



Article:

Divinlal Harilal, et al.

SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission.

Clin Chem 2020; 66:1450-58 <https://doi.org/10.1093/clinchem/hvaa187>

Guest: Dr. Ahmad Abou Tayoun, Director of the Genomic Center at Al Jalila Children's and an Associate Professor of Genetics at Mohammed Bin Rashid University of Medicine and Health Sciences in Dubai, United Arab Emirates.

Bob Barrett:

This is a podcast from *Clinical Chemistry*, sponsored by the Department of Laboratory Medicine at Boston Children's Hospital. I am Bob Barrett.

As we record this in November 2020, the COVID-19 pandemic continues to inflict devastating human life losses and has imposed major social changes and costly global economic shutdowns. Several governments are sketching out plans for slowly re-opening economies and reviving social life and economic activity. However, robust population-based surveillance systems are essential to track viral transmission during any re-opening process. While reverse transcriptase real time PCR targeting SARS-CoV-2 RNA can be effective in identifying infected individuals for isolation and contact tracing, it is not useful for determining which viral strains are circulating in the community - are they already indigenous or newly imported ones? It is important to know the origin of the strains, which in turn influences public health policy decisions. In addition, it is vital to identify super-spreader events as they can be influenced by the type of virus strain. That is something that whole genome sequencing of SARS-CoV-2 can detect.

A paper, appearing in the November 2020 issue of *Clinical Chemistry* entitled "SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission" examines this issue. The senior author for that study is Dr. Ahmad Abou Tayoun. He is a is the Director of the Genomics Center at Al Jalila Children's, and an Associate Professor of Genetics at Mohammed Bin Rashid University of Medicine and Health Sciences in Dubai, United Arab Emirates. Dr. Abou Tayoun is our guest in this podcast. So doctor, just how is SARS-CoV-2 whole genome sequencing performed?

Ahmad Abou Tayoun:

Thank you for a great question. So, the SARS-CoV-2 whole genome sequencing starts like with the COVID-19 PCR test, it starts with obtaining a nasal swab from the patient highly suspected to have COVID-19. Of course, the nasal swab is then treated to extract the RNA from the sample and when you extract the RNA, you have a mixture of viral RNA plus the human nasal epithelial cells RNA after which the RNA is

transformed to the cDNA using RT-PCR and then it's amplified. And then based on the sequencing protocol, you could use different adapters or primers to generate libraries that are more amenable to sequencing using your platform of choice.

So in our case we use Illumina sequencing so we use library preparation kits that will basically allow those libraries to be sequenced using Illumina. Now, of course after you generate the sequencing data, you apply highly specialized bioinformatics tools that basically align back those generated sequences to the referenced Wuhan genome and it tries to call variants that are different from this reference genome and then based on this information, the virus extracted from the patient, its genome will be assembled and then compared to other viruses or other viral sequences.

Bob Barrett: Well, PCR is a nucleic acid-based technique. How is whole genome sequencing different from PCR testing?

Ahmad Abou Tayoun: So, if you have a patient highly suspected to have COVID-19, the way to confirm this is by using a PCR assay and basically actually it's an RT-PCR assay because we know SARS-CoV-2 is an RNA virus so you have to first reverse transcribe the RNA into cDNA and then amplify it. So, if the patient is positive -- and by the way with PCR, you have specific primers that target a specific region in the SARS-CoV-2 genome. And so if the patient is positive for SARS-CoV-2, you will generate a signal that you can measure and confirm that the this patient is positive for SARS-CoV-2 and is having COVID-19 disease.

Now this information obviously is important because then this patient can be isolated, we could track recent contacts and try to limit the spread of the virus. However, if you do whole genome sequencing, not only can you identify the SARS-CoV-2, the virus and its genome, but you'll be able to use this genomic information that will tell you if it is similar to genomes that have been spreading say in Europe or Italy versus genomes of viruses that have been just inoculating within the region where the patient comes from. And the different scenarios will tell you about whether it's a case of community-based transmission versus it's the case where there's an outside external introduction of the virus.

Bob Barrett: How many strains of the SARS coronavirus-2 are there and how does genomic sequencing help in population surveillance?

Ahmad Abou Tayoun: So like I mentioned, when we generate the sequence of the virus, there is now thousands of those genomic sequences that are available at public databases. And so when we

generate a viral sequence from a patient now, we could basically align the sequence to viruses from all over the world and within a certain country, and so we can then see if this virus has certain mutations that are similar to any of those viruses that have been sequenced elsewhere. So, we know that the virus as it moves between patients, it acquires mutations at the rate of one mutation every two weeks. So most of those mutations are benign and do not necessarily affect the pathogenesis but they can be used to basically track the virus.

So, if we have sequenced a bunch of patients in a certain region and we have accumulated or understood the different type of mutations that are present in this population at a certain time, and then say three months later, we identify a new patient, if we do a PCR test, you basically will identify if this patient is positive or negative and if positive, you could apply certain measures where you can do the tracing and tracking. But with genome sequencing, you're not only identifying that the patient is positive, you could also see if this virus has mutations that suggests that it is coming from certain outside countries due to similarities in the mutations they carry or is due to mutations that have been characterized within this region and it's most likely to be community-based transmission.

