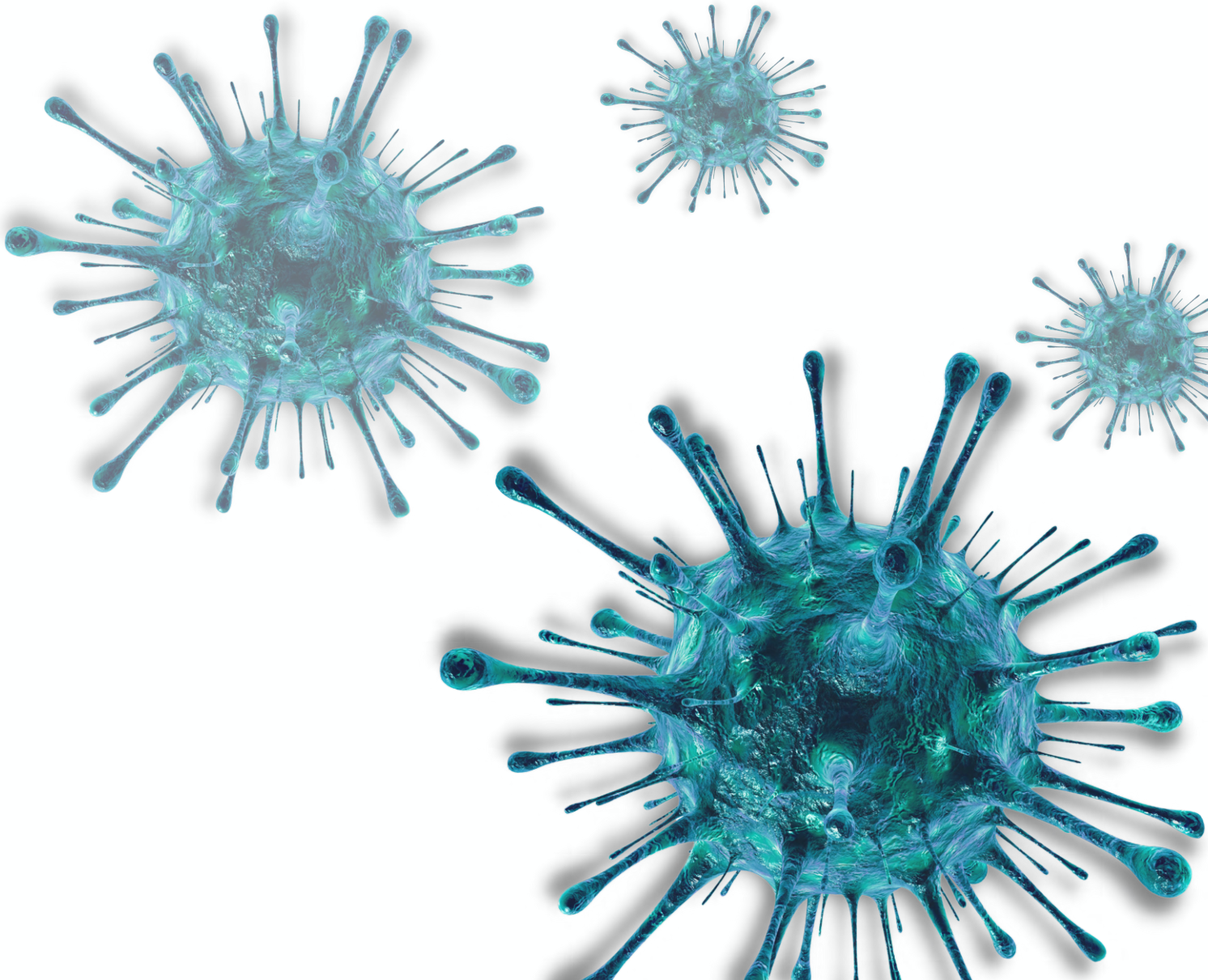
FIND

Because diagnosis matters

Next generation sequencing for SARS-CoV-2



Next generation sequencing for SARS-CoV-2/COVID-19

Acknowledgements

The PHG Foundation and FIND are grateful for the insight provided by the individuals consulted during the course of this work. Full consultee acknowledgements are listed in Appendix 1

Lead writers

Chantal Babb de Villiers (PHG Foundation)

Laura Blackburn (PHG Foundation)

Sarah Cook (PHG Foundation)

Joanna Janus (PHG Foundation)

Emma Johnson (PHG Foundation)

Mark Kroese (PHG Foundation)

Concept development and lead review

Swapna Uplekar (FIND)

Reviewers

Devy Emperador (FIND)

Jilian Sacks (FIND)

Marva Seifert (FIND)

Anita Suresh (FIND)

Report production

PHG Foundation

July 2020; Update March 2021

This report is the result of PHG Foundation's independent research and analysis and is not linked to a third party in any way. PHG Foundation has provided occasional analytical services to Oxford Nanopore Technologies (ONT) as part of a consultancy agreement.

Contents

Photo by Viki Mohamad on Unsplash

1

Introduction

Introduction

In the past few years, pathogen genome sequencing has emerged as a tool to support understanding of the molecular epidemiology of disease outbreaks, complementing and in some cases supplanting more established techniques. Recent advances in sequencing technologies have shown their applicability for research use in outbreak situations, for example, sequencing has been deployed in recent years to support understanding of the epidemiology of Ebola virus disease [1], Zika virus [2] and Lassa fever [3]. Today, sequencing of the novel coronavirus SARS-CoV-2 is supporting ongoing vaccine development efforts, and increasing knowledge of the origins and progression of the ongoing COVID-19 pandemic. However, lack of standardisation in testing and analysis protocols, the choice of reagents and the complexity of current methods can significantly complicate the use of next generation sequencing (NGS) technologies as routine surveillance tools.

FIND (Foundation for Innovative New Diagnostics) aims to provide a comprehensive set of standardised sequencing protocols and workflows in order to support the implementation of NGS technologies and tools for outbreak pathogens in low resource settings. To facilitate this effort, investigation of the scientific landscape of pathogen sequencing as applied to outbreaks is necessary, to identify the most promising applications and to identify any existing workflows and protocols.

This report provides an overview of how sequencing technologies are being used to understand and mitigate the ongoing COVID-19 pandemic. The first version of this report was completed in June 2020; an information update was carried out in February and March 2021 with an emphasis on global genomic surveillance and. An addendum focussing on recently identified variants of concern was also developed [4].

The following issues are specifically excluded from the project scope:

- COVID-19 disease details, treatment, detailed disease statistics
- Gap analysis for what needs to be addressed to take this technology to a validated, standardised tool for routine use in country surveillance programmes as well as outbreak response and management

1.1 Methods

This report summarises current knowledge on technology developments, protocols and global best practice. It is based on desk-based research and analysis informed by official publications, grey literature, peer-reviewed and pre-print literature, to summarise current knowledge on technology developments, protocols and global best practice.

Where appropriate, in-depth interviews (via telephone or video conference) have been conducted with experts and other relevant stakeholders to better understand the enablers and barriers to implementation and adoption of the techniques in appropriate settings. These experts are acknowledged in Appendix 1.

This first chapter provides a synthesis of the findings of the report and highlights key areas of research. Subsequent chapters in this report provide an overview of:

Chapter 2 – Applications of sequencing in global surveillance, including details of national and international sequencing efforts and consortia

Chapter 3 – Overview of SARS-CoV-2 sequencing for diagnosis of COVID-19, and current non-

sequencing nucleic acid tests

Chapter 4 – The broader sequencing research landscape for SARS-CoV-2

Chapter 5 – Overview of current sequencing technologies, assays, and associated tools for SARS-CoV-2

Chapter 6 – Sequencing implementation including workflows, summary of technologies and techniques in use, examples of available protocols

Chapter 7 – Conclusions

1.2 SARS-CoV-2

1.2.1 Emergence of SARS-CoV-2

There are currently seven known human coronaviruses, three of which have caused widespread concern to global public health. Severe acute respiratory syndrome coronavirus (SARS-CoV) emerged in 2002 in the Guangdong province of China and was responsible for an outbreak in 2003 affecting 26 countries, killing roughly 10% of the approximately 8000 people who were infected [5, 6]. Middle East respiratory syndrome coronavirus (MERS-CoV) emerged in 2012, and infected around 2,500 people, mostly in regions of the Middle East, with a fatality rate of 30% [5]. Most recently the SARS-CoV-2 coronavirus, so named because of its resemblance to SARS-CoV, emerged in Wuhan, China in December 2019; it is responsible for the current pandemic of 'coronavirus disease 2019' (COVID-19) [5, 7].

While the terminology 2019-nCoV is also used to describe the virus, the notation SARS-CoV-2 is used throughout this report.

1.2.2 Infection and mortality rates

SARS-CoV-2 causes respiratory disease in humans, and is transmitted in droplets or aerosols released via talking, coughing or sneezing that are then inhaled, or via fomites (contaminated surfaces). The virus can survive on surfaces for up to 72 hours, although the extent to which this contributes to transmission is still uncertain [8]. SARS-CoV-2 appears to have a lower mortality rate than the SARS and MERS coronaviruses, and a mean basic reproductive number (R0) of between 2-3 [7, 9, 10]. This compares to an R0 of around 3.0 for SARS-CoV and an R0 for MERS-CoV which ranged from 0.45 in Saudi Arabia to 8.1 in South Korea [7].

Despite having a similar R0 to SARS-CoV, SARS-CoV-2 has infected far more people, potentially due to it being spread by individuals who are asymptomatic or have very mild symptoms. As of 22nd March 2021 there have been over 122 million cases worldwide confirmed by the WHO, affecting more than 200 countries with over 2.7 million confirmed deaths [11]. This currently suggests a mortality rate of around 2.2%. The actual figure is uncertain due to different reporting strategies in different countries and alterations in testing strategies as the pandemic has progressed – for example, earlier in the pandemic many countries only tested cases presenting with severe illness.

The time period between the development of symptoms and death can also cause a lag in mortality rate reporting. Mortality from other causes is also expected to increase as an indirect impact of the virus, due to the disruption caused to economies and healthcare systems by measures to manage the pandemic.

1.3 SARS-CoV-2 genome sequencing

Coronaviruses are a diverse group of enveloped viruses belonging to the subfamily Coronavirinae of the family Coronaviridae, all of which have large genomes ranging from 26-32 kilobases (kb) in length, consisting of positive sense, single stranded RNA [5, 12]. SARS-CoV-2 has a genome that is approximately 30kb in length and reported to contain 14 open reading frames (ORFs) encoding 27 proteins [13].

Sequencing technologies have been widely used since the beginning of the pandemic to facilitate understanding of virus biology and epidemiology and to inform research efforts, with over 830,000 sequences shared via the Global Initiative on Sharing All Influenza Data (GISAID) as of 22 March 2021 [14].

Next generation sequencing has considerable utility as a tool in infectious diseases management, since it provides the highest resolution information available about pathogen genomes, allowing full nucleotide sequences to be read and discovery of novel genomic variation at scale. During the current pandemic, sequencing is being used in the following areas (further information in chapters 2-5 and in the WHO's latest report on SARS-CoV-2 sequencing [15]):

Surveillance

- Identifying disease origins, both zoonotic origins and of outbreaks within human populations
- Outbreak management; when used in conjunction with more traditional epidemiology techniques such as contact tracing, the higher resolution information provided by sequencing can help to unravel complex transmission events, disprove suspected transmission events, and help to identify disease 'super-spreaders'
- Transmission dynamics; tracking person-person or person-environmental transmission, in which situations transmission is occurring (community/hospital/environmental) and local/national/international transmission routes
- Examining viral population structure, to monitor viral evolution within or between people, geographical regions, or through time
- Tracking disease prevalence (the frequency of infections in a population at a particular point in time)

Diagnosis

- Pathogen identification, particularly early in the outbreak when less was known about the viral genome
- Genotyping to identify regions of the genome for use in diagnostic testing, and also to monitor genetic changes in these regions to ensure that diagnostic tests remain effective
- Detection of mixed infections/co-infection, particularly those infections with similar clinical presentations e.g. influenza, bacterial pneumonia

Research and development

- Development of novel diagnostics, therapeutics, and vaccines, including detecting possible drug resistance, understanding virus susceptibility to treatment, evaluating vaccine effectiveness
- Characterisation of virulence factors and resistance markers

- Improving understanding of disease biology, including mutation phasing, viral mutation rates and rare variant detection
- Understanding the link between host genomics and disease severity and susceptibility

1.4 SARS-CoV-2 biology

Genomic data in online databases is frequently used alongside biochemical and structural studies to understand the origins, biology and pathogenicity of SARS-CoV-2. Sequencing of SARS-CoV-2 early in the pandemic, and subsequent analysis, revealed that it is closely related to other SARS-like viruses in the Coronaviridae family; it is a member of the genus Betacoronavirus and subgenus Sarbecovirus, along with SARS-CoV [16-19].

1.4.1 Zoonotic origin of SARS-CoV-2

Both SARS-CoV and MERS-CoV originated in bats and were then transmitted to humans via a palm civet and a camel, respectively [5]. SARS-CoV-2 is also thought to have originated in bats and been transmitted to humans via an intermediate animal host, potentially undergoing genetic recombination to enable it to infect humans, although the source of the zoonotic transmission is not yet clear [7].

Based on the whole genome sequence of the virus, SARS-CoV-2 is most closely related to the SARS-like bat coronavirus lineage from which SARS-CoV descended, sharing over 87% sequence identity with two bat lineages compared to ~79% with SARS-CoV and ~50% with MERS-CoV [20]. Multiple research groups have used sequencing data to confirm that a bat coronavirus, RaTG13, shares the highest sequence identity to SARS-CoV-2 across the entire genome [18, 21, 22]. The closeness in genomic relationship between the viruses means that researchers can use the understanding of the biology of RaTG13 and other human coronaviruses, such as SARS-CoV, as a reference.

The general ecological separation of bats from humans suggests that an intermediate viral host exists, and multiple research groups have used pre-existing sequencing data to determine sequence similarity between human SARS-CoV-2 and other coronaviruses in mammalian hosts. For example, a highly cited paper by Ji et al [21] used pre-existing genomic data from various host species for different coronaviruses related to SARS-CoV-2 to investigate the possible virus reservoir, concluding that SARS-CoV-2 may be a recombinant virus derived from a bat coronavirus and another coronavirus of unknown origin.

Multiple groups have sequenced viruses infecting pangolins which has provided evidence that they are a potential intermediate host. One finding thought to be of particular importance is that the receptor binding motif (RBM) of the spike protein, which is essential for infecting human host cells, shows strong similarity between pangolin coronaviruses and SARS-CoV-2, suggesting the animal as a possible intermediate host [23, 24]. Researchers at the University of Shantou and the University of Hong Kong used metagenomic sequencing to identify SARS-CoV-2 related coronaviruses in Guangdong pangolins suggesting recombination among SARS-CoV-2, RaTG13, and the coronaviruses isolated from pangolins [25]. Another group of researchers from the South China Agricultural University, Guangzhou and Guangzhou Zoo also carried out metagenomic sequencing of pangolins to provide evidence that they were an intermediate host [26].

1.4.2 SARS-CoV-2 structure and function

SARS-CoV-2 viral particles (virions) are around 50-200 nanometres in diameter and are comprised of a viral envelope and RNA genome. The viral envelope is made up of three structural proteins: spike (S), envelope (E) and membrane (M). The fourth protein, nucleocapsid (N), encloses the RNA genome. These proteins are key to viral biology including its replication and infectivity and as such are the focus of structural studies and drug discovery efforts [27].

Three identical copies of the spike glycoprotein form each of the viral spikes that cover the outer envelope of the virus [28]. The spike glycoprotein consists of two functional subunits, of which one mediates the binding of the virus to the host cell receptor, and the other of which is responsible for the fusion of the viral and cellular membranes.

The human ACE2 receptor has been identified as the host cell receptor that binds SARS-CoV-2; it is found on lung cells as well as other cells in the body [28]. Studies by Wan et al [29] and Wrapp et al [30] using sequence and structural comparisons suggest that the SARS-CoV-2 spike receptor binding domain (RBD) is well suited to binding the human ACE2 receptor to gain entry into host cells. In addition, Letko et al [31] showed that host protease processing during viral entry enables SARS-CoV-2 to infect human cells.

Comparison of RaTG13 and SARS-CoV-2 sequences has shown a significant difference in the furin cleavage site between S1 and S2 subunits of the spike protein, which is essential for viral entry into the host cell [32]. In addition, the RBD of the SARS-CoV-2 spike glycoprotein and RaTG13 are around 85% similar, sharing just one of the six critical amino acid residues for binding host ACE2 receptors [33]. Whilst there is a general agreement in the literature that the jump from an animal host to humans was probably enabled by the introduction of the S1/S2 furin cleavage site and alterations in the RBD of the spike protein, how these changes came about is still uncertain, and sequencing efforts continue in a bid to solve this vital debate [33].

Once fused with the cell, the virus particle injects the viral genomic RNA, which is translated by the host cellular machinery into protein chains. These are then cleaved into functional units by the main protease to make new proteins required for viral replication [34, 35]. One essential viral protein is the enzyme RNA-dependent RNA polymerase (RdRp), which replicates the viral genome and transcribes viral RNA to allow translation into new virus proteins by the host cell [27, 36]. The new viral proteins and viral genome RNAs assemble into new virions, which are packaged into vesicles and released from the cell, after which they can infect further cells. Drugs that target any of the key proteins for viral replication could potentially be used to treat COVID-19. For example, the drug Remdesivir inhibits RdRp, and has been authorised for treatment of COVID-19 in several countries after showing some promise in trials [37].

1.4.3 Viral evolution

Analysis of SARS-CoV-2 genomes between 24 December 2019 and 25 March 2020 suggested that mutations accumulated in the SARS-CoV-2 genome at a relatively consistent and modest rate [38]. A recent estimate of the viral evolution rate is $\sim 1.1 \times 10^{-3}$ substitutions per site per year. This equates to 33 changes per year on average across the virus genome, which is within the range of evolutionary rates estimated for other human coronaviruses [39]. This compares to a rate of $\sim 4 \times 10^{-3}$ substitutions per site per year for the HIV virus [40].

As more viral genome sequences became available, researchers attempted to map which regions of the genome are subject to greater selection pressure – i.e. are targeted by host immune systems, meaning that changes in these regions could enable effective immune evasion by the virus – and have high mutation rates, compared with those that are integral to the infectivity of the virus and are therefore highly conserved.

This has important implications for the development and effectiveness of therapeutics, vaccines and diagnostics, for which there is a need to target conserved, unchanging parts of the viral genome. Some groups are investigating this by analysing the sequences in international databases uncovering mutational 'hotspots'. Most notably, Wang et al [41], Ceraolo et al [42], Pachetti et al [43] and Korber et al [44] identified areas of high mutation frequency in the RdRp and S genes and in open reading frame (ORF) 8 and ORF1a.

Early in the pandemic researchers began to perform phylogenetic analyses on available sequencing data in an attempt to classify the virus into different sub-groups based on the genetic similarity of observed mutations. Many of the viral lineages observed were likely to have arisen by chance rather than through specific evolutionary pressures, and it was uncertain if any were associated with an alteration in virulence or transmission. However it became clear that mutations were clustering in critical genes for the virus, including those that are the targets of vaccine and drug design (e.g. RdRP and S genes) [38, 45].

A number of variants of interest and variants of concern, where the virus has demonstrated altered biology and epidemiology, have been identified and monitored since December 2020. Further details of these variants are described in the addendum for this report [4]. Ongoing studies are critical to understand if variants have a functional effect upon the virus and to track their spread through different populations.

1.4.4 Intra-host variation of SARS-CoV-2

The viral isolates sampled early in the pandemic from Wuhan patients had relatively little genetic diversity, which was unsurprising given their recent common ancestry. However, due to the strong immunological pressure in humans, researchers started to investigate how SARS-CoV-2 accumulates mutations within individual hosts. A number of studies with small numbers of patients have taken place looking at this intra-host viral variation.

A collaborative research group led by the Beijing Institute of Genomics (BGI) used metatranscriptome sequencing of the virus to determine the intra-host diversity of the viral population [46]. They observed that viral mutation within hosts was common, but their results also revealed large differences in the variation of SARS-CoV-2 genomes within different hosts. Another study led by the University of Melbourne - the largest study of this kind, with 98 patients - found minimal intra-host genomic diversity [47]. Karamitros et al [48] sequenced SARS-CoV-2 genomes from three patients and they noted genome regions that are prone to alterations and a potential recombination hot-spot in the spike (S) gene.

Several studies investigating intra-host viral mutation have looked at patients experiencing persistent infection [49-51]. Results from these studies suggest there is intimate host and pathogen interaction during persistent infection resulting in an accumulation of changes in genes coding for viral proteins that influence the host immune response.

Notably, one study found that intra-host variation increased in an immunocompromised patient after receiving convalescent plasma as a treatment [51]. Another study has demonstrated the ongoing evolution of SARS-CoV-2, including a mutation shared with a variant of concern, in an immunocompromised patient treated with convalescent plasma [52]. It has been hypothesised that some variants of concern currently being monitored arose in immunocompromised patients, however direct evidence for this is not available. These studies highlight the importance of monitoring viral evolution within these individuals during the course of their illness. A host's immune status and/or treatment regime may accelerate intra-host diversity with possible implications for therapeutic strategies and public health decision making.

1.4.5 Transcriptome

Much of the understanding about the fundamental biology of SARS-CoV-2 including the annotation of genes and their translation into proteins, was based on the assumption that its biology is comparable to previously characterised coronaviruses such as SARS-CoV. Various groups have attempted to experimentally verify these assumptions and to fully characterise the viral transcriptome.

Understanding the expression of the viral genome is important to characterise the viral replication mechanisms and host-viral interactions involved in pathogenicity, and a fundamental step for the functional characterisation of the viral proteins.

For example, after sequencing the genome of a patient who tested positive for COVID-19 in Korea, researchers from Seoul National University and Korea's Centre for Disease Control and Prevention investigated the transcriptomic architecture [53]. The authors used a combination of Illumina short read sequencing, Beijing Genomics Institute (BGI) & MGI Tech nanoball sequencing and direct RNA sequencing using an Oxford Nanopore Technologies (ONT) sequencing platform to understand the life cycle and pathogenicity of SARS-CoV-2. Results from this study found that SARS-CoV-2 produces transcripts encoding unknown open reading frames and possesses RNA modification sites on viral transcripts, showing that there is more to be discovered and understood about the virus.

Taiaroa et al [54] also used direct RNA sequencing to further investigate the transcriptome and epitransciptome (the biochemical modifications of the RNA) of the virus. Results from this study confirmed the existence of previously assumed features of the SARS-CoV-2 genome.

Davidson et al [55] combined the use of transcriptomics and proteomics to produce a correlated transcriptome and proteome map of SARS-CoV-2 in Vero E6 cells (the cells widely used to propagate the virus in the laboratory). They found that when grown on a large scale, the virus spike glycoprotein undergoes alteration and therefore suggested that the viral genome should be monitored in laboratory-grown viruses because many of the attempts to develop treatments and vaccines target this protein.

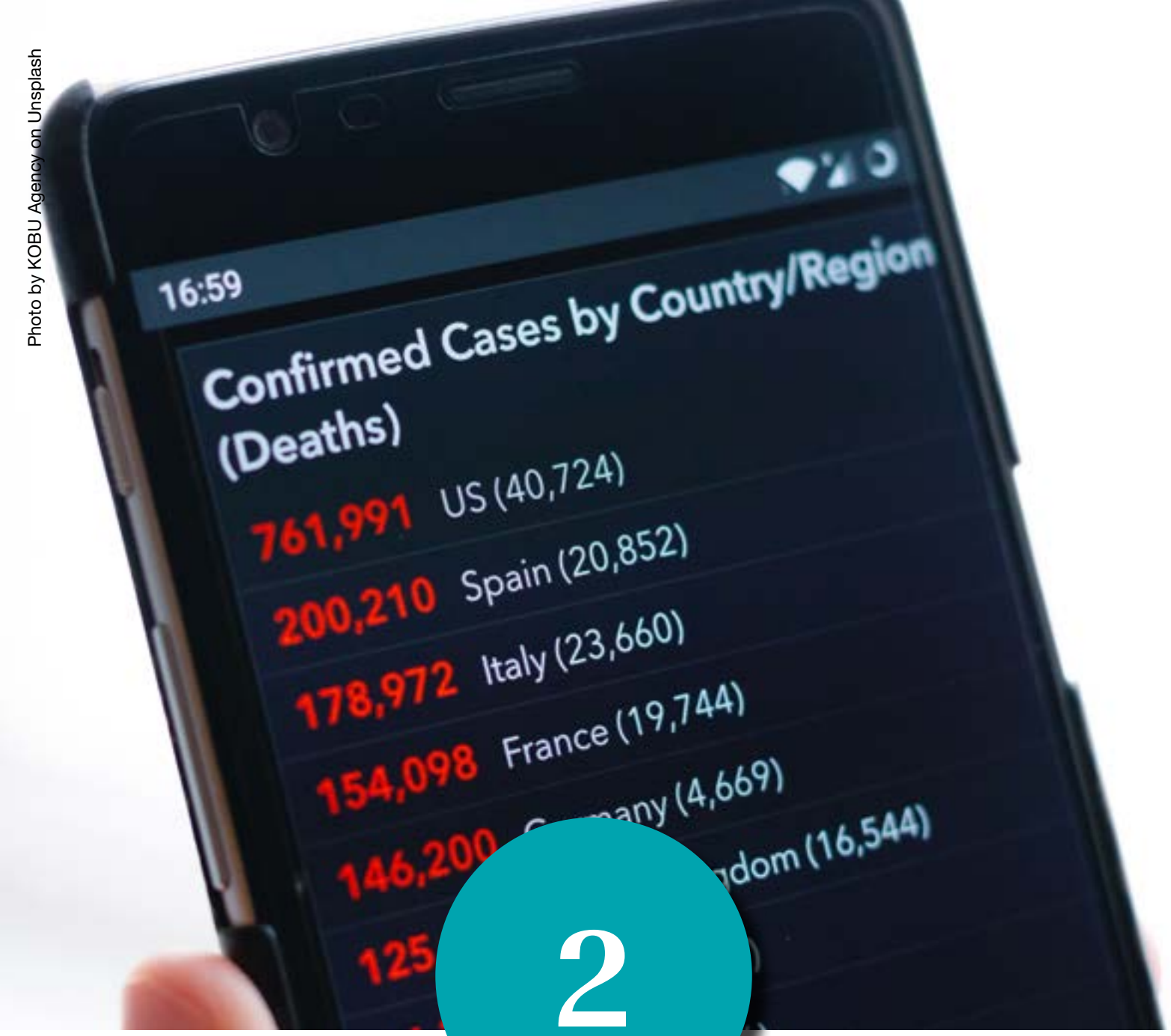


Photo by KOBU Agency on Unsplash

2

Global surveillance:
applications of sequencing technologies and techniques

2 Global surveillance: applications of sequencing technologies and techniques

As outlined in Chapter 1, there are a number of benefits to the sequencing of pathogen genomes, and next generation sequencing technologies are being applied in many different areas to facilitate understanding and management of the COVID-19 pandemic.

The importance of sequencing tools and studies to monitor genotypic change and to support the development of products to improve clinical care and public health measures has been emphasised in the WHO document *A coordinated global research roadmap: 2019 novel coronavirus report* [56]. The main steps have been identified as:

- Immediate steps: Surveillance studies to characterise virus sequence evolution, including maintenance of existing platforms (i.e. GISAID) and support for mechanisms to share information and materials
- Mid- to long-term steps: Harmonisation of metadata related to virus sequence and disease phenotype; functional assays for essential virus features related to human adaptation (receptor affinity, cell tropism, immune interaction, virus isolation and replication studies including reverse genetics)

Near real-time analysis of data has directly impacted the public health response and a recent WHO report, *SARS-CoV-2 genomic sequencing for public health goals*, has explored the impact of sequencing on public health [57].

Table 1: The public health objectives of SARS-CoV-2 genomic sequencing [57]

Activities that require a limited effort and once achieved might need either no sequencing or occasional sequencing for follow-up	Activities that require sustained sequencing activities over a long period of time	
Identify SARS-CoV-2 as the causative agent of the disease Develop diagnostics for SARS-CoV-2 Support the development of therapies and vaccines Investigate date of introduction into humans and investigate SARS-CoV-2 origin (ongoing) Reinfection: • Evaluate and improve understanding of this phenomenon • On the individual level, differentiate between prolonged infection and reinfections	SARS-CoV-2 evolution and its impact on: • Change in viral behaviour (phenotypic change) e.g. transmissibility or pathogenicity • Immunity (from vaccines or natural infection) • Diagnostics (i.e. molecular, serology, antigen assays) • Therapeutic interventions (e.g. monoclonal antibodies)	Montior viral movement and activity: • Investigate geographic spread and reintroduction between populations • Investigate outbreaks in specific settings and populations (e.g. in hospitals) • Track zoonotic reintroduction in both directions over the species barrier • Monitor environmental and waste water • Support classical surveillance by quantifying the period of transmission and evaluating drivers, and by estimating the transmission level in the population

2.1 Overview of surveillance

Integrated epidemiological and virological surveillance is playing a significant role in:

- Monitoring trends in COVID-19 disease and deaths at national and global levels
- Monitoring the spread and evolution of SARS-CoV-2 and monitor impacts on disease
- Enabling rapid detection, isolation, testing, and management of cases
- Supporting subsequent interpretation of observations of COVID-19 related respiratory disease and its epidemiology
- Understanding the co-circulation of SARS-CoV-2, influenza and other respiratory viruses
- Providing epidemiological information to conduct risk assessments at the national, regional and global level
- Evaluate the impact of the pandemic on health care systems and society
- Guide the implementation and adjustment of targeted control measures
- Supporting the development and updating of diagnostic tests
- Informing drug and vaccine development
- Detect and contain clusters and outbreaks, especially among vulnerable populations
- Identify, follow-up and quarantine contacts

Considering the potential for rapid exponential growth of COVID-19 cases, the WHO states that essential surveillance for COVID-19 should include identification, reporting, and the inclusion of data in epidemiological analysis within 24 hours for new cases. This means that national authorities should consider including COVID-19 as a mandatory notifiable disease with requirements for immediate reporting [58, 59].

Surveillance systems should be geographically comprehensive and include all persons and communities at risk, including enhanced surveillance for vulnerable or high-risk populations. This will require a combination of surveillance systems including contact tracing in the entire health care system, at the community level, as well as in closed residential settings and for vulnerable groups. The WHO document Public Health Surveillance for COVID-19 Interim guidance (16 December 2020) provides details on surveillance systems for COVID-19 [59].

Comprehensive COVID-19 surveillance approaches can consider using, adapting, strengthening and maintaining existing surveillance systems. To ensure comprehensive representation of populations a combination of surveillance systems can be used (table 2). Of note, systems do not need to be restricted to clinical settings and different levels of organisations can enact such approaches, from local municipalities or cities, to states or provinces to national programmes. For example, surveillance at a primary care level can detect cases and clusters in communities, whereas virological sentinel surveillance of COVID-19 using clinical samples from hospitals will play a significant role in monitoring the spread and evolution of SARS-CoV-2.

Given that influenza and COVID-19 are both respiratory viruses with similar clinical presentations, existing respiratory disease surveillance systems and associated networks are playing an important role in monitoring the spread of SARS-CoV-2 and will be relied on if comprehensive active case finding is challenging in countries with community transmission.

Table 2: Surveillance systems across different sites and contexts for COVID-19 [59]

	Surveillance system					
Site/Context	Immediate case notification	Contact tracing	Virologic surveillance	Cluster investigation	Mortalilty surveillance	Serologic surveillance
Community	X	X		X	X	X
Primary care sites (non-sentinel influenza-like illness/acute respiratory infections)	X		X	X		
Hospitals (non-sentinel influenza-like illness/ acute respiratory infections)	X		X	X	X	X
Sentinel influenza-like illness/ acute respiratory infections/severe acute respiratory infections sites	X		X			
Closed settings*	X	X		X	X	X
Healthcare associated SARS-CoV-2 infection	X	X		X	X	X
Travellers at points of entry	X	X		X		

** Including but not limited to long-term living facilities, prisons and dormitories*

The WHO Global Influenza Surveillance and Response System (GISRS) is a well-established network of over 150 national public health laboratories, including 144 WHO designated National Influenza Centres (NICs) in 125 countries, about 60% of WHO member states. GISRS monitors the epidemiology and evolution of influenza viruses and disease. Leveraging the GISRS system is an efficient and cost-effective approach to enhancing COVID-19 surveillance. Notably, by 25 March 2020, approximately 85% of more than 220 national public health laboratories currently testing for COVID-19 globally were laboratories closely associated with GISRS [60].

