A pandemic in the age of next-generation sequencing

Angela H. Beckett, **Kate F. Cook** and **Samuel C. Robson** (University of Portsmouth, United Kingdom)

Since December 2019, the world has found itself rocked by the emergence of a highly contagious novel coronavirus disease, COVID-19, caused by the virus SARS-CoV-2. The global scientific community has rapidly come together to understand the virus and identify potential treatments and vaccine strategies to minimise the impact on public health. Key to this has been the use of cutting-edge technological advances in DNA and RNA sequencing, allowing identification of changes in the viral genome sequence as the infection spreads. This approach has allowed a widespread 'genomic epidemiology' approach to infection control, whereby viral transmission (e.g. in healthcare settings) can be detected not only by epidemiological assessment, but also by identifying similarities between viral sub-types among individuals. The UK has been at the forefront of this response, with researchers collaborating with public health agencies and NHS Trusts across the UK to form the COVID-19 Genomics UK (COG-UK) Consortium. Genomic surveillance at this scale has provided critical insight into the virulence and transmission of the virus, enabling near real-time monitoring of variants of concern and informing infection control measures on local, national and global scales. In the future, next-generation sequencing technologies, such as nanopore sequencing, are likely to become ubiquitous in diagnostic and healthcare settings, marking the transition to a new era of molecular medicine.

Whole-genome sequencing to track infectious disease outbreaks

Molecular biology is commonly employed for the characterization of infectious disease agents, with the polymerase chain reaction (PCR) having been used regularly in diagnostic laboratories for decades. During the COVID-19 pandemic, reverse transcription-PCR (RT-PCR) assays have remained the gold standard approach for detecting SARS-CoV-2 from nasopharyngeal swab samples. This process involves the amplification of short, conserved regions of the SARS-CoV-2 genome, providing a result for the presence or absence of SARS-CoV-2. RT-PCR has enabled rapid diagnosis of infection to inform patient isolation procedures, but it provides a limited picture of pathogen biology compared to that offered by whole-genome sequencing (WGS).

Sequencing, the process of determining the sequence of nucleotide building blocks (adenine, cytosine, guanine and thymine) that constitute a genome, can provide a detailed insight into the biology and evolution of populations, given large enough numbers of samples for comparison. Since the first full genome (a 5 kb bacteriophage – phi X 174) was generated 45 years ago, sequencing technology has rapidly evolved, allowing for large-scale sequencing of even the human genome at nearly a million times that length. Alternatively, this can allow rapid and large-scale sequencing of smaller genomes, such as those of pathogenic microorganisms. As pathogens spread from person to person, the comparatively rapid rate of mutation can lead to changes in the genome sequence that may act as a marker for direct transmissions. Thus by comparing genome similarity of these pathogens between hosts, and linking with epidemiological data, direct transmission events can begin to be identified. This so-called genomic epidemiology can help us to understand how these infectious diseases are transmitted. In particular, this approach captures the full genome, as opposed to RT-PCR which targets small genomic regions, providing a more complete picture of the landscape of mutations.

Phylogenetic trees (which represent evolutionary relationships between sequences) of related viral samples allow us to visualize transmission pathways in various environments (Figure 1). This can be from individual healthcare settings to small communities, and even at national and global scales, provided enough samples are sequenced. In 2013, researchers at the University of Cambridge sequenced methicillin-resistant *Staphylococcus aureus* (MRSA) isolates from 26 patients during a hospital outbreak of the disease and discovered that a staff member had likely been carrying the pathogen, enabling it to persist in the hospital for a long period of time. As the technology has improved over time, modern next-generation sequencing (NGS) technologies have been used at ever larger scales. For example, during the 2014–2016 Ebola virus epidemic in West Africa, genome data from ~1500 viral samples were used to demonstrate that there had been a single introduction of the virus into the human population, with human-to-human transmission

