ILSI 2021
ANNUAL SYMPOSIUM
April 13-15 & 20-22
ACCESS TO A HEALTHY, SAFE AND SUSTAINABLE DIET
www.ilsi.org/annualmeeting

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

Transcript of the presentation, Foodborne Disease Outbreak Detection and Surveillance Using Whole Genome Sequencing, Kelley Hise, MPH, Centers for Disease Control and Prevention (CDC), United States

So, I'm just going to go through detecting foodborne disease outbreaks using whole genome sequencing, and then some of the tools that we utilize. I'm going to start with an introduction to PulseNet. I realize not everyone may be familiar with PulseNet. If you are, this will be a refresher. Okay. So, what is PulseNet? We're a national network of... I'm going to specifically talk about PulseNet USA although I will give a brief introduction into what PulseNet International is. We're a national network of 82 state and local public health/food and regulatory agency laboratories, and we're all coordinated by CDC and APHL.

We detect foodborne disease case clusters that may become larger outbreaks, provide real-time molecular surveillance of the most important bacterial foodborne diseases. These include STEC, Salmonella, Shigella, Listeria, Campylobacter, and Vibrio. We collaborate with our foodborne epidemiologists in investigating these outbreaks. And some of the ways we do this are by separating out outbreak associated cases from sporadic cases, providing case definition, and assist in rapidly identifying sources of outbreaks. And overall, as a network, we act as a rapid and effective means of communication between the public health laboratories. So, this is a map of the US. We have the different colors. Each color represents a region, and you'll notice a star in each of the regions and these are our area laboratories.

So, our area laboratories help provide troubleshooting assistance, search capacity testing, as well as some other things for the labs in their region. And this just lets you know that every state is involved in PulseNet along with Puerto Rico. And so, we have also several local city and county labs that are involved as well. So, this is PulseNet International divided into seven regions and 89 countries. As far as where PulseNet International is, USA, Canada, and Europe have fully transitioned to using WGS for PulseNet surveillance activities. But other regions, the situations are more bearable. Some are just starting PulseNet activities for the first time. Others are exclusively using PFGE, which was an older subtyping method that we used to use, and a few were preparing to make the switch from PFGE to WGS.

Some key priorities in the PulseNet International network are offering technical assistance, resources, and training for partners, facilitating communication, data sharing and regular meetings, and developing short and long-term visions for global PulseNet surveillance activities. So, there's basically three basic elements of PulseNet. You have your data acquisition. So that's a picture of a MiSeq instrument that produces the data that goes into our system. We use BIONUMERICS as a software analysis system, and all this information is shared with epidemiologists in SEDRIC. SEDRIC is the System for Enteric Disease Response, Investigation, and Coordination. It pulls in data from the PulseNet national databases into the system on an hourly basis. What this allows for is direct communication with our epidemiologists without them having to have the BIONUMERIC system.

Another way of exchanging data is through the PulseNet SharePoint site. And that's just where we post comments about ongoing investigations as well as troubleshooting and SOPs. So, this just shows you a little bit of how the network works. So, in each individual PulseNet laboratory, they do their actual sequencing there, and they analyze the data, and then they upload that information to our national databases at CDC in Atlanta, Georgia. And what they're able to do there is they're also able to pull back information into their local databases so they can do local cluster detection. They're also able to connect to the national databases and see what's going on at the national level. And then we are able to do national cluster detection and looking at the data on a daily basis.

So, this may look a little complicated. It's really not. It just breaks down what all is involved with our data analysis workflow, just so you get an idea of how complicated it can be. Again, you have your MiSeq, you have your raw sequence data the public health lab has produced. All the information is stored somewhere. We don't require the states to store them anywhere except for NCBI, but they will store it on base space, or their Cloud storage, or maybe even in external hard drive at their local location. That sequence data goes into a reference ID database. A reference ID database is where you just put all your sequence data. It doesn't matter if it's Salmonella, E. coli, Listeria, all goes in this one database. They submit their data to our CDC calculation engine, and what they get back is species identification and an initial quality check for contamination.

