want to leverage the knowledge,

expertise and data among agencies. We bring certain levels of expertise in different areas together, and that makes us stronger and certainly bringing our data together makes what we do more effective.

And then set up a structure for our collaboration that is efficient, guided by strategy and prioritize communications and stakeholder input over time.

So the Interagency Collaboration on Genomics for Food and Feed Safety, or Gen-FS, was established in 2015 to strengthen the federal collaboration on the use of whole genome sequencing in foodborne pathogen analysis and investigation.

Multiple federal agencies are involved and most recently, we've had ARS and the Animal and Plant Health Inspection Service, APHIS, at USDA join this collaboration.

So really Gen-FS is meant to support the implementation of a shared vision of coordinated networks for genomic sequencing. We want to use flexible tools and analyses and communications needed by the respective agencies to harmonize procedures

and standards where we should, and really these two large networks are the ones that we've concentrated on, making sure that we are, you know, collaborating and harmonizing as much as we can.

So we have some targets for what we are prioritizing for our development, coordination and harmonization, including some of the things that we've already talked about, the system tools and pipelines and methods; the analytic procedures, protocols and standards; sharing data and data availability; harmonizing some of the ways that we manage the networks with proficiency testing and training; how we use this data in surveillance, investigation and research; and then the external communications and partnerships.

So we have a draft charter that is actually undergoing the process of having our Agency heads sign it as we speak. We have a steering committee with representation from each agency and then we've set up these four workgroups. One's data standards, analytics, comparisons and interpretation, and I'll show you some of the output from some of our work,

interagency training, the network workflow harmonization and then communications.

So one of the things that we've had as a strategy and priority from the beginning is really sharing our work, our structure and our strategy, and one of the things that we've done from the beginning is worked to make sure that the DNA sequence and metadata that we produce are publicly available. So the data, as you've heard, is uploaded to NCBI and GenBank. That includes all organisms undergoing whole genome sequencing in PulseNet and GenomeTrakr. There are clinical, food and environmental isolates, and we must do so with the protection of personal and commercial information.

We also make he tools that we use publicly available either as open source or commercially available software and then publish our methods and validation analyses.

So the standards and validation, what that workgroup has really accomplished is establish the quality standards that are monitored for all submissions of all genome sequencing to GenBank and

to then develop and publish benchmark datasets which I'll show you a little bit more about in the next slide.

We'll do the validation studies. Some have already been mentioned having to do with SNPs. There's more to be done, I think including when we validate the whole genome MLST and really do some of the more careful cross comparisons of MLST and SNP analyses in individual pipelines and then that cross validation.

And then AMR genotype/phenotype comparison, you've hard quite a lot about already, but there is another publication out of this workgroup that's pending.

So these benchmark datasets I think as we were just talking before, is I think a great resource for anybody. Certainly it's what we need internally to be able to validate a number of our analyses and pipelines but being able to then provide them publicly allows those datasets to be used in validation and research studies for anybody.

So we have five DNA sequence datasets

consisting of 10 to 31 well characterized outbreak and unrelated isolates, that have been developed so far for *Listeria*, *Campylobacter*, *E. coli,* and *Salmonella*. And with each of those, there is publications that detail the outbreaks that they're associated with and these datasets are available for download at this GitHub site. So we are making those publicly available.

Harmonization across the networks, of course, this is important because a lot of the laboratories are both GenomeTrakr and PulseNet laboratories and they can't be doing two procedures for the same purpose. We really try to harmonize our procedures for the laboratories that participate.

So for the training, the training is provided for public health and regulatory program partners and PulseNet and GenomeTrakr networks. They're both CDC and FDA sponsored courses. Staff from each agency participates in the training as training faculty and training certification applies to both networks. So I think we've done a lot to integrate our training.

And the same is true for standard operating procedures in laboratories for whole genome sequencing procedures, for sample and library preparation, for the sequencing procedure itself, for the data management and upload to NCBI, and then incorporating new and changing technologies because for those of you who do this, you know that new chemistries come out periodically and they need their own SOPs to change accordingly.