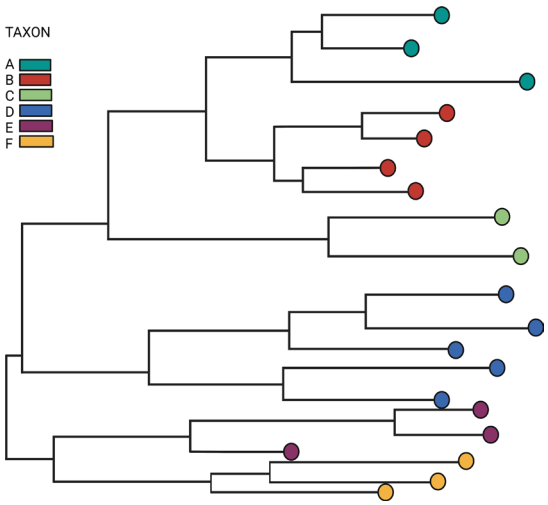

Figure 1. A mock phylogenetic tree displaying six different taxa of genetically related individuals. Figure created using BioRender.com.

occurring afterwards. As well as elucidating the origin of the outbreak, these analyses enabled the identification of cases whereby the infection had been maintained long term in individual hosts, causing outbreaks later down the line.

The use of sequencing technology during the Ebola and Zika epidemics, among others, built a foundation of knowledge that was essential to the large-scale uptake and usage of NGS throughout the COVID-19 pandemic. SARS-CoV-2 genome data has allowed scientists and public health agencies (PHAs) to monitor the current pandemic and provide rapid feedback to healthcare providers to help understand and prevent outbreaks.

The revolution in sequencing technologies

First-generation sequencing, which was developed in the late 1970 s by biochemist Fred Sanger, was based on knowledge of the chemical composition of DNA and DNA replication gained from analytical chemistry. While the method is highly accurate and still widely used today for targeted clinical sequencing, it is not efficient for high-throughput WGS.

Following the success of the 'second-generation' shotgun approach to sequencing, popularized by Illumina, whereby DNA is first broken down into short fragments prior to sequencing, recent 'third-generation' advances in NGS such as those from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) allow sequencing of full-length DNA molecules. Central to nanopore sequencing from ONT are 1-nm-wide perforations known as 'nanopores', which lie within an electro-resistant membrane on a flow cell. Double-stranded DNA (dsDNA) is unwound to single-stranded DNA (ssDNA) and facilitated through the nanopore by a protein helicase at a translocation speed which allows the sequence to be read one nucleotide at a time (Figure 2). Owing to the unique chemical structure of the four nucleotides and the effect that the electrical charges have on the ionic flow through the nanopore, the electrical current detected by the membrane is modified each time a nucleotide crosses it, whereby the signal can be instantly decoded by the computer in a process known as base-calling. Nanopore sequencing therefore offers the benefit of real-time base-calling, along with improved portability through systems such as the pocket-sized MinION device, providing benefits to pathogen detection.

Genomic surveillance of SARS-CoV-2

COVID-19 is the first pandemic in history where the pathogen has been monitored in near real time using high-throughput WGS technologies. SARS-CoV-2 has a single-stranded RNA (ssRNA) genome roughly 30 kb in size contained within a lipid envelope (Figure 3). To date more than 3 mn genome sequences have been uploaded to the Global Initiative on Sharing All Influenza Data (GISAID) website. This resource allows publicly available viral genome data to be visualized and diversity of the virus to be observed over time (Figure 4). The SARS-CoV-2 genome is now one of the most highly sequenced genomes of any organism on the planet, as scientists investigate how the virus is changing over time.

During the pandemic, PCR-positive samples from across the UK were submitted for sequencing to identify variants and routes of transmission. This has been particularly important during times when case numbers were too high for epidemiological data to provide clear evidence of transmission routes, or in situations involving 'super-spreaders'. Transmission information derived from genomic data has been used to inform infection control measures, such as introducing greater levels of testing and PPE in healthcare settings or limiting the number of wards that individual healthcare workers are working on. Similarly, advice incorporating these data has been used to inform decisions by policy makers on non-pharmaceutical interventions such as lockdowns, physical distancing and the wearing of face coverings.

