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

Transcript of the presentation, Advances in Metagenomic Approaches to Detect Foodborne Pathogens, Henk den Bakker, PhD, University of Georgia, United States

So today, I'm talking about advances in metagenomic approaches to detect foodborne pathogens. Advances in metagenomic approaches to detect foodborne pathogens or metagenomic approach, just to get to know more about the ecology of foodborne pathogens. And so, this is the overview of my talk. I will have a short introduction about metagenomic approaches. So different types of microbiome sequences, and while discussing those things, I will also stress the advances that have been made in microbiome sequencing approaches over the last couple of years. If we're talking about metagenomic approaches, we have to think of a trifecta. So, it's a trifecta of the data we collect, sequencing technology and how we analyze the data. So, my next topic will be about the introduction of long read sequencing, I think in the last couple of years in microbiome applications. And last but not least, I'm going to discuss some real-life data and to show how we can leverage actually the microbiome or pathogen detection.

So, if we're talking about microbiome sequencing, there are two main ways of sequencing microbiome, one of them is amplicon sequencing. So that means that we use a highly conserved, yet variable part of the genome, and about 300 to 500 base pair, dependent on the sequencing technology.

Now that we used to sequence of all or the preferred part of the organisms that we're interested in of a sample. And traditionally, those are rRNA genes. So 16S for archaea and bacteria. But one thing to keep in mind is that chloroplasts also have 16S, and 18S in ITS for eukaryotes.

So, a quick overview, the wet lab part of the sequencing is basically doing a PCR and preparing to PCR in a library prep and sequencing it on your sequencing system of choice. And a bioinformatics step where we do OTU classification. So, this is a part of microbiome research where we've... Have seen a lot of advances over the last couple of years, which consist of data filtering, filtering out cross contaminants, false index-pairing, et cetera. And one of the things is if you do this kind of sequencing, so amplicon-based sequencing is to put in a lot of controls, so negative controls and positive controls.

So, one of the advantages of amplicon sequencing is that you can start with small amounts of starting material, which is especially if you talk about an environmental sampling, for instance, in food processing plants, very important. Based on the primers that you use, only the organisms of choice. So only bacteria will be sequenced, and it's highly efficient. So, because you only look at one gene at a limited number of reads is enough for a microbiomes survey, which translates to, if you multiplex 96 samples on an illumina MISeq run dependent on how... What you pay for your consumables, you can push the price down to less than $40 a sample. The disadvantages of using amplicon sequencing are that all of the ways... Most bioinformatics applications for amplicon sequencing workflows are reliant on

database searches. So, you have to have a really good database with representatives of the organisms you expect in your database to classify your organisms.

The reliance again on primers makes it also possible that you are missing organisms that are important for further analysis because your primers didn't catch that diversity. And one of the things that I found is a disadvantage of RNA amplicon sequence, is that it also amplifies case 16S organelles of eukaryotes. So, if you want to look for the microbiome of living plants for endosymbionts, most of your reads will go to 16S of the chloroplast of the plant.

So, the other big, big way of sequencing microbiomes is shotgun metagenomics. So, in that case, all DNA in a sample is sequenced without any enrichment or PCR amplification. So, you treat the sample as if it was one large organism with a very big genome, of course, that genome consist of multiple genomes of microorganisms. So, there are two main approaches for analysis, you can do read classification. So, in that case, the bioinformatics workflow predicts perfect track to match it to what organism it belongs. And that's again, very database dependent. And then there is the De Novo assembly and binning approach. And that's a very interesting approach.

And here we have cartoon. Basically, what it does is here, we have our original microbiomes, and the circles are individual genomes. We sequence them and basically, we don't sequence genomes in long read. Usually, we sequence them in little puzzle pieces of 300 base pairs, 250 base pairs basically use different sequencing technologies. We can have it up to like 10,000 base pairs, but we never sequenced those genomes entirely. What we do then, we sequenced them, we assembled those genomes and tried to get to re-assemble those pieces of genomes and longer pieces. And then we tried to classify those pieces based on abundance. So, statistics and all sequence characteristics. So, in this case, this sample is Meta BAT, and we use a Tetranucleotides, so four base pair worth frequencies to make individual bins. And then we make those bins and try to reconstruct the original genomes again. But this is a very interesting application. So, what we get in the end from this workflow are MAGs, so metagenome assembled genomes.