And then the proficiency testing, so the same proficiency panels, the same analysis, the same reporting is used for both networks.

So communications: We've really tried to be able to communicate what is happening in the networks, how the data is being used in surveillance, investigation and regulatory functions. We have some industry collaborative forums. For instance, the Institute for Food Safety and Health has had a couple of meetings, and we've also had some forums in collaboration with the University of Georgia Center for Food Safety.

There's been many presentations and

discussions at food safety and scientific conferences, and you see some listed there, and then we are putting together a white paper for publication on the use of whole genome sequencing in food safety.

So looking to the future, and this is what I'm really concentrating on, how we're going to be using this information on a day-to-day basis in our agencies. Whole genome sequence technology will replace traditional methods for routine microbiologic characterization of foodborne pathogens for use in surveillance, investigation, and agency action.

Now, people ask, you know, it's still important to do some traditional microbiology because that gives you still additional information and, yes, there are times when that's going to be appropriate, but for our routine purposes, we think that whole genome sequencing will give us the information that we need.

Of course, to use these tools in this way, they have to be validated with comparable results no matter where the testing is done.

We want to have shared tools, standards and

data for all stakeholders, so that public health, regulatory partners can use these standards tools, but in addition, those in industry, academia and our international partners can also particular in using some of these same tools in order to be able to compare data among partners.

And then the other thing that we're really trying to do is come up with some tools and methods that will be simplified for WGS analysis because, you know, to do the research and to develop these does take a lot of computing power, and it takes a lot of bioinformatics expertise. Not every health department or other institution is going to have that kind of computing power and expertise, and so we do need to develop tools that are more simplified in order to have many more participants in these networks to be able to participate. So in those cases, there would not be a requirement for high performance computing and advance bioinformatics expertise in order to use these tools.

And I would like to thank the Genomics and Food Safety members, and especially the

communications workgroup that's worked on this and other presentations, USDA-FSIS for hosting this meeting that is meant as a meeting for all the Gen-FS members, and your interest and input. Thank you very much.

DR. EVANS: Thank you, Dr. Braden.

So now we'll have a question and answer session. So if all of the speakers from this Federal/State Collaboration session could come up to the stage, and they can all have the microphone. And if I could ask the panelists to state your name when you answer the question for our folks online.

Okay. We'll start with a question in the room.

DR. BOOREN: Great. Thank you. Betsy Booren with OFW Law. First of all, thank you all for your time today. I found the presentations fascinating and as I've been sitting here, and we've been discussing some at break, sort of the totality of what's been presented, and I have a question, and I'm not sure who's the best person to ask it, and it may be coming from a place of ignorance.

But as I'm looking and hearing what is being talked about, the number of isolates that have been sequenced, one of the questions that I have in my head is what, and perhaps it should be focused on GenomeTrakr, of the pie of isolates that have been sequenced, how much of that is clinical samples? How much of that is other isolate sources? And how much of that is regulatory? Because as someone who represents the food industry, I'm focused on the regulatory and trying to better understand, if I see a slide that says, papayas or meat and poultry have a high sense, what does that mean in the whole scheme of things.

And so what I'm trying to get at is for industry to look at trying to do research or other groups, in that regulated area, how much of those are regulatory surveillance? How much of those are for cause?

And as I talk with my industry partners, better understanding some of that information may inform research decisions that can help when they say why is an industry adding to the database? We want

it to be targeted and want to understand what's going on in those isolates.

I'm not sure whose got the right answer or if that's -- again, maybe there's a report out there that I haven't seen, but I think it's really important as when we've done other risk assessment type of work, certain agencies, certain industries have an abundance of data compared to others, and does that mean there's a true risk there? What does that data tell us? So I'd be curious in your thoughts as we try to, particularly from an industry standpoint, better understand what that data means.

DR. BRADEN: Thank you for your question. My name is Chris Braden from CDC. I think you have to be careful in the database to know that this is not a statistically representative sampling scheme. It does contain, you know, especially on the food and environmental side, those samples were collected for lots of different reasons and actually I'm not sure that we have the data to break out what sampling for cause and what sampling according to some assignment or routine sampling that might be done.