The WHO has provided guidance on how to leverage the existing GISRS influenza capacities and mechanisms for SARS-CoV-2 surveillance through influenza-like illness (ILI), acute respiratory infections (ARI) and severe ARI (SARI) sentinel surveillance systems [60, 61]. The WHO recommends active surveillance, with focus on case finding, testing and contact tracing in all transmission scenarios [62]. In addition, the European Centre for Disease Prevention and control (ECDC) has developed a strategy for COVID-19 surveillance at national and EU/EEA level [63], while the Africa Centres for Disease Control and Prevention (Africa CDC) has developed a protocol for enhanced SARI and ILI surveillance for COVID-19 in Africa [64].

Due to the integration of COVID-19 surveillance into existing SARI/ILI systems, sampling strategies use normal in-country SARI/ILI programme sampling. Sites collect and report data through the country's usual influenza reporting system, and should continue to report ILI and SARI data to the GISRS system. No additional samples beyond those collected during routine, sentinel SARI and ILI surveillance are required. The sample populations for surveillance are therefore the same as those of the member states' influenza programme.

2.2 Overview of genomic surveillance

Viral genome sequencing is used in clinical microbiology services and national public health laboratories to detect, monitor and control viral disease, such as influenza. Ongoing sequencing efforts as the pandemic develops allows for monitoring of the disease at a global and national level.

2.2.1 Selecting samples for sequencing

Continuous monitoring of virus evolution through sequencing of representative samples is considered essential for monitoring changes in the virus, particularly those that could have an impact on virus transmission or virulence.

It has been recommended that a representative subset of COVID-19 positive specimens based on geographic location, age, sex, and disease severity be selected for sequencing. Specimens or RNA extracts that test positive for SARS-CoV-2 with a decent amount of starting material (real-time PCR cycle threshold value (Ct value) <30), are considered good material for sequencing the whole or partial genome of the virus [60, 61, 63]. WHO guidance for sequencing of COVID-19 virus is now available and includes details on sampling [15, 57, 61, 65].

The WHO is encouraging countries to expedite genomic sequencing of SARS-CoV-2 to a minimum of 15 samples per week from sentinel surveillance systems and to share the genetic sequence data through a publicly accessible database [61].

Viral epidemiology studies require representative samples in order to consider issues of:

- Population inclusion and initiative location
- Breadth of genomic data
- Data access – sequencing and other metadata (including clinical data)
- Integration into clinical care and research
- Data quality assessment

Sequencing is also used in genomic surveillance of the coronavirus in animal populations [23, 25, 26, 65], not only to monitor coronaviruses but also in an attempt to determine the origin of current human SARS-CoV-2, as well as to investigate other potential sources of zoonotic infection.

At the beginning of the pandemic, diagnostic testing tended to be restricted to reference laboratories, which could select representative samples as required for genomic surveillance. As non-sequencing diagnostic testing has become more widespread, and is being carried out outside reference laboratories, it is important to consider the mechanisms by which a representative cohort of positive samples are selected for sequencing by reference laboratories to support ongoing genomic surveillance efforts.

The emergence of new variants, particularly variants of concern with altered disease biology or epidemiology, means that these have been prioritised for sequencing. This can result in disproportionate representation of certain variants within international sequence databases, particularly if it has not been indicated whether the sequencing was targeted or due to random sampling for surveillance. For example some regions have prioritised sequencing based on diagnostic S-gene target failure tests [4, 67].

2.3 Sequencing initiatives across the globe

Numerous initiatives and consortia have been established to coordinate sequencing efforts. These initiatives connect sequencing resources to public health programmes, and academia to public health expertise, by providing organisational and data management support that can also provide a foundation for surveillance and research efforts.

The sequencing initiatives have similar aims and objectives. The primary intention of all associated research projects is to monitor the virus at national and global levels.

Broadly, their strategic goals are to:

- Support national efforts to coordinate the work of healthcare, public, private and academic organisations to sequence and analyse the spread and evolution of the SARS-CoV-2 virus and how it affects patients
- Maximise the quality, quantity and usefulness of SARS-CoV-2 sequence data
- Facilitate and strengthen collaboration between partners and members
- Generate open data for public health and basic research by supporting metadata collection, data integration and data visualisation

Members are a combination of federal/regional agencies and laboratories, local public health laboratories, academic institutions, corporations and non-profit public health or research institutes, with sequencing resources frequently found within academic and research institute groups. There is therefore some overlap between research efforts (described in Chapter 4) and national public health programmes. For example, the national effort by Iceland included the sequencing of 643 SARS-CoV-2 samples as part of a population study [68]. Another example is a Brazilian study that sequenced 427 spatially representative samples from 84 different municipalities across 18 of the 27 federal units [69].

Numerous nations have initiatives running; these include the Irish Coronavirus sequencing consortium [70], Austria [71, 72], Finland [73], Japan [74], Switzerland's Swiss SARS-CoV-2 Sequencing Consortium (S3C) [75], and the German COVID-19 OMICS Initiative (DeCOI) [76]. City or region specific initiatives include Coronavirus Sequencing in Quebec (CoVSeQ) [77]. Some of these are part of surveillance programmes and routinely offer pathogen sequencing in response to outbreaks, such as the National Institute of Infectious Diseases (NIID) in Japan [74].

The scale of the initiatives varies in terms of participant numbers, the capacity of sequencing infrastructure, resources and level of national prioritisation. They may operate differently, some being a crowdsourced effort (e.g. SPHERES) and others a centrally managed structure (e.g. COG-UK).

The partnerships also allow groups to share insights and discoveries to drive understanding of the pandemic as it changes over time. The response of these consortia to the emergence of variants of concern is described in the addendum to this report [4]. Some of the major initiatives and consortia currently in operation are (see also table 3):

2.3.1 SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance (SPHERES)

The United States (US) Centers for Disease Control (CDC) advanced molecular detection (AMD) programme in the USA was launched in 2014, and integrates the latest NGS technologies with bioinformatics and epidemiology expertise across the CDC and the US to help find, track, and halt the spread of disease-causing pathogens. It has brought together NGS use in all CDC infectious disease laboratories, over 50 public health departments across the US and more than 40 additional public health laboratories. It is using NGS to detect disease outbreaks earlier – for example foodborne diseases, measles – and to monitor transmission of diseases such as tuberculosis and influenza.

On the 1May 2020 the CDC AMD programme announced it was leading the SARS-CoV-2 'sequencing for public health emergency response, epidemiology and surveillance' (SPHERES) consortium. This is a national effort to coordinate SARS-CoV-2 sequencing across the US, to help accelerate the use of near-real time pathogen genomic data for the pandemic response, through data standards and sharing [78, 79].

SPHERES is a consortium of the US public health and scientific communities that include Federal Agencies and Laboratories, state and local public health laboratories, academic institutions, corporations, non-profit public health and research laboratories, and international collaborations. Notable participant organisations are the Food and Drug Administration (FDA), National Institute of Allergy and Infectious Diseases, Office of Genomics and Advanced Technology, New York and Washington state laboratories, Stanford University, Yale University, Broad Institute, and Scripps Research. A complete list of members can be found on their website [78]; as of 20 May 2020, 442 scientists and 109 organisations were partners.

Another aspect of sequencing efforts in the United States is a framework and blueprint for dramatically expanding genomic surveillance, as well as an implementation framework, developed by the Rockefeller Foundation [80-82]. As part of the funding commitment the Foundation will be collaborating with government, academia, and the private sector in the US and abroad to develop, analyse, and share the data needed to support robust genomic surveillance.

2.3.2 COVID-19 Genomics UK (COG-UK)

The COVID-19 genomics UK (COG-UK) consortium was established on 23 March 2020 to deliver large-scale SARS-CoV-2 genome sequencing and analysis capacity to hospitals, regional National Health Service (NHS) centres and the Government [83]. The goal is for data generated by the consortium to be combined with epidemiological and clinical information, which will inform interventions and policy decisions. There are also five key areas that the initiative aims to support:

1. Scaling up NHS swab testing for those with medical need and, where possible, the most critical workers
2. Mass swab testing for critical key workers in the NHS, social care and other sectors
3. Mass-antibody testing to help determine if people have immunity to coronavirus
4. Surveillance testing to learn more about the disease and help develop new tests and treatments
5. Spearheading of a national diagnostics to build mass-testing capacity at scale

Consortium members include the four UK Public Health Agencies, multiple regional University hubs, and large sequencing centres such as the Wellcome Sanger Institute [84]. The list of consortia members and partners can be found on their website [85]. One member, the Centre for Genomic Pathogen Surveillance (CGPS), has a number of existing IT platforms that are being used to facilitate collection of samples and metadata from NHS testing laboratories as well as to generate real-time displays of the transmission and spread of the virus at the national and local level [86].

COG-UK members are outputting the data in standardised formats to open-access repositories, namely GISAID [14] and MRC-CLIMB (Cloud Infrastructure for Microbial Bioinformatics) [87]. A £1.2m funding boost announced in January 2021 will support the development of further computing and bioinformatics infrastructure for CLIMB [88].

A series of COG-UK regional sequencing laboratories exist, each with their own sequencing platforms. However to facilitate testing from the Lighthouse Laboratory National Testing Centres – four laboratories that receive diagnostic swabs from across the country for SARS-CoV-2 diagnostic testing – and from other diagnostic services in parts of the UK that are not covered, the Sanger Institute is providing a centralised service for large-scale genome sequencing [86]. This national sequencing hub also serves as a backup to take pressure off the regional sequencing labs during periods of high demand.

COG-UK is also conducting a number of studies to understand more about the transmission and evolution of the SARS-CoV-2 virus over time, such as the COG-UK Hospital Onset COVID-19 Infection (HOCI) Study, which will use sequencing as part of surveillance of infection spread in hospitals [89]. The GenOMICC study – a partnership between the NHS, NHS-owned company Genomics England and Illumina Inc. – aims to sequence the genomes of thousands of patients severely ill with COVID-19 globally [90], to understand susceptibility to the virus. The genome data from each individual will be linked to the virus genome data provided via COG-UK.

2.3.3 Canadian COVID Genomics Network (CanCOGeN)

Genome Canada launched the Canadian COVID Genomics Network (CanCOGeN) [91] in April 2020, along with C$40 million (US$28.5 million) in federal funding. The network, led by Genome Canada, in partnership with the six regional Genome Centres, national and provincial public health labs, genome sequencing centres through CGEn (Canada's national platform for genome sequencing and analysis), hospitals, universities and the private sector. CanCOGeN is coordinating and scaling up existing genomics-based COVID-19 research in Canada.

CanCOGEN is overseeing the sequencing of genomes of up to 150,000 viral samples and 10,000 patients to inform clinical and public health strategies. It also is establishing and managing a framework for data sharing, coordination, and analysis across Canada.

National priorities for genomic surveillance include:

High priority: Retrospective and prospective targeted genomic surveillance with a history of travel and close contact; S-gene dropouts for multi-gene RT-qPCR diagnostic assays; geographical sampling in regions with a pronounced increase in the case notification rate; and continued nationwide random sampling.

Medium priority: severe acute COVID-19 in individuals younger than 50 without significant co-morbidities; vaccinated individuals with subsequent laboratory confirmed SARS-CoV-2 infection; suspected reinfections and suspected or known super-spreading events.

2.3.4 Informed public health decision making in the Netherlands

The Dutch COVID-19 response team is a collaboration between 34 health organisations, including hospitals and public health departments. The team have been using a combination of real-time whole genome sequencing alongside data from the National Public Health response team to inform next steps in the public health response, embedding sequencing as part of the response plan. By 15 March 2020, 190 SARS-CoV-2 viruses from the Netherlands had been sequenced, which the response team noted at the time represented 27% of the total number of full genome sequences produced worldwide [92]. The information gained from sequencing enabled a more precise understanding of transmission patterns and contributed to decision making around the implementation of lockdown measures on 12 March 2020 to prevent further disease spread [92].

2.3.5 Africa CDC Pathogen Genomics Intelligence Institute

Africa CDC is supporting a continent-coordinated approach to maximize the benefits of new technologies for more effective disease prevention and control in Africa. In December 2019, the Africa CDC launched the Pathogen Genomics Intelligence Institute (PGII) with the goal of creating a network of laboratories across the continent with capacity to sequence pathogen genomes [93] that could support public health efforts through improved disease prevention, detection and response, and bioinformatics activities.

The vision of the institute is deeply rooted in the Africa CDC mission to support member states in strengthening health systems and institutions for improved prevention, detection, and response to public health threats. Rather than setting up the capacity in every country, countries without the ability to sequence genomes can send their samples to one of the regional laboratories. PGII links with the Africa CDC's five Regional Collaborating Centres (RCCs), Regional Integrated Surveillance and Laboratory Network (RISLNET) and Member States National Public Health Institutes (NPHIs).

In September 2020 the WHO and the Africa CDC launched a network of laboratories to reinforce genome sequencing of SARS-CoV-2 in Africa. Twelve specialised and regional reference laboratories in the network provide sequencing, data analysis and other technical support services to the countries where they are located as well as to neighbouring countries and countries in their sub-regions [94]. The following were selected as the specialised continental reference sequencing research laboratories for emerging pathogens: Redeemer's University African Centre of Excellence for Infectious Diseases (ACEGID), Nigeria; South African National Bioinformatics Institute (SANBI); and Kwazulu-Natal Research Innovation and Sequencing Platform (KRISP) in South Africa.

Genomes of SARS-CoV-2 from Africa have been sequenced in laboratories in the Democratic Republic of the Congo (DRC), Kenya, Egypt, Gambia, Ghana, Nigeria, Senegal, South Africa, Tunisia, and Uganda (see appendix 2 for list of laboratories). Algeria sent its samples to a laboratory in Paris for sequencing (28 May 2020) [95]. By May 2020, the Institut National de Recherche Biomédicale (INRB) in DRC contributed nearly 60% of the SARS-CoV-2 genome sequences from the African continent [95, 96]. It built up this capacity during the ongoing Ebola outbreak in the eastern part of the country. Further information on sequencing capacity for genomic pathogen surveillance in Africa has now been reported [97].

2.3.6 COVID Network for Genomics Surveillance South Africa (NGS-SA)

NGS-SA was launched in June 2020 when five of the largest laboratories of the South African National Health Laboratory Services (NHLS) and their associated academic institutions in Bloemfontein, Cape Town, Durban, Johannesburg and Tygerberg were awarded a grant from the Department of Science and Innovation and the South African Medical Research Council (SAMRC) to respond to the pandemic [98]. Their goal is to expand the network to also include private diagnostic laboratories and other academic institutions in South Africa and abroad. The five sequencing facilities in South Africa producing genomic data are:

- The Division of Medical Virology at NHLS Tygerberg Hospital, Stellenbosch University (SU)
- The Division of Medical Virology at NHLS Groote Schuur Hospital, University of Cape Town (UCT)
- The Division of Medical Virology at NHLS University Academic Laboratories, University of the Free State (UFS)
- The Division of Medical Virology at NHLS Inkosi Albert Luthuli Central Hospital (IALCH), University of Kwazulu-Natal (UKZN)
- National Institute for Communicable Diseases (NICD), Centre for Respiratory Diseases and Meningitis (CRDM)

2.3.7 SeqCOVID

SeqCOVID is Spain's national initiative for investigating genetic epidemiology of SARS-CoV-2 [99]. It is a nationwide group of infectious disease, genomics, bioinformaticians and clinical researchers from over 45 hospitals and research centres with the goal of sequencing between fifteen and twenty thousand samples [100]. The data produced is being stored in public repositories, as well as on the global NextStrain platform, of which a Spanish section has been created (nextspain.uv.es).

2.3.8 India Department of Biotechnology (DBT) Autonomous Institutions: PAN-INDIA 1000 SARS-CoV-2 RNA Genome Sequencing Consortium

The Department of Biotechnology (DBT) Autonomous Institutions Consortium, India announced a 1000-genome sequencing project to better understand the viral and host genomics of the COVID-19 outbreak on 29 April 2020 [101-103]. India's Council for Scientific and Industrial Research (CSIR), which undertook a 1008-human genome sequencing project in 2019, has been leading the sequencing efforts in India. This study will sequence 1000 SARS-CoV-2 genomes from clinical samples; it is being coordinated by National Institute of Biomedical Genomics (NIBMG), Kalyani with active participation from Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad; the Institute of Life Science (ILS), Bhubaneswar; National Centre for Cell Science (NCCS), Pune; and InStem, Bengaluru. Other DBT laboratories will contribute to sample collection and sequencing.

2.3.9 COVID-19 Network Investigations (CONI) Alliance, Thailand

Genomic surveillance of SARS-CoV-2 was implemented during March 2020 at a major diagnostic hub in Bangkok, Thailand [104]. It is using genomic technology to investigate and monitor COVID-19 infections in the country and a preprint details information on the sequencing of 27 anonymised samples from Ramathibodi Hospital in Bangkok during March 2020 [105].

At the time of writing this report the Thailand and India initiatives were the only consortium identified in Asia, and no consortia were identified in South America.

2.3.10 Danish Covid-19 Genome Consortium (DCGC)

The Danish Covid-19 Genome Consortium (DCGC) assists public health authorities to monitor SARS-CoV-2 spread [106]. It was established in March 2020 through the coordinated efforts of Aalborg University, Statens Serum Institute, Hvidovre Hospital, and the Aalborg University Hospital. Large-scale SARS-CoV-2 sequencing capacity was initially established at Aalborg University and local sequencing capacity at Statens Serum Institute and Hvidovre Hospital.

Since June 2020, additional local sequencing nodes were established at Aalborg University Hospital, Aarhus University Hospital, Slagelse Hospital, Rigshospitalet, Sygehus Lillebælt, and Odense University Hospital. The consortium receives SARS-CoV-2 positive samples and to limit potential sampling bias there is no selection based on Ct values. Since early 2021 the consortium aims to sequence all positive COVID-19 cases in Denmark.

2.3.11 Collaborations between initiatives

Partnerships are also being developed between the different consortia on an international level. For example the COG-UK consortium and CanCOGeN launched a new partnership on 4 May 2020 [107]. COG-UK is supporting two key genomic projects at CanCOGEN: Sequencing SARS-CoV-2 to understand how it functions and is evolving, and analysing people's genomes to understand why they experience such different health outcomes. These two consortia in turn are also collaborating with SPHERES, Africa CDC PGII and PHA4GE.

Table 3: Examples of global SARS-CoV-2 sequencing consortia

Name	Country/region	URL
Irish Coronavirus sequencing consortium	Ireland	https://www.teagasc.ie/food/research-and-innovation/research-areas/food-bioscience/irish-coronavirus-sequencing-consortium/
German COVID-19 OMICS Initiative (DeCOI)	Germany	https://decoi.eu/
Coronavirus Sequencing in Quebec (CoVSeQ)	Quebec, Canada	https://covseq.ca/
The COVID-19 host genetics initiative	Global	https://www.covid19hg.org/
COG-UK Project Hospital-Onset COVID-19 Infections Study (COG-UK HOCI)	UK	https://www.cogconsortium.uk/studies-publications/national-studies/the-hoci-study/
Mutational Dynamics of SARS-CoV-2 in Austria	Austria	https://www.sarscov2-austria.org/
SPHERES - sequencing for public health emergency response, epidemiology and surveillance consortium	USA	https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html

Table 3: Examples of global SARS-CoV-2 sequencing consortia

Name	Country/region	URL
Genetics Of Mortality In Critical Care - The GenOMICC Study	UK-based	https://genomicc.org/
Canadian COVID Genomics Network (CanCOGeN)	Canada	https://www.genomecanada.ca/en/cancogen
Africa CDC Institute for Pathogen Genomics	Africa	https://africacdc.org/africa-cdc-institutes/africa-cdc-institute-for-pathogen-genomics/
Netherlands sequencing efforts at RIVM (National Institute for Public Health and the Environment)	Netherlands	https://www.rivm.nl/en/news/update-on-spread-of-uk-coronavirus-variant-voc-20201201-in-netherlands
COVID Network for Genomics Surveillance South Africa (NGS-SA)	South Africa	http://www.krisp.org.za/ngs-sa/ngs-sa_network_for_genomic_surveillance_south_africa/
SeqCOVID	Spain	http://seqcovid.csic.es/
COVID-19 Network Investigations (CONI) Alliance	Thailand	https://coni.team/
COVID-19 Case in Cambodia	Cambodia	https://public.idseq.net/
Public Health Alliance for Genomic Epidemiology (PHA4GE)	Global	https://pha4ge.org/
COVID-19 High Performance Computing (HPC) Consortium	Global	https://covid19-hpc-consortium.org/
ARTIC network	Global	https://artic.network/
PAN-INDIA 1000 SARS-CoV-2 RNA Genome Sequencing Consortium	India	https://www.biorxiv.org/content/10.1101/2020.08.03.233718v1.full
Danish Covid-19 Genome Consortium (DCGC)	Denmark	https://www.covid19genomics.dk/home
National Institute of Infectious Diseases	Japan	https://www.niid.go.jp/niid/en/
Swiss SARS-CoV-2 Sequencing consortium	Switzerland	https://bsse.ethz.ch/cevo/cevo-press/2020/05/first-data-for-genomic-surveillance-of-sars-cov-2-in-switzerland-made-available.html

2.4 Sequencing data repositories and data sharing

GISRS provided the backdrop for the successful establishment in 2008 of the Global Initiative on Sharing All Influenza Data (GISAID) [14, 108] which has become an integral component of the collection, analysis and timely sharing of essential influenza virus data.

GISAID is a public-private partnership that promotes international sharing of all influenza virus sequences and associated clinical data in the case of human viruses. The principles underlying GISAID's data sharing mechanisms in a public health emergency are closely aligned with those outlined by the Global Research Collaboration for Infectious Disease Preparedness (GloPID-R) [109, 110]. The wider benefits that flow from access to the latest data are evident in the advanced tools and platforms available for analysing and interpreting sequence data. For example, it has ensured availability of the latest data for the biannual WHO influenza vaccine consultation meetings.

GISAID integrates sequence data with other clinical, virological and epidemiological data. SARS-CoV-2 genome sequence data are being uploaded to the GISAID database, a process that is being strongly encouraged globally. GISAID is the primary database being used globally to monitor the progression of the disease and provides insight into global sequencing efforts. Due to the data use agreements with GISAID, many of the data releases on this platform are occurring before or without academic publications.

In addition to GISAID, other existing genome sequence databases, which have been in place for some time, have been adapted and adopted to store and collate SARS-CoV-2 genome sequence data. The most commonly used databases are:

- International Nucleotide Sequence Databases (INSD) – comprising GenBank, the DNA Databank of Japan (DDBJ), and the European Molecular Biological Laboratory (EMBL), which together provide the principal repositories for DNA sequence data. The National Center for Biotechnology Information Sequence Read Archive (NCBI-SRA) [111] is included in these databases
- Medical Research Council Cloud Infrastructure for Microbial Bioinformatics (MRC CLIMB)
- China National GeneBank Sequence Archive

Sharing of sequence data allows analysis to be conducted by research groups with different specialisms and temporal, geographical and numerical expansion of datasets beyond that possible from a single sample collection point. However, there is currently inconsistency in the use of databases between different national and international organisations, and between research groups.

These databases have different data sharing policies and are also capable of storing different forms of the data, such as:

- Unprocessed, raw electrical signal files
- Raw reads
- Processed data
- Metadata
- Short reads
- Genome assemblies

Submission of data is voluntary, but many consortia are strongly encouraging or in some circumstances mandating data sharing to members. Selection of the database in which to release can be determined by a range of factors, including:

- The regulatory framework
- Sequencing data available
- Preferred form of data to upload
- Familiarity with the database

Data can also be released via multiple platforms, and in different forms. It is not always obvious which identifiers link data between databases when a sequence is uploaded to more than one database. In addition, not all globally generated sequence data is being uploaded to these public databases, for numerous reasons [112], therefore no database is completely comprehensive and is only representative of the sequences that have been uploaded. For the SARS-CoV-2 pandemic the most commonly used database is GISAID; another resource provided by the China National Center for Bioinformation [113] collects information from five sources – GISAID, National Microbiome Data Collaborative (NMDC), Genome warehouse, China National Genebank Database (CNGdb) and NCBI GenBank.

Data releases can also occur on other platforms. For example, the first African sequence – from Nigeria – was released on virological.org and made available on GitHub [114] before it was uploaded onto GISAID, and Cambodia has a dedicated website that contains eight of the country's data releases [115]. Other sequences, in contrast, may only be released after a research paper has been published.

Additional packages, such as nextstrain.org, provide a series of modular bioinformatics tools, to carry out data curation, collate and present data in an accessible way using data from existing databases, primarily GISAID.

More information on sequence databases can be found in the 2020 UN Convention on Biological Diversity report entitled *Combined study on digital sequence information in public and private databases and traceability* [116]. It provides a detailed overview of how sequence databases are operated and managed, such as the fact that 95% (705 out of 743) of nucleic sequence data databases directly link to or download nucleic sequence data from the INSD.

2.5 SARS-CoV-2 genomic data releases by country

At the time of writing thousands of sequences had been submitted to a number of databases from over 140 countries (figure 1). Figure 2 shows the number of sequences uploaded to GISAID in each global region, while figure 3 shows daily new COVID-19 cases reported by region over the same time period (end December 2019-end February 2021) [14, 108].

By 1 March 2020 2,248 SARS-CoV-2 sequences had been uploaded onto GISAID; this increased to over 22,000 by the end of March, and 54,539 as of 26 June 2020. There was a significant increase in sequence uploads in the second half of 2020 and into 2021: by 26 January 408,384 sequences were shared via GISAID, increasing to 834,259 by 22 March 2021 [14, 108].

The largest proportion of sequences come from Europe (63%), then North America (27%), Asia (5.8%), Oceania (2.3%), Africa (1.3%), with the fewest from South America (1.1%). Countries that are contributing the largest volume of sequences are the UK (37%), USA (24%), Denmark (6%) and Germany (4%). The other leading countries contributing to sequencing are Canada, Japan, Switzerland, Australia, Netherlands, Italy, Spain and France.

Submissions from African countries primarily come from South Africa (41% of African sequences), Mayotte (an overseas Department of France; 7%), Kenya (7%), Nigeria (6%), and The Gambia (5%). Egypt, Democratic Republic of the Congo, Ghana, Zimbabwe and Rwanda have each contributed approximately 3-4% of African sequences each.

Japan is the main contributor from the Asian region – accounting for 46% of all samples. This is followed by India (13%), Israel (8%), South Korea (6%), China (including Hong Kong) (6%), Singapore (4%) and the United Arab Emirates (4%). The following countries have each contributed around 2% of the sequences from Asia: Bangladesh, Saudi Arabia, Thailand and Indonesia.

For the South American region Brazil has contributed almost half of the sequences followed by Chile (14%), Peru (11%), Argentina (7%) and Colombia (6%).

Figure 1: Map showing numbers of SARS-CoV-2 genome sequences uploaded to GISAID by 28 February 2021 ([14], data downloaded 22 March 2021).

The boundaries used on this map do not imply the expression of any opinion whatsoever on the part of FIND concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

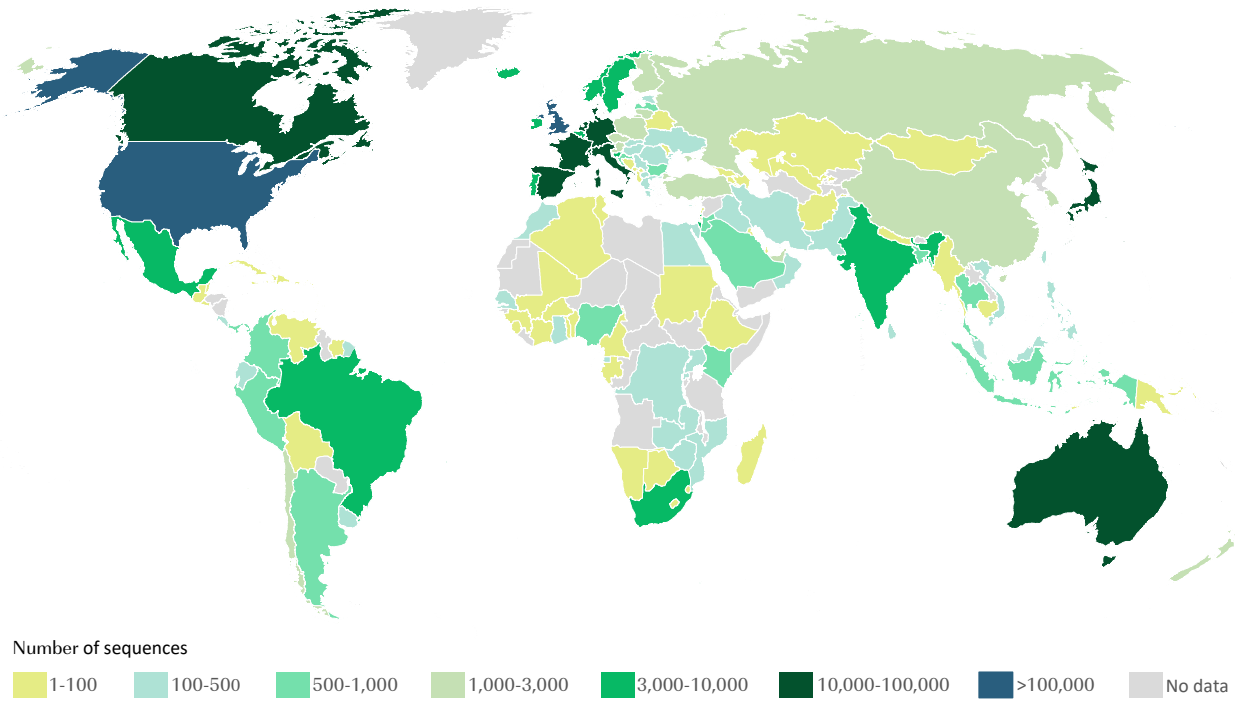


Figure 2: GISAID SARS-CoV-2 sequences released by month and region

Number of sequences released per month per region using GISAID data ([14[downloaded 22 March 2021; total sequences between December 2019 and February 2021 799,316). As the UK represents >35% of the sequences in the database they have been grouped together, the remainder of Europe is labelled 'Rest of Europe'.

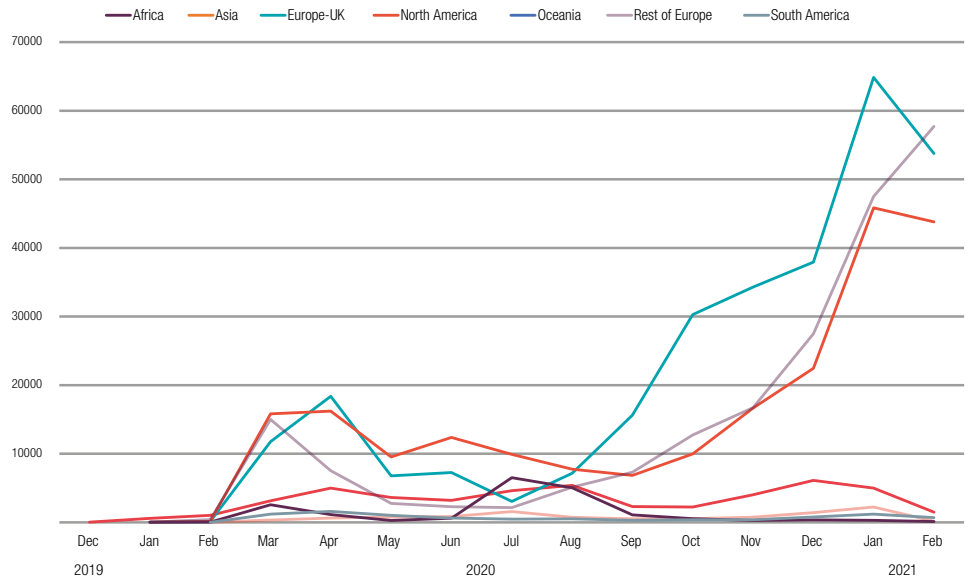
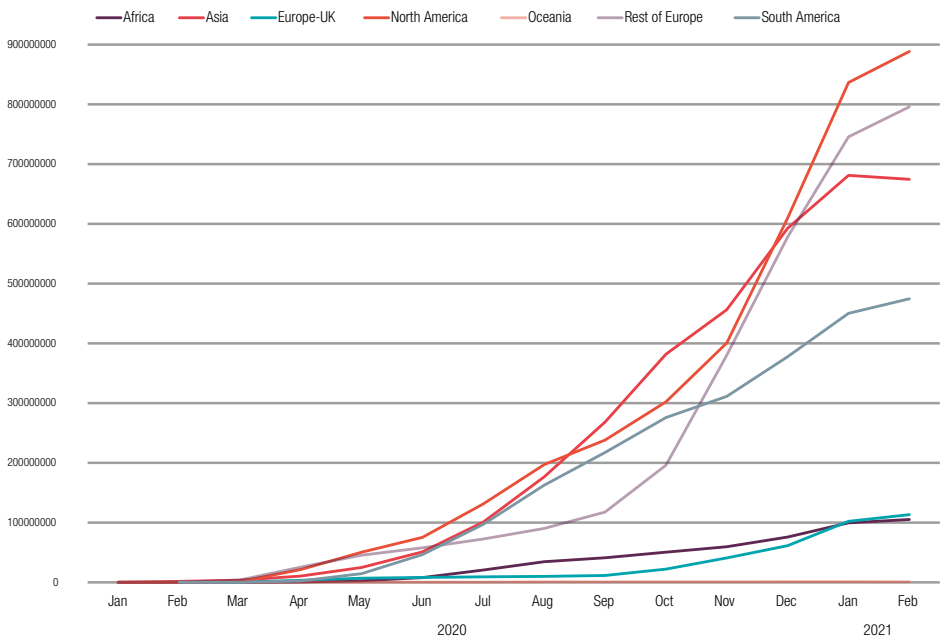


Figure 3: New SARS-CoV2 cases reported by month and region

Number of new SARS-CoV-2 cases per month per region using data from 'Our World In Data' 31 December 2019 - 28 February 2021 (downloaded 22 March 2021 from ourworldindata.org). To allow for comparison of sequences in Figure2 the UK has been separated out and the remainder of Europe is labelled 'Rest of Europe'.



