Know Thine Enemy: Viral Genome Sequencing in Outbreaks

Katherine Colvin

Editorial Assistant

CONTAINING a viral outbreak with public health measures firstly requires identification of the causative virus, followed by more detailed understanding of viral features. Genomic sequencing provides exhaustive insight into viral features that may help predict outbreak behaviours, assist in diagnosis and tracking, and shape treatment and vaccination strategies. When coupled with epidemiologic study of outbreak data, viral genomic sequencing can be used to direct public health measures and increase the speed of understanding compared to epidemiology alone.

Community spread of cases can be used to guide mathematic models and contact tracing of viral outbreaks for public health response. However, epidemiologic data alone better suits responses to low-prevalence and less-widespread outbreaks. Where pathogens have a longer latency period or spread affects rural and remote communities, features of the virus itself must be considered in determining the response. Genotypic and phenotypic characteristics, identified using molecular biology tools, can clarify the type and strain of a virus responsible for an outbreak, and inform and improve case diagnosis, treatment options, and vaccine development, as well as improve tracing accuracy.[1]

VIRAL GENOMIC SEQUENCING

Genomic analysis has developed to the point of near-real-time whole genome sequencing, evident in the identification and publication of the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) whole genome sequence on 11th January 2020, 12 days after the first announcement of the original cluster of cases in Wuhan, China, with a diagnostic test made available 2 days later.[2] This is a significant improvement in the speed of sequencing compared to the 2002–2003 SARS-CoV (SARS) outbreak, where the first case was identified on the 16th November 2002 and the viral genome sequence published on the 30th May 2003.[3]

Different sequencing methods provide specific insights into viral outbreaks. Amplicon-based sequencing rapidly duplicates viral nucleotide fragments via reverse transcriptase polymerase chain reaction (RT-PCR) and is helpful for the initial detection and study of viruses, as it can amplify even small fragments of viral material and provide fast results. However, its reliance on

Sanger sequencing means that it is unlikely to identify low-frequency variants, and amplification depends on the availability of PCR primers which requires pre-existing knowledge of the viral sequence, introduces bias, and limits the ability to undertake metagenomic analysis. It can also have limited utility in providing full genome sequences, as sample degradation prevents full-length amplicon production and high sequence variation in viruses means it is difficult to design primers that will produce full-length genomes.[1]

High-throughput sequencing, also referred to as second-generation sequencing, is more reliable for determining genome sequences of viral fragments or partially degraded samples. It can also detect low-frequency, within-host variants, can be used to sequence new or unknown pathogens, and can be used in metagenomic analysis. High-throughput sequencing combines selective RNase H-based digestion of contaminating RNA (mainly host ribosomal RNA) with sequence-independent primer amplification and was utilised during the 2014–2015 Ebola epidemic in western Africa.[1]

Features of a virus identified by genomic sequencing can determine the species responsible for an infection; clarify diversity within outbreaks, both genotypic and phenotypic; and provide understanding of the evolutionary history of the virus, which may aid in treatment and vaccine development. Construction of a phylogenetic tree, that maps sequenced samples against one another by comparing nucleotide substitutions, is used to track the evolutionary history of a new virus or outbreak. A phylogenetic tree was used in the 1997 avian flu outbreak in Hong Kong to identify the overlap between the human influenza A H5N1 virus in terrestrial poultry and a similar virus in quail.[1] Analysis of the phylogenetic tree mapping for the Ebola outbreak revealed high variation between outbreaks but low variation within outbreaks, which may suggest an animal reservoir and single zoonotic transmission for each outbreak of the virus.[1] The utility of a phylogenetic tree is affected by many factors, including accurate timestamping of collected samples and features of the virus itself, such as viral recombination during replication.[1]

EPIDEMIOLOGIC MEASURES AND MATHEMATICAL MODELLING

Epidemiology is the study of patient and community data to analyse population spread and behaviour of pathogens within populations. In the case of viral outbreaks, epidemiologic study provides insight into factors that affect distribution of infections within a community, including risk factors for contracting or transmitting a virus, and severity of infection; the prognosis of the viral illness; and the success of prevention and treatment strategies. Historically, epidemiology was dependent upon case-based data collection and analysis with deductive and inductive reasoning. However, the field has grown since the first application of mathematical modelling to population health by mathematician Daniel Bernoulli in 1760 tracking the effectiveness of an early smallpox vaccine. In modern epidemiology, use of statistical assessment and mathematical modelling has greatly expanded the reliability and applicability of epidemiological data.[3]

R_0 is the basic reproduction number of a virus, revealing the speed at which a virus spreads through a population by describing how many new cases can arise from one infected person. It is a theoretical parameter, in that it cannot be directly measured but instead is estimated based on epidemiological data including infection and recovery rates, viral transmissibility, and population size.[4] It is a valuable epidemiologic parameter but is imprecise, as factors contributing to R_0 can vary from person to person and the R_0 of a virus will change during the course of an outbreak. Pairing R_0 with genomic insights about viral evolution is part of a new field called phylodynamics, an emerging strategy for studying the activity and spread of viral outbreaks and helping to determine appropriate public health strategies.[1]