So that I think is one thing that you need to understand when you're analyzing that as a limitation to the data.

There are more food and environmental isolates than there are case isolates. I think there always will be because food and environmental isolates just depend on how much you test and there's only so many cases. So that's going to be the case going forward.

But nonetheless, even with those limitations, I think that there's lots that one can learn about the breadth of genomic variation in any number of ways in these databases that is helpful to answer some questions but won't be able to say, to break it down by, for instance, sampling for cause or for surveillance.

DR. TILLMAN: This is Glenn Tillman from FSIS as well. To follow up, one of Dr. Goldman's slides had that we've sequenced around 11,000 uploads into NCBI since inception. Around half of those are part of our non-regulatory cecal program, and those have been kind of moved to a different bio project

which could provide some opportunity to look at them, those strictly within that bio project, and one of my slides had all of the ones that we contribute to, and they're actually called a NARMS cecal.

The other 5,500 were collected as part of a regulatory initiative. So we do have kind of a half mix there, and again the NARMS cecal are moved into their own bio project. So that would be a good place to start on understanding what each one of those bio projects that NCBI might have within it.

DR. BESSER: John Besser from CDC. It hasn't come up today, but there's actually something called VoluntaryNet at University of Georgia which could be used by industry to anonymize the sequence data, and actually as a firewall between public health and VoluntaryNet. So it could be used to compare the food and environmental isolates submitted by industry to clinical cases to assess risk.

But we can't, the CDC, FDA, USDA, can't specifically query that database. We can ask them if there's matches. They can then ask their membership as to whether or not whoever submitted it would be

interested in sharing that data, but the idea is that there's concern in industry that the information would be used in a punitive manner and volunteering that is a safe place to submit data that can be used for risk assessments. It's not exactly answering your question, but it's a related topic.

DR. EVANS: Jorgen.

DR. SCHLUNDT: My name is Jorgen Schlundt, and I'm from Singapore. I love saying that I'm from Singapore. Okay. Especially with a Danish accent.

But it's fantastic to be in the U.S. discussing something like next-generation sequencing because this is clearly an area where U.S. is leading. I understand that in some other areas, U.S. is stepping back from leading, but actually in this area, you are really leading. We don't need to discuss the integration between the different agencies in U.S. Maybe it's a good thing that you have many different agencies because then you can move in different speed, and we have seen that also today.

But I have one question in relation to

maybe unifying the forces not only within the U.S., but also with the rest of the world. So I was looking at resistance, and I understand that CVM is moving forward with a Resistome Tracker, and I understand that NCBI is move forward with something else, with very long name that I didn't really get. And I understand that FSIS is using the WHO collaborating center thing from Denmark, the ResFinder.

Wouldn't it make sense if there was sort of a concerted effort to try to produce consistent, uniform methodology and also tools so that the rest of the world could really get a benefit out of U.S. leading in this area?

DR. McDERMOTT: Thanks, Jorgen, for the question. And it's a topic we discuss quite a lot. The need for harmonization in allele cause is essential. If we're going to take seriously antibiotic resistance as a global challenge, it must be addressed globally. We need a harmonized method just to have a common language.

So I think that the discrepancies are being

worked through between say ResFinder, NCBI, and Bill can tell us the latest on that, but I think just in general, it's essential that it be done so the outputs are the same. There might be tools with different interfaces that work with BioNumerics, for example, or different types of alleles, but as long as the output is either the same or can readily be translated into a common language, then that will be an important objective, but I do think that it's essential.

DR. KLIMKE: So, Jorgen, we were just at the ASM Genomics conference, and we had a roundtable with NCBI, the CARD database in Canada, and Ole(ph.) from DTU was there. Although he's technically not ResFinder, he's at least Danish, which is close enough. And we agreed that we would at least discuss harmonization or curation efforts because the content in the database should be the same. The methods to apply them I think will have to follow after that because we'll have to do comparisons to see if we get similar answers and the extent of application of related sequences is something we need to look at

but, yes, I would agree with what you're saying.