In contrast, by 26 June 2020 11,536 sequences had been uploaded into the NCBI-SRA, compared to 54,539 onto GISAID. By 22 March 2021 SRA held 254,626 sequences compared to 829,275 in GISAID. Over 45 countries have submitted data to the NCBI-SRA database with the UK, USA, Australia, India, Netherlands, Ireland, South Africa, Qatar, Italy, Canada, Denmark and Brazil submitting the largest number of sequences.

Given the speed at which sequencing data has been generated and uploaded in response to need in the early stages of the pandemic, there is heterogeneity in the data submitted that affects possible analyses. Variation is found in date formats attached to data, sequencing vs. sample collection dates, bioinformatics pipelines used to assemble genomes, and in under- or over-representation of data from different geographical areas [117]. There is also variability in how data submitters categorise which organisation or laboratory collected a sample and which laboratory or organisation carried out sequencing and/or submitted the data.

The current timeframes from data generation to data upload vary, sequences would preferably be uploaded as soon as possible after they are generated to reduce reporting burden, e.g. the ECDC has a minimum reporting time frame of once a week, which includes the submission of viral sequence data to GISAID [63]. The WHO code of conduct for pathogen genetic sequence data states the timeframe for data generation and release should not exceed 21 days from sample receipt, although even greater speed is highly desirable in the context of rapidly evolving outbreaks [118].

In the report *Risk related to spread of new SARS-CoV-2 variants of concern in the EU/EEA*, the ECDC commented on the speed at which sequencing data was being released, finding that sequencing workflow turnaround times varied between nations. They found that the UK had the quickest turnaround, with 5.3% of cases sequenced and published with an average delay of 23 days, in contrast to delays of over 80 days recorded for Croatia. Of all EU/EEA Member States, only Denmark and Norway have sequenced and published more than 1% of cases, and only eight countries have sequenced and published more than 0.1% of cases since 1 September 2020 [119].

2.6 Additional sequencing related initiatives

2.6.1 Public Health Alliance for Genomic Epidemiology (PHA4GE)

In October 2019, the Bill and Melinda Gates Foundation funded the launch of a global initiative called the Public Health Alliance for Genomic Epidemiology (PHA4GE) [120]. PHA4GE was established to champion the support for the development of bioinformatics and data standards for public health. PHA4GE is working with the US CDC's AMD Programme to coordinate the SPHERES consortium.

The Global Alliance for Genomics and Health (GA4GH) has aligned with the PHA4GE to guide efforts to share pathogen genomic data. PHA4GE models itself on GA4GH in goals and structure, and aims to develop an open source, community-supported ecosystem for bioinformatic software development, implementation and validation. Through PHA4GE, these standards will be disseminated into the broader international community.

PHA4GE is a global coalition that is actively working to:

- Establish consensus standards
- Document and share best practices
- Improve the availability of critical bioinformatic tools and resources
- Advocate for greater openness, interoperability, accessibility and reproducibility in public health microbial bioinformatics

They have ten recommendations for supporting open pathogen genomic analysis in public health settings, covering data management and programming; management and stewardship of data; accessibility, usability and reproducibility of bioinformatics; infrastructure development; genomic epidemiology; and data sharing [117].

2.6.2 COVID-19 High Performance Computing (HPC) Consortium

There is an increasing need for high performance computing and storage systems to enable next-generation sequencing technologies and analysis of the data generated. As data volumes increase and clinicians need ultra-fast results, the importance of computing capability is also growing. Absolute computing performance becomes a secondary concern compared with stability, security, storage flexibility, application portability and management [121].

Launched on 22 March 2020, the COVID-19 High Performance Computing (HPC) Consortium is a unique private-public effort spearheaded by the White House Office of Science and Technology Policy, the US Department of Energy and IBM to bring together federal government, industry, and academic leaders who are volunteering free compute time and resources on their world-class machines [122]. It is said to have supported 59 research projects already, which are all running on high performance computing machines [123,124]. The UK joined this Consortium on 29 May 2020, making it the first European super-computing partner to join [124], with Switzerland following shortly afterwards.

Although primarily focused on drug development, the Consortium will benefit SARS-CoV-2 research efforts through sharing resources and services, while complementing the work in the European Union being organised via PRACE, the Partnership for Advanced Computing in Europe [125].

2.6.3 ARTIC network

The ARTIC network [126] is a collaborative project between various academic and research institutions established before the COVID-19 pandemic to provide end-to-end sample processing for sequencing in viral outbreaks. It now provides a collection of resources for SARS-CoV-2; this includes primer lists for amplicon sequencing, and protocols for many aspects of Oxford Nanopore sequencing which have been widely used during the pandemic.

Photo by Jakayla Toney on Unsplash

3

Diagnostics

and the role of sequencing

3 Diagnostics and the role of sequencing

Beyond facilitating understanding of the evolution of the SARS-CoV-2 virus genome during the pandemic, sequencing has been used to support diagnostic efforts:

- Through direct use as a diagnostic assay
- Through the use of sequence information to identify viral genomic targets for molecular diagnostics
- To monitor the specificity or sensitivity of diagnostic tests, and inform design of new tests

3.1 Sequencing for diagnosis

NGS-based diagnostic tests for COVID-19 become available from June 2020 and are not widely used. As the COVID-19 pandemic progresses, both knowledge of the disease and virus, and strategies for managing infection rate and reducing transmission, are evolving. Shifting priorities are encouraging the development of different tools including the development of new diagnostics.

3.1.1 NGS-based diagnostics compared to other molecular diagnostic tests

There are a number of advantages to using NGS in diagnosis, including:

Increased testing capacity - NGS offers the potential to facilitate considerable scaling-up of testing due to the number of samples that could be processed in a single sequencing run. Despite the wide availability of qPCR instruments in both clinical and research laboratories today, capacity is less than that which could be offered through alternative methods. According to one estimate, a 384-well qPCR thermocycler can process a maximum of 640 patient tests per day if run in two 10-hour shifts. Increasing testing volume would require additional instruments [127]. In comparison, one sequencing instrument can provide greater capacity in a single run e.g. Illumina's COVIDSeq run on the NovaSeq platform is suggested to be capable of providing 3072 individual results in 12 hours [128]

Higher resolution information - Alongside providing diagnostic information, NGS-based tests can provide more detailed genomic information to assist with surveillance of the disease from a single sample and test run.

High sensitivity - NGS is a highly sensitive technique and could be used to detect rare variants in the viral genome which might otherwise cause problems or be difficult to detect using standard PCR-based testing. The ability to rapidly diagnose certain concerning variants is also valuable. NGS diagnostics could be used to help do this at scale, especially if PCR test strategies to detect specific variants are not available.

Examination of co-infection - As seasonal shifts occur, the potential for co-infection with primarily winter pathogens such as influenza viruses changes. Diagnosis of infection with alternative respiratory pathogens such as influenza viruses could lead clinicians to incorrectly determine this to be the sole cause of respiratory difficulties, where SARS-CoV-2 may also be present [129]. A number of countries and organisations are now focusing on the development of tests for the detection of multiple or different pathogens and sequencing could provide a means to facilitate this.

Diversifying testing strategies - In some situations, having NGS as an alternative approach to testing could overcome supply chain issues for specific test reagents and kits, as has been the case for heavily in-demand PCR tests. However, as many similar reagents are used in both PCR and NGS testing, this may not always provide a solution.

Although the potential advantages listed above are important, some of the logistical and practical aspects of NGS-based testing mean it is unlikely to be suitable in a number of situations and will not replace many simpler molecular-based diagnostics. NGS-based testing is often slower, more expensive, and more complex to conduct than cheap, quick, and generally reliable PCR-based testing and has therefore not been a priority for diagnosis in the COVID-19 pandemic thus far. Sequencing is currently demonstrating value in research and as an aid to surveillance. However, whilst sequencing provides broader and higher resolution information about SARS-CoV-2 and the particular patient case in question, the higher monetary and labour demand of sequencing compared to molecular methods make it less practical or possible in many circumstances, and it is not currently considered necessary for confirming diagnosis.

3.2 Current status of NGS-based diagnostics

Several different types of NGS-based diagnostics are in development, with some recently gaining regulatory approval. The current use of NGS diagnostics is very limited, although they may become more widely used in future if they gain wider regulatory approval from multiple nations, and upon use are shown to have utility over other test methods.

3.2.1 Approaches to NGS-based diagnostics

Testing for COVID-19 using NGS can use a number of approaches. The first broad approach is targeted sequencing – this is similar to approaches utilised in existing qPCR tests, in that it involves the examination of a small number (one or a couple) of short regions in the viral genome, to determine whether the virus is present or absent. One such method, called SwabSeq, has been developed by synthetic biology start-up Octant [130]. Others include LAMP-Seq, developed by researchers at the Broad Institute [127], Dx-Seq developed by a group at Pennsylvania State University, and Oxford Nanopore Technologies LAMPore test. These tests have the capability to sequence thousands of samples simultaneously, producing results with 24 hours. Tests which make use of loop-mediated isothermal amplification (LAMP) facilitate rapid amplification of viral RNA without the need for a thermocycler, as required for PCR-based tests. It is likely that for high-throughput targeted tests to fulfil their potential, automated laboratory equipment would be required to process the high volumes of samples.

NGS-based approaches may also be developed to sequence the entire SARS-CoV-2 genome, the approach taken by Illumina's COVID-seq test. Other tests may also include genomes of other respiratory pathogens (viruses and bacteria) in a panel. Owing to the greater total amount of genomic material examined, these tests cannot provide the same throughput as sequencing assays targeting shorter regions of the viral genome. However, given that they provide more genomic information, they can be used to gain valuable insights into the origin of an infection for surveillance purposes, or about co-infections. This approach is being investigated by several research groups and commercial companies including Baylor College of Medicine, Fulgent Genetics, Helix, HudsonAlpha, and Illumina in partnership with IDbyDNA [127].

3.2.2 NGS-based tests approved for diagnostic use

COVIDSeq is a test provided by Illumina, and was the first sequencing test to be approved for diagnostic use. It gained Emergency Use Authorization from the US FDA (10 June 2020) for the purpose of qualitative detection and diagnosis of COVID-19 [131]. The test utilises Illumina amplicon-based NGS to sequence amplicons that cover the entire SARS-CoV-2 genome, in a similar way to the commonly used ARTIC workflow (see Chapter 6). It is compatible with the most common types of

respiratory samples. Using the NovaSeq, the highest capacity sequencer available from Illumina (see Chapter 5), it is possible to examine just over 3,000 samples in 12 hours [128]. The accompanying workflow provides a near-complete process for detection of SARS-CoV-2 from RNA extraction to report generation, including use of the DRAGEN COVIDSeq Test pipeline, as discussed in chapters 5 and 6. As suggested previously, the test will enable both diagnosis and viral genome sequencing from the same sample and procedure. The test is only available as a diagnostic in the US, and regional availability of the test may vary. Illumina intend to provide their own testing service in the US using the test [132].

LamPORE is a new diagnostic provided by Oxford Nanopore Technologies. The LamPORE assay was CE marked for in vitro diagnostic use in October 2020, for the detection of the SARS-CoV-2 virus, using the Nanopore GridION device. Further regulatory approvals are being pursued in other countries, including Emergency Use Authorization in the United States, and with Nanopore's partner G42 in the United Arab Emirates [133]. The system uses LAMP for amplification of viral RNA, which is then sequenced using either the GridION or MinION platforms. This approach facilitates massive barcoding and multiplexing of samples, potentially enabling the simultaneous examination of up to 1,152 patient samples in three hours or 96 samples in one hour. The information retrieved will not be equivalent to that retrieved by whole genome sequencing, as the approach uses only two regions of viral RNA to determine whether the virus is present in the sample. This system is expected to be suited for high-throughput screening situations in which an answer is required quickly across a large number of samples, even where there is expected to be a low positive rate. Recent studies in the UK have found LamPORE has a similar sensitivity to RT-PCR tests in symptomatic patients, and also support the use of LamPORE for the testing of asymptomatic people [134]. In addition it was found saliva samples worked well, providing further opportunities to streamline testing.

SwabSeq is an amplicon-based NGS platform provided by the start-up Octant, and is available under the Open COVID License for anyone to use and develop further [135]. It is claimed to overcome a number of logistical limitations associated with other sequencing approaches, and may be suitable where reagents are limited, although there are several limitations [130]. It has a turnaround time of 12-24 hours, and is capable of processing up to 150,000 samples per run using a NovaSeq S2 flow cell [136]. The company Helix developed their COVID-19 NGS Test based on the SwabSeq protocol, gaining Emergency Use Authorization from the US FDA in August 2020 for use in authorised laboratories. A SwabSeq based test was also approved for Emergency Use Authorisation by the US FDA in October 2020, and only for use in two Los Angeles laboratories deemed by the FDA to meet the requirements to perform high complexity tests [137].

3.2.3 NGS-based diagnostics in development

Examples of different types of NGS-based diagnostic tests in development for COVID-19 are provided below; this is a non-exhaustive list.

Explify is a metagenomics tool developed by IDbyDNA which is to be co-marketed by Illumina for use with its NGS systems. The platform is a computational tool which includes applications for the workflow management, analysis and reporting of microbial metagenomic data. The aim is to provide a combined platform for clinical NGS for detection of pathogens, including SARS-CoV-2. The system uses a metagenomics approach for the detection of a range of pathogens, making the system potentially useful for the identification of co-infections. While it does not have FDA approval, Explify is a validated laboratory developed test and so can be used for clinical purposes. It has been adapted for the identification of SARS-CoV-2 alongside potential co-infection with 35 additional respiratory pathogens. A workflow developed by Illumina and IDbyDNA includes the Illumina respiratory virus oligo probe (RVOP) enrichment panel, with the aim of providing a sample-to-result process [138].

Fulgent SARS-CoV-2 NGS service is provided by Fulgent, a genetic testing company that provides a range of molecular testing and sequencing services. Alongside their FDA-approved RT-PCR test, they have developed an NGS based assay for which approval as a diagnostic is being sought. The assay relies on Illumina sequencing platforms and is provided as a service through Fulgent. Like other NGS methods, the assay also provides more information about the viral genome. Test outputs will be 'positive', 'negative' or 'intermediate'. Testing takes 2-4 days from sample receipt and samples can be retested if the result is uncertain.

Pathogen-Oriented Low-Cost Assembly & Resequencing (POLAR) is an NGS-based diagnostic in development by researchers at the Baylor College of Medicine. The assay uses viral enrichment and whole genome sequencing on short-read sequencing platforms. The methodology, which has yet to undergo peer review (9 February 2021), uses an enrichment protocol available via the ARTIC network [139]. It is suggested that the assay provides very high sensitivity, enabling diagnosis at even very low viral load. Current publications include low number patient sample comparisons with PCR-based tests developed by the CDC.

LAMP-Seq is a targeted NGS test developed by researchers at the Broad Institute. It utilises loop-mediated isothermal amplification (LAMP) with NGS and promises massive scalability, in the realm of hundreds of thousands of samples per day per sequencing facility [127, 140].

Dx-Seq is a targeted NGS test under development by a group at Pennsylvania State University. The test is currently (as of May 2020) being validated with collaborators using clinical samples. It uses a combination of reverse transcription, PCR, and sequencing to test up to 19,200 patient samples in a single workflow [127].

Other organisations including Guardant Health, Clear Labs, and Helix [141, 142] have also suggested that they aim to develop and distribute NGS-based diagnostics for COVID-19. Currently, there are no diagnostic tests authorised for the broad screening of SARS-CoV-2 in asymptomatic individuals [143], although it is often at the healthcare provider's discretion as to whether to offer testing to asymptomatic individuals on an individual basis [144]. Other new assay methods for NGS tests are constantly in development, often with the view of not just being applicable to the COVID-19 pandemic, but to help manage future pathogen outbreaks. In addition to the tests themselves, platforms to better analyse and interpret the results of highly multiplexed sequencing based diagnostics are also in development [145].

3.3 Sequencing for development and quality control of other diagnostics

Whilst it may not frequently be used as a diagnostic method itself, sequencing and the data produced has been key to informing the development of other diagnostic tests, especially nucleic acid based diagnostic tests. An overview of these tests, focusing on molecular test techniques, is provided in Section 3.4. As SARS-CoV-2 continues to evolve, the role of sequencing in monitoring and quality control of diagnostics has become increasingly important, as well as informing the design of new diagnostics.

3.3.1 Development of diagnostics

The release of the first sequenced SARS-CoV-2 whole genomes enabled the rapid development of a diagnostic test based on RT-PCR by Corman et al within days of the sequence data being made available [146]. Metagenomic sequencing was performed by various groups when SARS-CoV-2 was first discovered e.g. Wu et al and Zhou et al [17, 18], and when investigating related pathogens in other species e.g. Lam et al [25]. Since then, many publications have described various methods for diagnosis using probes based on genomic sequences.

During the early stages of the COVID-19 outbreak, sequence information from the closely related SARS virus was used to facilitate development of early RT-PCR assays for the detection of SARS-CoV-2. NGS data from SARS-CoV-2 positive samples allowed for further development and refinement of these techniques to make them specific to the novel virus. RT-PCR tests for the diagnosis of COVID-19 are based on several SARS-CoV-2 genomic regions, including the RdRP gene in the ORF1ab sequence, the E gene, N gene and S gene [20]. In March 2020, the WHO released a document containing a collection of available RT-PCR assays used by different health organisations internationally, which included a summary of the various gene targets.

Table 4: Genes used in RT-PCR tests

Adapted from WHO List of RT-PCR protocols document [147]

Institute	Gene targets
China Centre for Disease Control, China	ORF1ab and N gene
Institut Pasteur, Paris, France	Two targets in RdRP
United States Center for Disease Control, USA	Three targets in N gene
National Institute of Infectious Disease, Japan	Pan-corona and multiple targets, Spike protein
Charité, Germany	RdRP, E gene and N gene
Hong Kong University, Hong Kong SAR	ORFb-nsp14, N gene
National Institute of Health, Thailand	N gene

Commercially-developed kits also vary in which genes they include for RT-PCR testing. A comparison of seven commercial kits published in July 2020 lists gene targets for kits produced by each of: Altona Diagnostics; BGI; CarTes BioTec; KH Medical; PrimerDesign; R-Boipharm AG; and Seegene. These include, in various combinations, the E gene, ORF1ab, N gene, RdRp, and the S gene [148].

Sequencing data will be important in informing the ongoing development of new diagnostics. As described in the addendum to this report on variants of concern [4], FIND's report on SARS-CoV-2 variants, as new VOCs arise it is beneficial to be able to diagnose and track them at scale, which can be achieved through national testing strategies. The availability of sequence data for these variants has allowed the design of new PCR tests specifically designed to detect key variants.

Furthermore, as described below, the instability of certain regions of the SARS-CoV-2 genome means that some diagnostic test targets may be unreliable. Monitoring of SARS-CoV-2 sequences means that there is now greater understanding of which regions of the virus genome are more stable, and so are likely to make good test targets. Future tests can therefore be designed to avoid sole targeting of regions such as the S-gene, which are known to be more unstable. Indeed, WHO interim guidance on diagnostic testing for SARS-CoV-2 recommends the need for at least two independent test targets for diagnostic tests [149].

3.3.2 Monitoring effectiveness of diagnostics

The genomic evolution of RNA viruses such as SARS-CoV-2, and the fact that SARS-CoV-2 is a recombinant virus (such that the virus could undergo recombination with other human coronaviruses) means that current diagnostics may not remain fit for purpose. There is a need to minimise loss of performance of existing assays due to variants in the viral genome. This remains true for commercially manufactured kits, which may not be as rapidly adaptable as in-house PCR tests and may be less likely to have published primer/probe sequences. For example, PCR assay compatibility could change over time due to changes in probe or primer binding sites. This has already occurred on more than one occasion during the pandemic; most notably a Δ69/70 mutation in the S-gene prevents certain PCR tests with S-gene targets from working, widely known as S-gene target failure, or S-gene drop out (see addendum for further details [4]).

Other types of diagnostic test based on detection of viral antigens and antibodies may also be affected by mutations, if the mutations result in the structure of the target antigens or antibodies changing. The threat of reduced diagnostic effectiveness is minimised by creating diagnostics based on conserved regions of the genome, which are relatively stable.

It has been shown that there are now numerous mutations across virtually all virus genes, and therefore all current PCR test targets [150]. As a result, viral genome sequences are being monitored by groups around the world, including global consortia, to detect any genomic evolution that would cause the current diagnostics to produce false negatives. Several tools for monitoring relevant mutations are now available, as described in the WHO guidance for diagnostic testing of SARS-CoV-2 [149]. The US FDA has also published recommendations for test manufacturers, particularly of molecular tests, to routinely monitor test performance by comparing primer/probe pairs to publicly available genomes [151].

3.4 Non-sequencing molecular diagnostics for COVID-19

A considerable number of COVID-19 diagnostic tests have been developed or adapted from existing tools in the months since the beginning of the pandemic. COVID-19 diagnostic tests fall broadly into antibody, antigen or nucleic acid tests. Nucleic acid tests (NATs) are also called molecular tests, and focus on the detection of SARS-CoV-2 RNA. Whilst sequencing is being extensively used for surveillance and research surrounding the virus and the development of COVID-19, the molecular diagnostics currently used mostly rely on non-sequencing methods. Non-sequencing NATs can use different methods to detect specific nucleic acid sequences; they have different qualities and may be more or less useful in different environments. Reverse-transcription and real-time polymerase chain reaction (RT-PCR and qPCR) testing currently represent the gold-standard methods in SARS-CoV-2 diagnostic testing.

3.4.1 Non-sequencing diagnostic methods

Broad methods for nucleic acid testing for SARS-CoV-2 include [152]:

PCR-based tests provide a cheap, specific method for the identification of SARS-CoV-2 specific sequences, and the equipment required is available in many laboratory and clinical settings. The most commonly used methods are reverse-transcription and real-time polymerase chain reaction (RT-PCR and qPCR). qPCR uses rounds of thermal amplification of nucleic acids measured in real time to provide quantitative or qualitative analysis. It represents the current standard for the confirmatory diagnosis of COVID-19 internationally. qPCR systems have been developed by many different companies and come in a variety of forms, ranging from large, high sample load machines to mobile, low sample units and enclosed systems which reduce the need for interactions with the sample.

Endpoint PCR (EPCR) diagnostic tests are another possible testing strategy. These tests still use reverse transcription, but instead of quantifying viral load in real time, they quantify the amplified product at the end of the assay to produce a single negative or positive measurement. It is anticipated that use of EPCR tests could increase testing speed and expand national testing capacity. The UK government recently announced they would be switching from qPCR to EPCR tests [153].

Clustered regularly interspaced short palindromic repeats (CRISPR) is best known as a key component and class of genome editing tools. During the pandemic, CRISPR-based diagnostics have been developed for the detection of SARS-CoV-2. These tests utilise guide RNAs for the accurate detection of specific nucleic acid sequences complementary to the synthesised guide RNA. CRISPR-associated cutting enzymes then carry out cleavage of a reporter molecule, providing the potential for diagnostic capability.

In May 2020 the FDA approved a COVID-19 CRISPR-based test from Sherlock BioSciences for emergency use in authorised laboratories, the first time this type of test had been approved for use in healthcare [154]. CRISPR-based diagnostic assays have several advantages, being fast, cheap, and requiring minimal equipment, making them suitable for use as point of care tests. In addition new tests can be rapidly redesigned and deployed if needed.

Loop mediated isothermal amplification (LAMP) is a method of amplification that can be utilised prior to analysis. LAMP uses isothermal amplification to provide a fast method that does not require thermal cycling, relying on one-step, lower-temperature amplification compared to qPCR, removing reliance on thermocycling equipment. There is potential for point-of-care tests to be developed on this basis. Techniques using isothermal amplification include nicking endonuclease amplification reaction (NEAR) which provides ultra-fast amplification and analysis at much lower temperatures (<10 min, temperatures of 37-42°C).

3.4.2 Test characteristics

NATs have different capabilities and characteristics by which they can be further categorised. These characteristics may be more or less desirable or necessary in differing circumstances. These characteristics include:

Point of care (POC) testing can be carried out near to or at the site of the patient and within a relatively short period of time. These are tests (sample collection, analysis and result) conducted outside the laboratory setting, allowing a test result to be obtained whilst the patient is there. These tests have several advantages including speed, which may be important when diagnosing an infectious disease and for patient peace of mind, and can be conducted by individuals outside a laboratory i.e. a broader range of staff or non-experts may be trained to perform them. Examples include the ID NOW COVID-19 test [155].

Qualitative or quantitative – Tests for SARS-CoV-2 may be qualitative, quantitative, or semi-quantitative. Qualitative tests permit the detection of SARS-CoV-2 only, whereas quantitative tests facilitate quantification of viral load. Quantitation requires careful consideration to achieve accurately [156]. Tests for the diagnosis of COVID-19 rely primarily on qualitative tests.

Enclosed testing is testing performed in a closed unit, where many of the steps following sample collection are performed within one device requiring little to no further direct processing from the user. A number of SARS-CoV-2 tests have been developed to work using platforms which allow for sample preparation, extraction, amplification and detection within an enclosed unit. This reduces the need for higher level laboratory biosafety and provides a simplified workflow e.g. SAMBA II [157, 158].

Multi-pathogen detection is possible with a number of tests. A number of previously established syndromic panels have been adapted to incorporate SARS-CoV-2 testing capacity [129]. Panels can facilitate simultaneous detection of SARS-CoV-2 and other common respiratory pathogens (bacterial or viral) which is useful for determining the cause of disease where symptoms are broad or in the case of co-infection e.g. BioFire SARS-CoV-2 Respiratory Panel [159, 160].

3.4.3 Open versus closed platform testing

A range of NATs are now available for the diagnosis of COVID-19. These include both commercially developed PCR kits, platform-specific tests and laboratory protocols using a range of different reagents and equipment, as previously discussed. During the pandemic, many commercially developed tests received emergency use authorisation (EUA) or an equivalent from various regulatory bodies.

Commercially available test kits are validated on specific equipment, some include a range of equipment and can be considered 'open platform'; for example, tests such as the US CDC's EUA-authorised RT-PCR test specify a platform on which the testing should be performed, but the FDA extends the authorisation to include several alternative platforms, where the preferred equipment is not available [143]. Other commercially available kits, such as the TaqPath COVID-19 CE-IVD RT-PCR Kit which is only validated for use on some of the ThermoFisher range of qPCR cyclers, are generally considered 'closed platform' or 'platform-specific'; these often include proprietary equipment and/or reagents and may incorporate sample processing and PCR in the same machine [161].

Alongside specific assays, a range of molecular equipment such as small form qPCR machines – such as the Mic qPCR cycler (Bio Molecular Systems), and MyGo PCR systems (IT-IS Life Science Ltd) – are being used to increase diagnostic capacity in both clinical and research settings (personal communication).

An extensive, though not comprehensive, list of available and in-development molecular tests can be found on the FIND website [162]. As of 9 February 2021, this database included 426 molecular assays, however this list is frequently updated owing to the current pace of change in the field. Frequently updated records are also published by Massachusetts General Hospital online [152] (including lists of FDA cleared NATs), and by the European Commission as part of their COVID-19 In Vitro Diagnostic Devices and Test Methods Database [163]. The FDA also maintains its own list of EUA-authorised COVID-19 diagnostics [164].

4

SARS-CoV-2
research landscape

4 SARS-CoV-2 research landscape

This section provides an overview of the research landscape on sequencing of the SARS-CoV-2 virus. Given the rapid pace of change in research, and the large volume of papers published on a weekly basis, we have not summarised large numbers of research findings, but rather provided an overview of key research areas and emerging themes in terms of results, highlighting any areas where there is not a clear consensus and indicating where future research activity can be expected.

The outbreak of a novel coronavirus, SARS-CoV-2, was identified in December 2019 and the first SARS-CoV-2 genomes were isolated and sequenced from the Hubei province of China by January 2020 [16-19]. Since then there has been a significant increase in scientific research activity around SARS-CoV-2, which has led to the publication of tens of thousands of papers in the first few months of 2020. Since then, research activity into SARS-CoV-2 and COVID-19 has increased dramatically. LitCOVID is an NIH literature resource for tracking up-to-date scientific information about the pandemic in published articles listed on the PubMed resource [165]. Additionally, the WHO has an online literature resource that has over 170,00 publications listed (as of 17 March 2021), including preprints [166].

Research groups worldwide have sequenced the genomes of SARS-CoV-2 and related coronaviruses to:

- Inform molecular epidemiology and genomic surveillance efforts
- Increase understanding of the basic biology of the virus.
- Support the development of vaccines, treatments and diagnostics
- Improve understanding of the factors influencing symptom severity in patients

Genome sequencing of patients infected with the virus, sometimes in parallel with viral sequencing, is being used to support efforts to understand patient susceptibility to disease and develop new tests, treatments and vaccines. Whilst the primary focus of this chapter is viral sequencing, some key themes arising in this area are also indicated. The viral genome sequences produced are being curated in large, online, publicly available datasets to allow secondary use of the data by other researchers (as described in Chapter 2). These pre-existing datasets are being used to:

- Gain further insights into the transmission and spread of the virus throughout the human population
- Trace the genomic lineage of the virus to uncover its zoonotic origin
- Understand the biology and pathogenicity of the virus in human and animal hosts

The research themes outlined in this chapter are not exclusive, as many studies are cross-cutting with multiple investigations being described in the same study. In addition, the substantial research efforts to better understand viral and disease biology, and the zoonotic origin of the disease via utilisation of existing datasets and/or original sequencing efforts have been outlined in Chapter 1. Many publications described here are in preprint form, so have not yet been peer reviewed.

4.1 Molecular epidemiology and genomic surveillance

Many viral sequences produced during the early phase of the pandemic were not published, instead being submitted directly to databases such as GISAID. Of those early sequencing efforts described in the literature, many describe studies with the primary function of confirming if a patient had COVID-19, usually in situations where the patient was among the first suspected cases in a country or region. These included attempts to trace the introduction and transmission of the virus throughout populations, supported by genomic data available in databases such as GISAID. Table 5 outlines some of the early sequencing efforts in China once the epidemic was established.

Table 6 contains international examples of papers describing SARS-CoV-2 genome sequencing early in the pandemic and attempts to use phylogenetic analysis to determine origin and transmission. In many cases the papers describe sequencing of the first confirmed cases in a particular country, but as mentioned above, many of these first sequencing efforts were not published. Phylogenetic analysis allows researchers to gain insights into the relatedness of SARS-CoV-2 isolates to each other and to other viruses in the coronavirus family.

There has been some debate in the scientific community around the variation in approaches and methods used in phylogenetic analysis and interpretation. In particular, the use of different in silico modelling approaches on existing datasets, used to understand the virus circulating in human populations and in comparing these viruses with related coronaviruses isolated from different species; the latter being used to elucidate the zoonotic origin of SARS-CoV-2.

Early in the pandemic, genomic surveillance efforts concentrated on generating viral sequences to understand viral evolution characteristics and transmission dynamics. As the pandemic has progressed, additional projects have been established to build on ongoing genomic surveillance efforts and optimise use of viral sequencing data.