Contact tracing and mathematic analysis of case spread can provide insight into viral transmission and track success of public health measures. Contact tracing practices can clarify mode of transmission for new or unknown viruses, including whether a viral infection is vector-borne

"In the case of viral outbreaks, epidemiologic study provides insight into factors that affect distribution of infections within a community"

or capable of human-to-human transmission. Mathematical modelling to determine cost-effective and high-impact strategies to reduce viral prevalence during an outbreak are hindered by the fact that populations are heterogenous; population density, travel behaviours, and individual susceptibility or risk factors for illness vary throughout the population. Most epidemiologic mathematical modelling is undertaken as compartmental modelling to assess population subgroups, generally divided by risk factors such as age or by dividing populations into those at-risk, those infected, and those recovered from infection. This poses challenges for then attempting to scale insights into making public health recommendations across the full population.

The incubation period of a virus describes the time from infection to displaying symptoms, while the latent period of a virus describes the time from infection to becoming infectious to others. These periods vary greatly between viruses, but both impact the infection rate during an outbreak. They also have a significant impact on public health measures, as interventions such as isolating infected individuals may be challenging to implement in cases with long incubation periods. The case fatality rate is another epidemiologic measure tracked during an outbreak. However, determining the percentage of infections that are fatal is difficult in the presence of heterogenous viral phenotypes, where some infected cases are asymptomatic or mild community cases.[3]

VIRAL MUTATION RATES AND VIRULENCE FACTORS

Other strategies for pairing epidemiologic and genomic data can provide further comprehension of viral outbreaks. Determining mutation rate of a virus and within-host substitution rate can further clarify transmissibility factors, direct treatment strategies, and determine viability of vaccine strategies. This makes use of both genetic sequencing and longitudinal case sampling, which requires many cases over longer-term periods and is currently most utilised in studying chronic viral infections such as HIV, although applications in acute outbreaks are developing.[1]

The mutation rate of a virus reflects the evolutionary changes occurring during and between outbreaks and is dependent on properties of the virus including whether it contains RNA or DNA, the fidelity of its polymerase, and the speed of replication of its own genome. Usually, RNA viruses mutate faster than DNA viruses. The nucleotide substitution rate is a measure of nucleotide mutation accumulation over a virus' lineage, and is determined by mutation rate, effective viral population size, and natural selection. The nucleotide substitution rate is most useful during an outbreak because it can develop understanding of the selection pressures and can be calculated from the phylogenetic tree and sampling dates. During outbreaks, the calculated substitution rate may be falsely elevated; the natural course of viral evolution includes deleterious substitutions that will be selectively removed from viral populations over time but are still present at the time of outbreak analysis. Overall, most models for analysing genetic changes in viruses, particularly to identify selection factors for host infection and viral survival, were developed to compare across viral species rather than for dynamic analysis of a single species during an outbreak, so this branch of genomic analysis is still limited in its application to outbreaks.[5]