DR. BESSER: I appreciate your Danish honesty. That's great. You know, I agree with everything that's been said that we're moving towards that. I think what Chris Braden presented on the data quality is the really key harmonization point. If we share data quality, we can explore these different mechanisms for allele database curation for different systems.

And as with any new technology, there has to be sort of a creative period, and I think we're in that now, and so no one system has emerged yet as the winner. But as long as we share this core of data quality, I feel that we're all moving towards global standardization and curation. We just haven't gotten there yet, but I think we all agree that that's the direction we're moving.

DR. BRADEN: This is Chris Braden again, and you raised the issue having to do with resistance determinant databases, but I think the same can be said for other allele databases that we all should be coordinating and the types of databases we use, even

if the management is distributed in some way.

DR. BESSER: Thanks, Chris. At least CDC, we're envisioning the global curation of these allele databases for subtyping purposes. That's a really big issue, but the Pasteur Institute in France has done curation for *Salmonella* serotyping forever. We're envisioning a global collaborative curation of some sort, but I think it has to happen on a global level ultimately.

DR. EVANS: We have a question from the web.

DR. ALVARADO: The question from the online group is do you worry about the education training that may be required so that people can adequately interpret WGS data findings?

DR. BRADEN: So this is Chris Braden. Is the question -- is this a priority of some -- could you repeat please?

DR. ALVARADO: Sure. The question is do you worry about the education training that may be required so that people can adequately interpret the WGS data findings?

DR. BRADEN: So, yes, we do worry about that. It is difficult. Not only is it difficult to learn some of the bioinformatics that may be necessary to do this type of analysis in order to, you know, appropriately interpret it in a laboratory, but there's a whole other realm of disciplines out there that also needs some training on how to interpret this data and so probably the next outreach is with the epidemiologists really to do that genomics for epidemiologist type of training and there are a number of courses and I know Martin Wiedmann was here before and he's led some of those courses. And so that certainly is the case, but I think as we use this more, it's going to be, you know, there's going to have to be more training, you know, with more disciplines for instance in the regulatory agencies or with risk assessors and so forth.

MR. BOXRUD: So Dave Boxrud from the Minnesota Department of Health. So from a public health lab standpoint, we have additional challenges. We're a smaller group, and traditionally we've had

experts in microbiology identification, that sort of thing. Many of our staff simply don't have the background in sequencing or have the understanding of how to interpret sequencing and how it works.

We've been able to hire a number of new people that have that background but trying to bridge that knowledge gap between traditional subject matter experts and then this new generation sequencing group, there's a real challenge, and we have to work on it constantly to try to bridge that gap, and I agree with Chris that educating our epidemiologists is also really a vital thing so that everyone is always on the same page.

DR. TILLMAN: I just want to add one quick thing. This is Glenn Tillman. So a lot of the tools have really come a long way since the 2015 ASM Genomic Pipeline meeting in Washington. A lot of the online tools have come about and really moved things forward, BioNumerics and with the applied math has really moved forward and really made things a lot more streamline, allowing users to really be able to use that stuff. So I think training is a very big

part. I think the development of these newer analytical tools in more of a web-based form and/or just any type of GUI-based form have really been beneficial.

DR. ALLARD: Marc Allard, FDA. Once again I'm going to do a comment and then another question.

I just want to say that many people in the room are actively involved in international and global training, as well writing white paper documents on why you should adopt whole genome sequencing and the specific case studies that they're best used.

And so this is collaborations where the World Health Organization, Food and Agricultural Organization, and the OIE, and so there's ongoing efforts for global training because of global trade in food.

So my question is, it's a little off the topic while we're here at USDA, but I know that the Department of Health for New York, they don't just sequence the four or five foodborne pathogens. They're actually responsible and have been sequencing

efforts in almost 20 human pathogens. So this question I think starts with Dave Boxrud at Minnesota is, do you see a similar expansion of sequencing for other species? And what do you see as growth? And I guess then the question is, you know, are you seeing equivalent support from your federal partners, both in databasing, etc.?