At the start of the pandemic, scientists from academic institutes across the UK with experience and expertise in NGS sequencing collaborated with PHAs and the NHS to form the COVID-19 Genomics UK (COG-UK) Consortium. COG-UK was developed as a distributed network of sequencing hubs, providing expertise and infrastructure to local NHS Trusts as well as rapid reporting to aid in healthcare infection management. As research labs closed as the country

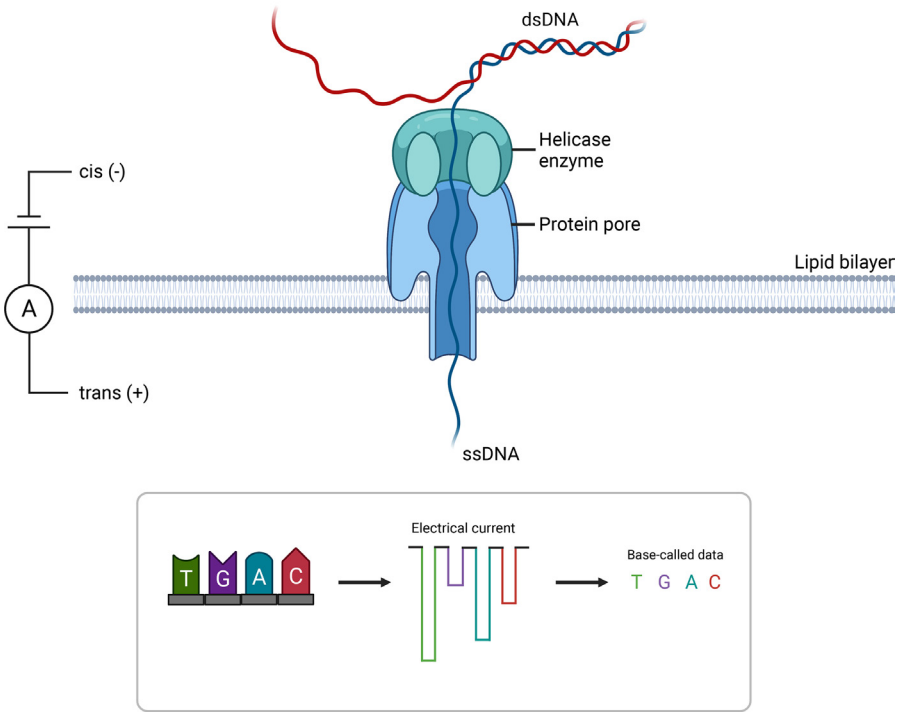

Figure 2. A schematic diagram of the mechanism of Oxford Nanopore Technologies (ONT) sequencing. Double-stranded DNA (dsDNA) is unwound by a helicase enzyme, and a single strand of DNA (ssDNA) is pulled through the nanopore. Each nucleotide base causes a characteristic change in the ionic flow through the nanopore, which can be computationally decoded to determine the underlying DNA sequence. Figure created using BioRender.com.

entered its first lockdown, scientists volunteered their time and expertise to tackle the pandemic, with close working relationships formed between researchers who may never have met in person. The logistics of distributing samples between clinical testing sites and sequencing sites was a monumental feat, with the goal to ensure that the network had a standardized system despite the distributed nature. In particular, the Wellcome Sanger Institute worked closely with NHS Test and Trace to develop an infrastructure pipeline for processing and sequencing the huge number of samples passing through the so-called Pillar two testing (symptomatic testing from the community).

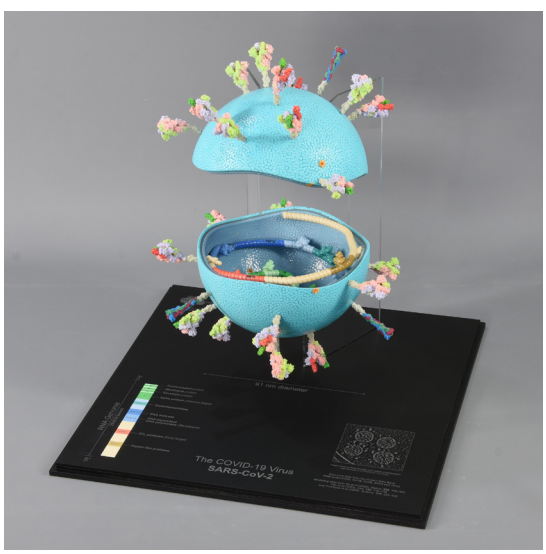

Figure 3. A 3D printed model of SARS-CoV-2, displaying the spike proteins attached to the outside of the viral envelope and the genomic RNA inside. Model created by Darren Gowers (molecmodels.co.uk).