For example, the COG-UK Hospital Onset Covid-19 Infection (HOCI) Study aims to evaluate the benefit of rapid COVID-19 genomic sequencing on infection control in preventing the spread of the virus in hospitals in the UK [89]. COG-UK are also contributing to the Genetics Of Mortality In Critical Care (GenOMICC Study) Study where host genome data will be linked with viral genome data provided by COG-UK [90]. This project aims to provide unique insights into how patient and virus genetic factors influence a patient's response to the infection.

The international effort towards generating and sharing SARS-CoV-2 genomic data has been of major scientific benefit. It has enabled the monitoring of SARS-CoV-2 evolution in nearly real time and on a global scale.

There are a number of publicly available online resources for tracking and interpreting changes in the SARS-CoV-2 genome. For example, the NextStrain website provides ongoing updates of publicly available data alongside powerful analytic and visualisation tools for use by the research community. Variants and mutations of concern can be investigated further via the associated interactive website CoVariants [167].

Another website, cov-lineages, also aims to capture and convey the genetic relationships and history of virus genomes to help understand disease epidemiology and support tracking of the virus [168]. This web interface is linked with GISAID data and when virus genomes are uploaded to GISAID they are automatically assigned a lineage. Researchers are also developing online tools to help the community track and analyse mutations arising in SARS-CoV-2 genome.

For example, Korber et al developed an open access, online bioinformatics pipeline to track changes in the SARS-CoV-2 genome, with a specific focus on the spike glycoprotein [169, 170]. The pipeline explores mutations in SARS-CoV-2 geographically and over time to identify variants that are concomitantly increasing in frequency in different geographic locations. More information and links to these data visualisation and analysis tools can be found in the addendum on SARS-CoV-2 variants [4].

Table 5: Examples of the first descriptions of SARS-CoV-2 viral isolation and sequencing in the literature from China

Paper	Details	Purpose of study	Sequencing approach	Shared data
Wu et al (2020) [17]	SARS-CoV-2 isolated from a patient who was working at the wet market implicated in the origin of the pandemic	Describing the virus in relation to other coronaviruses	Metagenomic sequencing using Illumina MiniSeq	GenBank
Zhou et al (2020) [18]	5 patients in early stage of the outbreak	Describing the virus in relation to other coronaviruses	Metagenomic sequencing using BGI MGISEQ2000 and Illumina MiSeq 3000 sequencers	GISAID, GenBank
Zhu et al (2020) [19]	4 samples from patients with pneumonia of an unknown cause	Describing the virus in relation to other coronaviruses	Metagenomic sequencing by Illumina and ONT	GISAID, GenBank
Lu et al (2020) [16]	SARS-CoV-2 genome sequencing from 9 patients, 8 of whom visited the Huanan seafood market in Wuhan	Describing the virus in relation to other coronaviruses	WGS using BGI sequencing, ONT, Illumina and Sanger	China National Microbiological Data Center and China National GenBank
Chan et al (2020) [22]	Sequenced SARS-CoV-2 from 2 patients in the same family	First description of familial transmission	WGS using ONT supplemented by Sanger sequencing	GenBank

Table 6: Examples of publications from research groups around the world describing the sequencing of SARS-CoV-2 from confirmed positive patients

Paper	Country/ region	Details	Purpose of study	Sequencing approach	Shared data
Seemann et al (2020) [47]	Australia, Victoria	Integrated epidemiological and genomic data of 903 samples	Trace transmission chains, assess impact of social restrictions	WGS using ARTIC version 1 or 3 primers with Illumina sequencing	GenBank
Caly et al (2020) [171]	Australia	SARS-CoV-2 sequencing from the first patient diagnosed with COVID 19 in Australia	Describe isolation and sequencing methods	Combination of ONT and Illumina short-read sequencing	GenBank, GISAID
Goes de Jesus et al (2020) [172]	Brazil	SARS-CoV-2 sequencing of the first 6 patients with confirmed COVID 19	Origin and local transmission of the virus	ARTIC protocols using ONT	
Wu et al (2020) [24]	China	11 SARS-CoV-2 genomes from patients in China	Determine the molecular evolution of SARS-CoV-2	WGS using a combination of Sanger, Illumina and ONT	
Böhmer et al (2020) [173]	Germany, Bavaria	SARS-CoV-2 sequencing of 15 patients in a localised outbreak	Tracing transmission events, incubation period and secondary attacks	WGS using Illumina NextSeq and MiSeq as well as RT-PCR product sequencing on ONT MinION. Gaps were filled by Sanger sequencing	
Gudbjartsson et al (2020) [68]	Iceland	Sequenced SARS-CoV-2 from 643 positive patients. This was part of the deCODE initiative	Origin and local transmission of the virus	ARTIC protocols for WGS using Illumina MiSeq	
Yadav et al (2020) [174]	India	The first 2 SARS-CoV-2 genomes sequenced in India	Determine origin of transmission in the country	WGS by Illumnia MiniSeq	GISAID
Capobianchi et al (2020) [175]	Italy	Sequencing of SARS-CoV-2 from the first patient in Italy	Determine origin of transmission in the country	Ion Torrent S5	GISAID and GenBank
Park et al (2020) [176]	Korea	Sequenced SARS-CoV-2 genome from the first patient confirmed to have COVID-19 in Korea	Relationship with other SARS-CoV-2 genomes	Illumina Nextseq 500 platform	

Table 6: Examples of publications from research groups around the world describing the sequencing of SARS-CoV-2 from confirmed positive patients

Paper	Country/ region	Details	Purpose of study	Sequencing approach	Shared data
Sah et al (2020) [177]	Nepal	Sequenced SARS-CoV-2 isolated from a Nepalese patient who contracted the virus in Wuhan	Produce sequencing data on SARS-CoV-2	Illumina MiSeq	GenBank and GISAID
Kim et al (2020) [178]	South Korea	Sequenced SARS-CoV-2 from putative patients	Determine sequence homology with other SARS-CoV-2 genomes and related species	Illumina MiSeq	GISAID
Kanteh et al (2020) [179]	The Gambia	Sequenced SARS-CoV-2 from the first six cases in The Gambia	Determine origin of SARS-CoV-2 in the country	WGS via Illumina MiSeq and ONT GridION	GISAID
Meredith et al (2020) [180]	UK, East of England	Sequenced SARS-CoV-2 in healthcare associated COVID-19 infections. They integrated epidemiological and clinical data along with genome sequences	Investigated the utility of rapid viral sequencing to inform infection control measures within a hospital environment	Sequenced by ONT with validation by Illumina sequencing. Primers were designed by ARTIC network V3 protocol and analysed by the ARTIC network assembly pipeline	
Salazar et al (2020) [181]	Uruguay, Montevideo	Sequenced SARS-CoV-2 from 10 positive samples (around 10% of the confirmed cases) in the first week	Determine origin of SARS-CoV-2 in the country	ONT MinION	
Kujawski et al (2020) [182]	USA	The COVID-19 Investigation Team sequenced SARS-CoV-2 from the first 12 patients with confirmed COVID 19 in the USA	Determine the natural history of SARS-CoV-2	WGS using ONT and Sanger sequencing	GISAID and GenBank
Gonzalez-Reiche et al (2020) [183]	USA, New York	The COVID-19 sequenced 90 virus isolates from 84 patients in New York city	Original and local transmission of the virus	ARTIC primers, sequenced by Illumina MiSeq and PacBio	GISAID
Bedford et al (2020) [184]	USA, Washington State	Sequenced 346 SARS-CoV-2 genomes from positive patients	Origin and local transmission of the virus	Metagenomic analysis using Illumina MiSeq or NextSeq	GISAID and GenBank

All of these papers sequenced the entire SARS-CoV-2 genome and performed phylogenetic analysis with the genomes they sequenced alongside other genomes from GISAID and the reference SARS-CoV-2 from genome from GenBank. The information provided in the table was sourced from the publications.

4.1.1 Environmental sequencing

Examples of sequencing of environmental samples to inform molecular epidemiology include samples taken from surfaces of the Wuhan wet markets implicated in the origin of the outbreak. Analysis revealed high sequence similarity to the first confirmed cases in Wuhan patients [185]. Whether the samples tested originated from infected humans or animals remains unknown, though analysis of the genomic data by Zhan et al [186] led the authors to conclude that the virus isolated from the wet market was likely to be of human origin due to the high sequence similarity with that seen in patients.

Other environmental sequencing efforts include the sequencing of virus isolates found in wastewater and rivers. The first report of analysing wastewater for the virus was by Lodder et al [187] who used RT-PCR to confirm presence of SARS-CoV-2. Rimoldi et al [188] also used RT-PCR to confirm the presence of SARS-CoV-2 in waste water and rivers in Italy but also sequenced the virus found in the samples to perform phylogenetic analysis. These efforts are part of wastewater-based epidemiology approaches to help identify outbreaks and estimate infection prevalence in the population.

4.1.2 Lineages, strains and variants

There was debate in the scientific community early in the pandemic around whether a more virulent or transmissible strain of the virus had emerged. A paper by Tang et al [185] proposed that SARS-CoV-2 could be split into two main lineages, defined by two single nucleotide polymorphisms. The group analysed 103 SARS-CoV-2 genome sequences alongside several related species from online databases and concluded that there were two distinct strains of SARS-CoV-2 circulating. The two strains were dubbed S and L, with S supposedly arising from an ancestral lineage and L having evolved from the former. In the preprint version of the paper they speculated that the L strain of SARS-CoV-2 was more transmissible than S. These claims were removed in the final published version of the paper, however they concluded that their analysis found the L lineage was more prevalent in the samples they tested and advised further investigations were needed to determine the effect on transmission and virulence.

The paper prompted strong debate, highlighting the disagreement around what constitutes a new strain, with some virologists adamant that there must be biological differences resulting from the genetic difference to justify designation as a different strain. For example, MacLean et al [189] highly critiqued the paper by Tang et al suggesting their claims that there were two strains of the virus were unsubstantiated due to the difficulty in demonstrating a functional impact of viral mutation, i.e. a change in the biology of the virus that means a new sub-type or strain has emerged. MacLean et al. also discredited any claims that one strain is more transmissible than other due to the founder effect.

Researchers at Fundan University in Shanghai sequenced 112 viral genomes and used phylogenetic analysis to determine that although there were two major lineages of the virus with differential exposure histories in the early stages of the pandemic, there was no difference in virulence and clinical outcomes in the population they sampled [26].

Another notable paper by University of Cambridge researchers [190] used GISAID data in a phylogenetic network analysis of 160 complete SARS-CoV-2 genomes isolated from humans. Their conclusions were that there are three distinct variants distinguished by amino acid changes in viral proteins. Others critiqued the study, for example Mavian et al [191] and MacLean et al [189] whose main objections are that the use of the bat coronavirus as a root was erroneous and the functional impact of the changes were not explored.

The debate continued with several papers in preprint suggesting there was more than one strain, or that SARS-CoV-2 was mutating sufficiently to confer differences in its biology and hence its infectivity.

For example, a paper by Korber et al [44] suggested that mutations in the spike protein, in particular D614G, may confer enhanced transmission. The frequency of this mutation was analysed by another team in the UK, which found no evidence of the emergence of a more transmissible strain [192]. However, the impact of the D614G on infectivity and viral load is still under investigation.

From December 2020, however, several variants of concern (VOCs) that alter virus phenotype were identified and are the subject of ongoing investigation – the role of genomic surveillance in identifying these variants and their impact on currently available diagnostic tests, public health measures and vaccines are covered in the addendum to this report [4].

There are discussions and ambiguities around the terminology and naming systems used to describe the genetic diversity of SARS-CoV-2. The terms 'variant', 'strain' and 'lineage' have been used somewhat interchangeably. Though the term variant has prevailed there is some debate about how appropriate the use of the term variant is with some suggesting that a description of the 'constellation' of mutations is more important to define than the variants themselves.

Several naming systems for variants, particularly VOCs, have been proposed. Public health authorities have taken to naming VOCs by the date of their identification. For example, Public Health England dubbed the variant they identified that was enhancing transmission, Variant of Concern (VOC) 202012/01, after the date it was first identified (this is also known as VOC B.1.1.7).

Researchers in South Africa who identified a VOC named it after a defining mutation at the 501st amino acid site, 501Y.V2 [193]. The V2 part of the name pertains to it being the second variant identified with the mutation at the 501st amino acid site, the first variant with this mutation being B.1.1.7. However, naming a variant after a mutation might cause confusion since it is not always clear when first identified which mutations may be of concern, and it is possible that a combination of several mutations is more important in terms of defining the VOC.

A naming system proposed by researchers at NextStrain uses a year and letter system to define different SARS-CoV-2 clades [194]. In April 2020, Rambaut et al proposed a lineage naming system for SARS-CoV-2 [195]. This naming system is currently the most commonly used when referring to current known variants and VOCs e.g. B.1.1.7, B.1.351 and P.1. This is the naming convention followed in this report and the addendum on VOCs [4]. Discussions initiated by the WHO Virus Evolution Working Group are underway to produce a standardised nomenclature for SARS-CoV-2 [196].

On 25 February 2021, the WHO released a document outlining working definitions of VOCs and variants of interest (VOI), including recommended actions for member states if a VOI or VOC is identified [197].

4.2 Host genomics

There are dozens of studies ongoing globally collecting data on how host genomics relates to disease severity. Sequencing efforts are being applied to understand the variation in response to infection, interactions between host and viral genomics, and pathogenicity of the virus in humans. A useful interactive map of these many trials, projects and consortia are detailed on the COVID-19 Host Genetics Initiative (HGI) partners page [198].

The COVID-19 HGI brings together the network of human genetics researchers from numerous international groups to share and analyse sequencing data, with the aim to better understand the link between host genetics and severity of disease.

The latest GWAS results from the COVID-19 HGI identified regions in seven different chromosomes regions associated with patients experiencing severe COVID-19 [199]. These include regions in chromosomes 3, 6, 9, 12, 19, and 21 that harbour genes that regulate immunity or are associated with lung diseases. Three of these were identified by the Genetics of Mortality in Critical Care (GenOMICC) study [200].

Multiple groups are looking at using single cell sequencing of host cells to understand the infectivity of SARS-CoV-2 in humans [201-204]. Zhao et al [204] used single cell sequencing to determine ACE2 expression in normal human lungs and found that it was expressed in high concentration in type 2 alveolar cells. They also found other genes being expressed that positively regulate viral entry, reproduction and transmission. Another notable study by researchers in the HCA Lung Biological Network analysed the expression of viral entry associated genes from multiple tissues from healthy human donors [202].

Others are using host transcriptomics to determine host immune response to infection with SARS-CoV-2. For example, researchers at Wuhan University used sequencing to examine the transcriptome of bronchoalveolar lavage fluid and peripheral blood mononuclear cells from three healthy patients (free from COVID-19 infection) and three patients with confirmed COVID-19 [205]. Results from this study revealed a distinct inflammatory cytokine profile in patients infected with SARS-CoV-2, along with increased cell death in lymphocytes, helping to explain the cause of patients' lymphopenia (low levels of lymphocytes in the blood).

There are various projects worldwide that are investigating the interaction between the host and viral genetic factors – 56 are listed on the COVID-19 HGI. Data collection is still underway for the majority of these studies and results are not yet published. In addition, sequencing the microbiome of a collection site (e.g. nasopharyngeal) could provide insights into the underlying causes of severe disease and may also be useful to direct clinical management by uncovering primary and secondary infections [206, 207].

4.3 Development of vaccines and treatments

Researchers are utilising the large sequencing datasets produced by international sequencing efforts for in silico analysis to aid the development of vaccines and treatments.

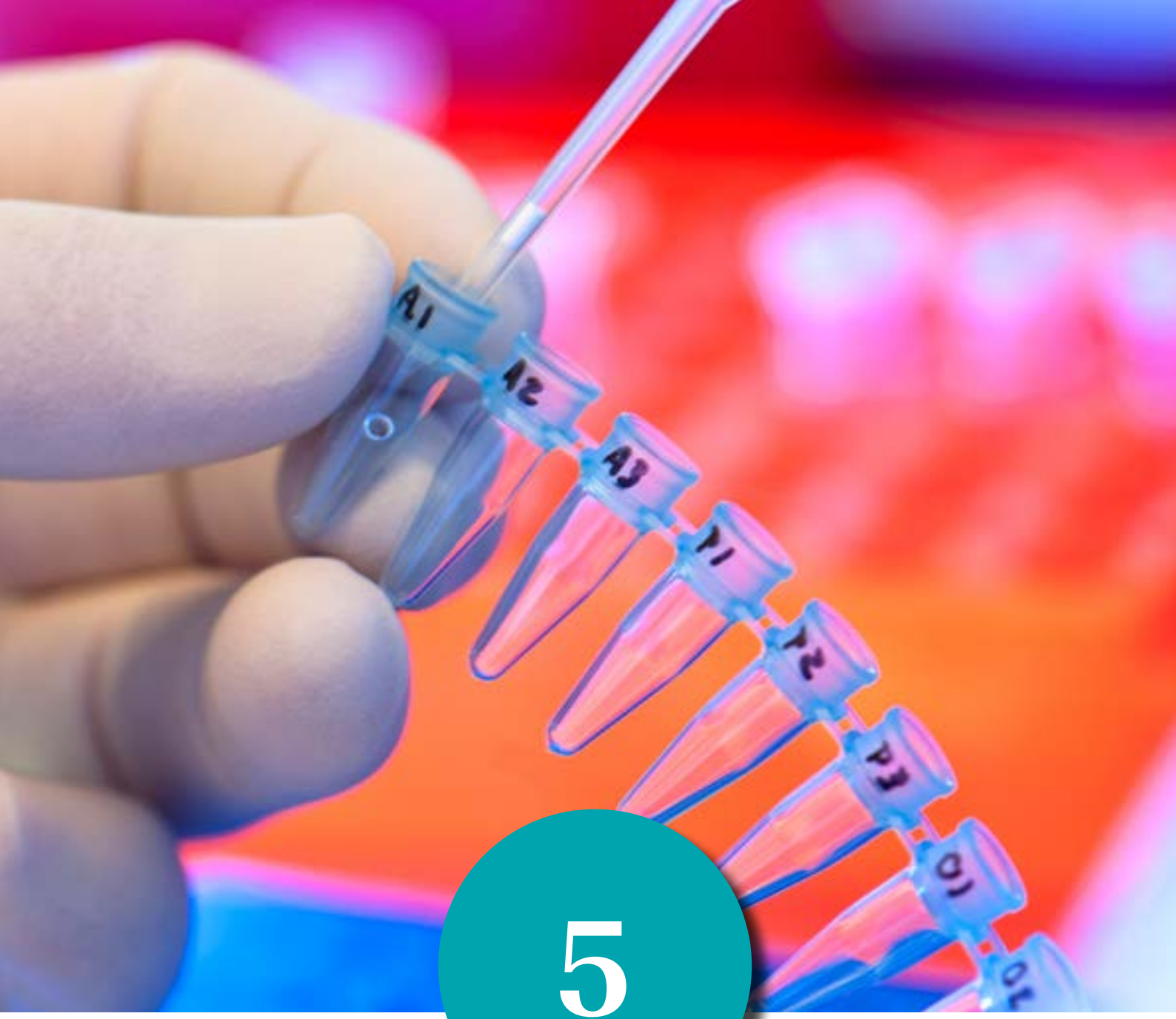
Understanding the human immune response to SARS-CoV-2 including the potential B and T cell epitopes is an area of increased attention due to the insights offered for vaccine and serological diagnostic design. For example, researchers have used the genetic similarity between SARS-CoV-2 and other coronaviruses, such as SARS-CoV, to leverage data produced from immunological studies on these viruses [208, 209].

Zhou et al [210] used available viral sequencing data in their analysis to uncover potential drug repurposing to treat COVID-19. Meanwhile, researchers from the University of Melbourne have developed an online resource, COVID-3D, to explore the structural distribution of genetic variation in SARS-CoV-2 and its implication on therapeutic development [211].

As described in Chapter 1, there have been various investigations into which regions of the viral genome are undergoing higher rates of mutation compared with more conserved regions. The results from these analyses have relevance to vaccine development and therapeutics. For example, Pachetti et al analysed available virus genomes to determine mutational hotspots and found relatively high mutation in the RdRp gene [43].

This is important as several treatments target the RdRp enzyme so the authors advise careful monitoring to detect potential resistance that could arise to treatments. Additionally, many monoclonal antibody therapies have been found to be ineffective against variants of SARS-CoV-2 carrying specific mutations [212].

Sequencing has proven to be not only essential in the design and development of vaccines but also in monitoring the impact of viral genetic variation on their efficacy. There are currently eleven vaccines approved for use to protect against SARS-CoV-2 [213]. Section five of the addendum to this report provides further information including on the impact of VOCs on vaccines [4].



5

Sequencing
technologies

5 Sequencing technologies

Sequencing is the process by which the nature and order of nucleic acids in a sample are converted into data that can then be analysed. There exist a number of different ways that this process can be achieved and a number of different technologies that can facilitate it.

5.1 Global sequencing efforts and the role of sequencing technology developers

As we described in Chapter 2, the emergence of COVID-19 and subsequent global spread has seen the establishment of public-private partnerships (some as consortia), both formal and informal, across the globe. The partnerships include technology companies responsible for the development and provision of sequencing systems. All major sequencing companies have been involved in global SARS-CoV-2 sequencing efforts (Section 5.3), and many have built on already established relationships with national and international health organisations to expedite use of their technologies during the pandemic. SARS-CoV-2 products (table 7) available from each are discussed in more detail later on in the chapter.

Table 7: Examples of SARS-CoV-2 sequencing-related products developed or in development by major sequencing technology companies

**RUO = Research Use Only*

Company	Product	Platform	Sequencing approach	Status
Illumina	COVIDSeq	NovaSeq Can also be performed on the NextSeq for RUO*	SARS-CoV-2 specific targeted amplicon	Emergency Use Authorisation by FDA
	AmpliSeq for Illumina SARS-CoV-2 Research Panel	iSeq 100, MiSeq, MiniSeq, and NextSeq systems	SARS-CoV-2 whole genome amplicon	For use in research
	SARS-CoV-2 NGS Data Toolkit	Various	Various	For use in research
Thermo Fisher Scientific	Ion AmpliSeq SARS-CoV-2 Research Panel	Ion Torrent Genexus, Genestudio S5	SARS-CoV-2 whole genome amplicon	For use in research and in development for Genexus system
BGI & MGI Tech	SARS-CoV-2 testing for surveillance	DNBSEQ-T7	Various	Emergency use approval by NMPA Available for research in the EU through Ares Genetics
Oxford Nanopore Technologies	LamPORE diagnostic assay	MinION GridION	SARS-CoV-2 targeted amplicon	CE marked for in vitro diagnostic use. Further regulatory approvals being pursued in other countries

5.2 Approaches to sequencing

Alongside the different technologies that can be used for viral genome sequencing, there are several approaches to sequencing viral samples. Use of these is determined based on the project or experiment aims and resources available. These involve the use of different sample and library preparation techniques prior to actual sequencing, to retrieve a different depth, scale or type of sequence information. Changes in bioinformatics pipelines often accompany these choices. The major approaches are discussed below.

5.2.1 Techniques

Amplicon sequencing – amplicon sequencing relies on the prior amplification of regions of the viral genome in order that these regions exist in greater concentration in the processed pre-sequencing sample than the original raw sample. This allows for deep sequencing of these regions (where the region is sequenced repeatedly). This approach is often used to detect rare variants that may be present in only a small proportion of the genomes in a sample, and when conducting surveillance or species identification studies. Amplicon-based sequencing can be used across the whole viral genome or selected parts of it dependent upon primer design. The ARTIC network and others have shared primer designs and protocols for amplicon-based sequencing of SARS-CoV-2 which have been widely used during the COVID-19 pandemic.

Bait capture/target enrichment – this technique involves nucleic acid capture through the use of 'bait' molecules, facilitating the selection and targeting of specific regions of the genome. Several companies have produced target enrichment panels for SARS-CoV-2 sequencing. For example, Arbor Biosciences has produced a hybridisation capture kit which it has offered to researchers for free and is compatible with both short- and long-read sequencing systems. It has been developed using available genome sequences in the NCBI-SRA database. Bait capture is another method of isolating viral sequence from contaminating sequences (such as human RNA), or other viral sequences which may also be present in clinically derived samples.

Direct-RNA sequencing – the direct sequencing of RNA molecules removes the need for amplification and conversion of RNA into cDNA. Direct RNA sequencing can only be performed using nanopore sequencing techniques and has both advantages and disadvantages. Removal of an amplification steps means that some bias that would normally be introduced through PCR is avoided, potentially reducing error. It also allows base modifications (such as methylation) to be detected on the unconverted molecule. However, in order to achieve sufficient coverage of the viral genome, a high quality and RNA-rich sample is required.

5.2.2 Genomic target

Whole genome sequencing – whole genome sequencing approaches aim to retrieve as much of the viral genome as possible. The proportion of the viral genome for which genomic reads have been retrieved to an adequate depth (enough sequencing reads from a particular region to expect bases to be called accurately) is often expressed as a percentage e.g. >98% coverage. Amplicon, target enrichment and direct RNA sequencing can all be used to undertake whole genome sequencing of the SARS-CoV-2 genome.

Targeted sequencing – Targeted sequencing may be used to explore specific regions of the genome, when whole genome sequencing is not required, or not possible. It is most suited to use in sequencing based diagnostics, where only a few specific regions of the viral genome need to be sequenced to determine if the virus is present (see Chapter 3).

Metagenomic sequencing – the metagenome includes all nucleic acids present in a sample at one time – this may be restricted to nucleic acids of one type i.e. RNA or DNA. It may involve amplification of shared target 'barcoding' regions such as 16S region in bacteria or the internal transcribed spacer (ITS) region in fungi, which enable discrimination between different organisms whilst sequencing only a selected portion of the genome of each. Alternatively, it may rely on 'shotgun sequencing' whereby all available regions of all nucleic acids are targeted. Metagenomic sequencing in the context of SARS-CoV-2 represents the sequencing of all RNA (and sometimes DNA) available in a sample. It can be especially useful following the emergence of a new pathogen, for which there is little genomic information, meaning effective targeting is not possible. The approach was heavily used at the start of the COVID-19 outbreak in China to identify and increase knowledge of the SARS-CoV-2 genome. It is also useful where co-infection is possible, enabling the identification of multiple pathogens from a single sample.

5.3 Sequencing systems

A number of different sequencing platforms are being used to sequence the SARS-CoV-2 genome. These include broad categories of sequencing technologies, often referred to in generations: first (Sanger and others), second (high-throughput) and third (long-read) generation sequencing technologies. The term 'next generation sequencing' (NGS) is used to refer to sequencing techniques and technologies belonging to the second and third generations. These techniques include:

- Sequencing by synthesis (Illumina)
- Ion torrent semiconductor sequencing (Thermo Fisher Scientific)
- DNA nanoball sequencing (Beijing Genomics Institute and MGI Tech)
- Single molecule real-time sequencing (Pacific Biosciences)
- Nanopore sequencing (Oxford Nanopore Technologies)

Each of the above techniques, which are broadly aligned to specific companies, can be performed on a range of instruments produced by the above-named developers.

A number of other NGS platforms and techniques exist but are not listed above. These include instruments such as Roche 454 sequencing platforms (pyrosequencing) and SOLiD sequencing systems from Thermo Fisher Scientific (sequencing by ligation). These are no longer sold by the manufacturer but are still used in some research settings. These systems are less widely used than the above key systems and have not been extensively employed for the sequencing of SARS-CoV-2.

Lesser known technologies such as BioelectronSeq 4000 produced by CapitalBio have also been used to sequence the SARS-CoV-2 genome, but to a much lesser extent than the above key systems. Sanger sequencing has been used during the COVID-19 pandemic to increase understanding of the virus in combination with NGS sequencing technologies. For this reason, a brief overview of Sanger sequencing is included in this chapter.

Sequencing may either be provided as a service through a provider, or the equipment may be purchased by research groups and institutions with systems support provided by the developer. Many producers operate a mixed model.

5.3.1 *Illumina sequencing by synthesis*

Sequencing by synthesis represents the basis of the most widely used NGS methods. Sequencing using Illumina systems provides high throughput short-read sequencing and is widely used.

Sequencing by synthesis nucleotide identification occurs as modified nucleotides are incorporated into newly forming DNA. Fluorescently tagged (modified) bases are detected as they are incorporated. Unlike early chain-termination methods (Sanger sequencing) fluorescently tagged bases do not cause DNA synthesis to stop. Each time a base is incorporated, the attached fluorescent tags are washed away after detection, allowing for more modified bases to be added after this point. The process is repeated until the maximum number of cycles (and therefore sequence length) is reached.

This method requires the prior conversion of RNA molecules into cDNA (performed during library preparation) for bridge amplification, which is a key step in the process, to take place.

Illumina produces a range of platforms which cover a large range of sequencing applications; these differ in size, capacity, and cost. Illumina's key sequencing platforms are described below (table 8). In addition to those listed below, several other systems are or have been available to purchase from Illumina, including a large range of HiSeq systems and related systems utilising different flow cell capacities (e.g. NovaSeq 5000 and 6000).

Illumina NGS sequencing has several advantages, primary amongst these is high throughput. The highest throughput system, the NovaSeq, can produce around six terabytes of sequence data per run, on a par with MGI's highest output system. The NovaSeq is both very large and heavy and must be supported by reinforced floors. System cooling is required whilst the sequencer is active.

Table 8: Information on Illumina sequencing platforms

Sourced from illumina.com

	iSeq 100	MiniSeq	MiSeq	NextSeq 550	NextSeq 2000	NovaSeq 6000
Run time	9.5–19 hr	4–24 hr	4–55 hr	12–30 hr	24-48 hr	13 - 44 hr (flow cell dependent)
Maximum sequence data output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	300 Gb	6Tb
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 x 250bp
Description	Lower cost, lower capacity, smallest form benchtop sequencer available from Illumina. Released 2018.	Lower cost, lower capacity benchtop sequencer. Released 2016.	Mid-range benchtop sequencer providing longest reads available on Illumina platforms. Released 2011.	Mid-range benchtop sequencer, providing greater throughput than cheaper devices. The related NextSeq 550 Dx is diagnostic tool for specific clinical applications only [214]	Higher cost, high capacity benchtop system. Released 2020. As with the smaller capacity NextSeq, a Dx version of this machine is also available.	Highest and massive capacity, very large form free-standing system. Capable of sequencing multiple human genomes in one run. Released 2017.

Use of Illumina technologies and available tools for SARS-CoV-2 sequencing

Illumina systems are used across an extremely broad range of sequencing applications, including pathogen sequencing, and have featured prominently in SARS-CoV-2 sequencing efforts (figure 4).

AmpliSeq – Illumina have developed a custom panel through its AmpliSeq system designed for the efficient sequencing of the SARS-CoV-2 genome. The panel contains 247 amplicons providing more than 99% coverage of the genome [215].

NGS Data Toolkit – a recently released SARS-CoV-2 NGS Data Toolkit, which is a collection of previously developed apps: DRAGEN RNA Pathogen Detection App, DRAGEN Metagenomics App, (NCBI-) SRA Import App, and GISAID Submission App. The apps have been made available free of charge but are only available through Illumina's data management portal, BaseSpace. The aim of this toolkit is to provide integrated workflows for consistent and simple analysis of SARS-CoV-2 sequence data [216, 217].

COVIDSeq – this was the first NGS-based test to receive FDA approval for the diagnosis of COVID-19. The test utilises respiratory samples and can allow for the collection of both diagnostic and further genomic information from up to just over 3,000 samples per sequencing run (the maximum is highly dependent upon which platform is used). This test is further discussed in Chapter 3 and Chapter 6.

Illumina technology is being used in the COVID-19 pandemic in over 10,000 laboratories across 155 countries. Technologies and protocols for the examination of the SARS-CoV-2 genome include those for shotgun metagenomics to detect and characterise coronavirus variants, and targeted sequencing panels to detect the virus and genomic mutations at a higher resolution [218].

An Africa CDC-Illumina partnership was formed on 27 April 2020, under which Illumina donated sequencing systems and related consumables to the African Union Commission through the Africa CDC to strengthen SARS-CoV-2 sequencing capabilities and capacity in 10 African countries [219].

Advantages of Illumina sequencing

- Comparatively low-cost sequencing at high throughput, where many samples are being processed
- SARS-CoV-2 specific protocols and tools are available
- High accuracy
- One of the more commonly used systems for high resolution genomic analysis - used in around 155 countries – meaning collaborative development of expertise and advancements is easily possible, many genetic or research laboratories may already possess these systems, and bioinformatics pipelines are relatively well-established
- High levels of sample multiplexing are possible, meaning a high number of samples can be run at once

Limitations

- Longer time for sequencing runs
- Most platforms are large and costly to purchase, some require specialised infrastructure for safe use
- Relatively short reads may limit accuracy in certain genomic regions

5.3.2 Ion Torrent sequencing

Thermo Fisher Scientific supplies semi-conductor sequencing through its Ion range of systems. Sequencing utilises a semi-conductor chip and a bead-based system which provide a physical platform for sequencing through DNA synthesis. This system relies on the detection of hydrogen ions released as nucleotides are incorporated.

