Step 1.7 How was the novel coronavirus identified?

In this mini-lecture Professor Martin Hibberd discusses the process that led to identification of the novel coronavirus and how genome sequence data can be used, for example in developing the initial PCR diagnostic and tracking the evolution and diversity in the genome as it spreads (recorded 1st May 2020).

In the first run of the course, whilst many people enjoyed this step, others found it a bit too technical. In this new format, we have tried to explain the more technical language and there are also glossaries in the See Also links below which you may find helpful.

But if you find it too technical, don't worry, you'll probably find the ones which follow OK!

Video transcript:

PROFESSOR MARTIN HIBBERD: The virus was originally isolated from lung wash from patients with pneumonia of unknown aetiology in Wuhan, China. They used tissue culture to amplify the virus and allow for electron microscopy. They also used standard multiplex PCR-type assays. These are molecular probes that recognise all conventional viruses. And these days we also have the chance to use deep sequencing to look at the entire material and identify anything unexpected or novel in that sample.

And here we can see pictures of the electron microscopy results showing that virion - those little spikes around the outside. And those are typical of coronaviruses. Multiplex PCR interestingly showed similar results identifying a coronavirus, but couldn't identify a specific type of coronavirus. So it was novel. And our deep sequencing identified something that was unexpected in humans and not seen before.

The sequencing technology is rather complicated and difficult, and moving it from that clinical sample into a nice conventional genome that we can all look at is not easy. But in China, they did that very rapidly. And luckily they published that online so that we could all see it.

And this is a little process of that sequencing technology going from the original genome, which is fragmented, small pieces of sequence that short read, and then recompiled into this genome. And this is a 30,000 base per genome that you can see, which is pretty large for viruses. And if you look along all those gene names, you might recognise the spike protein which is the receptor binding and how the virus enters human cells.

When we have that genome, we can now start to put it together with all the other viruses. And here you can see all the coronaviruses that we know about put together, and you can see SARS CoV-2 there along with the bat viruses not too far away from the SARS coronavirus. We can see that zoonotic viruses that occurred since 2003 is the MERS FutureLearn 2 coronavirus, SARS coronavirus, and now the SARS coronavirus-2, which is the cause of COVID-19.

We can also see in this tree the so-called human coronaviruses, which are commonly circulating in humans. And those probably originated from rodents in the historical past. If we look more carefully at those viruses most similar to SARS CoV-2, you can see that they originated in horseshoe bats. And there is quite a diversity of these coronaviruses in the horseshoe bats. And SARS CoV-2 fits right into that.

However, there is a little bit of evidence that some of the SARS CoV-2 virus looks like a pangolin coronavirus sequence. And that suggests that potentially the virus passed from the horseshoe bats into the pangolin, and then from there into humans.

But now the virus is really in humans, and there's a diversity starting to build up. From the more than 3 million infections that we know about - and there's probably more than that - we've managed to a whole genome sequence more than 5,000 of those. And here we have the tree of 5,000 genomes. You can see that there's a little bit of clustering of those, and we've clustered one, the original, in green here, and cluster two is in the reds and orange and browns. But these genomes are not very far apart from each other yet. That scale is one single nucleotide polymorphism, so it's only one SNP apart. And most of these genomes are just a few SNPs apart. However, this genetic diversity is extremely important. We are expecting more diversity to generate over time. But even right now those diagnostic primers that we designed to that very first genome may not match all of these 5,000 genomes that we have or the diversity of genomes which are out there already. And over time, these might even build up potentially resistance or divergence in their response to therapy, and even to the vaccine.

In this picture, we have a Bayesian time data tree. And you can see back in January the number of viruses was relatively small. But by March, this tree has 1,200 genomes, and you can see that diversity increasing rapidly. And we coloured it by the different continents. And you can see there's some strains are more common, for example, in North America and some in Europe by the different colorings. And this might mean that the diagnostics, for example, might need to be a little bit different depending on where you are.

Luckily, there's a worldwide effort to keep track of all of this diversity. And we look forward to seeing how that generates in the future. And we can all keep an eye on it. I'm very confident that we won't see too much diversity as this virus seems very well adapted to humans, but we need to keep an eye on it.

See Also

Whole genome sequencing: what is whole genome sequencing used for now?

https://www.youtube.com/watch?v=34VvcS97QP0

LSHTM Viral podcast S1E4 Diagnosis... coronavirus?

https://anchor.fm/lshtm

Whole genome sequencing for infectious disease outbreaks

https://www.youtube.com/watch?v=C45-ZjE8gLg

Lab tests blog

http://www.labtestsblog.com/glossary-of-laboratory-diagnostic-terms/

SARS-COV-2 (severe acute respiratory syndrome coronavirus 2) sequences

https://www.ncbi.nlm.nih.gov/sars-cov-2/