The additional challenges of SARS-CoV-2 testing and sequencing included the unpredictability of the number of samples and cases to be received, the subsequent difficulty in maintaining reagent stocks, planning staffing around the ever-present possibility of staff having to self-isolate and the development of workflows in an ever-changing environment. A collegiate atmosphere within the consortium ensured that the effort was a success, with research groups collaborating closely on technical support, sharing expertise as well as sending supplies to neighbouring sites and offering troubleshooting suggestions as required.

The Sequencing and Bioinformatics Group at the University of Portsmouth was one such group and was a good example of how general biological expertise

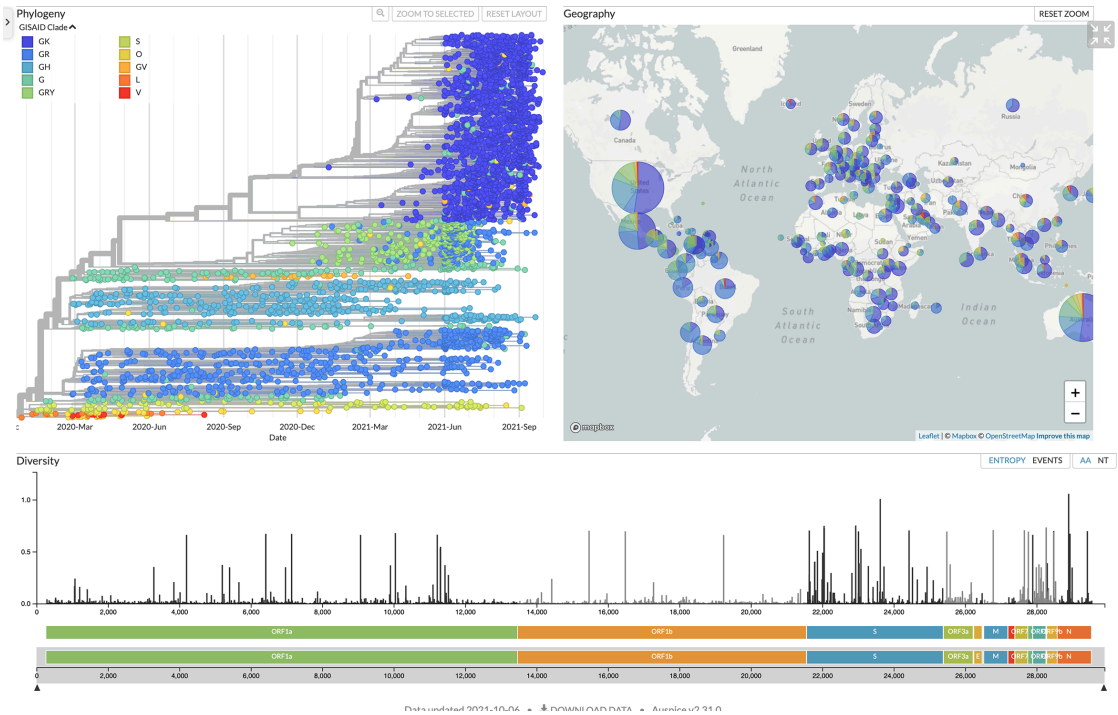

Figure 4. A screenshot from the Global Initiative on Sharing All Influenza Data (GISAID) website, highlighting the diversity of SARS-CoV-2 lineages globally over time. Top left: The current phylogeny of global cases, showing evolution of the virus through time. Top right: Geographical distribution of SARS-CoV-2 throughout the world. Bottom: Diversity of specific mutations and their location within the SARS-CoV-2 genome.

was repurposed to aid in the global crisis. Like many groups involved, we had no formal background in virology, but we did have expertise in molecular biology and bioinformatics and were able to apply our skills and knowledge to monitoring the pandemic. Before the closure of the University laboratories in the first lockdown, we transported our sequencing equipment to a lab within our local NHS Trust following the string research collaboration between the Trust and the University. We initially set out to perform sequencing on 500 SARS-CoV-2 samples from the hospital to aid with their infection control procedures within the first wave, but our capacity has since grown substantially, and we have sequenced over 15,000 samples from multiple NHS sites across the South coast. This is a picture seen around the UK and worldwide.