So differentiating between imported virus transmission or outside introduction of the virus versus community-based transmission can have different public health measures. So if you have a community-based transmission, so public health decision makers might look into lockdowns of schools or work areas and institute the remote work versus if you have more of the introductions that are coming from outside the country, then there can be more measures that have to be more strict around travel hubs like airports for example. So, genomic sequencing can generate more data that is very helpful in not only identifying the patient and tracing or tracking them, but also in identifying the most relevant public health measures at the population level.

Bob Barrett:

From a practical standpoint, is it realistic to adopt this approach? For example, compared to PCR, is it scalable to large populations and have a similar turnaround time? And what about the cost compared to other techniques, is it cost effective?

Ahmad Abou Tayoun:

This is a very good question and obviously if something -- we talked about the advantages of doing whole genome sequencing but of course for it to be practical, it has to be scalable and cost-effective. And so like I mentioned earlier, the nasal specimen, the nasal swab when you extract the RNA, you have total RNA from humans and the virus -- actually, 99% or more of the RNA comes from human

epithelial cells and only 1% or even less comes from SARS-CoV-2. So if you sequence all the RNA in that specimen, you're left with little amount of data that you could use to generate the whole genome sequence for the SARS-CoV-2.

So because of this small amount of the starting material, what you end up doing here is you have to sequence at the higher depth, which means in a single batch you have to use less number of samples and which means the cost would be higher, right? So, when you have less samples, you're using the same reagents for less samples so the cost is going to be very high. And in this case, you have to sequence at really high depth in order to obtain meaningful data to assemble the viral genome.

However, what we published in this paper is a method where we actually amplify this 1% so we describe and show the advantages of a target amplification protocol where we use 26 overlapping PCR products to amplify the SARS-CoV-2 genome. We then fragment those amplicons and pull them into a single tube then sequence this and then obtain the viral genome at the very high sensitivity. We simply just amplify the 1% of the extracted RNA from the nasal specimen and then therefore this enables us to multiplex and use a much higher number of samples in a sequencing run which really drops the cost significantly. We calculated the cost and just to give you an example: if you do total RNA sequencing using what we call a shotgun total RNA sequencing approach, you have to generate around four gigabytes of data to have an average coverage of say 45x.

However, with the target amplification protocol that we described, you only need 0.02 gigabyte, so this is 200-fold less data to generate a genome at 400x. So tenfold more coverage. So you get tenfold more coverage at 200-fold less data and this equates basically, if we do the cost analysis, the cost of doing whole genome sequencing at this specification is around \$87, which is comparable to that of the PCR given that we're getting additional information that helps in guiding public health measures. So yes, whole genome sequencing is very practical if we're using the right protocol like we show in this paper, and this protocol consists of target amplification first and then followed by sequencing.

Bob Barrett:

Well, finally doctor, can you give an example where the advantages of using whole genome sequencing outweigh the shortcomings?

Ahmad About Tayoun:

Yes, absolutely. I mean, there are many examples and I think the most famous is the first case that was documented in the United States which I think was in mid-January in Seattle area and this was when the virus was sequenced,

that genome had three mutations that were similar to a genome that was sequenced in China. And apparently, this patient had a recent travel history to visit family in Wuhan, so it was clear that this is most likely to be a direct introduction from Wuhan. Now, five weeks later, I think in February 24th, another high school student in Snohomish County in Washington developed flu-like symptoms. When a nasal swab was obtained the patient was confirmed to have COVID-19, by PCR of course.

But now, again, this is an example where the PCR does not tell you much, it just tells you it's positive and we have to isolate and maybe track interactions and trace, but it doesn't tell you where the virus came from. So when the virus was sequenced in this patient, it seemed that he had the same or she had the same mutations as the first index case in January, five weeks ago, in addition to a few other mutations. This patient did not have any travel history.

So what this tells us is that this is an example of community-based transmission and most importantly, this virus has been traveling around for five weeks silently without any strict measures because in the U.S. I think the measure was basically to focus on patients with direct travel from China, so those patients were silently spreading around, and the estimate that was given by genomic epidemiologists back then that during these five weeks, the virus could have reached or have spread among 600 or a thousand patients. So, had there been more strict public health measures at least to limit community-based transmission that were implemented early on rather than just focusing on travel restrictions, this could have limited the early spread of the virus in Washington. So this is a clear example where genomic information can be clearly utilized to understand the transmission of the virus.

Bob Barrett:

That was Dr. Ahmad Abou Tayoun, Director of the Genomic Center at Al Jalila Children's and an Associate Professor of Genetics at Mohammed Bin Rashid University of Medicine and Health Sciences in Dubai, United Arab Emirates. He has been our guest in this podcast on whole genome amplification and sequencing for SARS-CoV-2. He is a co-author of a paper describing that approach that appears in the November 2020 issue of *Clinical Chemistry*. I'm Bob Barrett. Thanks for listening.