They verify that and then what it does is it pieces it out. So, all your Listerias go here, your STEC, your Salmonellas, and it knows which organism specific database to go to. So, my Salmonellas are going to go to my organism specific database. And then once they're in there, they submit that to the calculation engine again, and what they get back are their allele calls and genotyping results. And genotyping results will be serotype, AST data, virulence and other information. They're going to verify their quality one more time, and then we ask them to upload this information to the national within 24 hours of results. And then they can perform surveillance in their local database. You'll also see that we have them submit their raw sequence data to NCBI for either storage, or just so that those outside the PulseNet network have access to the data.

Okay. So now, that was a brief intro of PulseNet. We're going to get into whole genome sequencing and allele codes. So, I decided to throw this slide in because I realized I was talking about all this stuff with alleles and didn't really explain what that was. So, when you push your raw sequence data information to the calculation engine, what you get back are allele calls at different loci. And you'll see that this is actually an example of a Listeria cluster in our national database. So, the allele calls at each locus are basically compared within each other, and you get thousands of these. So obviously, you're not doing this by hand. You need an analysis system to do this, and that's what we use BIONUMERICS for. If you pulled all of these into one window, it pulls up a phylogenetic tree, and it will tell you where those differences lie.

So those differences go into the calculation or creation of allele codes. These are also used when we're doing cluster detection, and we look at these allele differences to determine if what we see is potentially a cluster worth investigating. So, allele codes. So, it's just a code, and I'm going to break that code down for you in another slide. But it's assigned to every sequence that passes quality in the PulseNet national databases, and these can be downloaded back into the local databases. What this allows for is then you now have a number or a code that you can search on in the whole database without having to do all these calculations over and over and over again. Allele codes provide a hierarchical name to show relatedness between isolates, and these are based off of the core genome or cgMLST. They offer a compact view of the entire population structure for an organism database such as Salmonella, E. coli.

Therefore, you need to think about these allele codes in terms of population and not outbreaks. And remember that allele codes are not the same as outbreak codes. So super busy slide. And I thought about just breaking this into two. What we can do is start from the top and work down. So, the top is a sample breakdown of an allele code. The first part will tell you the organism. So, this is LMOs. So that tells me it's Listeria monocytogenes. The version is 1.1. So, we've already had a change. So, it went from 1.0 to 1.1, and then each number as you go further into the allele code, you'll notice the number of alleles goes down. And this just means that if you're looking at your example here of sequence A and sequence B, they match up to the third number. So that just means that sequences A and B are between approximately... There's something popping up on... Nope.

Approximately 36 and 19 alleles of each other. And you can see that up at the top in the code where the third digit... If they're at a third digit and they don't have the fourth digit, there's somewhere in between those two. They're not exact, but it's just a tool that gets you closer to seeing which ones are more alike. And then I have a table that just summarizes this for Campylobacter, STEC, Listeria, and Salmonella. The number that is highlighted in purple is the digit in which we recommend local users to do their cluster searches on. So, for Campy, it's the fourth digit, same for STEC, same for Salmonella. For Listeria, it's actually the fifth digit. This is a tool that's in our BIONUMERICS software. And you can see up at the top, you just choose which digit you want to look at. And like we just said, we're going to look at the fourth digit for Salmonella.

So, I'll do four digits. I can decide to include or not include non-human sources as well as historical isolates. I will say for initial cluster detection, we usually don't include those at first. We identify our cluster size. We typically say that you should probably have three or more in your cluster. And then how many days do we want to look back? So, we typically look back 60 days. In smaller states, maybe they want to look back further. And then your result set when you hit recalculate, you'll see the allele code, you'll see the serotype, and then you'll see an outbreak code that we have identified at the national level that's been downloaded in the local database if there is one. And that's always good to know too because maybe you have a new sequence in your database and you say, "Oh, well, it matches this ongoing outbreak investigation. I need to let my epidemiologists know."