In addition to transmission information, one of the main advantages of such large-scale sequencing is that the evolution of SARS-CoV-2 has been tracked and new variants of the virus have been identified and characterized quickly. The UK's genomic surveillance system has developed such that information about new variants can be made available to policy-making authorities rapidly following their appearance, meaning that action can be quickly taken to respond to the evolving virus.

What is a variant?

SARS-CoV-2 variants have become household names over the course of the pandemic, present in news headlines and discussed among the public and scientists alike. While there is a widespread understanding that there are variants of concern (VOCs) that may be more transmissible, increase disease severity or evade treatment or vaccination, many people may not know exactly what constitutes a variant. It is worth noting that there is no universal method for naming variants and that scientists themselves have used a range of naming conventions throughout the pandemic.

A variant is a virus whose genome sequence differs from that of a reference virus as a result of mutations over time. The reference genome often used for SARS-CoV-2 is that from one of the earliest COVID-19 cases detected in Wuhan, China, in December 2019, known as Wuhan-Hu-1. New variants arise owing to changes in the nucleotide sequence that occur as the virus replicates. The point at which a distinct set of mutations is considered a new variant in the various naming conventions may differ. The majority of SARS-CoV-2 mutations do not result in any phenotypic change; however, occasionally they generate an advantageous phenotype such as increased replicative fitness or the ability to evade neutralising antibodies. In the UK,

variants with the potential for such phenotypic changes are classified as Variants Under Investigation (VUIs) by PHAs and are closely monitored. VUIs may be reclassified to VOCs as additional studies are performed based on their likelihood to impact on disease severity, therapeutic efficacy and replicative fitness.

Naming conventions for SARS-CoV-2 have posed additional challenges as there have been multiple classification systems in use throughout the course of the pandemic. Publicly available SARS-CoV-2 databases and tools such as GISAID, Pangolin and Nextstrain are all used for variant tracking and initially used slightly different nomenclature within their classification systems. This lack of consistency, along with the fact that some variant names were confusing, made it challenging to track VOCs and VUIs using public data during the first year of the pandemic. In particular, it was often found that VOCs had similar designations, for instance B.1.177 (prevalent in the UK through summer 2020) and B.1.1.7 (prevalent in the UK, and later globally, in winter 2020). In June this year, the WHO introduced a global naming scheme in which letters of the Greek alphabet were assigned to different VOCs and VUIs, with the aim of providing a standardised system.

The future of NGS

Moving forward, the vast knowledge and expertise acquired during this pandemic as well as the incorporation of sequencing technologies into routine testing within clinical settings is likely to put sequencing at the forefront of innovative healthcare and research. Utilization of this technique will become commonplace to monitor the spread of hospital-acquired infections, antibiotic-resistant bacteria, as well as future epidemics. Sequencing has the potential to quickly diagnose rare genetic disorders that are difficult to diagnose otherwise. It also offers the ability to refine cancer diagnoses by characterising the mutations present and thus guiding therapeutic approaches. It is possible that in the future WGS of human genomes may result in a healthcare reform whereby personalized medicine approaches are used instead of standardized treatment procedures.

NGS technologies have advanced rapidly over the past 50 years and the current pandemic has highlighted what the most advanced of these technologies can offer in terms of monitoring and managing pathogen transmission when used in large scale. This pandemic may mark a shift in the application of NGS from being used predominantly in research towards widespread, rapid use in healthcare, saving time and resources in diagnostics and treatment measures. As technology continues to develop, and the capacity of such machines increases, the possibilities remain endless for furthering our understanding of the processes underpinning life on this planet, and ultimately mitigating the impacts and burden of disease among the human population.