MR. BOXRUD: Yeah, it's a great question. But PulseNet and Foodborne is definitely in many ways the leader in sequencing. They're kind of the bull because there's so much of it, but we are involved with a lot of different types of pathogen testing, *Legionella*, *Strep pneumo*, MRSA. I believe it's about 10 different projects that we're involved with, and many of these, we've reached out with our collaborators at CDC and said we're interested in this organism. We have background in this. Let's work together and they've been very, very willing to do that. And that's been really key for us.

One of the real challenges with using sequencing with foodborne disease surveillance is while it's an awesome technique, it's an awesome

method, if you batch it and do it once a month, you're really slowing down the process.

And so we fill in our runs with other organisms, with other pathogens. The *Legionella* group from CDC has a really nice partnership, and they have created -- I think there's about six laboratories that are doing a study with them. The TB group is doing, from CDC, I believe they funded one site to do essentially all or most of the clinical TB cases throughout the country. There's a flu group, we're not involved with this, but there's three sites that are doing a lot of influenza testing.

So there is a lot going on but some of the public health labs are not always embracing the new technology for different organisms. So I think hopefully as sequencing becomes more routine at all public health laboratories, that they will consider doing other pathogens because I do think, you know, this is a technology change that is not going to go away and it's going to continue to provide additional important information for different pathogens.

DR. EVANS: Question from the room.

DR. WIEDMANN: Martin Wiedmann, Cornell University. So we talked a lot about standardization of database as alleles cause, quality control. Where are we with regard to standardization of metadata? And I'm going to start out with just even food data. I think porcine, pork, different names, to call the same thing. If we move to environmental, you know, data, how do you describe environmental sample? It gets even more challenging.

And I think there's an importance to it that hasn't been mentioned. I think when we try to interpret SNP differences, it obviously depends on the environment the organism is in. If *Salmonella* is in a dry environment, it might only replicate 10 times a years, and therefore accumulates SNPs at a much, much slower rate than if the same *Salmonella* sits in a poultry house where it might replicate every hour because it's actively infecting one chicken after the other. And so if you don't have that information in some sort of standardized way, we sometimes will run into challenges how we interpret

SNP differences.

So is there parallel effort as part of Gen-FS to also standardize those data or is that sort of the next step?

DR. KLIMKE: I can only tell you about the things I know, and that is that the global microbial identifier and part of a ISO working group on next-generation sequencing for food safety, that there is an effort to use some metadata ontologies to standardize. This is called Genna GO (ph.). It's been worked on by the Canadians. CFSAN has said they will look at using that in their metadata, but I should tell you that what we would rely on then is probably the submitters to apply those terms in a standard way because a lot of people will say that ontologies are the solution to everything. They're not, if they're not applied in a standardized way by all the people contributing.

Since we are the people who are integrating all the data, if everyone in the U.S. goes a certain way and then the Europeans do it differently, we're sort of left at the mercy of that, where we would

have to either invest time and effort on our side to standardize the metadata once it gets indicated and submitted to us, or we just leave it alone and only worry about the standardized metadata in the U.S.

So I know someone from CFSAN can probably mention that they're looking at this. You could probably talk to them. I don't know -- if USDA's doing that, they probably will. You guys can probably talk about that.

DR. EVANS: We got a question from the web.

DR. BRADEN: Chris Braden again. You know, we started by trying to standardize what the metadata fields would be in the Gen-FS agencies, and it's a place to start, but I think definitely needs to expand to be able to have all partners be able to participate with some standard expectations.

DR. EVANS: Okay. We have a question from the web.

DR. ALVARADO: Can you say more about providing safe harbor for producers, manufacturers, others in the industry, to (a) gain familiarity with WGS and (b) provide useful information to regulators

without bringing regulatory response down on their heads?

DR. BRADEN: This is Chris Braden. I'm not sure we have the representation here, but I don't think I really understood the question. If you could ask it again.

DR. ALVARADO: Sure, I can repeat the question. Can you say more about providing safe harbor for producers, manufacturers and others in the industry (a) to gain familiarity with WGS and to provide useful information to regulators without bringing regulatory response down on their heads?