Virulence factors specific to a virus are also important features for guiding the public health response to an outbreak. Virulence factors are features of the virus that determine the harm that it can do to the host, usually meant in terms of mortality. High virulence is often associated with high viral load, as the interaction of viral and host factors optimises virus survival and replication. However, in cases where viral features trigger an inappropriate immune response in the host, virulence may be high in the absence of high viral load. Vector-borne diseases often have higher virulence than viruses transmitted host-to-host. This is because, evolutionarily, host-to-host transmission and viral survival is dependent on host mobility and behaviour, although this virulence is increased in the case of viruses capable of prolonged environmental survival, i.e., survival outside of a host or vector carrier.[6,7] Virulence factors detected in laboratory assessment may be further understood with genomic analysis, such as cell surface receptor virion attachment, replication rate at different temperatures and in different inflammatory conditions, and virus tissue specificity or tropism.[8]
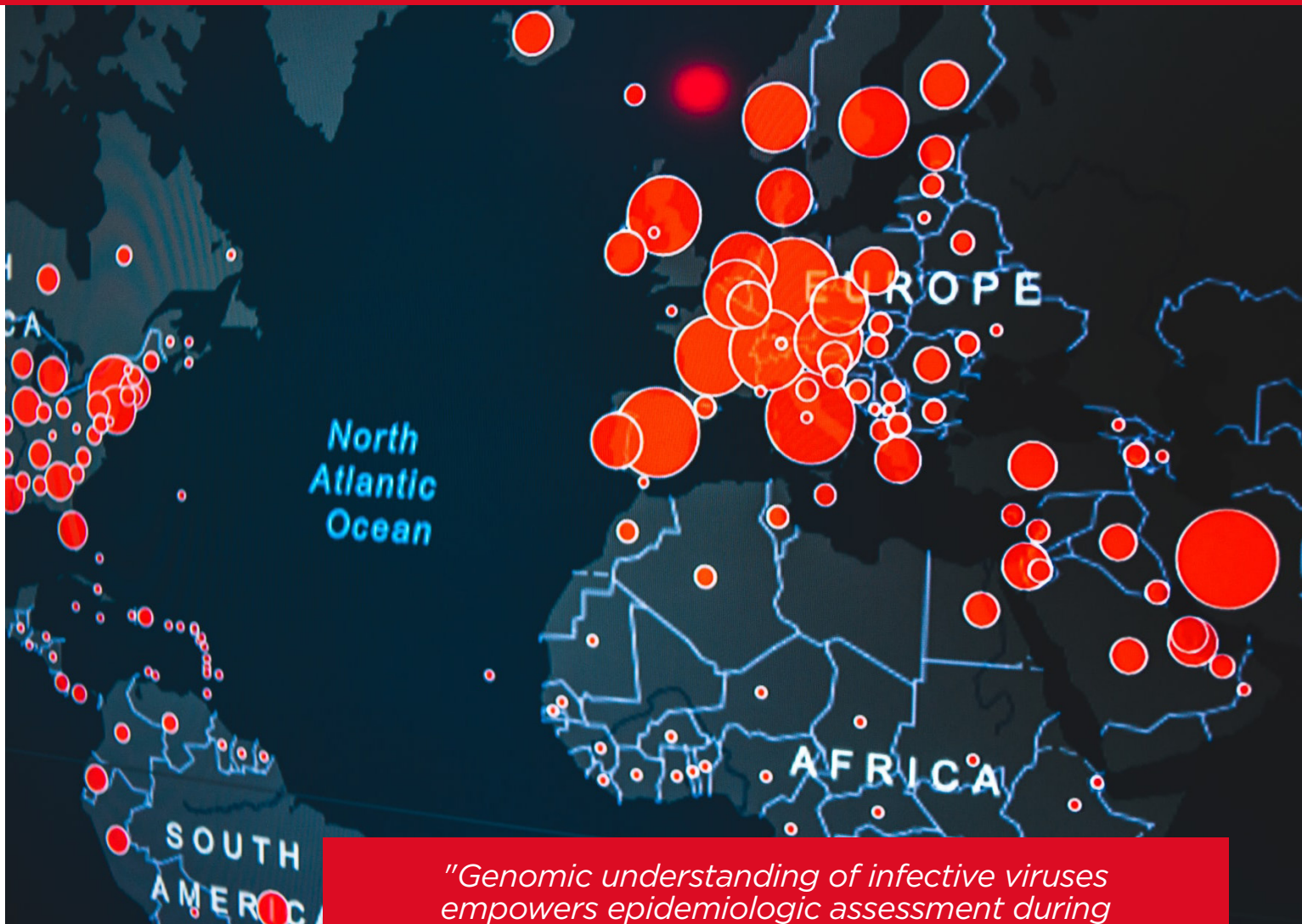
"Genomic understanding of infective viruses empowers epidemiologic assessment during outbreaks and provides specific insights to aid diagnosis, treatment, and vaccine strategies"

CONCLUSION

Genomic understanding of infective viruses empowers epidemiologic assessment during outbreaks and provides specific insights to aid diagnosis, treatment, and vaccine strategies. The improvement in genomic sequencing techniques has meant that public health action advances much more rapidly during an outbreak. The complexity of virus–host interactions and the heterogeneity of human populations make public health interventions difficult, but viral genomic sequencing provides a valuable contribution to planning whole-population and individual-level responses.

References

1. Wohl et al. Genomic analysis of viral outbreaks. Annu Rev Virol. 2016;3:173-195.

2. World Health Organization (WHO). Diagnostic detection of Wuhan coronavirus 2019 by real-time RTPCR. 2020. Available at: https://www.who.int/docs/default-source/coronaviruse/wuhan-virus-assay-v1991527e5122341d99287a1b17c111902.pdf. Last accessed: 1 May 2020.

3. White P, Enright M, "Chapter 5 - Mathematical models in infectious disease epidemiology," Cohen J et al. (eds.), Infectious Diseases: Volume 1 (2010) 3rd edition, Philadelphia: Mosby Elsevier, pp.70-5.

4. Ridenhour B et al. Unraveling R0: Considerations for public health applications. Am J Public Health. 2014;104(2):e32–41.

5. Morse SM, Khan AS. "CHAPTER 8 - Epidemiologic Investigation for Public Health, Biodefense, and Forensic Microbiology," Breeze RG et al (eds.), Microbial Forensics (2005), San Diego: Academic Press, pp.157-71.

6. Longdon B et al. The causes and consequences of changes in virulence following pathogen host shifts. PLoS Pathog. 2015;11(3): e1004728.

7. Brown NF et al. Crossing the line: selection and evolution of virulence traits. PLoS Pathog. 2006. Available at: https://doi.org/10.1371/journal.ppat.0020042. Last accessed: 1 May 2020.

8. Baron S et al. "Chapter 45 Viral Pathogenesis," Baron S (ed.), Medical Microbiology (1996) 4th edition, Galveston: University of Texas Medical Branch at Galveston. Available at: https://www.ncbi.nlm.nih.gov/books/NBK8149/. Last accessed: 1 May 2020.