The Ion GeneStudio S5 is the newest and highest capacity NGS system in this range. Older systems include the Ion Proton system and Ion Personal Genome Machine (PGM). The Ion Torrent Genexus, a new automated 'specimen to report' system providing rapid sequencing with reduced sample contact, became available in late 2020 [220]. A number of sequencers in the Ion range are described below (table 9).

Table 9: Information on Ion sequencing platforms

Sourced from thermofisher.com

	Ion PGM system + Ion 318 chip	Ion Proton system + PI chip	Ion GeneStudio S5 System + Ion 540 chip	Ion GeneStudio S5 Prime System + Ion 550 chips	Ion Genexus
Total sequencing and analysis time at max. throughput	7.2 hr (400bp)	4 hr	19 hr	6.5 hr	24 hours
Max. throughput/day	2Gb	10Gb	15 Gb	50 Gb	Unknown
Read length	400bp	200bp	200bp	200bp	Unknown
Description	Large benchtop sequencer providing the longest reads available from the Ion sequencing range. Older system with low throughput.	Fast, modest throughput benchtop sequencer.	Lowest capacity and throughput of the Ion GeneStudio range of sequencers. Low capacity in comparison to many other technologies. Small to medium profile desktop machine.	Fast and highest capacity Ion sequencing system.	Specimen to report automated system, claiming 10 minutes of 'hands-on' time. Permits variable throughput.

Use of Ion Torrent technologies and available tools for SARS-CoV-2 sequencing

Ion AmpliSeq SARS-CoV-2 Research Panel - Thermo Fisher has developed the Ion AmpliSeq SARS-CoV-2 Research Panel for use with the GeneStudio S5 system. This is a targeted NGS panel which covers more than 99% of the SARS-CoV-2 genome and facilitates sequencing of up to 80 samples per chip.

Although less commonly used than Illumina sequencing, Ion sequencing has also previously been used for sequencing of viral genomes during outbreaks. This includes recent applications in outbreaks of Ebola virus disease [1]. Ion systems have been used during the COVID-19 pandemic to facilitate viral sequencing, sometimes in combination with other sequencing technologies, including some of the first descriptions of the SARS-CoV-2 genome in Italy [175].

On 7 May 2020, Thermo Fisher Scientific announced a SARS-CoV-2 Global Access Sequencing Program [221]. Under this programme, the company will provide 50 units of the Ion Torrent Genexus System at a subsidised price to support global collaborative COVID-19 research. They claim the Genexus System is the world's first easy to use and automated NGS solution that is designed to deliver results in a day with five minutes of hands-on time in a decentralised laboratory setting. Launched in November 2019, the company's newest sequencing platform automates the targeted NGS workflow and can deliver results in around 14 hours. The platform can be combined with the Ion AmpliSeq SARS-CoV-2 Research Panel to provide a specimen-to-report workflow. Optimisation and validation of the research panel on the Genexus System is underway in collaboration with Thermo Fisher customers [222].

Advantages of Ion Torrent sequencing

- Relatively inexpensive at low throughput
- Comparatively quick sequencing runs enable faster return of results
- SARS-CoV-2 specific protocols are available
- Low substitution error rate
- More commonly used sequencing technology in some countries
- Some systems facilitate a highly automated workflow for easy adoption and consistent application of sequencing

Limitations

- Lower throughput in comparison to other NGS technologies, therefore comparatively expensive at high throughput
- Shorter reads than are possible with other NGS technologies

5.3.3 BGI and MGI Tech DNA Nanoball sequencing

DNA Nanoball sequencing (DNBSEQ) platforms produced by MGI Tech are available through the Beijing Genomics Institute (BGI). MGI's proprietary DNBSEQ technology enables flexible, high throughput, short-read sequencing performed on one of a range of instruments.

DNBSEQ utilises circularised reads which are repeatedly amplified using rolling consensus amplification to create a single long strand of DNA. A barcode and primers, which also enable circularisation, are attached to the target sequence during library preparation. The sequence is then massively amplified forming what are known as DNA 'nanoballs' (DNBs). The DNBs are then loaded onto a flow cell with embedded wells which facilitate detection of nucleotide integration through light detection in a similar manner to sequencing by synthesis. The repeated nanoball sequence is read to generate high accuracy consensus sequence data.

Table 10: Information on BGI and MGI Tech sequencing platforms

Sourced from bgi.com

	DNBSEQ-G50	DNBSEQ-G400	DNBSEQ-T7
Run time – dependent upon several different parameters	10-66 hr	13-109 hr	20-24 hr (including loading and base calling)
Max. throughput/run	150 Gb	1.44Tb	6Tb
Max. read length	2 x 150bp	2 x 200bp	2 x 150
Description	Medium throughput desktop system, suitable for smaller genome sequencing and panels	Very high throughput, flexible, desktop sequencing system	Massive throughput, fast sequencing system. Very large form factor. The latest system to be released from MGI

Use of BGI and MGI Technologies and available tools for SARS-CoV-2 sequencing

DNBSEQ has been used for the sequencing of the SARS-CoV-2 genome, primarily in East Asia and from an early stage of the outbreak [18, 53]. Additional collaborations have been established with partners in Europe and other parts of the world. Recent applications have involved the combined use of DNBSEQ and other NGS techniques for the examination of the viral transcriptome [53].

The BGI Group subsidiary MGI Tech produces a range of sequencing equipment and associated products (table 10). By April 2020, BGI was providing real-time non-sequencing SARS-CoV-2 tests to over 80 countries [223]. Based in China, they were one of the first technology companies to respond when the virus emerged. Within a week (by 5 February 2020) a 2000 square meter laboratory was established, providing qPCR and sequencing testing in Wuhan [224, 225].

DNBSEQ-T7 sequencing platform - In late January 2020, the DNBSEQ-T7 received emergency use approval from the Chinese National Medical Products Administration (NMPA), for the surveillance, discovery and identification of unknown infectious diseases, including SARS-CoV-2 [226]. The sequencer and analysis software now have NMPA certification as a Class III medical device to support future epidemic prevention and control [227].

BGI & MGI Tech will reportedly work with Ares Genetics, their long-term strategic partner in Europe, and the Curetis Group to make a testing portfolio for SARS-CoV-2 available to public health institutions and hospitals for outbreak monitoring, infection control, and epidemiology. Ares Genetics will provide NGS services for SARS-CoV-2 out of its NGS laboratory in Vienna, Austria, for infection control and tracking pathogen evolution from February 2020 based on MGI's DNBSEQ sequencing platform [227, 228]. Researchers at BGI are maintaining an online resource integrating the genomic and proteomic data, and associated metadata from multiple databases. One key feature is the 2019 Novel Coronavirus Resource from the China National Center for Bioinformation (2019nCoVR, https://bigd.big.ac.cn/ncov) [229].

Advantages of BGI and MGI technologies

- Flexible sequencing including range of run times, reads lengths and output
- High throughput
- Linear amplification reduces error accumulation during amplification

Limitations

- Shorter reads than are possible with other NGS technologies
- Highest throughput systems are very large

5.3.4 Long read sequencing

Long read single molecule sequencers use distinct base technologies to read longer contiguous strands of DNA than other NGS sequencing platforms. Reads of 10,000 – 100,000 base pairs in length are produced, with the potential for molecules of more than 100,000 base pairs to be read contiguously. The two main providers of non-synthetic 'true' long read technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT).

Alongside advantages and disadvantages associated with the specific sequencing platforms, there are inherent advantages to producing longer sequence reads. As read lengths increase, it is more likely that a read will be distinct from other reads. This allows them to be computationally reassembled with less ambiguity. This is particularly useful for sequencing of highly polymorphic or highly repetitive genomes. Some systems also offer the option to perform amplification-free sequencing, potentially removing some amplification bias and facilitating the examination of epigenetic modifications.

5.3.5 Pacific Biosciences

Pacific Biosciences' (PacBio) sequencing technologies use single molecule real-time (SMRT) sequencing to produce high accuracy long-read sequence data.

Target nucleic acid molecules are individually immobilised in microscopic pits called zero-mode wave guides (ZMWs) embedded in SMRT flow cells. The many embedded ZMWs contain a fixed DNA polymerase and are open to a pool of free labelled nucleotides. Sequence information is collected through the high-precision detection of miniscule emissions of light which are produced as complementary nucleotides are incorporated into the target molecule.

PacBio offer SMRT sequencing on their newer Sequel II system and older Sequel system (table 11). The Sequel II system enables high-throughput, high fidelity sequencing, with accuracy reported to exceed 99% and output of around 160GB per SMRT cell. PacBio's systems offer high fidelity (HiFi) sequencing, achieving higher accuracy than other long read systems.

PacBio sequencing is conducted through sequencing service providers. A list of certified service providers is available through the PacBio website [230]. Established service providers exist in many countries including the USA, Saudi Arabia, and South Korea for example.

SMRT sequencing has been used in research predominantly to aid in the assembly of whole genome sequences but has also been used to investigate microbial genomes.

Table 11: Information on PacBio sequencing platforms

Sourced from PacB.com

	Sequel System + 1M SMRT cell	Sequel II System + 8M SMRT cell
Max run time per SMRT Cell	20 hrs	30 hrs
Average read length	Up to 30kb	Up to 15kb (high fidelity reads)
Max output per SMRT cell	20Gb	160Gb
Description	Large, mid-range long read sequencing platform capable of producing very high accuracy reads	High cost, large high-throughput long read sequencing platform capable of producing very high accuracy reads

Use of PacBio technologies and available tools for SARS-CoV-2 sequencing

Protocols for using PacBio systems for investigating the virus are available, having been developed through customer collaboration and can be found through the PacBio website [231].

PacBio announced on 8 April 2020 that it would be working with commercial, academic and government research teams that are investigating SARS-CoV-2 [232]. LabCorp, UC San Diego and the National Institute of Allergy and Infectious Diseases (NIAID) are among the organisations utilising PacBio's long-read sequencing technology to study SARS-CoV-2 and the related immune response in patients.

LabCorp, a healthcare laboratory diagnostics group in the USA, is supporting the response to COVID-19 in the US and globally through its diagnostics and drug development businesses, launching its internally developed molecular test on 5 March 2020. LabCorp will work closely with PacBio to sequence a large number of SARS-CoV-2 viruses from de-identified positive samples, the goal of which is to look for rare variants using longer amplicons to understand regional and temporal differences in the virus.

Scientists at the NIAID Vaccine Research Center are using the Sequel® II System to study samples collected from infected individuals [233]. A number of studies are using PacBio technology including a longitudinal study of individual patients from hospital admission until after discharge [234].

Advantages of PacBio sequencing

- Capable of high throughput, equivalent to that of Illumina sequencing platforms
- Capable of producing very high accuracy consensus reads – HiFi sequencing reads around 15,000 bases in length at over 99% accuracy
- Produces long reads
- Errors are random, not systematic, and can be overcome with deeper sequencing
- Sequences read in real-time – allowing for termination when user determines enough reads have been generated

Limitations

- Relatively expensive to run compared to other NGS technologies
- Systems are large and more costly than some alternatives

5.3.6 *Oxford Nanopore Technologies*

Oxford Nanopore Technologies (ONT) is a UK-based company that produces a range of sequencing systems based on nanopores. ONT's systems are designed to be relatively mobile, generate ultra-long reads and be more accessible to those with less experience and expertise. The systems are relatively low cost and are provided primarily through equipment purchase and customer support.

Extracted nucleic acids are prepared for sequencing by ligation of a motor protein and adapter sequence at the ends of each strand. RNA may either be amplified and converted to cDNA prior to sequencing as occurs with other sequencing systems, or the RNA can be read directly (direct RNA sequencing) without prior amplification or conversion to cDNA.

During sequencing, tagged, single stranded DNA or RNA molecules are fed through a membrane-bound protein pore – a 'nanopore' – by a motor protein. As each DNA or RNA nucleotide is fed through the nanopore, it interrupts the electrical current that exists across the pore and these signals are detected by the sequencing system. This pattern of disruption can be read to determine the base sequence of the molecule.

Much like PacBio, ONT produces systems capable of producing very long sequencing reads, up to around 800,000 bases in length. Sequencing is also fairly rapid: one nucleic acid can be read by each pore at a given moment, and each molecule is read at approximately 400 bases per second.

ONT produce a range of sequencing systems (table12) which provide different capacity, throughput and mobility, and cover a wide range of price points.

Table 12: Information on ONT sequencing platforms

Sourced from nanoporetech.com

	Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Run time – flexible dependent upon data required	1 min - 16 hrs	1 min - 48 hrs	1 min - 48 hrs	1 min - 48 hrs	1 min - 72 hrs	1 min - 72 hrs
Maximum output	2Gb	50 Gb	50 Gb	250 Gb	5.2 Tb	10.5 Tb
Read length	Dependent on length of target molecule Max. to date > 2Mb.					
Description	Lowest cost, reduced capacity adapter for MinION sequencer	Low cost, mobile, long read sequencer	Mobile long read sequencer and analysis platform in one unit	Medium capacity desktop long read sequencer with onboard analysis platform	High capacity desktop long read sequencer	Higher cost, high capacity desktop long read sequencer

Use of ONT systems and available tools for SARS-CoV-2 sequencing

ONT's sequencing systems, primarily the MinION, have been used extensively for microbial sequencing over the past five years prior to the COVID-19 pandemic. The portability of the MinION and associated equipment has facilitated the development of the 'lab in a bag' used during Ebola and other outbreaks for surveillance and, more recently, for diagnostic testing [235]. It has also seen use in hospital settings for monitoring the spread of nosocomial infections [236, 237].

Early descriptions of the SARS-CoV-2 genome were produced using Nanopore sequencing in combination with one or more short-read technologies or Sanger sequencing [16, 19, 22]. ONT is working with public health laboratories and researchers around the world to support the current COVID-19 pandemic response. ONT is maintaining a timeline of use of their technology relating to SARS-CoV-2, supported by both collaborating researchers and consortia [238]. The company also announced development of LamPORE, a rapid, scalable assay for the detection of SARS-CoV-2. This is discussed further in Section 3.2.

The ARTIC network has published a series of protocols for the preparation and sequencing of SARS-CoV-2 samples using nanopore sequencing [239]. These protocols have been used internationally and have been subsequently adopted by ONT as their SARS-CoV-2 protocol of choice for nanopore sequencing.

Advantages of Oxford Nanopore Technologies sequencing

- Rapid and flexible - particularly useful for sequencing smaller genomes. Sequences read in real time – allowing for termination when user determines enough reads have been generated
- Smaller sequencing units can be purchased at low cost
- Relatively inexpensive at low throughput
- Mobile sequencing – the small size and high portability of some systems means that these can be used in the field
- Many reagents do not require cold storage meaning they can be safely stored in environments where refrigeration is not possible or unreliable
- Simple user interface and analysis platforms – although knowledge is still required, the need for expertise for many applications of this technology is not required
- Well-established sample/library preparation and sequencing protocols for SARS-CoV-2 have been developed and distributed by researchers
- Widely used platform for SARS-CoV-2 sequencing that has been adapted and improved upon by the research community
- Simultaneous examination of methylation possible using direct RNA sequencing

Limitations

- Limited barcoding means this approach is more expensive than other approaches for sequencing at high throughput (a high number of samples). Currently the mobile sequencing units are not capable of providing the same level of multi-plexing as other next generation sequencing technologies
- Raw signal output files are very large – this makes files difficult to store. As software and pipelines for analysis evolve rapidly, it is useful if not essential for these files to be available for subsequent, quality analysis of the data
- Higher error rate in homopolymeric regions

5.3.7 Sanger sequencing

Sanger sequencing emerged in 1977, and was the primary method used for genome sequencing until the development and improvement of NGS technologies from 2006 onwards. Sanger produces highly accurate sequence reads of up to 900bp in length [240]. Although older than many approaches, Sanger sequencing still provides extremely high accuracy sequencing and is still cost effective for several small scale, small target sequencing applications. Sanger sequencing has been used to sequence the SARS-CoV-2 genome alongside, and complementary to, higher-throughput technologies [16, 22].

Automated sequencing systems for Sanger sequencing include Thermo Fisher Scientific (previously Applied Biosystems) SeqStudio and the 3500 and 3730 Series Genetic Analyzers.

5.4 Extent of use of different sequencing systems

Significant resources have been invested in a short span of time in order to develop and implement specialist tools and protocols to make sequencing of SARS-CoV-2 more efficient.

Whilst a number of developers have experienced an increase in demand for sequencing systems and products, much of the sequencing of SARS-CoV-2 samples has taken place using instruments already in place (personal communications). This is probably due in part to the need to act quickly, and also to the often-significant costs involved in establishing sequencing equipment, skills, and associated tools. This means that to some extent at least, system usage reflects previously established capacity.

Up to 26 June 2020, the most commonly used technologies listed in GISAID were from Illumina, with more than 65% of the sequences (over 37,000 samples), followed by Oxford Nanopore Technologies representing just over a quarter of the samples (nearly 14,000 sequences), and Ion Torrent (Thermo Fisher, just over 1,000 sequences). BGI was listed for over 180 samples, BioelectronSeq 4000 for 12 samples (a single group submitted these) and PacBio for three samples. Sanger sequencing was used for almost 500 samples [14] (figure 4).

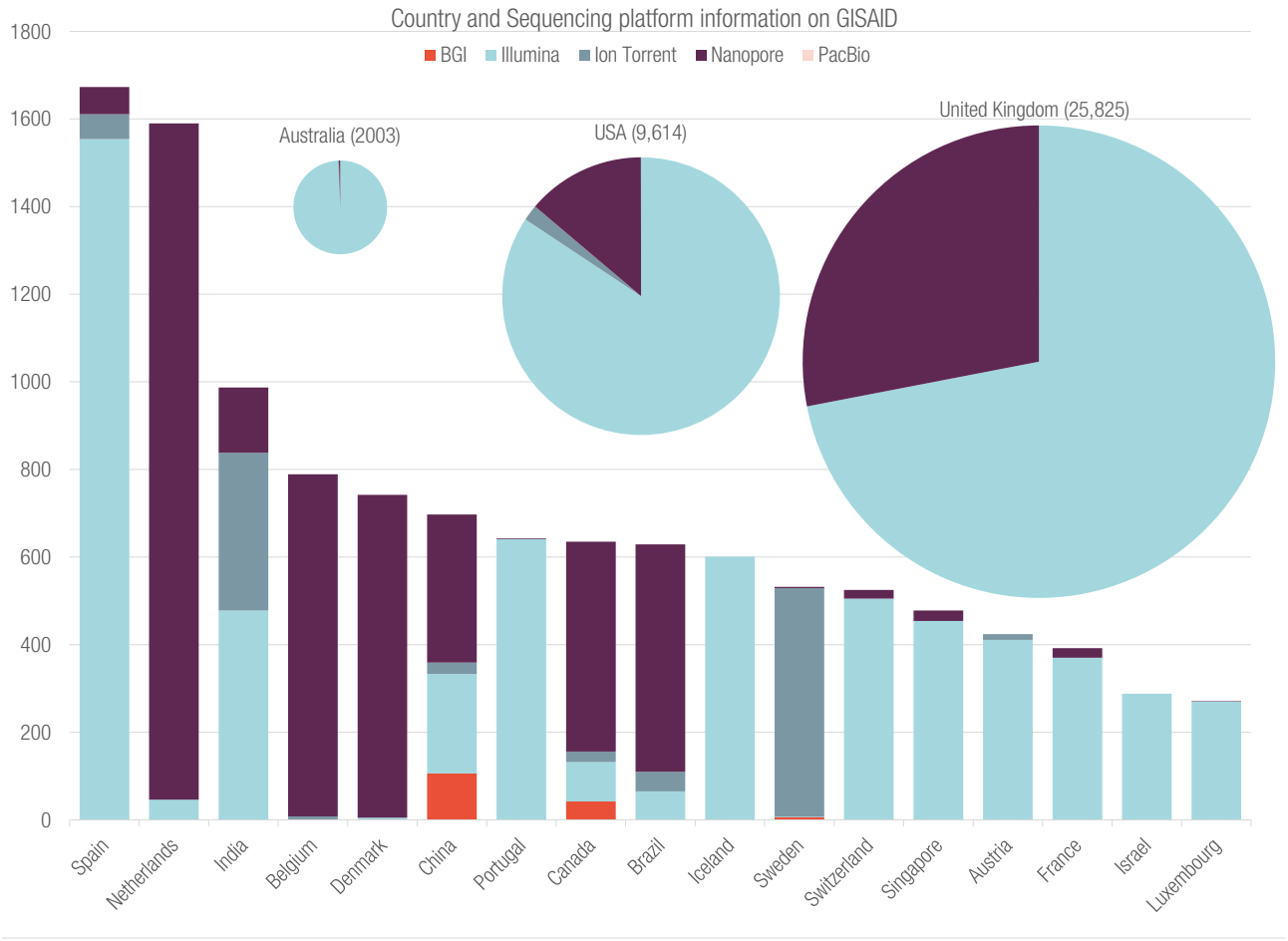
Within GISAID in June 2020 a wide variety of bioinformatics tools were listed: Burrows-Wheeler Aligner Tool (BWA), iVar, SAMtools, ARTIC, minimap, bcftools, CLC genomic workbench, bowtie, Geneious, V-pipe, medaka, Iterative refinement meta-assembler (IRMA), SPAdes, nanopolish, megahit and assembly trinity [14]. Due to the volume of information available to download on GISAID, an update to the information above and figure 4 is not currently possible, as of March 2021.

The NCBI-SRA database showed Illumina (over 85%) and ONT (over 12%) technologies are most commonly used, with Ion Torrent (ThermoFisher), BGI and PacBio used on the remainder. Similar proportions were seen in June 2020, February 2021 and March 2021, with Illumina still the most frequently used followed by ONT.

Sequence databases do not provide a perfect representation of sequencing performed as not all sequences are uploaded to public databases, but rather an indication of commonly used tools and systems.

Figure 4: The top twenty countries for highest number of sequences released to GISAID [14] and the technology platforms that were used as of 26 June 2020

Data from UK, USA and Australia are represented separately due to the large numbers of sequences released from these countries (not to scale). For some countries, the systems used could be reflected by the preferred equipment used in a small number of submitting laboratories (circles are not to scale).



5.5 Method development

The type of sequencing approach used varies between individual groups and countries, and is largely based on the available technology and the aims of the experiments. Research efforts have included method development to optimise sequencing strategies for specific purposes. Table 13 contains some examples of research publications that describe sequencing method developments and protocols specifically for SARS-CoV-2 to overcome various challenges, particularly those around diagnosis and tracking viral transmission. These are only representative of the few method developments that have been published. Individual labs may well have developed new methods, but have not published them.

Table 13: A selection research publications describing novel methods for sequencing SARS-CoV-2

Paper	Sequencing technology	Approach	Primary purpose	Challenge	Novelty in protocol
Resende et al (2020) [241]	ONT, Illumina and/or Sanger	Whole genome sequencing	Describing three protocols using a unique primer set designed to recover long reads of SARS-CoV-2 genomic material by direct RNA sequencing	Obtaining complete viral genomes from clinical samples in a timely manner	Tiling multiplex PCR method. Recovered 2kb reads and decreased number of primers required
St Hilaire et al (2020) [139]	Illumina NextSeq500	Whole genome sequencing	Describing a low cost, high throughput method for diagnosis of SARS-CoV-2 that amplifies the entire SARS-CoV-2 genome	Rapid diagnosis, overcoming limitations of RT-PCR diagnostics (e.g. false negatives)	SARS-CoV-2 enrichment method developed by the ARTIC Network with short-read DNA sequencing and de novo genome assembly
Gohl et al (2020) [242]	Illumina, MiSeq	Targeted amplicon, long read sequencing	Development of a scalable method of sequencing for diagnosis and viral tracking	Sequencing SARS-CoV-2 at scale is limited by the cost and labour associated with making sequencing libraries	All amplicon-based method for sequencing SARS-CoV-2, which bypasses costly and time-consuming library preparation steps. Benchmarked against the ARTIC v3 approach
Wang et al (2020) [243]	ONT	Targeted amplicon, long read sequencing	Describing a rapid method for diagnosing SARS-CoV-2 infection along with screening for 9 other respiratory viruses	Current diagnostics based on TR-PCR exhibit false negatives, low sensitivity and do not detect other respiratory infections	Used primers developed in house and combined the advantages of target amplification with long-read, real-time sequencing

Table 13: A selection research publications describing novel methods for sequencing SARS-CoV-2

Paper	Sequencing technology	Approach	Primary purpose	Challenge	Novelty in protocol
Moore et al (2020) [207]	ONT	Targeted amplicon and metagenomic sequencing	Describing a metagenomic approach using long read sequencing	The potential of coinfection with other pathogens causing severe COVID-19 disease	Used amplicon based long read sequencing with sequence independent single primer amplification (SISPA)

All of these publications were in preprint at the time of writing and therefore not peer reviewed.

Note: It is likely that many developments in sequencing methods for sequencing of SARS-CoV-2 are not published in the literature.

6

Sequencing implementation

6 Sequencing implementation

This section describes how sequencing technologies are being used in practice for the genomic analysis of SARS-CoV-2, illustrating the main components of a sequencing workflow. The choice and design of workflow will depend on the needs of a particular sequencing project and the resources available; key considerations surrounding workflow choice are highlighted. The various challenges to successful implementation of sequencing are also addressed, some of which are particularly relevant in low- and middle-income countries. These challenges often relate to the broader context of sequencing use in disease management, a reminder that whilst it may be possible to implement sequencing in terms of generating a sequence from a sample, the success and utility of sequencing for disease control efforts often depends on other factors, such as the ability to share data and the resources available for analysis. Finally, specific examples of the different types of workflows in use for SARS-CoV-2 sequencing are provided.

A number of recently published documents provide further advice when considering sequencing implementation. The WHO recently published the document *Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health*, which provides detailed technical guidance for laboratories, concerning sequencing implementation and method choice for different sequencing applications [15]. A further WHO document provides higher level policy considerations for use of sequencing to achieve public health goals [57]. Finally, the European Centre of Disease Control (ECDC) has produced a technical note to provide guidelines to laboratories and relevant stakeholders in making decisions on establishing sequencing capacities and capabilities [244].

6.1 The basic sequencing workflow

Producing readable genomic data from a physical sample requires completion of several steps (figure 5). Each of these steps impacts upon the type or quality of information retrieved from the original sample, and the ability to interpret something useful from that information. Sequencing workflows can be highly flexible – the steps may be specified in one complete or part-complete workflow or tool (dependent upon how the technology is developed), or produced as individual parts that can be combined with other technologies or processes at different stages of the workflow to provide something highly customised.

The complexity of the steps varies depending upon a number of factors, including: sequencing equipment selected, sample type and information required. Although interpretation is key to drawing appropriate conclusions from data, the importance of preparatory steps for assuring good quality data and appropriate subsequent interpretation should not be underestimated.

The steps both prior to and following the sequencing step are substantially variable – there are many options for how these can be conducted, and many tools and consumables exist with broad or exclusive application to different types of sequencing and sequencing equipment. Logistical and environmental limitations also impact upon the steps that can be taken; where optimal conditions and resources are not available, there are often alternative ways of completing a specific step.

Key steps in a basic workflow can be summarised as follows:

Sample collection and preparation

Includes the steps from collection of a sample, such as a throat swab, from a patient through to the storage and transportation of that sample, to the extraction of viral RNA prior to further processing. It can also extend through to conversion of RNA to cDNA, depending upon the workflow.

Library preparation

Transforms the retained nucleic acid portion of a collected sample into a prepared sample library ready for sequencing. This may include fragmenting or size selection of nucleic acids (dependent upon application), addition of sequencing adapters and quantification and quality control of resulting libraries.

Sequencing

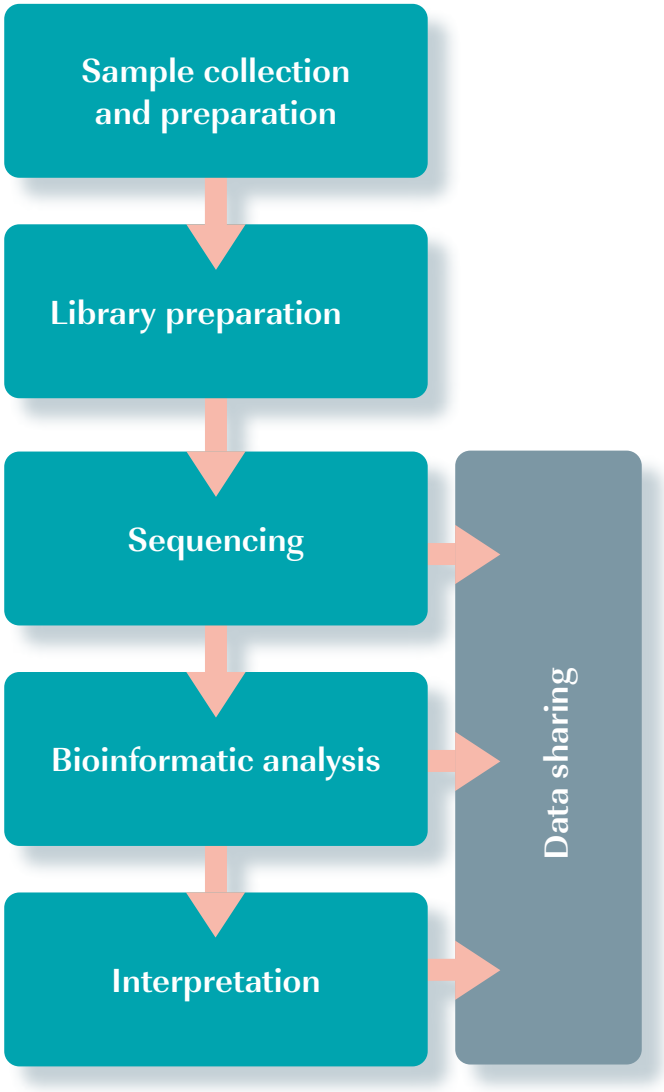
The process by which the sequence of bases in a series of nucleic acids is detected by one of a number of methods to provide readable data – raw sequence reads. This process covers entry of a prepared sample library into a sequencing system to retrieval of raw sequence information.

Bioinformatic analysis

This includes the processing and conversion of raw data that is produced during sequencing into one of several formats that is suitable for ongoing analysis or interpretation.

Data sharing may also be considered part of this workflow. The approaches adopted towards data sharing, including the databases used, file type, and attachment of metadata, will impact upon the ability of others outside of the original sequencing group to effectively use the information.

Figure 5: A basic sequencing workflow



6.2 Sample collection and preparation

Sequenced genomes have typically been captured from SARS-CoV-2 positive swabs following diagnostic testing. Samples must be appropriately processed before they can be run on sequencing equipment.

6.2.1 Sample collection

The US CDC lists five different types of acceptable swabs for COVID-19 diagnostic testing. These are:

- A nasopharyngeal (NP) specimen
- An oropharyngeal (OP) specimen
- A nasal mid-turbinate swab
- An anterior nares (nasal) swab
- Nasopharyngeal wash/aspirate or nasal wash/aspirate (NW) specimen

A list of sample types from WHO can be found on their website, including information on transport and storage [149].

Following collection of a swab from the patient, samples can be preserved in one of several different media, these include:

- Viral transport medium (VTM)
- Viral inactivation medium
- Amies transport medium
- Sterile saline
- Viral lysis buffer
- DNA/RNA Shield or similar viral inactivation agent

Working with live pathogenic viruses (especially respiratory pathogens) requires processing of samples to take place in facilities with specified biocontainment capabilities. The level of biocontainment a facility is capable of may be referred to as a ‘biosafety level’ e.g. in the US, or ‘containment level’ e.g. in the UK and Canada.

Higher levels such as containment level 3 (CL-3) laboratories (the level required in the UK for handling SARS-CoV-2 samples prior to viral inactivation, e.g. for viral culture) have strict standards for their design and build, and are often in short supply. Lower containment levels are simpler to achieve in a laboratory setting, as they do not impose the same strict requirements on the facilities or staff using them, and are therefore more common. In order to safely use lower level laboratories for SARS-CoV-2 testing, and to rapidly increase testing capacity, some groups and companies have devised mechanisms of virus inactivation. There are accordingly several approaches to sample collection, transportation, and initial processing.

