Review of: "Can the definitions of SARS-Cov-2 and Covid-19 stand up to epistemological scrutiny?"

David W. Ussery[1]

1 University of Arkansas for Medical Sciences

Potential competing interests: No potential competing interests to declare.

Review of "Can the definitions of SARS-Cov-2 and Covid-19 stand up to epistemological scrutiny?"

As someone who has been working in the area of 'genomic epidemiology for more than a dozen years, I think this article raises some good questions. In contrast to the main conclusion of this article, I DO think that 'genomic epidemiology' of the Covid-19 pandemic has been successful in some areas, particularly in keeping track of new waves of infections, related to novel genomtypes. This is kind of at a 'high level' of the natural kinds referred to in the article. By this I mean in broad strokes the Alpha wave that peaked in March, 2021, followed by the Delta wave that peaked in October, 2021, and then the sudden rise in the Omicron variant, which is still the dominant strain circulating around the world. Mapping the relative proportion of these variants over time shows the clear selective pressure of the virus, and can be related to different clinical outcomes, with fatalities much more pronounced in the Delta wave, and less in the Omicron wave. (see for example, FEMS Microbiology Reviews, 46:1-19, 2022). The CDC has been keeping track of novel SARS-CoV-2 variants as they come and go over time (https://covid.cdc.gov/covid-data-tracker/#variant-summary), and I think that this is maybe useful 'epidemiology' in terms of relating specific sequence variations to temporal and geographic locations.

Having said that, I think that there are two major problems with the massive amount of data generated from the Covid-19 pandemic. The first is substantial quality problems with much of the data, and the second (related) is a 'big data' problem - how to deal with massive amounts of noisy, fuzzy data.

Quality problems with SARS-CoV-2 genomes. As I see it, there are two aspects here. First, there are problems with trying to keep track of the various new strains of Covid-19. In January of 2021, there was a wonderful short news article in Nature magazine, with the title "A bloody mess: confusion reigns over naming new COVID variants" (Nature, 589:339, 2021). I often cite this article in lectures about trying to do genomic epidemiology of the SARS-CoV-2 virus. There are multiple naming schemes, and the thousands of names make keeping track of 'new variants' difficult. There are way too many names to be useful for most applications.

The second problem is the massive amount of data available. The data being sampled is enormously diverse; every infection results in a fuzzy mixture of different genomic RNA sequences. For an individual patient, a 'consensus' genome sequence is produced from a fuzzy, diverse set of sequences. There are currently more than 15 million SARS-CoV-2 genome sequence available, but few of these are 100% complete, containing only 4 bases along the complete genome, from one end to the other. So there is a data quality issue, with many sequences containing long stretches of 'NNNN's,

where that region was not amplified by the primers. Further, even if 100% of the roughly hundred different segments were successfully amplified, there's still a bit of boundary on either end of the viral genome that does not get amplified. But even if we had 15 million complete viral genome sequences, these still would not fully capture the diversity, since each sequence represents one sequence out of a cloud of similar but different sequences.

Single molecule sequencing is a disruptive change within genomics. With the continual drop in cost of sequencing and increase in quality and read lengths (now more than 4 million bp), it is becoming possible to map the enormous diversity within biological systems, and that diversity is far greater than many have thought. On top of this, it is now possible to map base modifications (epigenetics) in both DNA and to directly sequence RNA (and its modified bases), allowing for additional 'epigenetic' sequence information to be obtained. If done properly, in the right context, this explosion of information can be linked to epidemiological information, and allow true genomic epidemiology to provide useful information for use in hospitals and Public Health. But I agree with the authors of this paper that just a genome sequence on its own does not provide any epidemiological information, without context.