DR. BRADEN: So, yes, safe harbor. I think that's where is that safe place to be able to start to implement this kind of technology in order to learn more about your particular producing environments if you're in the industry, for instance. I think it is an important point. We have had a number of discussions with industry members about what might work if there's a third party that could be responsible for holding the key, for instance, and not releasing any of the identifiable information and

datasets.

I would hope that we can find such partnerships to be able to do so. John Besser had mentioned one at the University of Georgia called VoluntaryNet which is an industry collaboration. I know that at the Institute for Food Safety and Health, they are also considering being able to provide that kind of third party resource. There may be other resources out there, and I think, for instance, you know, IEH has said that, you know, they can certainly provide the service and provide that kind of data back that would be helpful to industry members to understand how to use this information.

So I think there are some resources out there. It hasn't been widespread to my knowledge. Maybe IEH is maybe used the most but, you know, I think it's certainly worth exploring for industry members.

DR. ALLARD: Marc Allard, FDA. I can comment just a little bit about this. We've been doing outreach to industry. FDA primarily works with the Institute for Food Safety and Health, IFSH, which

is out at Illinois Institute of Technology in the Moffett Center in Chicago. And essentially from 3 or 4 years speaking to food industry, probably the most feasible path to learn new things about this technology is to work with a third party provider, essentially do this on line.

Steve Musser touched on this earlier today. There's many, many companies that are involved in this, as well as many academic folks like Andy Benson at University of Nebraska and Martin Wiedmann at Cornell, but there's companies like Eurofins, IEH, Ecolabs, NSF International. You can find these on the internet that provide genomic services and can assist you. This is an expanding market because there are a lot of people in industry that would prefer to do it as a third party provider as opposed to building their own laboratory.

So we can give you more information or reach out to your local lobbying group of GMA, United Fresh, SQF, IAFP. There's many speakers at these meetings that have a diversity of services to the food industry.

DR. EVANS: There's a question from the web.

DR. ALVARADO: There was no mention in today's talks by the different agencies on how WGS is being used in the detection of outbreaks and for carrying out outbreak investigations. How is the CDC using the information and how is the integration with what the FDA, FSIS, might have in their database taking place?

DR. BRADEN: So as I mentioned in my presentation, basically we will be transitioning on all of our traditional characterization techniques over to whole genome sequencing. So whole genome sequencing will be replacing pulsed-field gel electrophoresis in our cluster detection, and it will be replacing the assays that we use for isolate confirmation, serotyping, virulence typing, and resistance typing.

So all those separate assays will now be carried out with one assay using whole genome sequencing in the future. We're in the midst of that transition now. As David Boxrud had said, that it's

a difficult time because we're actually doing both traditional typing and whole genome sequencing and that makes it more expensive and more time consuming but we anticipate that in the long run, this will be time and cost saving.

So that's how we're looking at moving forward, and I think that I can talk for the other agencies to say that that's their plan, too.

As far as, you know, how we use this data together, well, we do talk a little bit about GenomeTrakr database and PulseNet database and so forth, but in reality, we're all submitting all of the data into a single database and we're just contributors. So PulseNet is contributing to a single database. GenomeTrakr is contributing to a single database. And we're all using all of that data in our analyses for outbreak detection, investigation, surveillance, and action. So that's how it comes together.

DR. EVANS: So I have a question. John mentioned that by the end of fiscal '18, I believe, that there would be all or close to all *Salmonella*

would be sequenced, and I'm just curious about any logistical issues that would be raised by that, and what do you expect at the end of fiscal '18 when we have all these new *Salmonella* sequencing real time?

DR. BESSER: I'm John Besser. That's a very good question, Peter. Yes, it's not going to be smooth sailing. We're really getting into a whole big data era. I think the experience of Public Health England would probably be helpful. They saw a dramatic increase in the number of clusters that needed to be investigated. Now, we saw this with our *Listeria* combined initiative as well.