So, these are just some screenshots of what people would see at the local level if they're using allele codes to do searches. So, this one, we're wanting to see if it's a match to an ongoing cluster. So, what they will do is they pull that isolate into a dendrogram, and you'll notice that three of these match to the sixth digit and then one is to the fifth. So, you want to still verify what the allele range is in a dendrogram, and here it's zero to one, which is extremely close. So, we would notify our local epidemiologist about this match. And then, this is another one. Let's say we're doing a cluster detection. We look in the past 60 days and we find these three that appear to be close at least by allele code, and we see that their range is two to seven alleles. So, we would report this new cluster to our epidemiologist. This is one of the biggest tools, I think, for allele codes, is being able to look at historical matches.

So, I can just query the database for something text. That way, I don't have to tell the database, "Okay. I want to compare all these allele calls in this one isolate to everything." That takes a lot of computer power. It takes time. All you've got to do is have a network glitch and everything that you were searching for went away. So, this is way easier. Not only that, but this information is shared within SEDRIC, that system I told you that the epidemiologists use. So, you can actually just go there and do a quick text query and be able to look up everything that had this allele code in the past. Now, I do want to verify this in a dendrogram, and I look at these. And so, I would let my epidemiologist know that there was a chicken isolate back from 2019 that matches this ongoing cluster. This could potentially lead them in the direction for finding the source of this current outbreak. And then this one just shows as you get... You'll notice that not all of these allele codes are exactly the same.

The top few share up to the sixth digit. So that's like, "Okay. These are zero alleles. They're all indistinguishable. I know this is a cluster." But then I have this other one that only goes to the fifth digit, and then another one that doesn't even. It only goes to the fourth digit, but this is why we ask them to look at the fourth digit because now, you're including these and you're only going out to eight alleles. So, we would include all of these in our cluster search or in our cluster report. All right. So, we've gone over PulseNet, and we've talked about allele codes, but I really want to talk about how this is used in cluster detection and outbreak investigations. So overall, we use annotated dendrograms and similarity matrices to investigate clusters. These are communicated back to the public health laboratory scientists as well as epidemiologists. And sometimes, this shows that clusters are close, but genetically different enough that they could be investigated separately.

And an example of that is below, where you have what looks to be... They're all very close, but they are indistinct groups. So cgMLST is the primary cluster detection method. And in general, as a guideline, we asked for local clusters to be sequences within 10 alleles. And then at least two of those sequences should be within five alleles. And all of this, you still have to keep in mind, allele differences within a cluster may be larger or smaller depending on the organism and the epidemiologic data. So, this is switching a little bit. We are showing allele differences here, but we're also showing a resistance pattern. So here, the WGS analysis showed that a strain of Salmonella Typhi was associated with travel to Iran and Iraq, and this was genetically distinct from a multi-drug resistant Salmonella Typhi associated with travel to Pakistan. So, the top clade you'll see has the travel to Iraq and Iran, and the bottom clade is travel to Pakistan. And we separated these by their allele differences.

And another interesting point is that the different clades also show different resistance profiles, which is something interesting we've seen. This just shows where WGS was used in outbreak investigation of Salmonella Newport that ended up being associated with melons. The isolates with the earlier isolation dates formed one clade. And then as we got more and more, we noticed that another separate clade was forming. Ultimately, because of the epidemiology, it was investigated altogether, and the epi information indicated a common supplier for the melons and a traceback information from FDA indicated one single farm. But what this spurs is some interest in looking at enhanced surveillance for this particular melon-growing region. So, the WGS analysis will be used to compare strains from the region over time and how those might change. So, switching gears to E. coli. So, this is something that was a local cluster. So, Arkansas did a cluster search in January, and they noted that this particular allele code had been seen four times in the local database in the past 60 days.