So that strategy of shotgun metagenomics is that we don't have any reliance on primers, everything is sequenced. So, we don't have to worry that our primers didn't catch all of our diversity and whole genomes are sequenced. So, it makes it possible to study gene abundance, presence of gene families of interest. So, you can predict potentially virulence genes or anti-microbial resistance genes in your microbial population. And the ability to study the uncultured, the microbial dark matter. So, it's better than amplicon-based approaches.

And I also like to show this picture. This is a tree of life and all those red dots here, including this large new candidate violet are new organisms that have been discovered for [inaudible 00:42:06] with that assembly-based approach.

So, these advantages are, again, it's almost like what's an advantage is also a disadvantage because we don't rely on primers. Everything is sequenced. So, if we go to a meat processing plant and we sequence our DNA, 99% of our DNA or of the sequence reads may go to actual animals that are processed in that plant. So, in this case, chicken or turkey. And because we need a lot of... We don't look at it in a single gene, but all genomes, it's also expensive. So, to get enough coverage for certain communities, especially complex communities and an entire sequencing run is sometimes necessary. So instead of $40 that we spent per sample for 16S sequencing, if you have a super complex microbiome, you may spend like a thousand to $2,000 per sample.

So, one of the advantages that... So, there are a lot of different variants of these two ways of sequencing microbiomes, getting a picture of what's in the microbiome. And one of the advances, I think in respect of... With respect to foodborne pathogens that has been pursued in by the Deng lab, at the Center for Food Safety at UGA and by the FDA is the introduction of quasi-metagenomics. So, one thing to keep in mind with foodborne pathogens, if things didn't get out of hand that they usually make up a very minor component of the microbiome. So, if you want to do shotgun metagenomics, you needed a lot of sequence capacity even to find those foodborne pathogens. So, they can easily go and detect it in traditional microbiome analysis. So quasi-metagenomics uses a number of amplification and selective steps to sequence organisms of interest.

So, here's an overview and this was a paper that was published in 2018, by Xiangyu Deng's group. In this case, they took three different products, black pepper, lettuce, and chicken breasts. And what they did was first, they put them through a traditional enrichment. In this case, they looked at Salmonella enterica so some Salmonella enterica enrichment, then instead of going through the whole enrichment, they could stop the enrichment at the very... After a very short time, like 12 or 24 hours. And then they enriched the final sample with immuno-magnetic beads that were specific for Enterobacteriaceae. And to top it off, they created more DNA after the enrichment immuno-magnetic beads with a whole genome amplification step.

So, if the phi29 DNA polymerase fetch polymerase. So, what you can do then, you can do real-time PCR to go straight for your organism of interest and also sequence your organism with your sequence platform of choice. Interesting thing here is that they spiked the samples with known strains, and they were able to recover the entire genome at accurate enough level to do strain identification and do the kind of strain level subtyping that was discussed in previous talk.

So, the other advantage, if we talk about sequencing is the introduction of long read sequencing technologies into metagenomics. So, a lot of what we do as bioinformaticians, and I specifically say we, because I consider myself a bioinformaticians, a lot of uncertainty in microbiome datasets has to do with the fact that we use short reads. So, Illumina reads are typically like between 150 and 300 base pairs these days they're [inaudible 00:47:22], but they're very short. So, all our approaches we were dealing with small pieces of genome, and they just contain less taxonomic information. So, we have to do a lot of bioinformatics tricks to name them.

So, replacing or including reads obtained with long read technologies can greatly improve the accuracy of some of those microbiome analyses like I showed you the MAG workflow. So, the assemblies are better, so we can better recall for whole genomes from, from microbiomes, but we also have more data to make more accurate predictions of what is in our microbiomes. So, the two different analysis workflows just get better. The only problem is that long read technologies as they are now, and again, that advances almost every day is that they have higher sequencing error rate as compared to, for example, Illumina reads.

So here is an example of one of the early players of long read technology. So Pacific Biosciences with their smart, single molecule real time sequencing. So, it's a long-read platform. And one of the things to keep in mind is that it has a fairly large footprint, even their smaller sequences have large footprint, and you need a lot of specialized equipment for library preps. So, kind of the new kid on the block, at least a couple of years ago, it was Oxford Nanopore. And they've produced these really small sequencing devices that you can plug into a laptop and do your sequencing in-house. And here we have minION and

here we have, I still haven't seen it in real life smidgION that they developed as a sequencing device that you could plug into your iPhone. And this is an interesting, great long read technology. So, there are competitions of users to get the longest reads. So usually, the long reads are between a hundred thousand base pairs, but recently I've heard people that could get long reads that were close to half a E. coli genome.