Acknowledgements

The authors declare no competing interests. Angela H. Beckett and Kate F. Cook have contributed equally to this article. Angela H. Beckett and Kate F. Cook have contributed equally to this study and should be considered as co-first authors. The authors gratefully acknowledge all members of the COVID-19 Genomics UK (COG-UK) Consortium for their contributions to generating the data described in this article. COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, and the study protocol was approved by the Public Health England Research Ethics Governance Group (reference number R&D NR0195). SR and AB are additionally funded from Research England's Expanding Excellence in England (E3) Fund.▪

Further reading

- NHS England » NHS Genomic Medicine Service. https://www.england.nhs.uk/genomics/nhs-genomic-med-service/ [Accessed 23 September 2021]
- What do virologists mean by 'mutation', 'variant' and 'strain'? COVID-19 Genomics UK Consortium. https://www.cogconsortium.uk/what-do-virologists-mean-by-mutation-variant-and-strain/ [Accessed 23 September 2021]
- Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ [Accessed 23 September 2021]
- GISAID - Initiative. https://www.gisaid.org/ [Accessed 7 October 2021]
- Severe acute respiratory syndrome coronavirus two isolate Wuhan-Hu-1, co - Nucleotide NCBI. https://www.ncbi.nlm.nih.gov/nuccore/1798174254. [Accessed 23 September 2021]
- The COVID-19 Genomics UK (COG-UK) (2020) An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100. DOI: 10.1016/S2666-5247(20)30054-9

(Continued)

Further reading (*Continued*)

- Alm, E., Broberg, E.K., Connor, T. et al. (2020) Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill*. **25**, 2001410. DOI: 10.2807/1560-7917.ES.2020.25.32.2001410
- Graham, M.S., Sudre, C.H., May, A. et al. (2021) Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Health* **6**, e335–e345. DOI: 10.1016/S2468-2667(21)00055-4
- du Plessis, L., McCrone, J.T., Zarebski, A.E. et al. (2021) Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712. DOI: 10.1126/science.abf2946
- Aggarwal, D., Peacock, S., Hamilton, W. et al. (2021) The role of viral genomics in understanding COVID-19 outbreaks in long term care facilities. *Lancet Microbe*. DOI: 10.1016/S2666-5247(21)00208-1
- Wenlock, R.D., Tausan, M., Mann, R. et al. (2021) Nosocomial or not? A combined epidemiological and genomic investigation to understand hospital-acquired COVID-19 infection on an elderly care ward. *Infect. Prev. Pract*. **3**, 100165. DOI: 10.1016/j.infpip.2021.100165
- Konings, F., Perkins, M.D., Kuhn, J.H. et al. (2021) SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nat. Microbiol*. **6**, 821–823. DOI: 10.1038/s41564-021-00932-w

Angela H. Beckett is a Specialist Research Technician at the University of Portsmouth, with an MSc in medical microbiology and a background in infectious disease and high containment. She helped to establish and maintain the South Coast COVID-19 Genomics UK Consortium (COG-UK) sequencing hub, for SARS-CoV-2 sequencing, and is involved in a range of projects using next-generation sequencing technologies as part of the Sequencing and Bioinformatics Group and the Centre for Enzyme Innovation (CEI).

Kate F. Cook is a research associate at the University of Portsmouth who has been heavily involved in the sequencing of SARS-CoV-2 within the South Coast COG-UK sequencing hub. She has a background in pathogen genomics and is currently working on the analysis of SARS-CoV-2 whole-genome sequence data to understand transmission dynamics within a hospital setting and genomic associations with patient outcomes.

Dr Samuel C. Robson is a Reader in Genomics and Bioinformatics at the University of Portsmouth and the Bioinformatics Lead at the Centre for Enzyme Innovation. His research focuses on the use of bioinformatics and high-throughput sequencing technologies in a wide range of fields, including environmental microbiology, viral genomics, clinical research and palaeogenomics. Recently, he has led the South Coast COG-UK sequencing hub, utilizing rapid sequencing of tens of thousands of clinical COVID-19 samples to help understand transmission of the virus in healthcare settings and the community across the UK. Email: samuel.robson@port.ac.uk