This was unique to their database, and then they queried the national database to see if there were matches from other labs. And they confirmed that the allele range was zero, so indistinguishable. So, it was definitely something they wanted to investigate. Turns out this was part of a larger national outbreak investigation. We gave it an allele code. Is part of a REP strain. I will go over REP strains in just a moment, and the allele range remained zero alleles throughout the investigation. So that just gives you an example of a cluster with a single allele code. And I don't have any dendrogram pictures for this, but an outbreak was identified in 2019 that was linked to ready-to-eat chicken product. The WGS helped to link the cases and the ready-to-eat product, and what's very interesting about this and so interesting about Listeria is that it spanned over three years, which is something that we might not have been able to figure out with PFGE alone. Canada also had human matches and they performed sampling and found Listeria in ready-to-eat chicken product.

And then this is just a screenshot from the recall. So, I just talked about a REP code. And what is that? So now, I'm going to talk about what that is, what REP strains are and how WGS is used for surveillance. So, REP is an acronym. It just stands for reoccurring, emerging, and persisting. So reoccurring strain is just a strain that periodically causes substantial number of illnesses, typically in outbreaks, separated by periods in which is not isolated from people, or it causes very few illnesses. So, it might have this kind of bar graph, up and down. Emerging is a strain that causes illnesses that have increased in frequency or have potential to increase in frequency over time. And persisting is a strain that causes illnesses consistently over time, something that we might consider a very common strain that we're just always seeing. We see a lot of that with enteritis. So, for your information, PulseNet started assigning REP codes in December of 2019, and we've been tracking these in the PulseNet national databases, and these are also viewable to our public health partners via SEDRIC.

So, our PulseNet database managers assign these codes along with additional input from our reference lab as well as epidemiologists. And this just gives you an example of what a REP code looks like if you were to happen to see one. So, it tells you immediately it's a REP code because it says REP at the beginning, and then it has the serotype code followed by a number. So, this REPEXH01 just indicates that it was the first coded REP of E. coli O157. And then it may or may not have a dash number, and that would just indicate a particular clade within that REP code. And this really just helps us identify and look over time what we're seeing with our WGS data, and we've already talked about future studies that would include source attribution as well as looking at the evolution of the sequences over time. Just some current REP strains we have are Campylobacter jejuni. That was a strain from the MDR outbreak linked to puppies, O157 linked to that Yuma romaine outbreak in 2018, Salmonella Kentucky, Salmonella Reading linked to various turkey products, and Vibrio parahaemolyticus linked to reoccurring outbreaks linked to raw oysters.

So, what this is a minimum spanning tree, and these are really good for helping us visualize a lot of data. So, this particular spanning tree includes over 850 isolates. And you'll see that there are five different subclades involved in this particular REP. This is the Yuma strain REP. And you'll notice, I mean, as you would expect, that each clade is a different or each subclade is a different color. So, your major one is in red, and then you have your other colors identifying the different clades. And this just shows that ultimately, they're related, but they do span off. There's been approximately 20 different outbreaks associated with this REP strain and sources have included beef, romaine lettuce, leafy greens, as well as a lake exposure. And sequences will be included or excluded from these clusters based on relatedness to other uploads, but not just the allele code.

And this is just a graph over time showing this particular REP strain. And as you can see, the largest number we had was during that 2018 Yuma romaine lettuce outbreak. I did want to note that we didn't fully move over to WGS as our subtyping method until July of 2019. So, any past numbers like 2016 looks very scant. It's just because we didn't have the data. So, in conclusion, PulseNet is a network that works collaboratively with food regulatory agencies and foodborne laboratory scientists and epidemiologists to detect, investigate foodborne disease outbreaks. WGS is used as a tool in detecting foodborne outbreaks as well as discovering reoccurring, emerging, and persisting strains of those REP strains. Allele codes, outbreak codes, and REP codes are all tools used to track and identify strains and outbreaks in the PulseNet national databases, and thresholds for cluster identification are ever changing as we collect more data and learn more. So, thank you for your time and I'll take any questions you have.