In addition, in order for automated RNA extraction to be performed (which normally takes place outside of biocontainment cabinets) the virus must first be thoroughly inactivated. Some RNA extraction buffers do not efficiently inactivate the virus, so an additional inactivation step may be required [246].

Methods of viral inactivation include submersion in a viral lysis buffer or component such as 5M guanidinium thiocyanate [247] or DNA/RNA Shield (Zymo Research) [247, 248], and thermoinactivation [250-251] which may provide a solution in settings where supply of some reagents is either unreliable, insufficient or inconsistent.

Although beneficial in some regards, viral inactivation requires that the sample be worked on quickly to avoid significant degradation, so proximity to testing facilities may prove a limiting factor. Viral inactivation buffers also contain components such as beta-mercaptoethanol or guanidinium thiocyanate, meaning they are inappropriate for where required facilities such as fume hoods are unavailable.

Depending on the sequencing equipment to be used, different sample preparation methods may be recommended for each sample type.

6.2.2 RNA extraction

RNA extraction is the removal of non-RNA substances from a sample containing RNA. RNA extraction involves a number of stages, each of which is important for retrieving a high-quality sample that is more likely to yield accurate and more complete results. Failure to remove contaminants can compromise results.

SARS-CoV-2 RNA extraction tools and methods

Extraction of viral RNA involves the following stages:

1. Lysis – breaking open of virus structures must first take place to enable access to RNA. Lysis is normally chemical, and performed through submersion in an appropriate buffer solution
2. Purification – this involves removal of other molecules and contaminants from the sample, including degradation of proteins and DNA
3. Washing and resuspension – RNA may be bound to beads or otherwise isolated while other components are removed through ethanol washes
4. Quality assurance – the quality of the retrieved extract may be assessed, and the RNA quantified using qPCR
5. Reverse transcription – for most sequencing methods, transformation of RNA into cDNA is required prior to sequencing

Many kits are available for the extraction of RNA. Extraction kits are normally compatible with all sequencing systems. Extraction kits are not mandatory for the successful extraction of RNA, extraction can be performed using independently acquired reagents. Examples of viral RNA extraction kits promoted for use with samples from COVID-19 patients are given below.

- QIAGEN QIAmp Viral Mini Kit [252]
- Quick-DNA/RNA Viral MagBead Kit from Cambridge Bioscience [253]
- Rapid Viral RNA Extraction Kit from NBS Biologicals [254]
- innuPREP Virus RNA Kits from Analytik Jena [255]

A small number of sequencing systems now provide integrated sample and library preparation steps for sequencing [220, 256], and several extraction-free protocols and tools have been developed for the detection of SARS-CoV-2 RNA [257-259], though test accuracy may be reduced in comparison to protocols utilising RNA extraction.

In addition, there are several challenges related to the extraction and purification of RNA. RNA extraction is one of the rate-limiting steps in viral sequencing; when working with harmful respiratory pathogens this step can present health and safety concerns and necessitate strict biocontainment measures accordingly.

6.3 Library preparation

During library preparation, nucleic acids are prepared for sequencing by the addition of identifiers and adapters that allow molecules to adhere to the sequencing flow cell. Library preparation is an essential step prior to sequencing using most systems.

The type of library preparation tools and techniques applied are normally closely tied to the sequencing equipment to be used, and the steps included in these workflows depend upon how the sequencing itself is conducted. Some steps are common to library preparation methods across several systems, as described below. Details of library preparation kits and methods for use with particular sequencing workflows are included in Section 6.7.

Library preparation may also include fragmentation or size selection of extracted nucleic acids (dependent upon approach and sequencing system to be used), amplification (involving the conversion of RNA into cDNA), quality control, and sample quantitation steps.

Quality assessment and quantitation of nucleic acid samples can be performed using one of a range of methods, including qPCR. These steps help to ensure the correct amount of sample can be loaded onto the sequencer and determine whether potentially disruptive contaminants are present that could impede sequencing.

Many library preparation kits are sequencing-platform-specific and sold by companies producing associated sequencing equipment. However, a range of kits are also available from other suppliers and some of the preparation may be performed without commercial kits. The extent and type of sample and library preparation required depends upon several factors, including sequencing platform and type of sequencing; amount and quality of starting material; the biocontainment level of available facilities; and time taken.

Selection and execution of appropriate sample and library preparation can subsequently impact upon overall sequence quality, genomic coverage and uniformity, error rate, selection of bioinformatic pipelines, and variant interpretation.

6.4 Sequencing, bioinformatics and data sharing

The different sequencing platforms and approaches used for different sequencing applications are discussed in Chapter 5, with examples of specific workflows provided in Section 6.7. As discussed below, there are a variety of practical considerations that will influence both the type of technology chosen and the following bioinformatics analyses.

There are a large number of different bioinformatics processes available for processing and interpreting sequence data (see section 5.4 for a list of bioinformatics tools used for SARS-CoV-2 early in the pandemic). Commercial software is often recommended by the technology providers or protocol developers, which allow for the majority of variants to be detected, and provide consensus sequences for uploading onto data sharing platforms. For example, the ARTIC protocol and the commercial kits for Illumina and Thermo Fisher platforms all come with recommended bioinformatics analysis pipelines.

Others develop more complex in-house pipelines, often using a range of software to provide a specific type of analyses for a specific application. In this case considerable expertise in bioinformatics is required, as well as rigorous internal validation of the pipelines.

For a more in-depth overview of different bioinformatics tools suitable for different sequencing applications, see the WHO document Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health [15].

Data sharing of sequences is essential to ensure that sequencing has the maximum impact at a global health level. As much metadata as possible, for example regarding the location, date, type of sample and preferably clinical information, should also be shared to allow the data to be placed into context and maximum information obtained [57]. Section 2.4 provides an overview of the sequencing repositories and data sharing frameworks being used across the globe.

A wide variety of open access tools and resources have been developed to make this data easily accessible to those requiring it for public health and research purposes. For example, the geographical and temporal trends of virus clades can be followed on the GISAID and NextStrain websites, whilst concerning lineages can be tracked on websites such as CovLineages [168].

Other web based applications provide specific information on viral mutations. For example, CoV-GLUE is a website powered by data uploaded to GISAID that tracks mutations resulting in amino acid variations, which is informative for those developing diagnostic tests or drugs [260]. For a more extensive list of data analysis and visualisation resources see Appendix 1 in the VOC addendum [4].

6.5 Practical considerations for implementation

A number of technologies and methods are in use for the sequencing of SARS-CoV-2 as described in Chapter 5. Most are being used in efforts to generate whole genome sequences or broad sequence coverage of the virus for a variety of surveillance and research applications. Section 3.2 describes the few NGS based tests approved for diagnostic purposes.

Many variations of sequencing workflows and protocols exist for the same types of sequencing methods, which is a reflection of the varied requirements and capabilities of the many sequencing efforts taking place worldwide. Even within consortia and collaborations currently sequencing the SARS-CoV-2 genome as part of coordinated national or international efforts, the approach has not been to mandate that the same sequencing technology and workflows be used by the entire consortium. Instead each contributing group has been using their preferred technology and protocols to generate the required data. There are many factors that influence workflow choice including:

Application type

The reason for providing a sequence and the type of application for the data will inevitably influence the type of sequencing chosen and broad category of workflow choice. For example, metagenomic sequencing may be useful for providing pathogen identification at the start of an outbreak, and in cases of co-infection with other pathogens. This type of sequencing may be less suited to tracking the development of novel variants over time, for which a targeted amplification approach may be more suitable.

The existing infrastructure and expertise of a sequencing facility

A facility that has already invested in a particular sequencing technology will be likely to choose workflows that make use of that equipment and their existing expertise and reagents, or create and adapt new protocols as necessary, rather than investing resources in procuring and learning to use a new technology. These choices will be influenced by the individual circumstances and interests of those sequencing; for example, labs with larger budgets or an academic interest in certain sequencing technologies may be more likely to purchase new equipment.

Technology throughput and turnaround times

The investigation of viral genomes for the purposes of guiding national public health strategies may involve sequencing samples from a large proportion of the population, and rapid and high throughput technology platforms may be required to produce timely results. In contrast, a project designed to characterise a few viral sequences at the start of an outbreak will need rapid but not necessarily high throughput technologies. For other purposes, such as research projects requiring highly accurate and detailed viral genomes, neither the speed nor throughput of the technology may be as important as its accuracy and resolution.

Cost

The estimated costs of a workflow, including any upfront infrastructure and equipment investments, reagent costs, training and personnel costs may impact upon the type of technology chosen and which variation of a workflow to use.

Accessibility and support

How accessible a workflow appears and the availability of support from the technology provider or from experienced colleagues could influence the choice of workflow. A well validated workflow with plenty of guidance available on its use may be more attractive than a newly developed workflow that needs further validation, especially to those with less experience in viral sequencing.

Project logistics

The logistics of the sequencing project often need to be taken into account when considering the most suitable technology and workflow to use. For example, the logistical issues surrounding the transfer of samples from labs to suitable sequencing facilities may play a role in the type of technology used. In cases where labs are located a long distance from sample collection points, or where there are few labs with the appropriate biosecurity, it may be easier to invest in portable sequencers that can be taken to the site of the sample rather than transport the sample to the lab. The biological sample type itself may also impact on which protocols are most suitable, for example samples with a lower viral load may require more sensitive sequencing methods that necessitate targeted sequencing.

6.6 Key challenges

There were a number of challenges facing the use of sequencing highlighted by our research, including feedback from interviewees. While all of the challenges set out below have an impact to a greater or lesser extent in all resource settings, many are likely to be particularly acute in lower resource settings. One vital consideration around the use of sequencing in lower resource settings is that there is a significant barrier in terms of resource and expertise availability in computing and bioinformatics.

In terms of the costs of sequencing pathways, absolute costs vary by jurisdiction and scale of sequencing application, e.g. small scale vs. large, centralised laboratories. We have instead provided comparative information on which platforms and approaches are more expensive than others and the relative advantages and disadvantages of each different sequencing application.

There are a number of 'pre-sequencing' considerations, including:

- **Sample type, collection and transport -** measures need to be taken to minimise contamination of the sample prior to sequencing, to ensure that samples are collected and stored in the appropriate manner and that samples reach the laboratory in a timely manner
- **Recording and use of sample metadata -** this is needed to support further analysis of the sample, including whole genome sequencing, but the collection and use of metadata varies depending on local regulations and infrastructure available to store and manage this data
- **Sample numbers -** a critical mass of samples that accurately represent the infected population is required. In many countries, there is an opportunity to make use of existing resources and processes put in place during past outbreaks to help gather samples. Further help might be required to establish infrastructure and protocols for appropriate sample collection in different settings
- **Priorities in the jurisdiction in which sequencing is taking place -** these can have an impact on the circumstances in which sequencing is carried out and may include: current measures put in place to manage epidemics; any other ongoing epidemics that require management/ surveillance (or may require more urgent attention); and vaccination cover (if a vaccine is available)

6.6.1 Challenges and considerations related to SARS-CoV-2 sequencing and data interpretation

Aiming to conduct informative pathogen sequencing studies of any kind raises a number of considerations, not all of them specific to SARS-CoV-2 sequencing:

Viral mutation rate - Compared to other viral genomes, coronaviruses (of which SARS-CoV-2 is a member) are comparatively large. As described in Chapter 1, SARS-CoV-2 has a relatively slow mutation rate compared to other viruses. This can make studies of molecular epidemiology challenging as fewer identifying genomic differences can be tracked through transmission chains.

Sample quality and preservation - As with other RNA viruses and RNA sequencing more broadly, sample preservation is critical. RNA is a notably unstable molecule, and highly susceptible to degradation. Where viral titre (concentration of virus in sample) is low, detection becomes difficult. The virus may not be detected even where it is present, especially without prior amplification or culture. In addition, pathogenic viruses require inactivation before sequencing – degradation can occur more quickly after inactivation.

Biohazard - There are significant challenges in working with a highly contagious and potentially dangerous respiratory pathogen. Higher level biocontainment facilities may be required than are widely available in order to conduct high volumes of testing or sample preparation for sequencing. In addition suitably trained personnel are required to use such facilities.

Availability of reagents and tools - The rapid and extensive nature of infection spread has meant that demand for specific reagents and tools related to viral sequencing and molecular diagnostics has increased dramatically across the globe in a short space of time. This has led to difficulties in the acquisition of certain reagents and lab equipment.

Resources for sample preparation - For several technologies, sequencing itself is not the rate-limiting step. Library preparation steps are often the bottleneck in the processing of samples. Technologies providing the ability to automate these steps, especially where a large volume of samples are being processed, are likely to be highly beneficial.

Cross contamination – The high sensitivity of NGS is beneficial for detecting viral RNA at low concentration and for discovering rare variants in the viral genome, however this high sensitivity means that extra care needs to be taken to avoid contamination of samples.

6.6.2 Computing and bioinformatics

Computing and bioinformatics were areas particularly highlighted by interviewees as having an impact on the delivery of sequencing in lower resource settings. The challenges can be outlined as follows:

- Lack of available computing engineering expertise, which is needed to establish the computing infrastructure and systems required to carry out large-scale bioinformatics. This includes considerations around the types of computing infrastructure to put in place and types of data storage systems that could be used
- A shortage of specific bioinformatics expertise, which is needed to run and develop bioinformatics analysis, this includes expertise in data management
- Local expertise to contribute to understanding of the wider landscape – i.e. locally trained experts who understand the wider cultural and social contexts in which sequencing is taking place
- Access to academic expertise to contribute to the training of computing experts and bioinformaticians. Lower resource settings can lack the 'critical mass' of experts both to carry out analysis and also support development of the next generation. Concerns were also expressed about a 'brain drain' of expertise to higher paid industries, particularly the commercial sector
- Insufficient capacity in terms of expertise and computing infrastructure to benefit from the global information available e.g. sequences deposited in international databases. Access to international information is necessary to put sequences collected locally in context, but the expertise and computing capacity to do this may be scarce in lower resource settings
- As sequencing methods evolve, so do the bioinformatic methods used to analyse raw genetic data. This can have an impact on data comparability through time and highlights the need for standardisation of sequencing and analysis methods

Although substantial challenges remain, there is a high level of awareness of these issues. Coordinated efforts and initiatives are underway to address them; for example, the Pan-African bioinformatics network, H3ABioNet [261] and the Public Health Alliance for Genomic Epidemiology (PHA4GE) consortium launched in 2019 [97, 262].

6.6.3 Data sharing

There are challenges and concerns around the issue of data sharing that are not only having an impact during the current pandemic, but also on pathogen genome sequencing efforts more generally.

While there is a desire to democratise sequencing and increase the accessibility of sequencing technologies and access to the data they generate, there are ongoing concerns in some countries about access to data and questions around data ownership. It should also be acknowledged that uploading genomic data to databases can be a time-consuming and difficult task, and the challenges outlined above – particularly in terms of access to infrastructure and expertise – can have an impact on efforts to achieve this.

With the ever-increasing volume of data being generated there is a need to develop standards and consensus on what is required to support sharing of pathogen genomic data and analysis. One group of researchers have developed recommendations for supporting open pathogen genomic analysis in public health [97] which are to:

- Develop guidelines for management and stewardship of genomic data
- Make bioinformatics pipelines fully open-source and broadly accessible
- Improve the reproducibility of bioinformatics analyses
- Develop best practices to support open data sharing
- Support new infrastructure and software development demands with an expanding technical workforce

6.7 Sequencing workflows in use by international consortia and others

6.7.1 Sources and types of workflows

It is important to distinguish between the overall project workflow for a sequencing project, which could include many elements from obtaining a sample to data sharing and visualisation, and the workflows provided specifically for the sequencing element of a project, which have the primary aim of generating sequence data.

As each sequencing project is different, there is no one project workflow that will fit all projects, and instead workflows and/or guidance tend to be available for each separate element and can then be used as required. For example, any sequencing project will inevitably require bioinformatics analyses, which are critical for interpreting sequence data.

However, bioinformatics methods may not always be described in a sequencing workflow, or if they are they may not be the type of bioinformatics methods required for the project. Separate workflows and guidance on bioinformatics are therefore often used to process raw sequence data, such as those produced by the Galaxy [263] and ELIXIR [264] projects. In addition to specific workflows some organisations are developing general guidance on areas of best practice such as quality control and data sharing.

This section describes key workflows in use for SARS-CoV-2 sequencing. The manufacturers of the sequencing technologies each provide their own recommended workflows and guidance. Organised efforts such as the ARTIC network have developed protocols for more specific methods such as tiled amplicon sequencing, or in the case of the US CDC have tailored workflows to suit the infrastructure of their national laboratories. Individual groups may also develop their own protocols to suit their needs.

The scientific community has made efforts to freely share the varied workflows, protocols and other guidance that has been developed through organisation and consortium websites, as well as online forums such as protocols.io, virological.org and GitHub. Of particular note is the US CDC page on GitHub, which has assembled a collection of sequencing protocols and workflows in international use [265]. The COG-UK consortium have also published all their protocols [266]. Protocols are also widely available in pre-print open access repositories such as bioRxiv, as well as published peer-reviewed papers or on network websites such as ARTIC.

Whilst acknowledging that variations to specific protocols exist, it is possible to summarise from these sources the main sequencing workflows in use based on both the sequencing method and technology used.

6.7.2 Workflows overview

There are several sequencing approaches and technologies in use by international consortia and others:

Sequencing approaches	Sequencing technologies
• Targeted approaches (i.e. amplicon & hybridisation capture based sequencing) • Metagenomic sequencing • Direct RNA sequencing	• Oxford Nanopore Technologies • Illumina • Pacific Biosciences • Ion sequencing • BGI/MGI DNA Nanoball sequencing

Most sequencing methods can be carried out using several different sequencing technologies in a 'pick and mix' approach, with different sequencing workflows and/or protocols existing for different combinations of method and technology. The exception is workflows used as diagnostic tests, which have to use the specific materials and equipment specified in the approved test protocol. For example, when used as a diagnostic test the Illumina COVIDSeq workflow has to use the NovaSeq 6000 S4 Reagent Kit and the Illumina NovaSeq Xp sequencing system.

Here we have grouped some of the most widely used workflows into broad categories based on sequencing approaches.

Group 1 - Workflows for research and surveillance sequencing

Note this list is non-exhaustive and other workflows are also in use or in development.

Category A - Amplicon-based sequencing workflows

- Tiled amplicon sequencing (ARTIC protocols) using ONT technology
- Tiled amplicon sequencing (ARTIC protocols) using Illumina technology
- Tiled amplicon (ARTIC protocols) using PacBio technology
- Amplicon sequencing using Ion Torrent (Ion AmpliSeq SARS-CoV-2 Research Panel)

Category B - Target enrichment-based sequencing

- Target enrichment sequencing panel (Nextera Flex) using Illumina sequencing
- SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian followed by a probe-based bait capture (SeqCap, Roche and xGen, IDT) using Illumina Sequencing
- DNA nanoball sequencing using MGI DNBSEQ technology

Category C - Metagenomic sequencing workflows

- Shotgun metagenomic sequencing using Illumina technology
- Metagenomic sequencing using ONT technology
- Metagenomic sequencing using BGI/MGI DNBSEQ technology

Category D - Direct RNA sequencing workflows

- Direct RNA sequencing using ONT technology

Group 2 - Workflows with regulatory approval for diagnostic sequencing

- Illumina COVIDSeq workflow
- ONT LamPORE workflow
- UCLA SwabSeq COVID-19 diagnostic platform workflow

6.7.3 Examples of specific workflows

Examples of published workflows in each of the categories listed above is provided in this section, as summarised in table 14. For each example the main elements of a specific published workflow are described, and as such these examples are not intended to be a general summary of all workflows available for a particular category.

The workflows have been selected as examples primarily on the basis of how widely they have been used internationally, as well as how complete and accessible they are. Many researchers have modified these workflows to meet their own purposes. As a result, some variations have emerged that have become relatively frequently used or meet a particularly urgent need; where these exist, they have been noted. As this list was compiled in July 2020, it is possible that new adaptations or completely new workflows now exist. However, all workflows described remain in use, and illustrate key steps and reagents required for sequencing.

Table 14: Summary of workflows of SARS-CoV-2 sequencing described

	Workflow / method	Sample	RNA extraction	Library prep	Sequencing	Data analysis
Group 1: Research and surveillance sequencing workflows	Category A: Tiled amplicon sequencing, ONT	Total RNA	General guidance	ONT provided	MinION plus rampant	Provided through ARTIC
	Category A: Tiled amplicon sequencing, Illumina	Total RNA	Not specified	PreprNEB NEBNext Ultra II DNA library prep kit (Illumina)	MiSeq or NovaSeq	None specified, but can use Illumina LRM resequencing module
	Category B: Nextera DNA Flex Enrichment with Illumina Nextera Virus Oligo Panel	Viral culture/ patient sample	Not specified	Illumina Nextera Flex for enrichment and IDT for Illumina Nextera DNA UD indexes	Any Illumina instrument (iSeq, MiniSeq, MiSeq are ideal)	LRM resequencing module, BaseSpace, IDbyDNA, Explify platform
	Category C: Shotgun metagenomics using Illumina technology	Viral culture/ patient sample	Extract viral RNA, remove human rRNA	TruSeq stranded total RNA	Any Illumina system (NextSeq or iSeq, MiniSeq or MiSeq are ideal)	LRM resequencing module and range of additional tools
	Category D: ONT Direct RNA Sequencing Protocol	RNA for viral culture	Not specified	ONT direct RNA seq kit	MinIon or GridIon plus MinKNOW	Albacore software and additional approaches
Group 2: Diagnostics	Illumina COVIDSeq test	Decontaminated patient swab sample	QIAmp viral RNA mini kit	Illumina provided-tagmentationt	NovaSeq	DRAGEN COVIDSeq test pipeline

6.7.4 Group 1: Workflows in use for research and surveillance purposes

Category A, example 1: Tiled amplicon sequencing (ARTIC protocols) using Oxford Nanopore Technologies (ONT) sequencing systems [267]

Workflow description	Provided by Oxford Nanopore Technologies: workflow for tiled amplicon sequencing based on ARTIC protocols
Starting material	Total RNA extracted from positive samples screened by a suitable qPCR assay. This specific workflow does not produce specific protocols for RNA extraction and quantification, but general guidance is available
Sample prep	• Reverse transcribe RNA into ctDNA • Multiplex PCR to generate 400bp tiled amplicons, overlapping by ~20bp. Primer set used is available to order
Library prep	• Add barcodes using Native Barcoding Expansions 1-12 (EXP-NBD104) and 13-24 • Add adaptors using Adaptor mix II
Sequencing method	• Sequence using ONT device e.g. MinION (~1 hour) • Monitor coverage in real time using rampant technology
Bioinformatics approach	Uses ARTIC network pipelines which take an alignment based approach to generate SNP and indel calls and subsequently a consensus sequence. Encourages data sharing
Throughput	From 1-24 samples per run on single flow cell, to 24-1000 samples per run if used on higher throughput ONT devices
Equipment and reagents listed in workflow	Materials • Input RNA • Native Barcoding Expansion 1-12 (EXP-NBD104) and 13-24 (EXP-NBD114) if multiplexing more than 12 samples • Ligation Sequencing Kit (SQK-LSK109) • Flow Cell Priming Kit (EXP-FLP002) • SFB Expansion (EXP-SFB001)

Equipment and reagents listed in workflow continued

Consumables

- Q5® Hot Start High-Fidelity 2X Master Mix (NEB, M0494)
- Random Primer Mix (NEB, S1330S)
- 10 mM dNTP solution (e.g. NEB N0447)
- SuperScript IV reverse transcriptase, 5X RT buffer and 100 mM DTT (ThermoFisher Scientific, 18090010)
- RNaseOUT™, 40 U/µl (Life Technologies, 10777019)
- COVID-19 primers (lab-ready at 100 µM, IDT)
- Nuclease-free water (e.g. ThermoFisher, cat # AM9937)
- Agencourt AMPure XP beads
- Freshly prepared 80% ethanol in nuclease-free water
- Qubit dsDNA HS Assay Kit (ThermoFisher Q32851)
- NEBNext Ultra II End repair / dA-tailing Module (E7546)
- NEBNext Ultra II Ligation Module (E7595)
- NEBNext Quick Ligation Module (E6056)
- DNA 12000 Kit & Reagents - optional (Agilent Technologies)
- 0.2 ml thin-walled PCR tubes
- 1.5 ml Eppendorf DNA LoBind tubes

Equipment

- Hula mixer (gentle rotator mixer)
- Magnetic separator, suitable for 1.5 ml Eppendorf tubes
- Microfuge
- Vortex mixer
- Thermal cycler
- P1000 pipette and tips
- P200 pipette and tips
- P100 pipette and tips
- P20 pipette and tips
- P10 pipette and tips
- P2 pipette and tips
- Ice bucket with ice
- Timer

Equipment and reagents listed in workflow continued	Optional Equipment • Agilent Bioanalyzer (or equivalent) • Qubit fluorometer (or equivalent for QC check) • Eppendorf 5424 centrifuge (or equivalent) • PCR hood with UV steriliser (optional but recommended to reduce cross-contamination) • PCR-Cooler (Eppendorf)
IT and software requirements	• Unless using a MinIT device with pre-installed software, sequencing on a MinION Mk1B requires a high-spec computer or laptop to keep up with the rate of data acquisition. Details are provided on their website • An internet connection is required for updates and telemetry • MinKnow software is required for all experiments • Optional software for further analysis capabilities are EPI2ME and Guppy
Costs	Sequencing equipment required for this protocol is relatively inexpensive compared to alternative technologies. Flow cells differ in cost dependent on sequencing systems used and quantity purchased.
Turnaround time	Approximately 7 hours from RNA sample to sequence, of which 1 hour is sequencing time, for this specific workflow as claimed by ONT [267].
Workflow and protocol availability and support	Openly available with full protocols on ONT website. Other support including an online protocol builder and selector tools to produce tailored guidance are also available, as well as extraction methods for different biological samples and bioinformatics guidance.
Important notes	Very important to include negative controls for RT-PCR step, easy to cross contaminate and the PCR reaction is extremely sensitive. Similarly recommend having a pre-and post PCR area preferably using hoods, and clean surfaces well.
Variations and related protocols	1. ARTIC network protocols -This workflow was developed based on the ARTIC network protocols, but the ARTIC network may recommend variations and updates [126] 2. CDC NCIRD/DVD ONT Sequencing Protocol - Protocol and guidance on multiplexing PCR followed by minion sequencing based on ARTIC/ ONT workflow. Developed by the Viral Discovery laboratory at CDC/ NCIRD, where it was used to generate the first 16 SARS-CoV-2 genome sequences from the United States [268]

Category A example 2: Tiled amplicon sequencing using Illumina technology [269]

Workflow description	Protocol for COVID-19 ARTIC v3 Illumina library construction and sequencing protocol V.4, developed by Wellcome Sanger Institute. Uses ARTIC protocols for producing tiled amplicons then sequences on Illumina.
Starting material	Total RNA extracted from positive samples screened by a suitable qPCR assay.
Sample prep	• Reverse transcribe RNA into ctDNA • Multiplex PCR to generate 400bp tiled amplicons overlapping by ~20bp. Primer set used is available to order • Purify amplified cDNA and quantify
Library prep	Use PreprNEB NEBNext® Ultra™ II DNA Library Prep Kit for Illumina • Add adaptors, either TruSeq or NEBNext • Purify and perform library PCR • Quantify library
Sequencing method	Sequence samples on an Illumina NovaSeq SP flow cell, using the XP workflow. Alternatively, samples may be sequenced on an Illumina MiSeq using either v2 (500 cycle) or v3 (600 cycle) reagent kits.
Bioinformatics approach	No specific approach stated in this protocol. However Illumina provide an LRM Resequencing module to perform on-instrument analysis (automated). The Resequencing module provides alignment, coverage, and small variant data as well as FASTQ, BAM, and VCF files for use in other data analysis pipelines, if desired. The BaseSpace Sequence Hub to upload sequencing data and perform additional in-depth analysis with cloud-based tools.
Throughput	• NovaSeq: plex up to 384 samples per NovaSeq SP lane • MiSeq: plex up to 96 samples per run, this could be increased further depending on coverage requirements (This is based on experience of protocol authors)

Equipment and reagents listed in workflow	• NEBNext Ultra II DNA Library Prep Kit for Illumina - 96 rxns #E7645L New England Biolabs • 2x Kapa HiFi Hotstart Readymix #KK2602 Kapa Biosystems • LunaScript RT SuperMix Kit #E3010L New England Biolabs • Illumina Library Quantitation Complete kit (Universal) #KK4824 Kapa Biosystems • NEB Q5® Hot Start High-Fidelity 2X Master Mix #M0494L New England Biolabs • AccuClear® Ultra High Sensitivity dsDNA Quantitation Kit with DNA Standards #31028 Biotium
Costs	The cost of sequencing systems varies substantially. The highest throughout system, NovaSeq, requires significant investment. Consumable costs are available from illumina.com
Turnaround time	Time from RNA to sequence generation can range widely from ~12 hours to ~63 hours depending on sequencing type used, and if automation of steps is used as in this protocol. Estimate based on around ~4 hours from RNA to producing amplicons [267], ~3 hours for library prep incubations not including manual pipetting time [270]. Sequencing time can range from 4-56 hours depending on platform used.
Workflow and protocol availability and support	This version is available on protocols.io and is open access as long as group is referenced [269]. Additional support available through ARTIC network and Illumina.
Important notes	• It is vital the cDNA setup is performed in a laboratory in which post-PCR COVID-19 amplicons are not present, to minimise any risk of sample contamination • Throughout the protocol it is indicated that liquid handling automation is in use at Sanger for specific parts of the process. However, these steps could be performed on alternative liquid handlers or manually
Variations and related protocols	1. ARTIC network protocols: This workflow was developed based on the ARTIC network protocols, but the ARTIC network may recommend variations and updates [126] 2. Nextera Flex library prep: An adapted version of this workflow is available which uses Nextera Flex DNA library preparation methods. This workflow appears to have been developed due to a lack of TruSeq or NebNext reagents in some countries and is also faster. In South Africa this has been shown to reduce hands on time to 3 hours compared to 12 hours with TruSeq library prep (in non-automated lab) [271] 3. CDC NCIRD/DVD Illumina NEBNext Protocol: Protocol and guidance on multiplexing PCR followed by Illumina sequencing using NEBNext library prep. Developed by the Viral Discovery laboratory at CDC/NCIRD [268]

Use case study of tiled amplicon sequencing using Illumina Technology: Sequencing positive cases of COVID-19 in State of Victoria, Australia [47]

Purpose: To integrate epidemiological and genomic data to help trace transmission chains and assess the impact of social restrictions in the State of Victoria in Australia (population 6.24 million).

Scale of sequencing: Of 1333 positive cases of Covid-19, 75% (903 samples) were sequenced after applying internal quality control parameters using ARTIC version 1 or 3 primers with Illumina sequencing.

Why was this method chosen? '*Key to this effort was high-throughput sequencing using an amplicon-based approach, which allowed us to process a large number of samples in a short period of time*'.

Important technical points:

- '*Stringent QC to ensure only high-quality consensus sequences entered the final alignment was particularly important when considering the minimal diversity in SARS-CoV-2 sequence data used to infer genomic clusters*'
- '*While use of a predefined Ct value to select samples for SARS CoV-2 genomic sequencing could be considered, our use of QC parameters, rather than a Ct value, enabled the inclusion of additional samples for genomic analysis*'

Information provided:

- 76 distinct genomic clusters were identified; these included large clusters associated with social venues, healthcare facilities and cruise ships
- Sequencing of sequential samples from 98 patients revealed minimal intra-patient SARS-CoV-2 genomic diversity
- Phylodynamic modelling indicated a significant reduction in the effective viral reproductive number (Re) from 1.63 to 0.48 after the implementation of travel restrictions and population-level physical distancing

Category B example: Nextera DNA Flex Enrichment with Illumina Respiratory Virus Oligo Panel [272]

Workflow description	Workflow provided by Illumina- described as a target enrichment sequencing workflow for highly sensitive detection and characterization of common respiratory viruses, including coronavirus strains. Note this workflow is provided as an example due to its comprehensive and accessible nature, but it is relatively new and not yet widely used.
Starting material	Designed for use with either viral culture or patient samples
Sample prep	• Extract total nucleic acids (viral RNA) • Reverse transcribe into cDNA
Library prep	Library prep performed using Illumina Nextera Flex for Enrichment and IDT for Illumina Nextera DNA UD Indexes: • Enrichment reactions performed with the Respiratory Virus Oligos Panel which features ~7800 probes designed to detect respiratory viruses, recent flu strains, and SARS-CoV-2, as well as human probes to act as positive controls • Libraries quantified and pooled
Sequencing method	• Prepared libraries can be sequenced on any Illumina instrument • The benchtop iSeq™ 100, MiniSeq™, and MiSeq Systems are particularly well suited due to the low read requirements for these samples • In the study example provided for this workflow, libraries were sequenced on the MiSeq System at 2 × 151 base pairs read length using MiSeq v3 reagents
Bioinformatics approach	Use LRM Resequencing module to generate sequencing data on instrument of choice (automated). The Resequencing module provides alignment, coverage, and small variant data as well as FASTQ, BAM, and VCF files for use in other data analysis pipelines, if desired The BaseSpace sequence hub can be used to upload sequencing data to perform additional in depth analysis with cloud based tools. The IDbyDNA Explify Platform can be accessed via the BaseSpace sequence hub and provides an easy-to-use solution for in-depth data analysis that features robust data quality control, standardized result interpretation, carefully curated databases, and custom report generation. Data analysis is based on k-mers and alignment steps, including protein-level detection of viruses, which increases the ability to identify novel and highly divergent viruses.

