# Analysis of Whole-Genome Data in a Public Health Lab

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

[Sarah Gregory] Today, I'm talking with Dr. Kelly Oakeson about bioinformatics and genome sequencing in a public health lab. Dr. Oakeson is a bioinformatics and genomics research analyst with the Utah Department of Health. Welcome Dr. Oakeson.

[Kelly Oakeson] Thank you. Happy to be here.

[Sarah Gregory] What are the functions of a public health lab and what part do you play in the Utah Public Health Laboratory?

[Kelly Oakeson] Well, public health laboratories have many core functions, for example, they assist in disease prevention, control, surveillance, they also help out with food safety, environmental health and protection, even in newborn screening for congenital disorders in newborn babies. We also provide support to local clinical laboratories to help confirm their test results, as well.

At the Utah Public Health laboratory, I oversee the area of the lab that performs Pulsed-field gel electrophoresis, also known as PFGE. I also do whole genome sequencing. And I take care of all of the bioinformatic analyses, as well here.

[Sarah Gregory] Ok, so what does sequence data in a public health lab do for us?

[Kelly Oakeson] Well, whole genome sequence data provides a very high-resolution tool for everything from identifying what types of microbes or viruses you might have in any given sample to determining antimicrobial resistance and identifying clusters of related microbes in the case of a foodborne illness outbreak.

[Sarah Gregory] And following on that, what is Next Generation Sequencing?

[Kelly Oakeson] Next Generation Sequencing is a very high throughput, massively parallel method for sequencing DNA. It allows us to quickly generate huge amounts of data and DNA sequence in a very short period of time. We can sequence all of the DNA in a bacterium at over fifty times depth in just under two days.

[Sarah Gregory] I understand that the ability of this sequencing to work optimally is dependent on bioinformatics. Can you explain this to us?

[Kelly Oakeson] Yeah, so Next Generation Sequencing generates a lot of sequence data that need to be translated into meaningful biological information. You know, what do we do with all of these As, Ts, and Cs? What do they all mean? In order to do this, we need to use computational tools. You know, with so much data that's being generated, we need to use computers to help us understand what's going on. So, bioinformatics combines computer science and biology in order to make heads or tails out of all of this sequence data we're making.

[Sarah Gregory] And what is the purpose of your perspective in the EID journal?

[Kelly Oakeson] The purpose of our perspective in EID was really to help public health laboratories around the country understand sort of basics of typical bioinformatic pipelines and analyses, right? To help illustrate that they can be flexible in terms of what types of computer technologies they can use and the types of analyses they can do.

[Sarah Gregory] And what are some of the obstacles to getting proper bioinformatics in labs?

[Kelly Oakeson] Well, I think there are several kinda large obstacles that are really preventing laboratories from doing proper bioinformatics. I think a general lack of understanding as to what bioinformatics is and how important it is, especially when dealing with next generation sequence data. Fear, you know, I think laboratories hear the term bioinformatics and it conjures up these visions of huge rooms full of computers and all of these guys sitting down in front of, you know, a tech screen writing code from scratch. I also think, you know, finding people who have the proper skill set to perform bioinformatics is a challenge, as well.

[Sarah Gregory] And are there tools to overcome these obstacles?

[Kelly Oakeson] Yeah, there are some tools to help overcome these obstacles. The CDC's Office of Advanced Molecular Detection is doing a great job of helping educate people on what bioinformatics is and how it's being applied. There are more and more options for commercial software that allow laboratories without people with computer science backgrounds perform these advanced analyses. There's even a handful of bioinformatics working in different states that are helping out each other. We formed our own kinda little community. StaPH-B is what we're calling it and it's a working group and a support group for state public health labs and bioinformaticians to help each other out. You know, we're an active group that was formed to help each other and advocate for bioinformatics at the state level.

[Sarah Gregory] Can you tell us what the public health impact of bioinformatics is?

[Kelly Oakeson] Yeah, bioinformatics has a wide-ranging impact on public health. There are many examples of bioinformatics and Next Generation Sequencing helping to quickly solve foodborne illness outbreaks and pinpoint the cause of where those microbes came from. Bioinformatics is playing a huge role in addressing the spread of antimicrobial resistance in bacteria by analyzing how these bacteria are sharing and swapping the genes responsible. And we're actually here in Utah trying to apply bioinformatics to newborn screening and developing new sequencing methods for detecting these congenital defects and metabolic disorders really early on. And there are many more examples, I could go on and on and on.

[Sarah Gregory] Ok, well tell us about the bioinformatics pipeline developed by the Utah Public Health Lab.

[Kelly Oakeson] The pipeline we developed here is essentially a set of programs that we kind of run one after another that takes the sequence data we generate and performs several different analyses on that data. We start with the raw sequence that comes off of our sequencers and we

perform quality control on that data. And this ensures that the As, Cs, Ts, and Gs that we generate, that we have confidence in those, that they really are an A, C, T, or G.

Next, we perform de novo genome assembly on those sequence reads for each of our isolates. This takes all of those short fragments of DNA that we sequenced and puts them back together to generate a complete genome sequence. With that genome sequence we then find all of the protein coding genes in those genomes and annotate them with the proper gene name and function. Now this allows us to look for genes involved in, let's say, antimicrobial resistance, or virulence genes, or toxin genes, and so on.

Now, after we know the inventory of protein coding genes in an isolate, we can compare those inventories between isolates. So, let's say we want to determine how related a group of isolates are, well, we can go in and extract all of those shared homologous protein coding genes, build a nucleotide alignment, and then build a phylogenetic tree using maximum likelihoods. And this allows us then to kind of compare the genomes to each other and see how related they are and see if they share a common evolutionary history. One really kind of unique thing about our pipeline is we are not relying on comparing our sequence data to any kind of database or previously generated whole genome sequence. We are reference free.

[Sarah Gregory] What do you see as the public health impact of the Pipeline?

[Kelly Oakeson] I think one of the biggest impacts will be how other states can see how what we're doing here in Utah and use that as an example so they can start developing their own pipelines. You know, this'll allow states to start immediately using all of their sequence data that they're generating and applying it to their own local public health issues, like solving their own local foodborne illness outbreaks. I think in a larger context, however, bioinformatics in public health can have positive impacts, not only on infectious disease, but also, again, in newborn screening. I think applying what we've learned about next generation sequencing and bioinformatics for microbes can be used collaboratively across all of public health.

[Sarah Gregory] Thank you, Dr. Oakeson. I've been talking with Dr. Kelly Oakeson about his September 2017, article, Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory, 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.*