And the cool thing about this data, this platform is that it's fast. So, we can do a sequencing and run in a day. And even after a couple of hours, you can start analyzing the data. So, while the sequencing is in progress. So, if you want to look for specific organism, you can start to analyze your data while it comes off the machine. Problem is that it has a high per read error rate. So [inaudible 00:50:43] have to deal with that and the applications as they are now largely overlap with PacBio, but with a much lower cost than upfront investment. And just as a something to note is that this is currently the sequencer of choice in a lot of sources [inaudible 00:51:08] to sequencing efforts.

So, then I come to my last part of the talk. So, we talked about looking for foodborne pathogens and method or microbiome datasets. We looked at how to amplify our pathogens of choice in a dataset, but can we just use the microbiome data as they are without much manipulation to predict the occurrence of pathogens. Again, to keep in mind is that you don't find a lot of reads in your datasets of foodborne pathogens. If we have those microbiome data, can we find higher abundance, microbial taxa, families, genera, species that can accurate... Are accurate predictors of the occurrence of foodborne pathogens. So, can we find novel indicator species with our microbiome data? Can we look at the whole community and associate, like certain community profiles with occurrence of foodborne pathogens. And I'm not going into detail on that... Into that question. But that's certainly something we can do, and we can if do it 16S data.

So here are some examples of a group of Tan state of the Kofax lab. And they looked at the microbiome of three apple and all their tree fruit packing facilities to look for. And they used amplicon-based methods 16S for bacterial organisms and fungal ITS2 communities to see if there was a relation to the presence of Listeria monocytogenes.

And they did find that the microbiota in this facility and specifically in the wet processing area was associated with the Listeria monocytogenes. And then highest Listeria monocytogenes occurrence was unique with those microbiomes were uniquely predominated by the bacterial family of the Pseudomonadaceae and fungal family of Dipodascaceae. So, the interesting thing here is that... This initial, and that's where bioinformatics comes in. Right now, we have information at the family level that we can use in our taxonomy. So, more precision in our taxonomic placement and new work for bioinformatics pipelines to go even deeper into the sequencing data, to home in on potential spaces, therefore associations, if possible, that can tell us more about those community. So, I used DADA2. So, I re analyzed their data set.

And that's a method to infer the exact sequence of templates of the amplicon and amplicon method. And it gives us amplicon sequence variants. So, we can use, we reconstruct the actual templates that gave us the PCR product. Next thing, analyze it with DeSeq2 and that's a statistical method, which tells us which of those sequence variants are significantly over or underrepresented in specific groups. So, I compared samples that were Listeria monocytogenes positive versus those that are Listeria monocytogenes negative. And next, I looked for a data file, genetic analysis of those partial six 16S sequence to figure out if, what the diversity, how far we could go to figure out what the species or subpopulations were that constituted to that Listeria monocytogenes associated taxa.

So here we have the output of these DeSeq analysis. It's a kind of a difficult picture here, but all of these dots are in the taxa that are significantly associated with the positive occurrence of Listeria or negative occurrence of Listeria monocytogenes. And so, one of the interesting things here is that the gene is that's most, mostly associated with Listeria here positively, but also negatively associated with Listeria is a subgroup of Pseudomonas_E. And you can see here that apparently not all Pseudomonas_E strains are created equal because some of them are positively associated. So, this is a phylogenetic tree of pseudomonas here, and some of them are negatively associated. And some of them are more abundant with... When you compare them between environments with Listeria monocytogenes and others. So, one of the things here is... notice that Pseudomonas fragi clade. So, a cold tolerant pseudomonas fits very good in biofilm formation, which seems to be highly associated with Listeria monocytogenes.

So those Pseudomonas fragi, if we look at... Could be like an indicator species that tells us more about harbored sites, potentially harbored sites of Listeria monocytogenes with how to have them to search for Listeria monocytogenes. So, this is just an example of how bioinformatics and improved bioinformatics methods that are improving over the years are contributing to our understanding of foodborne pathogens and their role relation to the microbiome.

So, in summary, I hope I've shown you that microbiome studies can help us to advance detection of foodborne pathogens, and at least they can help to advance understanding of the ecology of foodborne pathogens. And I think the main advances in microbiome research are made in sequencing approaches, advances in sequencing technologies and advances in bioinformatics. And what I want to mention last is that the public availability of those datasets even old datasets in repositories, like NCBI SRA make it possible to go back to some of those old datasets and retroactively apply some of those advantages, especially to the advances in bioinformatics to all datasets. And I want to thank you for your attention.