Throughput	Library prep: Pre-enrichment pooling of up to 12-plex is tested, workflows are available for 16 or 96 samples [273] Sequencing: Throughput depends on sequence platform used (table 8) There is potential for automation: Liquid handling robots can be used [273]
Equipment and reagents listed in workflow	RNA extraction: QIAGEN QIAmp Viral Mini Kit (QIAGEN, Catalog no. 52904) used for RNA extraction cDNA synthesis: Either Illumina TruSeq RNA reagents or Thermo Scientific Maxima H Minus Double-Stranded cDNA Synthesis Kit (Thermo Scientific, Catalog no. K2561). Library prep: • Nextera Flex for Enrichment (Illumina, Catalog no. 20025524) • IDT for Illumina Nextera DNA UD Indexes (Illumina, Catalog no. 20027213) • Respiratory Virus Oligos Panel (Illumina, Catalog no. 20042472)
Turnaround time	Library prep (from cDNA): total workflow time ~6.5 hours. Total hands-on time ~2 hours [273] Sequencing time can range from 4-56 hours depending on platform used, See table 8 for details.
Workflow and protocol availability and support	Full workflow and protocols available on Illumina website [272]
Important notes	• The recommended minimum RNA/total nucleic acid input for reverse transcription is 10 ng • For best results, reverse transcription should be performed on freshly extracted nucleic acid samples • Total DNA input recommended for tagmentation is 10–1000 ng As a general guideline, the read recommendation for this workflow is 500k reads per sample but these numbers can be variable and this is only a recommended starting point

Category C example: Shotgun metagenomics using Illumina technology [274]

Workflow description	Provided by Illumina- A comprehensive shotgun metagenomics workflow for detecting and characterising coronavirus strains using Illumina benchtop systems
Starting material	Designed for use with either viral culture or patient samples
Sample prep	• Extract viral RNA • Remove human cytoplasmic and mitochondrial rRNA
Library prep	TruSeq Stranded Total RNA protocol used: • RNA fragmentation • First- and second-strand cDNA synthesis • Adenylation • Adapter ligation • Amplification
Sequencing method	Prepared libraries quantified and pooled and loaded onto an illumina benchtop sequencing system. Any Illumina instrument can be used, but samples prepared directly from swabs or similar matrices are most suited for the NextSeq™ Series of Systems. Libraries prepared from viral culture using the same workflow are particularly well suited for the benchtop iSeq™ 100, MiniSeq™, and MiSeq Systems due to the lower recommended read count of 500,000 reads per sample.
Bioinformatics approach	• Use LRM Resequencing module to generate sequencing data on instrument of choice (automated). The Resequencing module provides - alignment to the reference genome, depth of coverage for the reference genome, a summary of the identified small variants, including single nucleotide variants (SNVs) and insertions and deletions (indels), a summary of the fragment length observed, duplicate information for library diversity • Additional commercial tools are available if cloud-based data analysis is not possible, but they will need to be evaluated by the end user • For cloud-based, in-depth analysis, the IDbyDNA Explify Platform enables comprehensive identification of more than 35,000 viruses using a proprietary database of curated DNA and RNA reference sequences
Throughput	Sequencing: Throughput depends on sequence platform used (see table 8)

Equipment and reagents listed in workflow	• RNA extraction (QIAGEN QIAmp Viral Mini Kit, PN 52904) • RiboZero Gold rRNA depletion protocol (Illumina, 48 samples, Cat no. 20020598, 96 samples, 20020599) • TruSeq Stranded Total RNA Library Prep Gold kit (Illumina, Cat no. 20020599) • IDT for Illumina TruSeq RNA UD Indexes (96 indexes, 96 samples) (Illumina, Cat no. 20022371) • Benchtop Illumina System • Illumina Local Run Manager (LRM) Resequencing Module for local analysis • Other commercial tools if cloud based analysis not possible (would require evaluating) • IDbyDNA Explify Platform for cloud based analysis
Turnaround time	Sequencing time can range from 4-56 hours depending on platform used, see table 8 for details
Workflow and protocol availability and support	Full workflow and protocols available on Illumina website [274]
Important notes	Virus titre, efficiency of human rRNA depletion, and the number of reads per sample impact the number of virus-specific reads obtained and coverage of the viral genome. A general guideline includes 10M reads for direct-from-patient samples and 0.5M reads for positive virus culture, but these numbers can be variable and are only a recommended starting point
Variations and related protocols	ACEGID used a protocol based on this workflow for first genome sequence of SARS-CoV-2 from Africa (complemented by separate Sanger sequencing of RdRp region) [114]

Category D example: ONT Direct RNA Sequencing Protocol [275]

Workflow description	This workflow is the provided by ONT [276] and has been used for direct RNA sequencing of SARS-CoV-2, in some cases with minor variations, in at least 3 publications by Kim et al [53], Taiaroa et al [54] and Viehweger et al [276]. In all cases this workflow has been used for in depth analysis of the genome and sub-genome of a single sample of SARS-CoV-2
Starting material	RNA extracted from viral culture
Sample prep	Not provided in this workflow, however details of sample prep are provided in publications cited above
Library prep	Library prep performed using ONT's Direct RNA Sequencing Kit (SQK-RNA001) • Reverse transcription adaptor ligation • Quantify reverse-transcribed and adapted RNA using a Qubit DNA HS assay
Sequencing method	• This workflow provides instructions for use on MinION but can also be used on GridION. Full details on how to prepare and load flow cell provided • Uses MiniKNOW to run protocol
Bioinformatics approach	• Workflow says to run the Albacore software to basecall reads but no further details given in this specific workflow. Further details provided on ONT website. Also see 'Important notes' below
Throughput	• Workflow designed for preparation of single samples i.e. not multiplexing • Throughputs obtained from sequencing of single RNA samples in publications range from ~0.237 Gb (225,000 reads) to ~1.9 GB (879,679 reads) [53, 54, 276]. Note the proportion of viral RNA varied, with some RNA obtained from human host and/or cell type used
Equipment and reagents listed in workflow	• Direct RNA Sequencing Kit (SQK-RNA001) • SpotON Flow Cell FLO-MIN106 • Agencourt RNAClean XP beds • SuperScript III Reverse Transcriptase RTA / RCS / ELB / WSB on ice • RRB / RMX in freezer until needed

Equipment and reagents listed in workflow continued	• Freshly prepared 70% EtOH Nuclease-free water (NFW) • 10 mM dNTP solution NEB Blunt / TA Ligase Master Mix T4 DNA ligase • NEBNext® Quick Ligation Reaction Buffer • Qubit RNA HS Assay Kit • Qubit dsDNA HS Assay Kit • Microfuge Vortexer and Hula mixer • Magnet for bead separation • Approx 20 DNA LoBind Eppendorf tubes • 0.2 ml thin-walled PCR tubes • Thermal cycler Pipettes • P2, P20, P100/200, P1000 pipette tips • P2, P20, P100/200, P1000 pipettes
Turnaround time	Library prep (from RNA): total time ~100 minutes for incubation, plus extra time for pipetting etc and setting up Minlon [277] Sequencing: 48 hours [277]
Workflow and protocol availability and support	Detailed example workflow including step by step instructions and checklist available from ONT [275]
Important notes	In practice users can use their own bioinformatics approach- Of note a resource [277] has been independently set up to allow uniform processing of all nanopore direct RNA sequencing data using the MasterofPores workflow, using raw FAST5 data. The workflow consists of 3 modules: • Module 1: NanoPreprocess. Performs base-calling, demultiplexing, quality-filtering, mapping, per-gene/per-transcript counting, mapping and reporting (HTML MultiQC report) • Module 2: NanoMod. Prediction of RNA modifications • Module 3: NanoTail. Prediction of polyA tail length estimations Protocols detailing viral culture and extraction, as well as some variations to sequencing and types of bioinformatics used are available in publications by Kim et al [53], Taiaroa et al [54] and Viehweger et al [276]

6.7.5 Group 2: Workflows with regulatory approval for diagnostic sequencing

Illumina COVIDseq test workflow [278]

Workflow description	This high-throughput NGS test detects SARS-CoV-2 in nasopharyngeal, oropharyngeal, and mid-turbinate nasal swabs from patients suspected of COVID-19. It detects 98 targets on SARS-CoV-2 for highly accurate detection using a modified version of the validated, publicly available ARTIC multiplex PCR protocol [128]. The workflow includes detailed instructions from viral RNA extraction through to analysis and report generation.
Starting material	Decontaminated patient sample- Nasopharyngeal (NP) swabs, oropharyngeal (OP) swabs, anterior nasal swabs, mid-turbinate nasal swabs, nasopharyngeal wash/aspirates, nasal aspirates, and bronchoalveolar lavage (BAL) specimens
Sample prep	• Guidance is provided on sample collection, transport and storage • RNA extraction using the QIAamp Viral RNA Mini Kit. • cDNA synthesis • PCR amplification to generate amplicons using two separate reactions to amplify cDNA, which are then pooled
Library prep	• Tagment PCR Amplicons (a process that fragments and tags the PCR amplicons with adapter sequences) • Amplification- The adapter-tagged amplicons undergo a second round of PCR amplification using a PCR master mix and unique index adapters. • Pool and clean libraries
Sequencing method	• Pooled libraries are clustered onto a flow cell, and then sequenced using sequencing by synthesis (SBS) chemistry on the NovaSeq 6000 Sequencing System using the NovaSeq Xp S4 flow cell workflow • Step by step instructions on how to sequence are provided in the workflow
Bioinformatics approach	• The Illumina DRAGEN COVIDSeq Test Pipeline analyzes sequencing results to detect the presence of SARS-CoV-2 RNA in each sample for diagnostic use under the FDA Emergency Use Authorization • For each result with at least 90 SARS-CoV-2 virus targets, the Illumina DRAGEN COVIDSeq Test Pipeline performs small variant calling and generates a consensus sequence in FASTA format for research use • Step by step instructions on data analysis are provided in the workflow

Throughput	The test can be scaled up or down to accommodate different numbers of samples. Up to 3072 results can be processed in 12 hours on the NovaSeq 6000 System using two NovaSeq 6000 S4 Reagent Kits with the Xp workflow.
Equipment and reagents listed in workflow	Reagents: • Illumina COVIDSeq Test (3072 Samples), # 20043675 • IDT for Illumina Nextera UD Indexes Sets A–D (384 Indexes, 384 Samples), # 20027217 • 13 QIAamp Viral RNA Mini Kit, Qiagen, # 52906 • QIAamp Viral RNA Mini Kit reagents. See QIAmp Viral RNA Mini Handbook (document #HB-0354-006) • Qubit dsDNA HS Assay Kit, Thermo Fisher Scientific, # Q32851 or Q32854 • The following NovaSeq 6000 Sequencing System reagents for 3072 samples:2 NovaSeq 6000 Sequencing System S4 Reagent Kit (200 cycles), Illumina, # 20027466 and 2 NovaSeq Xp 4-Lane Kit, Illumina, # 20021665 • 2 N NaoH • 400 mM Tris-HCl, pH 8.0 • Nuclease-free water • Ethanol, 100% (200 proof) of molecular biology grade, Sigma-Aldrich, # E7023 • Optional: DNAZap & RNaseZap Equipment: A full list of items including pipettes etc is provided in the workflow. Specific items are listed below • BioShake iQ QInstruments # 1808-0506 • DRAGEN Bio-IT Platform Illumina • QIAamp Viral RNA Mini Kit equipment See QIAmp Viral RNA Mini Handbook (document #HB0354-006) • Magnetic Stand-96 Thermo Fisher Scientific # AM10027 • One of the following magnetic stands: Dynabeads MPC-S (Magnetic Particle Concentrator) #A13346 or MagnaRack Magnetic Separation Rack # CS15000 • NovaSeq 6000 Sequencing System Illumina • NovaSeq Xp Flow Cell Dock Illumina, # 20021663 • Quibit Fluorometer 3.0 Thermo Fisher, catalog # Q33216, Q33217, or Q33218 • Bio-Rad C-1000 Touch thermal cycler Bio-Rad, Part # 1851197

Equipment and reagents listed in workflow continued	Consumables: A full list of items including pipettes etc is provided in the workflow. Specific items are listed below • Hard-Shell 96-Well PCR Plates Bio-Rad, # HSP-9601 • 1.7 ml LoBind microcentrifuge tubes Eppendorf, # 022431021 • 5 ml LoBind microcentrifuge tube Eppendorf, # 003012234 • Microseal 'B' adhesive seals Bio-Rad, # MSB-1001 • RNase/DNase-free Disposable Pipetting Resovoirs VWR # 89094-658 • Quibit dsDNA HS Assay Kit One of the following, depending on kit size: ThermoFisher Scientific # Q32851, ThermoFisher Scientific # Q32854 • Qubit Assay Tubes ThermoFisher Scientific # Q32856
Turnaround time	3072 results in 12 hours on the NovaSeq 6000 System
Workflow and protocol availability and support	Entire workflow and support available from Illumina [128]
Important notes	• As a quality feature, an internal control consisting of 11 human mRNA targets is included in every sample to monitor for errors • The Illumina COVIDSeq Test is intended for use by qualified and trained clinical laboratory personnel specifically trained in the use of the Illumina NovaSeq 6000 Sequencing System and next-generation sequencing workflows as well as in vitro diagnostic procedures • Testing is limited to laboratories certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA), 42 U.S.C. §263a, to perform high complexity tests • COVIDSeq has not been FDA cleared or approved. This test is authorised by FDA under an EUA for the duration of the declaration that circumstances exist justifying the authorization of emergency use of in vitro diagnostics for detection and/or diagnosis of COVID-19 • This product is available for Performance Evaluation Only (PEO) in European countries regulated by CE-IVD, or as Research Use Only (RUO) in other non-US countries

Photo by Julian Wan on Unsplash

7

Conclusions

7 Conclusions

Sequencing is being widely used in COVID-19 related research, and to support SARS-CoV-2 surveillance efforts. In contrast, diagnostics in current use are based on non-sequencing technologies, and are being managed via separate pathways and processes. Only one sequencing based diagnostic, COVIDSeq, has received emergency use authorisation from the US FDA.

However, the landscape of how sequencing technologies are being used continues to evolve rapidly. A large number of techniques and workflows are being used, reflecting the many different situations in which they are being applied, the specific expertise of the people using the technologies, and the infrastructure and equipment available to them. For example, long read sequencing using Oxford Nanopore Technologies is in wide use by researchers and consortia in the UK and dominates sequences being submitted from the UK to particular databases, whereas globally, Illumina short-read technologies are widely used.

Due to time constraints and need for rapid information availability during the pandemic, current resources have been being widely repurposed, with less focus on standardisation of workflows and protocols than might otherwise be expected.

As the pandemic has progressed, there has been increasing use of sequencing as a surveillance tool, including further upload of sequences to relevant databases. In particular, the use of genome sequence data to rapidly identify variants of concern (VOCs) in the UK, South Africa and Brazil in late in 2020, and the timely sharing of this data, demonstrated the value of genomic surveillance, allowing countries to put additional public health measures in place in response to this data. Genomic information also has ongoing value in terms of allowing vaccine and diagnostics developers to ensure that their products are effective against the VOCs. Work is ongoing to examine the efficacy of existing vaccines against the new variants, and to update vaccines, if required [4].

It remains to be seen whether sequencing will also become a routinely used diagnostic tool. At present there are a large number of existing tests using simpler methods that are fit for purpose. The value of using sequencing for diagnostics is questionable, particularly when a rapid diagnosis is likely to be required in a clinical context, as it is more expensive and has slower turn-around time. However, strengthening the infrastructure that has been put in place to support the collection of samples from diagnosed patients for sequencing, which can only build on established surveillance efforts, would have considerable value.

It will be some time before the clinical and epidemiological situation settles, and while some countries are beginning to emerge from pandemic lockdowns, many still have measures in place to reduce spread of COVID-19. The focus is still on generating viral sequences to support information gathering about the virus and surveillance, but in the mid- to long-term, health systems and other stakeholders will need to consider carefully the optimal sequencing strategy to pursue, in terms of how much sequencing to undertake and when, to support ongoing disease management efforts. Current influenza surveillance strategies, and in particular the use of sequencing, will provide valuable guidance on how SARS-CoV-2 surveillance could be refined in the future. For example, monitoring will be needed to detect the emergence of SARS-CoV-2 variants that could affect vaccine efficacy, and consideration will need to be given as to how best to strategically target sequencing to optimise knowledge gathering.

Another important factor is the collaboration likely to be needed between national and international sequencing consortia and other sequencing efforts, in order to standardise laboratory protocols and processes for the consolidation and sharing of information and data. For example, planned external quality assurance programmes will be useful in helping to determine the most suitable protocol for particular circumstances and sequencing technologies. In addition, there is a need for continued support of sequencing in lower resource settings, and the WHO has outlined the development of low-cost sequencing, along with underpinning digital connectivity, as a strategic innovation priority. This will include ongoing initiatives to support timely sharing of data on publicly accessible databases and other platforms [279].

While the sequencing efforts of the past year have been unprecedented, there is much to be learned about establishing new sequencing initiatives and protocols. This will have an impact on current pathogen surveillance efforts, and will also inform management of any future disease outbreaks and contribute to preventing or minimising pandemics.

8 Appendix

8.1 Appendix 1. Acknowledgements

We thank the following people for contributing their time and expertise to project information gathering:

- Dr Meredith Ashby - Director, Market Strategy for Infectious Disease, Immunology, and Microbial Genomics, Pacific Biosciences
- Dr David Bentley - Vice President and Chief Scientist, Illumina Cambridge Ltd
- Prof Judy Breuer - Professor of Virology, UCL; Lead of Clinical and Virology Working Group, COG-UK
- Dr Michael Chapman – Director of Health Informatics, Health Data Research UK Cambridge; Metadata and patient linkage/epidemiology/health informatics working group lead, COG-UK
- Prof Alan Christoffels - Director, South African National Bioinformatics Institute (SANBI); Public Health Alliance for Genomic Epidemiology (PHA4GE)
- Dr Andrea Ganna - EMBL-group leader, Institute for Molecular Medicine Finland (FIMM) / Massachusetts General Hospital/Harvard Medical School; COVID-19 host genetics initiative (COVID-19 hg)
- Dr Ewan Harrison - Project Manager and sample logistics working group lead, COG-UK; Career Development Fellow, Wellcome Sanger Institute
- Dr Ellen Higginson - Research Associate, Department of Medicine, University of Cambridge
- Dr Leila Luheshi - Associate Director for Clinical and Translational Research, Oxford Nanopore Technologies
- Dr Duncan MacCannell - Chief Science Officer, Office of Advanced Molecular Detection, US Centers for Disease Control
- Prof Joseph Sriyal Malik Peiris - Professor in Medical Science and Chair of Virology, Hong Kong University School of Public Health
- Prof Leo Lit Man Poon - Professor and Division Head, Hong Kong University School of Public Health
- Dr Josh Quick - Sequencing working group lead, COG-UK; Post-doctoral researcher, Institute of Microbiology and Infection, University of Birmingham
- Mr Neil Ward - Commercial Director North Europe, Illumina Cambridge Ltd.

GISAID: We gratefully acknowledge all authors from the originating laboratories responsible for obtaining the specimens/sequence data and the submitting laboratories where genetic sequence data were generated or submitted and shared via the GISAID Initiative, on which some of this research is based.

8.2 Appendix 2. African laboratories that have submitted more than 100 sequences to GISAID [14] as of 26 January 2021

Submitting lab	Number of sequences
South Africa KRISP, KZN Research Innovation and Sequencing Platform	2090
Kenya KEMRI-Wellcome Trust Research Programme/KEMRI-CGMR-C Kilifi	512
South Africa National Institute for Communicable Diseases of the National Health Laboratory Service	459
Gambia MRCG at LSHTM Genomics lab	427
Democratic Republic of the Congo Pathogen Sequencing Lab, National Institute for Biomedical Research (INRB)	353
South Africa NHLS/UCT	316
Nigeria African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria	213
Senegal Institut Pasteur de Dakar	136
Uganda MRC/UVRI & LSHTM Uganda Research Unit	133
South Africa National Health Laboratory Service (NHLS), Tygerberg, Cape Town	119

9 References

1. Arias, A., Watson, S. J., Asogun, D., et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. Virus Evol. 2016. 2(1): p. vew016
2. Giovanetti, M., Faria, N. R., Lourenco, J., et al. Genomic and Epidemiological Surveillance of Zika Virus in the Amazon Region. Cell Rep. 2020. 30(7): pp. 2275-2283 e7
3. Andersen, K. G., Shapiro, B. J., Matranga, C. B., et al. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell. 2015. 162(4): pp. 738-50
4. PHG Foundation and FIND SARS-CoV-2 variants. FIND 2021; Available from: https://www.finddx.org/covid-19/novel-variants/
5. Weston, S., Frieman, M. B. COVID-19: Knowns, Unknowns, and Questions. mSphere. 2020. 5(2): pp. e00203-20
6. SARS (Severe Acute Respiratory Syndrome). World Health Organisation. 2020; Available from: https://www.who.int/ith/diseases/sars/en/
7. Khan, S., Siddique, R., Shereen, M. A., et al. Emergence of a Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2: Biology and Therapeutic Options. Journal of Clinical Microbiology. 2020. 58(5): pp. e00187-20
8. Fiorillo, L., Cervino, G., Matarese, M., et al. COVID-19 Surface Persistence: A Recent Data Summary and Its Importance for Medical and Dental Settings. Int J Environ Res Public Health. 2020. 17(9): p. 3132
9. Liu, Yi., Gayle, A. A., Wilder-Smith, A., et al. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine. 2020. 27(2): p. taaa021
10. Petersen, E., Koopmans, M., Go, U., et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect Dis. 2020. 20(9): pp. e238-e244
11. Coronavirus disease (COVID-19) pandemic. World Health Organisation. 2020; Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019
12. Zumla, A., Chan, J. F., Azhar, E. I., et al. Coronaviruses - drug discovery and therapeutic options. Nat Rev Drug Discov. 2016. 15(5): pp. 327-47
13. Wu, A., Peng, Y., Huang, B., et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe. 2020. 27(3): pp. 325-328
14. The GISAID Initiative. 2020; Available from: https://www.gisaid.org/
15. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. World Health Organization. 2021; Available from: https://www.who.int/publications/i/item/9789240018440
16. Lu, R., Zhao, X., Li, J., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020. 395(10224): pp. 565-574
17. Wu, F., Zhao, S., Yu, B., et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020. 579(7798): pp. 265-269
18. Zhou, P., Yang, X. L., Wang, X. G., et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020. 579(7798): pp. 270-273

19. Zhu, N., Zhang, D., Wang, W., et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020. 382(8): pp. 727-733

20. Wang, H., Li, X., Li, T., et al. The genetic sequence, origin, and diagnosis of SARS-CoV-2. Eur J Clin Microbiol Infect Dis. 2020. pp. 10.1007/s10096-020-03899-4

21. Ji, W., Wang, W., Zhao, X., et al. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020. 92(4): pp. 433-440

22. Chan, J. F.-W., Yuan, S., Kok, K.-H., et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet. 2020. 395(10223): pp. 514-523

23. Zhang, T., Wu, Q., Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol. 2020. 30(7): pp. 1346-1351 e2

24. Wu, A., Niu, P., Wang, L., et al. Mutations, Recombination and Insertion in the Evolution of 2019-nCoV. BioRxiv. 2020. p. 10.1101/2020.02.29.971101

25. Lam, T. T., Shum, M. H., Zhu, H. C., et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature. 2020. 583(7815): pp. 282-285

26. Xiao, K., Zhai, J., Feng, Y., et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020. 583(7815): pp. 286-289

27. Scudellari, M. The sprint to solve coronavirus protein structures — and disarm them with drugs. Nature. 2020; Available from: https://www.nature.com/articles/d41586-020-01444-z

28. Walls, A. C., Park, Y. J., Tortorici, M. A., et al. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020. 181(2): pp. 281-292 e6

29. Wan, Y., Shang, J., Graham, R., et al. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020. 94(7): pp. e00127-20

30. Wrapp, D. , Wang, N., Corbett, K. S., et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020. 367: pp. 1260-1263

31. Letko, M., Marzi, A., Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat Microbiol. 2020. 5(4): pp. 562-569

32. Coutard, B., Valle, C., de Lamballerie, X., et al. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. 2020. 176: p. 104742

33. Andersen, K. G., Rambaut, A., Lipkin, W. I., et al. The proximal origin of SARS-CoV-2. Nat Med. 2020. 26(4): pp. 450-452

34. Zhang, L., Lin, D., Sun, X., et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science. 2020. 368(6489): pp. 409-412

35. Jin, Z. , Du, X. , Xu, Y., et al. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. Nature. 2020. pp. 10.1038/s41586-020-2223-y

36. Gao, Y., Yan, L., Huang, Y., et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. 2020. 368(6492): pp. 779-782

37. De Vito, N. J., Drysdale, H. M., Aronson, J. K. COVID-19 Clinical Trials Report Card: Remdesivir. The Centre for Evidence-Based Medicine, University of Oxford. 2020; Available from: https://www.cebm.net/covid-19/covid-19-clinical-trials-report-card-remdesivir/

38. Laamarti, M., Alouane, T., Kartti, S., et al. Large scale genomic analysis of 3067 SARS2 CoV-2 genomes reveals a clonal geo-distribution 3 and a rich genetic variations of hotspots 4 mutations. bioRxiv. 2020. p. 10.1101/2020.05.03.074567

39. Candido, D. S., Claro, I. M., de Jesus, J. G., et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020. 369(6508): pp. 1255-1260

40. Andrews, S. M., Rowland-Jones, S. Recent advances in understanding HIV evolution. F1000Res. 2017. 6: p. 597

41. Wang, C., Liu, Z., Chen, Z., et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020. 92(6): pp. 667-674

42. Ceraolo, C., Giorgi, F. M. Genomic variance of the 2019-nCoV coronavirus. J Med Virol. 2020. 92(5): pp. 522-528

43. Pachetti, M., Marini, B., Benedetti, F., et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020. 18(1): p. 179

44. Korber, B., Fischer, W. M., Gnanakaran, S., et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020. p. 10.1101/2020.04.29.069054

45. Pachetti, M., Marini, B., Benedetti, F., Giudici, F., et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine. 2020. 18(1): p. 179

46. Shen, Z., Xiao, Y., Kang, L., et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clin Infect Dis. 2020. p. 10.1093/cid/ciaa203

47. Seeman, T., Lane, C. R., Sherry, N. L., et al. Tracking the COVID-19 pandemic in Australia using genomics. medRxiv. 2020. p. 10.1101/2020.05.12.20099929

48. Karamitros, T., Papadopoulou, G., Bousali, M., et al. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. BioRxiv. 2020. p. 10.1101/2020.03.27.009480

49. Choi, B., Choudhary, M. C., Regan, J., et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. N Engl J Med. 2020. 383(23): pp. 2291-2293

50. Voloch, C. M., da Silva Jr, R. F., de Almeida, L. G. P., et al. Intra-host evolution during SARS-CoV-2 persistent infection. medRxiv. 2021. p. 10.1101/2020.11.13.20231217

51. Carabelli, A. M., Robertson, D. L., Peacock, S. Persistent SARS-CoV-2 infection and viral evolution tracked in an immunocompromised patient. COG consortium. 2020; Available from: https://www.cogconsortium.uk/news_item/persistent-sars-cov-2-infection-and-viral-evolution-tracked-in-an-immunocompromised-patient/

52. Kemp, S. A., Collier, D. A., Datir, R. P., et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021

53. Kim, D., Lee, J. Y., Yang, J. S., et al. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020. 181(4): pp. 914-921 e10

54. Taiaroa, G., Rawlinson, D., Featherstone, L., et al. Direct RNA sequencing and early evolution of SARS-CoV-2. bioRxiv. 2020. p. 10.1101/2020.03.05.976167

55. Davidson, A. D., Williamson, M. K., Lewis, S., et al. Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. bioRxiv. 2020. p. 10.1101/2020.03.22.002204

56. A Coordinated global research roadmap: 2019 novel Coronavirus. World Health Organization. 2020; Available from: https://www.who.int/blueprint/priority-diseases/key-action/Coronavirus_Roadmap_V9.pdf

57. SARS-CoV-2 genomic sequencing for public health goals: Interim guidance, 8 January 2021. World Health Organization. 2021; Available from: https://www.who.int/publications/i/item/WHO-2019-nCoV-genomic_sequencing-2021.1

58. Surveillance strategies for COVID-19 human infection. Interim guidance 10 May 2020. World Health Organization. 2020; Available from: https://www.who.int/publications-detail/surveillance-strategies-for-covid-19-human-infection

59. Public health surveillance for COVID-19: interim guidance, 16 December 2020. World Health Organization. 2020; Available from: https://apps.who.int/iris/handle/10665/337897

60. Operational considerations for COVID-19 surveillance using GISRS: interim guidance, 26 March 2020. World Health Organization. 2020; Available from: https://apps.who.int/iris/bitstream/handle/10665/331589/WHO-2019-nCoV-Leveraging_GISRS-2020.1-eng.pdf

61. Operational considerations to expedite genomic sequencing component of GISRS surveillance of SARS-CoV-2. World Health Organization. 2021; Available from: https://www.who.int/publications/i/item/WHO-2019-nCoV-genomic-sequencing-GISRS-2021.1

62. Global surveillance for COVID-19 caused by human infection with COVID-19 virus - Interim guidance 20 March 2020 World Health Organization. 2020; Available from: https://apps.who.int/iris/bitstream/handle/10665/331506/WHO-2019-nCoV-SurveillanceGuidance-2020.6-eng.pdf

63. Strategies for the surveillance of COVID-19. European Centre for Disease Prevention and Control. 2020; Available from: https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-surveillance-strategy-9-Apr-2020.pdf

64. Protocol for Enhanced Severe Acute Respiratory Illness and Influenza-Like Illness Surveillance for COVID-19 in Africa. Africa Centres for Disease Control. 2020; Available from: https://africacdc.org/download/protocol-for-enhanced-severe-acute-respiratory-illness-and-influenza-like-illness-surveillance-for-covid-19-in-africa/

65. Coronavirus disease (COVID-19) technical guidance: Laboratory testing for 2019-nCoV in humans. World Health Organization. 2020; Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance.