So I think there's going to be a need to develop new tools for cluster triage, what clusters are most likely to be productive but I think there's going to be more -- there already are more clusters to investigate than there are investigators to investigate them, sorry for the convoluted sentence there, but I think this is going to become a major issue and the investigative resources are going to be a problem.

I think we're working to streamline on the

laboratory side to make these things easier. For instance, you heard about nomenclature that's being developed which will allow for ready recognition, easier recognition of clusters and easier communication, but I think a lot more work needs to be done in order to prepare for this big data era that's rapidly approaching.

DR. EVANS: There's a question in the room.

MS. McGARRY: Sherri McGarry, CDC, and the question is mostly for David of the Minnesota Department of Health, but I'd welcome other panelists to weigh in. And it piggybacks somewhat what you were just talking about. So when PFGE first came on board, it took longer than it takes now, right? So there are efficiencies that were gained and techniques that were modified to make it faster. Where do you see some innovation at the state level to kind of speed things up a little bit? I know we're still at the early phase, but maybe this is also an innovative phase, too. So where do you see efficiencies to reduce the total amount of time?

MR. BOXRUD: Thanks. David Boxrud from

Minnesota. Yeah, there's a lot of ways to make the process more efficient, and that's something as we continue to do more sequencing, we're going to have to continue to look at these ways. We're going to have to adapt to new technologies. You know, if the work that we're doing with Illumina MiSeq expired 5 years from now, we may have a completely different technology than what we have.

For us, as we're really ramping up the amount of sequencing that we're doing, it's a lot of training, a lot of getting everyone on the same page within our laboratory, and as we're doing that, we're okay with a little bit of extra time to make sure that our quality is correct, and that we don't have any issues.

Once we really get this incorporated into a standardized workflow, then we're going to see what we can do to try to make it more efficient. We're in the process of getting a library prep instrument that will take some of the time and the labor of creating the library prep which is one of the more labor intensive parts of the process.

And also the kits that we're using, the technology is changing. The MinION could be a faster tool than what we have right now, and right now for PulseNet and various other things, it's probably not ready for prime time, but very soon it will be.

So I think all of us are going to have to continue to adapt, but what is great about sequencing is if you have that quality of sequence, if you use a new technology, they're very comparable going forward.

DR. BESSER: That's a good answer, Dave. This is John Besser, and I think Dave already mentioned the batching issue which is a big potential problem in turnaround time, but as he mentioned, there's also new technological developments, a new library prep and DNA prep, chemistry that just came out recently, and we'll be looking at all of those.

But I think what I didn't have time to talk about earlier today was some of the new tests. You asked about innovation. One in particular, amplicon sequencing, we're looking at a quasi whole genome MLST that can be done directly from a specimen, and

that actually could shave weeks off of the whole process if successful, but it's still highly experimental at this phase, but I think ultimately where we want to go towards is direct specimen testing which will have the biggest overall impact on turnaround time because the culture is the slowest part. Often culture happens in a clinical size at the state and sometimes even at the CDC and that can really slow down the whole process and we're hoping to bypass that at some point in the near future.

DR. EVANS: We have a question in the room.

DR. GOLDMAN: Yeah, David Goldman, FSIS. I just want to go back to the question that was asked online about outbreak investigation, and when the question was asked, I realized we really hadn't done a recent case study of an outbreak where whole genome sequencing was used.

And I think, you know, everyone knows that we rode PulseNet with PFGE very successfully, highly successful for 20 years and, you know, if two patterns were indistinguishable, we said, okay. Everyone agreed with that. We even dealt with one

band differences very well I think.

But now with whole genome sequencing, the picture is less clear as you've heard many speakers attest earlier, and I want to go back and emphasize something Martin Wiedmann said earlier which is that to me, where we are now is that the epi is more important than it was using pulsed field analysis as a way of judging and ultimately including cases in the case definition.

I'll just briefly reference a very recent outbreak in which we used both PFGE and whole genome sequencing and in this particular instance, the PFGE was done first by most of the state labs who were members of this outbreak, affected states, and what we found out was that after the PFGE seemed to include cases, a week or so later, we get this genomic sequence information which would exclude those cases.

