Using Genomic Data to Find the Source of Salmonella *enterica* Typhimuriums

[Announcer] This program is presented by the Centers for Disease Control and Prevention.

[Sarah Gregory] Hi, I'm Sarah Gregory and today I'm talking with Dr. Xiangyu Deng, an assistant professor at the University of Georgia. We'll be discussing the use of genetic data to identify the source of salmonella infection. Welcome, Dr. Deng.

[Xiangyu Deng] Thanks, Sarah. It's great to be here.

[Sarah Gregory] We hear about outbreaks of salmonella regularly, but I don't think most people realize that there's different kinds. Your study is about *Salmonella enterica* serotype Typhimurium. How is that different than others?

[Xiangyu Deng] Well, serotype Typhimurium is not necessarily different from other serotypes. What makes it important is that it's one of the most common serotypes and it's very diverse. There are more than 2,600 serotypes of *Salmonella* and Typhimurium is one of them. In many countries, Typhimurium is the most or second most prevalent serotype. According to the U.S. national surveillance data since 1960s, about a quarter of human infections of *Salmonella* in the country are caused by Typhimurium.

And within this serotype, there are diverse lineages and subtypes. Some of them have broad host range and are found in many species of mammals and birds. Others appear to show varying degrees of host adaptation and be associated with certain hosts.

[Sarah Gregory] Okay, so how's it spread?

[Xiangyu Deng] Well, it's a zoonotic pathogen that is spread between animals and humans. I think a notable example is the emergence and the global spread of a multidrug-resistant subtype called DT104. It was first isolated in 1980s in the United Kingdom. It's transmitted from cattle to other livestock in the country and then spread globally in 1990s. Now it's all over the world.

Another example of the recent emergence and the spread of the pathogen is the multilocus sequence type 313. It was estimated to have emerged about 40 to 50 years ago in sub-Saharan Africa. Its spread in Africa coincided with an HIV pandemic, which may explain its regional transmission and invasive symptoms among infected people.

[Sarah Gregory] Even though, as I said, there seem to be a lot of outbreaks lately, how common is it, actually?

[Xiangyu Deng] So, according to the Foodborne Disease Outbreak Surveillance System, there were nearly 3,000 reported outbreaks with a single confirmed etiology in the United States from 2009 to 2015. And that's all kinds of foodborne pathogens. Among these outbreaks, close to 900, or 30 percent, were caused by different serotypes of *Salmonella*, including Typhimurium. So far this year in the United States, we have had at least three outbreaks of Typhimurium or its close variant. And these outbreaks were linked to chicken, chicken salad, and dried coconut.

[Sarah Gregory] When there's a salmonella outbreak, researchers are usually able to pinpoint a source that caused the outbreak, such as eggs from a specific farm or, as you're saying, chicken, chicken salad, dried coconut. How does whole genome sequencing play into this?

[Xiangyu Deng] Well, to find out the source of the salmonella outbreak, both laboratory and epidemiologic investigations are needed. On the lab side, whole gene sequencing provides the ultimate resolution to differentiate closely related isolates. So, with whole gene sequencing, investigators can, first, better link cases of the same outbreak and, second, better match isolates from food or food processing environments to the isolates from sick people. And such lab evidence gives investigators more confidence to implicate a specific source that's behind the outbreak.

[Sarah Gregory] Your study talks about, and I'm quoting, "accumulation of lineage-specific pseudogenes after divergence from generalist populations." That's a mouthful, and what does that mean?

[Xiangyu Deng] So, pseudogenes are DNA sequences that used to be genes, still look like genes, but no longer function the way they did because of mutations. And usually such mutations are detrimental to the bacteria and, therefore, unlikely to be retained in the population. But as a lineage becomes adapting to a particular host and evolves from a host generalist to a host specialist, one theory is that the otherwise detrimental mutations would not affect the specialist lineage as much because the bacteria only need a set of essential genes to deal with the host. As long as the essential genes are functional, they are fine. So according to this theory, accumulation of pseudogenes can be a sign of host adaptation. In our study, we observed in multiple cases that divergence of the putative host-adapting lineage was accompanied by the accumulation of pseudogenes. And those pseudogenes are lineage specific, which indicates that those potentially host-adapting lineages may have emerged and evolved independently from each other.

[Sarah Gregory] About 95 percent of cases of foodborne illness occur sporadically, seemingly striking one person at random. Why is it harder for researchers to find the source of salmonella when you say "just one person is infected"?

[Xiangyu Deng] Well, finding the source of a foodborne infection of *Salmonella* requires epidemiologic investigation that takes time and resources. The sheer volume of these foodborne infections makes it impossible to investigate every single case of foodborne illness. Typically, only outbreaks get investigated. The source of the sporadic case is often unknown because there is no investigation.

[Sarah Gregory] Ah, I see. OK, well, you developed a machine learning classifier that uses these genetic differences to help identify the sources of salmonella infection. What's it called and how does it work?

[Xiangyu Deng] Well, thanks to routine use of whole gene sequencing in public health labs, we now have a large and expanding volumes of *Salmonella* genomes. For the particular case of Typhimurium, we had more than a thousand genomes at the time of our study. And many of those genomes came from major livestock animals and wild birds. So, we were able to put together a big collection of Typhimurium genomes from known sources. So, I will refer to this collection as a training set when we start talking about machine learning later on.