So there was this sort of sequential activity which proved quite challenging for us as a federal family with the state partners in trying to determine whether this outbreak was over, have we

fallen below the epidemic threshold or are we still above it? And so this was a very recent example, and I think while we're in this transition phase still, using both tools, PFGE and whole genome sequencing, we may find similar challenges going forward.

MR. HEINZELMANN: All right. Joe Heinzelmann from Neogen. One of the things I had hoped would be addressed in some of the talks today, and maybe it will be tomorrow, is around the question of the statute of limitations around isolates in a database. Specifically let's say you find an isolate in a facility, you take corrective actions. Does that absolve you from that type of isolate in the database? Are there things that people can do or use this database?

So I guess what I'm really trying to get to is what can people do with whole genome sequencing data once an isolate is in the database from a facility to say that they've made corrective actions or they've used this technology and how long is that data point still real and applicable throughout the life of the food?

DR. BRADEN: Yeah. So I'm not sure if we have somebody in the regulatory community that can address that.

DR. ALLARD: Asking sort of policy related questions is always the third rail of the FDA, but I'll just make a comment. And so the emphasis is that a genetic match is not a regulatory decision. A genetic match is an indication of a shared common ancestry, shared isolates and it's a presumed link, a potential link to a food and clinical, whatever gets matched, but inspection in epidemiology, you have to follow and, in fact, FDA will not do any regulatory action on genetic match alone. There must be an investigation.

So we've seen 5-year-old matches, 10-year-old matches. Essentially this leads us back to a facility, to a region, a country, a state, but it depends what an investigator finds and it depends how the company responds. There's a whole process that hasn't changed. We just have a new genetic tool that helps establish linkage.

And so if investigators go and inspect a

facility and it's clean as a whistle, then there's nothing to be done. There's no regulatory activity. If the inspectors come and they found positives for foodborne pathogens, then it's like what Dr. Musser said. It depends on whether there's an association with clinical or not, and so it depends.

But the databases last forever, or at least 30 years. I know I have data in the database that's been there for 30 years. So our full expectation is NCBI will not go away anytime soon, and that we'll continue to see and use these linkages.

DR. BESSER: This is John Besser. Marc, thank you for that response. I think that was an excellent response.

I just wanted to suggest that we might, because this technology is new, we're suddenly able with more specificity to connect current cases with past isolates in the database, that this may be less of a problem in the future because we'll be able to connect cases more or less in real time with isolates as they're being found in the environment.

I think this is a circumstance where we've

got the new meeting the old, and I think it may not be as much of a problem as people are concerned about.

And as Marc pointed out, this activity is not new. We've always compared to the historical database. What's new is the specificity with which we can make that connection.

DR. EVANS: Okay. So this is the last chance for questions on the web, in the room.

And seeing no questions, we're going to start up again tomorrow morning at 8:00 a.m. The first set of presentations will be on what's happening with our international organizations, Mexico, Canada, and also globally, and then we'll have some presentations by our stakeholders, the food industry and other stakeholders.

So I'm really looking forward to the presentations tomorrow, and I hope to see everybody here and again, we'll be starting at 8:00 a.m. You'll enter through the fifth wing, and if there are no other additions, then we'll break until tomorrow morning.

Thank you.

(Whereupon, at 4:17 p.m., the meeting in the above-entitled matter was continued, to resume the next day, Friday, October 27, 2017, at 8:00 a.m.)

C E R T I F I C A T E

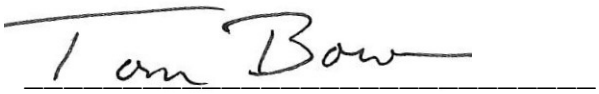
This is to certify that the attached proceedings in the matter of:

USE OF WHOLE GENOME SEQUENCE (WGS) ANALYSIS TO IMPROVE FOOD SAFETY AND PUBLIC HEALTH

Washington, D.C.

October 26, 2017

were held as herein appears, and that this is the original transcription thereof for the files of the U.S. Department of Agriculture.



TOM BOWMAN

Official Reporter