So, we then built a machine learning classifier. The classifier can predict the zoonotic source of a Typhimurium isolate by interrogating thousands of genetic features of its genome. And we trained the classifier to do this prediction over and over again using the genomes in the training set I just mentioned. And after the training, the classifier was able to learn how to predict major livestock and wild bird sources with a decent accuracy. It also figured out how important a specific genetic feature was for source prediction.

[Sarah Gregory] Your classifier relies on a large database of genetic sequences from different populations of *Salmonella*. When and where did these sequences come from?

[Xiangyu Deng] Alright, so most of these *Salmonella* Typhimurium genomes used in our study came from three major surveillance and monitoring programs in the United States. First, we selected human isolates from outbreak and sporadic cases in the United States over 65 years since 1949. And these isolates represented the genetic diversity of serotype Typhimurium according to the surveillance by the PulseNet, which is a laboratory network managed by CDC and focused on human isolates.

And second, we have genomes from the GenomeTrakr database as of January 2017. GenomeTrakr is a database managed by FDA and focused on foodborne pathogen genomes from food products in the environment. Our GenomeTrakr genomes came from various sources in the United States, Europe, South America, Asia, and Africa.

Last but not least, we took advantage of the retail meat isolates sampled by the FDA arm of the National Antimicrobial Resistance Monitoring System, or NARMS. So it gives us a lot of genomes from poultry, beef, and pork, which turned out to be very helpful for livestock source attribution.

[Sarah Gregory] So, is this classifier very accurate?

[Xiangyu Deng] Overall, the accuracy for zoonotic source attribution was about 83 percent. The classifier can flag its prediction as either precise or imprecise. So, when the prediction was precise, the accuracy was about 92 percent. We also retrospectively analyzed eight major zoonotic outbreaks in the United States from 1998 to 2013; seven of them were attributed to the correct livestock source.

[Sarah Gregory] So, if I get food poisoning, will I be able to go to my doctor and have the source identified?

[Xiangyu Deng] Probably not. In a future world, if your doctor requests whole genome sequencing of the pathogen that made you sick, and the pathogen turns out to be *Salmonella* Typhimurium, then the whole genome sequencing data could be analyzed by our machine learning tool and you might be able to make educated guess where the pathogen may have come from. But in our current system, the isolate is submitted to a local or state public health lab and then becomes part of the national surveillance. And like I said before, source identification requires epidemiologic investigation, which may happen only if you are part of an outbreak that gets investigated. So I think our tool has the potential to help outbreak investigations.

[Sarah Gregory] Is the problem of *Salmonella* in livestock getting worse?

[Xiangyu Deng] Well, actually, in recent years we've seen some positive trends. For example, according to the latest integrated report from the National Antimicrobial Resistance Monitoring System in 2015, the prevalence of *Salmonella* in retail poultry products continued to decline and reached the lowest levels since 2002. It was at 6.1 percent for retail chicken in 2015, down from its peak at over 20 percent in 2009.

[Sarah Gregory] Okay, are there any more important aspects of this study that you'd like to mention?

[Xiangyu Deng] We demonstrated that a small set of 50 genetic features, including SNPs, indels, and genes were enough for fast and robust zoonotic source attribution, instead of analyzing entire genome. So, this finding may lead to a rapid and a scalable source attribution tool because you don't have to analyze the entire genomes to do it. It would also be interesting to see if these features simply correlate with the source association or if they play any functional roles in host adaptation or preference.

We also developed a method to systematically screen for individual lineages that we can estimate their ages by evolutionary analysis. From a large global phylogeny of Typhimurium, we found two pig lineages and one chicken lineage. And these lineages originated recently in 1990s and quickly became circulating across the United States, possibly due to industrialized livestock production.

[Sarah Gregory] Tell me about your job at the University of Georgia and how it relates to salmonella infections.

[Xiangyu Deng] I run a lab at the University of Georgia Center for Food Safety. A focus of our work is what we called genomic epidemiology. So we take genomics, bioinformatics, and data science approaches to study the epidemiology of foodborne pathogens, especially *Salmonella*. For example, we developed a bioinformatics tool called SeqSero that can predict *Salmonella* serotypes from whole genome sequencing data. My interest in *Salmonella* started when I was an American Society for Microbiology Fellow working at CDC more than five years ago. During my fellowship at CDC, serotype Typhimurium was the second serotype I took on after Enteritidis. By the way, the Enteritidis paper was also published in EID about four years ago. I started with about 100 Typhimurium genomes and then whole genome sequencing started to take off and transform public health microbiology. Four years later, we ended up with so many genomes, whole genomes, that machine leaning became the natural choice to deal with all these data.

[Sarah Gregory] Thank you so much, Dr. Deng. Listeners can read the full January 2019 article, Zoonotic Source Attribution of *Salmonella enterica* Serotype Typhimurium Using Genomic Surveillance Data, United States, online at cdc.gov/eid.

I'm Sarah Gregory for *Emerging Infectious Diseases*.

[Announcer] For the most accurate health information, visit cdc.gov or call 1-800-CDC-INFO.