66. Mallapaty, S. The search for animals harbouring coronavirus - and why it matters. Nature. 2021. 591(7848): pp. 26-28

67. O'Toole, Á., Hill, V., Pybus, O. G., et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. Virological. 2021; Available from: https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592

68. Gudbjartsson, D. F., Helgason, A., Jonsson, H., et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med. 2020. 382(24): pp. 2302-2315

69. da Silva Candido, D, Morales Claro, I., Goes de Jesus, J., et al. Evolution and epidemic spread of SARS-Cov-2 in Brazil. Science. 2020. 369(6508): pp 1255-1260

70. Irish Coronavirus Sequencing Consortium - Tracking the genetics of the COVID-19 virus in Ireland. Department of the Taoiseach. 2020; Available from: https://www.gov.ie/en/publication/8d2282-covid-19-rapid-response-research-case-studies/#irish-coronavirus-sequencing-consortium-tracking-the-genetics-of-the-covid-19-virus-in-ireland

71. First genome sequences of SARS-CoV-2 viruses in Austria openly available. News Medical Life Sciences. 2020; Available from: https://www.news-medical.net/news/20200403/First-genome-sequences-of-SARS-CoV-2-viruses-in-Austria-openly-available.aspx

72. Mutational Dynamics of SARS-CoV-2 in Austria. Research Centre for Molecular Medicine of the Austrian Academy of Sciences. 2020; Available from: https://cemm.at/research/sars-cov-2-at/about/

73. Virus sequencing: Finnish SARS-CoV-2 sequences. University of Helsinki. 2021; Available from: https://www2.helsinki.fi/en/researchgroups/covid-19/virus-sequencing

74. National Institute of Infectious Diseases, Japan. NIID. 2021; Available from: https://www.niid.go.jp/niid/en/

75. First data for genomic surveillance of SARS-CoV-2 in Switzerland made available. ETH Zurich. 2020; Available from: https://bsse.ethz.ch/cevo/cevo-press/2020/05/first-data-for-genomic-surveillance-of-sars-cov-2-in-switzerland-made-available.html

76. Deutsche COVID-19 OMICS Initiative (DeCOI). 2020; Available from: https://decoi.eu/

77. New partnership to unravel genetic sequence of SARS-CoV-2 virus in Quebec. McGill University. 2020; Available from: https://www.mcgill.ca/newsroom/channels/news/new-partnership-unravel-genetic-sequence-sars-cov-2-virus-quebec-321890

78. SPHERES. SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance. A National Open Genomics Consortium for the COVID-19 Response. Centers for Disease Control and Prevention. 2020; Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html

79. CDC launches national viral genomics consortium to better map SARS-CoV-2 transmission. Centers for Disease Control and Prevention. 2020; Available from: https://www.cdc.gov/media/releases/2020/p0501-SARS-CoV-2-transmission-map.html

80. Implementation Framework: Toward a National Genomic Surveillance Network. Rockefeller Foundation 2021; Available from: https://www.rockefellerfoundation.org/wp-content/uploads/2021/03/Toward-a-National-Genomic-Surveillance-NetworkFINAL.pdf

81. Accelerating national genomic surveillance. Rockefeller Foundation. 2021; Available from: https://www.rockefellerfoundation.org/wp-content/uploads/2021/03/The-Rockefeller-Foundation_Accelerating-National-Genomic-Surveillance.pdf

82. Getting Ahead of the Pandemic: Accelerating National Genomic Surveillance. Rockefeller Foundation. 2021; Available from: https://www.rockefellerfoundation.org/report/accelerating-national-genomic-surveillance/

83. Introducing the COVID-19 Genomic UK (COG-UK) Consortium COG Consortium. 2020; Available from: https://www.cogconsortium.uk/wp-content/uploads/2020/05/C0367-Participating-in-the-COVID-19-Genomic-UK-COG-UK-Consortium-24-April-2020-1.pdf.

84. COVID-19 Genomics UK Consortium. COG Consortium. 2020; Available from: https://www.cogconsortium.uk/

85. COVID-19 Genomics UK Consortium Partners. COG-UK. 2020; Available from: https://www.cogconsortium.uk/about/

86. COVID-19. Wellcome Sanger Insitute. 2020; Available from: https://www.sanger.ac.uk/collaboration/covid-19-genomics-uk-cog-uk-consortium/

87. Cloud Infrastructure for Microbial Bioinformatics. CLIMB. 2020; Available from: https://www.climb.ac.uk/

88. 'World-class' CLIMB project receives £1.2 million funding boost from UKRI COG consortium. 2021; Available from: https://www.cogconsortium.uk/news_item/world-class-climb-project-receives-1-2-million-funding-boost-from-ukri/

89. COG-UK Project Hospital-Onset COVID-19 Infections Study (COG-UK HOCI). US NIH. 2020; Available from: https://clinicaltrials.gov/ct2/show/NCT04405934

90. The GenOMICC Study. 2020; Available from: https://genomicc.org/

91. Genomics on a mission: Meeting the COVID-19 challenge. GenomeCanada. 2020; Available from: https://www.genomecanada.ca/en/news/genomics-mission-meeting-covid-19-challenge

92. Oude Munnink, B. B., Nieuwenhuijse, D. F., Stein, M., et al. Rapid SARS-CoV-2 whole genome sequencing for informed public health decision making in the Netherlands. Nat Med. 2020. 26: pp. 1405-10

93. Africa CDC Pathogen Genomics Intelligence Institute. Africa Center for Disease Control and Prevention. 2020; Available from: https://africacdc.org/africa-cdc-institutes/africa-cdc-pathogen-genomics-intelligence-institute/

94. COVID-19 genome sequencing laboratory network launches in Africa. World Health Organization. 2020; Available from: https://www.afro.who.int/news/covid-19-genome-sequencing-laboratory-network-launches-africa

95. Jerving, S. Strengthening Africa's ability to 'decode' the coronavirus. 2020; Available from: https://www.devex.com/news/strengthening-africa-s-ability-to-decode-the-coronavirus-97319

96. Gámbaro, F., Behillil, S., Baidaliuk, A., et al. Introductions and early spread of SARS-CoV-2 in France. bioRxiv. 2020. p. 10.1101/2020.04.24.059576.

97. Inzaule, S. C., Tessema, S. K., Kedede, Y., et al. Genomic-informed pathogen surveillance in Africa: opportunities and chellenges. The Lancet Infectious Diseases. 2021. pp. 10.1016/S1473-3099(20)30939-7

98. COVID Network for Genomics Surveillance South Africa (NGS-SA). Kwazulu Natal Research Innovation and Sequencing Platform (KRISP). 2020; Available from: https://www.krisp.org.za/ngs-sa/who_we_are.php

99. SeqCOVID. 2020; Available from: http://seqcovid.csic.es/

100. Espadas, I. C., Candelas, F. G., Marin, A. M. G., et al. What do we know to date about the transmission of SARS-CoV-2? Archyde. 2020; Available from: https://www.archyde.com/what-do-we-know-to-date-about-the-transmission-of-sars-cov-2/

101. DBT launches study to sequence 1000 SARS Cov-2 genomes. BioSpectrum India. 2020; Available from: https://www.biospectrumindia.com/news/72/16345/dbt-launches-study-to-sequence-1000-sars-cov-2-genomes.html

102. Launch of 1000 Genome sequencing of SARS-Cov 2 Virus (section 2). DBT India. 2020; Available from: http://dbtindia.gov.in/sites/default/files/uploadfiles/DBT%20COVID%20Stories.pdf

103. Priyadarshini, S. Massive coronavirus sequencing efforts urgently need patient data. Nature India. 2020; Available from: https://www.natureasia.com/en/nindia/article/10.1038/nindia.2020.75

104. COVID-19 Network Investigations (CONI) Alliance. 2020; Available from: https://coni.team/

105. Joonlasak, K., Batty, E. M., Kochakarn, T., et al. Genomic surveillance of SARS-CoV-2 in Thailand reveals mixed imported populations, a local lineage expansion and a virus with truncated ORF7a. Virus Res. 2021. 292 p. 198233

106. Danish Covid-19 Genome Consortium DCGC. 2020; Available from: https://www.covid19genomics.dk/home

107. The COVID-19 Genomics UK (COG-UK) consortium and the Canadian COVID Genomics Network (CanCOGeN) launch new partnership. COG Consortium. 2020; Available from: https://www.cogconsortium.uk/news/the-covid-19-genomics-uk-cog-uk-consortium-and-the-canadian-covid-genomics-network-cancogen-launch-new-partnership/

108. Elbe, S., Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017. 1(1): pp. 33-46

109. GISAID's Comments on GloPID-R. Principles for Data Sharing in Public Health Emergencies. GISAID. 2017; Available from: https://www.gisaid.org/references/statements-clarifications/comments-on-glopid-r-principles-for-data-sharing-in-public-health-emergencies/

110. Global research collaboration for infectious disease preparedness (GLOPID-R). 2020; Available from: https://www.glopid-r.org/

111. Leinonen, R., Sugawara, H., Shumway, M., et al. The sequence read archive. Nucleic Acids Res. 2011. 39(Database issue): pp. D19-21.

112. Van Noorden, R. Scientists call for fully open sharing of coronavirus genome data. Nature. 2021. 590(7845): pp. 195-196

113. China National Center for Bioinformation. NGDC. 2021; Available from: https://bigd.big.ac.cn/ncov/release_genome

114. Oluniyi, P. First African SARS-CoV-2 genome sequence from Nigerian COVID-19 case. Virological. 2020; Available from: http://virological.org/t/first-african-sars-cov-2-genome-sequence-from-nigerian-covid-19-case/421

115. COVID-19 case in Cambodia. IDSeq. 2020; Available from: https://public.idseq.net/

116. Rohden, F., Huang, S., Droge, G. , et al. Combined study on Digital Sequence Information (DSI) in public and private databases and traceability. UN Convention on Biological Diversity. 2020; Available from: https://www.cbd.int/doc/c/1f8f/d793/57cb114ca40cb6468f479584/dsi-ahteg-2020-01-04-en.pdf

117. Black, A., MacCannell, D. R., Sibley, T. R., et al. Ten Recommendations for Supporting Open Pathogen Genomic Analysis in Public Health Settings. Nat Med. 2020. 26: pp. 832–841

118. WHO's code of conduct for open and timely sharing of pathogen genetic sequence data during outbreaks of infectious disease (DRAFT). World Health Organization. 2019; Available from: https://www.who.int/blueprint/what/norms-standards/gsdsharing/en/

119. Risk Assessment: Risk related to spread of new SARS-CoV-2 variants of concern in the EU/EEA. European Cenre for Disease Prevention and Control. 2020; Available from: https://www.ecdc.europa.eu/en/publications-data/covid-19-risk-assessment-spread-new-sars-cov-2-variants-eueea

120. Public Health Alliance for Genomic Epidemiology. 2020; Available from: https://pha4ge.org/

121. A New Sequence for Bioinformatics High Performance Computing. DELL Technologies. 2020; Available from: https://www.dellemc.com/resources/en-us/asset/customer-profiles-case-studies/products/ready-solutions/dell-ukhealth-cardiff-tech-case-study.pdf

122. White House Announces New Partnership to Unleash U.S. Supercomputing Resources to Fight COVID-19. Office of Science and Technology Policy. 2020; Available from: https://www.whitehouse.gov/briefings-statements/white-house-announces-new-partnership-unleash-u-s-supercomputing-resources-fight-covid-19/

123. Lemonick, S. Consortia lend their supercomputers to fight COVID-19. Chemical and Engineering News. 2020; Available from: https://cen.acs.org/physical-chemistry/computational-chemistry/Consortia-lend-supercomputers-fight-COVID/98/i15

124. UK joins US-led Covid-19 HPC consortium to tackle prevailing health crisis. Government Computing. 2020; Available from: https://www.hpcwire.com/off-the-wire/uk-joins-covid-19-high-performance-computing-consortium/

125. Partnership for Advanced Computing in Europe (PRACE). 2020; Available from: https://prace-ri.eu/

126. SARS-CoV-2. ARTIC Network. 2020; Available from: https://artic.network/ncov-2019

127. Karow, J. High-Throughput Diagnostic COVID-19 Sequencing Assays May Enable Large-Scale Testing. GenomeWeb. 2020; Available from: https://www.genomeweb.com/molecular-diagnostics/high-throughput-diagnostic-covid-19-sequencing-assays-may-enable-large-scale

128. Illumina COVIDSeq Test. Illumina. 2020; Available from: https://emea.illumina.com/products/by-type/ivd-products/covidseq.html

129. Lai, C. C., Wang, C. Y., Hsueh, P. R. Co-infections among patients with COVID-19: The need for combination therapy with non-anti-SARS-CoV-2 agents? J Microbiol Immunol Infect. 2020. p. 10.1016/j.jmii.2020.05.013

130. Octant SwabSeq Testing. 2020; Available from: https://www.notion.so/Octant-SwabSeq-Testing-9eb80e793d7e46348038aa80a5a901fd

131. Coronavirus (COVID-19) Update: FDA Authorizes First Next Generation Sequence Test for Diagnosing COVID-19. US Food and Drug Administration. 2020; Available from: https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-first-next-generation-sequence-test-diagnosing-covid-19

132. Febbo, P. Illumina's COVID testing service. Illumina. 2020; Available from: https://emea.illumina.com/company/news-center/feature-articles/illumina-covid-testing-service.html

133. LamPORE test for SARS-CoV-2 detection gains CE-IVD mark. Oxford Nanopore Technologies. 2020; Available from: https://nanoporetech.com/about-us/news/lampore-test-sars-cov-2-detection-gains-ce-ivd-mark

134. Large study in UK NHS labs shows gold-standard accuracy of Oxford Nanopore's COVID-19 test LamPORE for both symptomatic and asymptomatic patients. University of Birmingham. 2020; Available from: https://www.birmingham.ac.uk/news/latest/2020/12/large-study-in-uk-nhs-labs-shows-gold-standard-accuracy-of-covid-19-test.aspx

135. Open COVID License. Open COVID Pledge. 2020; Available from: https://opencovidpledge.org/licenses/v1-0/

136. Bloom, J. S., Jones, E. M. , Gasperini, M. , et al. Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. CaltechAUTHORS. 2020; Available from: https://resolver.caltech.edu/CaltechAUTHORS:20201119-132151980

137. Emergency Use Authorization (EUA) Summary: UCLA SwabSeq COVID-19 Diagnostic Platform. US Food and Drug Administration. 2020; Available from: https://www.fda.gov/media/142805/download

138. Illumina and IDbyDNA Partner to Deliver NGS Infectious Disease Solutions. Illumina. 2020; Available from: https://emea.illumina.com/company/news-center/feature-articles/illumina-and-idbydna-partner-to-deliver-ngs-infectious-disease-s.html

139. St Hilaire, B. G., Durand, N. C., Mitra, N., et al. A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. bioRxiv. 2020. p. 10.1101/2020.04.25.061499

140. Schmid-Burgk, J., Schmithausen, R. M., Li, D., et al. LAMP-Seq: Population-Scale COVID-19 Diagnostics Using Combinatorial Barcoding. bioRxiv. 2020. p. 10.1101/2020.04.06.025635

141. Ashford, M. Guardant Health Developing COVID-19 Test, Touts ctDNA Advantages for Pandemic Cancer Testing. GenomeWeb. 2020; Available from: https://www.genomeweb.com/business-news/guardant-health-developing-covid-19-test-touts-ctdna-advantages-pandemic-cancer

142. Karow, J. Clear Labs Raises $18M, Will Launch Diagnostic Nanopore Sequencing COVID-19 Assay System. GenomeWeb. 2020; Available from: https://www.genomeweb.com/business-news/clear-labs-raises-18m-will-launch-diagnostic-nanopore-sequencing-covid-19-assay-system

143. FAQs on Testing for SARS-CoV-2. US Food and Drug Administration. 2020; Available from: https://www.fda.gov/medical-devices/emergency-situations-medical-devices/faqs-testing-sars-cov-2

144. Johnson, M. FDA Provides Guidance on Pooled Samples, Asymptomatic Testing for SARS-CoV-2. GenomeWeb. 2020; Available from: https://www.genomeweb.com/pcr/fda-provides-guidance-pooled-samples-asymptomatic-testing-sars-cov-2

145. Booeshaghi, A. S., Lubock, N. B., Cooper, A. R., et al. Reliable and accurate diagnostics from highly multiplexed sequencing assays. Sci Rep. 2020. 10(1): p. 21759

146. Corman, V. M., Landt, O., Kaiser, M., et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020. 25(3): p. 2000045

147. Summary of available protocols. World Health Organisation. 2020; Available from: https://www.who.int/docs/default-source/coronaviruse/whoinhouseassays.pdf?sfvrsn=de3a76aa_2

148. van Kasteren, P. B., van der Veer, B., van den Brink, S., et al. Comparison of seven commercial RT-PCR diagnostic kits for COVID-19. J Clin Virol. 2020. 128: p. 104412

149. Diagnostic testing for SARS-CoV-2: interim guidance. World Health Organization. 2020; Available from: https://www.who.int/publications/i/item/diagnostic-testing-for-sars-cov-2

150. Wang, R., Hozumi, Y., Yin, C., et al. Mutations on COVID-19 diagnostic targets. Genomics. 2020. 112(6): pp. 5204-5213

151. Policy for Evaluating Impact of Viral Mutations on COVID-19 Tests. US Food and Drug Administration. 2021; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-evaluating-impact-viral-mutations-covid-19-tests.

152. Weissleder, R, Lee, H, Ko, J, et al. COVID-19 Diagnostics In Context. Center for Systems Biology, Massachusetts General Hospital. 2020; Available from: https://csb.mgh.harvard.edu/covid

153. Evaluation of endpoint PCR (EPCR) as a central laboratory based diagnostic test technology for SARS-CoV-2. Department of Health and Social Care. 2021; Available from: https://www.gov.uk/government/publications/evaluation-of-endpoint-pcr-epcr-as-a-diagnostic-test-technology-for-sars-cov-2/evaluation-of-endpoint-pcr-epcr-as-a-central-laboratory-based-diagnostic-test-technology-for-sars-cov-2

154. Emergency Use Authorization: Sherlock CRISPR SARS-CoV-2 Kit. US Food and Drug Administration. 2020; Available from: https://www.fda.gov/media/137747/download

155. ID NOW COVID-19 - Emergency Use Authorisation. . US Food and Drug Administration. 2020; Available from: https://www.fda.gov/media/136525/download

156. Han, M. S., Byun, J. H., Cho, Y., et al. RT-PCR for SARS-CoV-2: quantitative versus qualitative. Lancet Infect Dis. 2020. pp. S1473-3099(20)30424-2

157. UK Hospitals Evaluating DRW's Molecular POC Coronavirus Test, SAMBA II Platform. GenomeWeb. 2020; Available from: https://www.genomeweb.com/molecular-diagnostics/uk-hospitals-evaluating-drws-molecular-poc-coronavirus-test-samba-ii-platform

158. Collier, D. A., Assennato, S. M., Sithole, N., et al. Rapid point of care nucleic acid testing for SARS-CoV-2 in hospitalised patients: a clinical trial and implementation study. medRxiv. 2020. p. 10.1101/2020.05.31.20114520

159. BioMérieux's BioFire SARS-CoV-2 Respiratory Panel Wins FDA Emergency Use Authorization. GenomeWeb. 2020; Available from: https://www.genomeweb.com/regulatory-news-fda-approvals/biom-rieuxs-biofire-sars-cov-2-respiratory-panel-wins-fda-emergency

160. Press Release: BIOFIRE® Respiratory Panel 2.1 (RP2.1) with SARS-CoV-2 obtains FDA Emergency Use Authorization. Biomerieux. 2020; Available from: https://www.biomerieux.com/en/biofirer-respiratory-panel-21-rp21-sars-cov-2-obtains-fda-emergency-use-authorization

161. TaqPath COVID-19 Multiplex Diagnostic Solution (CE-IVD). Thermo Fisher Scientific. 2020; Available from: https://www.thermofisher.com/uk/en/home/clinical/clinical-genomics/pathogen-detection-solutions/taqpath-covid-19-diagnostic-kit.html

162. SARS-COV-2 Diagnostic Pipeline. Foundation for Innovative New Diagnostics (FIND). 2020; Available from: https://www.finddx.org/covid-19/pipeline/

163. COVID-19 In Vitro Diagnostic Devices and Test Methods Database. European Commission. 2020; Available from: https://covid-19-diagnostics.jrc.ec.europa.eu/devices

164. Emergency Use Authorization (COVID-19 in vitro tests). US Food and Drug Administration. 2020; Available from: https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization#covidinvitrodev

165. LitCovid. US National Insitutes of Health. 2021; Available from: https://www.ncbi.nlm.nih.gov/research/coronavirus/

166. COVID-19. Global literature on coronavirus disease. World Health Organization. 2021; Available from: https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/

167. COVID-19 variants. CoVariants. 2021; Available from: https://covariants.org

168. PANGO lineages. CovLineages. 2021; Available from: https://cov-lineages.org

169. Korber, B., Fischer, W. M., Gnanakaran, S., et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020. 182(4): pp. 812-827 e19

170. COVID-19 Viral Genome Analysis Pipeline. Los Alamos National Laboratory. 2021; Available from: https://cov.lanl.gov/content/index

171. Caly, L., Druce, J., Roberts, J., et al. Isolation and rapid sharing of the 2019 novel coronavirus (SARS-CoV-2) from the first patient diagnosed with COVID-19 in Australia. Med J Aust. 2020. 212(10): pp. 459-462

172. Goes de Jesus, J., Sacchi, C., Candido, D. D. S., et al. Importation and early local transmission of COVID-19 in Brazil, 2020. Rev Inst Med Trop Sao Paulo. 2020. 62: p. e30

173. Böhmer, M. M., Buchholz, U., Corman, V. M., et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. The Lancet Infectious Diseases. 2020. pp. 10.1016/S1473-3099(20)30314-5

174. Yadav, P. D., Potdar, V. A., Choudhary, M. L., et al. Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res. 2020. 151(2 & 3): pp. 200-209

175. Capobianchi, M. R., Rueca, M., Messina, F., et al. Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. Clin Microbiol Infect. 2020. 26(7): pp. 954-956

176. Park, W. B., Kwon, N. J., Choi, S. J., et al. Virus Isolation from the First Patient with SARS-CoV-2 in Korea. J Korean Med Sci. 2020. 35(7): p. e84

177. Sah, R., Rodriguez-Morales, A. J., Jha, R., et al. Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Isolated in Nepal. Microbiol Resour Announc. 2020. 9(11): pp. e00169-20

178. Kim, J. M., Chung, Y. S., Jo, H. J., et al. Identification of Coronavirus Isolated from a Patient in Korea with COVID-19. Osong Public Health Res Perspect. 2020. 11(1): pp. 3-7

179. Kanteh, A., Manneh, J., Jabang, S., et al. Origin of imported SARS-CoV-2 strains in The Gambia identified from Whole Genome Sequences. bioRxiv. 2020. p. 10.1101/2020.04.30.070771

180. Meredith, L. W., Hamilton, W. L., Warne, B., et al. Rapid implementation of real-time SARS-CoV-2 sequencing to investigate healthcareassociated COVID-19 infections. medRxiv. 2020. p. 10.1101/2020.05.08.20095687

181. Salazar, C., Díaz-Viraqué, F., Pereira-Gómez, M., et al. Multiple introductions, regional spread and local differentiation during the first week of COVID-19 epidemic in Montevideo, Uruguay. bioRxiv. 2020. p. 10.1101/2020.05.09.086223

182. Covid-19 Investigation Team. Clinical and virologic characteristics of the first 12 patients with coronavirus disease 2019 (COVID-19) in the United States. Nat Med. 2020. 26(6): pp. 861-868

183. Gonzalez-Reiche, A. S., Hermandez, M. M., Sullivan, M., et al. Introductions and early spread of SARS-CoV-2 in the New York City area. medRxiv. 2020. p. 10.1101/2020.04.08.20056929

184. Bedford, T., Greninger, A. L., Roychoudhury, P., et al. Cryptic transmission of SARS-CoV-2 in Washington State. medRxiv. 2020. p. 10.1101/2020.04.02.20051417

185. Tang, X., Lu, J., Cui, J., et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020. 7(6): pp. 1012–1023

186. Zhan, S. H., Deverman, B. E., Chan, Y. A. SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence? bioRxiv. 2020. p. 10.1101/2020.05.01.073262

187. Lodder, W., de Roda Husman, A. M. SARS-CoV-2 in wastewater: potential health risk, but also data source. The Lancet Gastroenterology & Hepatology. 2020. 5(6): pp. 533-534

188. Rimoldi, S. G., Stefani, F., Gigantiello, A., et al. Presence and vitality of SARS-CoV-2 virus in wastewaters and rivers. medRxiv. 2020. p. 10.1101/2020.05.01.20086009

189. MacLean, O. A., Orton, R. J., Singer, J. B., et al. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evol. 2020. 6(1): p. veaa034

190. Forster, P., Forster, L., Renfrew, C., et al. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A. 2020. 117(17): pp. 9241-9243

191. Mavian, C., Pond, S. K., Marini, S., et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. Proc Natl Acad Sci U S A. 2020. 117(23): pp. 12522-12523

192. van Dorp, L., Richard, D., Tan, C. C. S., et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. bioRxiv. 2020. p. 10.1101/2020.05.21.108506

193. Tegally, H., Wilkinson, E., Giovanetti, M., et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv. 2020. p. 10.1101/2020.12.21.20248640

194. Bedford, T., Hodcroft, E. B., Neher, R. A. Updated Nextstrain SARS-CoV-2 clade naming strategy. NextStrain. 2021; Available from: https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming

195. Rambaut, A., Holmes, E. C., O'Toole, A., et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020. 5(11): pp. 1403-1407

196. Callaway, E. 'A bloody mess': Confusion reigns over naming of new COVID variants. Nature. 2021; Available from: https://www.nature.com/articles/d41586-021-00097-w

197. Weekly epidemiological update - 25 February 2021. World Health Organization. 2021; Available from: https://www.who.int/publications/m/item/covid-19-weekly-epidemiological-update

198. The COVID-19 Host Genetic Initiative. 2020; Available from: https://www.covid19hg.org

199. Nasir, J., Wolford, B., Veerapen, K. COVID-19 HGI Results for Data Freeze 4 (October 2020). COVID-19 hg. 2020; Available from: https://www.covid19hg.org/blog/2020-11-24-covid-19-hgi-results-for-data-freeze-4-october-2020/

200. Pairo-Castineira, E., Clohisey, S., Klaric, L., et al. Genetic mechanisms of critical illness in Covid-19. medRxiv. 2020. p. 10.1101/2020.09.24.20200048

201. Qi, F., Qian, S., Zhang, S., et al. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. Biochem Biophys Res Commun. 2020. 526(1): pp. 135-140

202. Sungnak, W., Huang, N., Becavin, C., et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med. 2020. 26(5): pp. 681-687

203. Wen, W., Su, W., Tang, H., et al. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discov. 2020. 6: p. 31

204. Zhao, Y., Zhao, Z., Wang, Y., et al. Single-Cell RNA expression profiling of ACE2, the receptor of SARS-CoV-2. bioRxiv. 2020. p. 10.1101/2020.01.26.919985

205. Xiong, Y., Liu, Y., Cao, L., et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020. 9(1): pp. 761-770

206. Peddu, V., Shean, R. C., Xie, H., et al. Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. Clin Chem. 2020. 66(7): pp. 966-972

207. Moore, S. C., Penrice-Randal, R., Alruwaili, M., et al. Amplicon based MinION sequencing of SARS-CoV-2 and metagenomic characterisation of nasopharyngeal swabs from patients with COVID-19. medRxiv. 2020. p. 10.1101/2020.03.05.20032011

208. Grifoni, A., Sidney, J., Zhang, Y., et al. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe. 2020. 27(4): pp. 671-680 e2

209. Ahmed, S. F., Quadeer, A. A., McKay, M. R. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses. 2020. 12(3): p. 254

210. Zhou, Y., Hou, Y., Shen, J., et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. 2020. 6: p. 14

211. COVID-3D: An online resource to explore the structural distribution of genetic variation in SARS-CoV-2 and its implication on therapeutic development. University of Melbourne. 2020; Available from: http://biosig.unimelb.edu.au/covid3d/

212. COG-UK mutation explorer. COG-UK Consortium 2021; Available from: http://sars2.cvr.gla.ac.uk/cog-uk/

213. COVID-19 vaccine tracker. Regulatory Affairs Professionals Society. 2021; Available from: https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker

214. NextSeq 550Dx - The next level in diagnostic power for the clinical lab. Illumina. 2020; Available from: https://emea.illumina.com/systems/sequencing-platforms/nextseq-dx.html

215. AmpliSeq for Illumina SARS-CoV-2 Research Panel. Illumina. 2020; Available from: https://emea.illumina.com/products/by-brand/ampliseq/community-panels/sars-cov-2.html

216. Our software toolkit supports the fight against SARS-CoV-2. Illumina. 2020; Available from: https://emea.illumina.com/informatics/specialized-bioinformatics-applications/coronavirus-software.html

217. Illumina Makes Software Toolkit Available Free of Charge to Support Worldwide Efforts to Combat COVID-19. Illumina. 2020; Available from: https://emea.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=6a75de1d-9c92-4aee-9217-bf4d767aed56

218. Illumina Supporting COVID-19 efforts. Illumina. 2020; Available from: https://www.illumina.com/company/supporting-covid-19-efforts.html

219. Illumina partners with Africa CDC to strengthen sequencing capacity for COVID-19 surveillance in Africa. Africa Center for Disease Control and Prevention. 2020; Available from: https://africacdc.org/news-item/illumina-partners-with-africa-cdc-to-strengthen-sequencing-capacity-for-covid-19-surveillance-in-africa/

220. Ion Torrent Genexus System. Thermo Fisher Scientific. 2020; Available from: https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-torrent-genexus-system.html

221. Thermo Fisher Scientific Announces SARS-CoV-2 GlobalAccess Sequencing Program. Thermo Fisher. 2020; Available from: https://thermofisher.mediaroom.com/2020-05-07-Thermo-Fisher-Scientific-Announces-SARS-CoV-2-GlobalAccess-Sequencing-Program

222. Italian Researchers Identify New SARS-CoV-2 Gene Variants That Provide Clues to Coronavirus's Epidemiology. PR Newswire. 2020; Available from: https://www.prnewswire.co.uk/news-releases/italian-researchers-identify-new-sars-cov-2-gene-variants-that-provide-clues-to-coronavirus-s-epidemiology-851685069.html

223. BGI Group helping over 80 countries for timely COVID-19 detection and intervention. PR Newswire. 2020; Available from: https://www.prnewswire.com/news-releases/bgi-group-helping-over-80-countries-for-timely-covid-19-detection-and-intervention-301043895.html

224. BGI's Coronavirus Response? Build a Lab in Wuhan. Genetic Engineereing & Biotechnology News. 2020; Available from: https://www.genengnews.com/insights/bgis-coronavirus-response-build-a-lab-in-wuhan-in-a-week/

225. New Emergency Detection Laboratory Run by BGI Starts Trial Operation in Wuhan, Designed to Test 10,000 Samples Daily. BGI. 2020; Available from: https://www.bgi.com/global/company/news/new-emergency-detection-laboratory-run-by-bgi-starts-trial-operation-in-wuhan-designed-to-test-10000-samples-daily/

226. BGI Sequencer, Coronavirus Molecular Assays Granted Emergency Use Approval in China. GenomeWeb. 2020; Available from: https://www.genomeweb.com/regulatory-news-fda-approvals/bgi-sequencer-coronavirus-molecular-assays-granted-emergency-use

227. Curetis Group Company Ares Genetics and BGI Group Collaborate to Offer Next-Generation Sequencing and PCR-based Coronavirus (2019-nCoV) Testing in Europe. GlobeNewswire. 2020; Available from: https://www.globenewswire.com/news-release/2020/01/30/1977226/0/en/Curetis-Group-Company-Ares-Genetics-and-BGI-Group-Collaborate-to-Offer-Next-Generation-Sequencing-and-PCR-based-Coronavirus-2019-nCoV-Testing-in-Europe.html

228. EIT Health-supported Ares Genetics launches testing services for control and monitoring of COVID-19. European insitute of innovation & technology. 2020; Available from: https://eit.europa.eu/our-activities/covid-19-response/solutions/eit-health-supported-ares-genetics-launches-testing

229. Zhao, W.-M., Song, S.-H., Chen, M.-L., et al. The 2019 novel coronavirus resource. Yi Chuan. 2020. 42(2): pp. 212–221. Available at: https://bigd.big.ac.cn/ncov

230. PacBio - Certified Service Providers. Pacific Biosciences,. 2020; Available from: https://www.pacb.com/products-and-services/service-providers/

231. Understanding Coronavirus with PacBio Sequenicng. Pacific Biosciences. 2020; Available from: https://www.pacb.com/research-focus/microbiology/COVID-19-sequencing-tools-and-resources/.

232. Pacific Biosciences Powers SARS-CoV-2 Research at Commercial, Academic and Government Labs. Pacific Biosciences. 2020; Available from: https://www.pacb.com/press_releases/pacific-biosciences-powers-sars-cov-2-research-at-commercial-academic-and-government-labs/

233. Pacific Biosciences Shares Spike on SARS-CoV-2 Research Initiatives. GenomeWeb. 2020; Available from: https://www.genomeweb.com/sequencing/pacific-biosciences-shares-spike-sars-cov-2-research-initiatives

234. SARS-CoV-2 Genetics: Analysis of intrapatient SARS-Cov-2 genetic variation. National Insitute of Allergy and Infectious Diseases. 2020; Available from: https://www.niaid.nih.gov/research/sars-cov-2-genetics

235. Quick, J., Loman, N. J., Duraffour, S., et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016. 530(7589): pp. 228-232

236. Xu, Y., Lewandowski, K., Jeffery, K., et al. Nanopore metagenomic sequencing to investigate nosocomial transmission of human metapneumovirus from a unique genetic group among haematology patients in the United Kingdom. J Infect. 2020. 80(5): pp. 571-577

237. Rhodes, J., Abdolrasouli, A., Farrer, R. A., et al. Genomic epidemiology of the UK outbreak of the emerging human fungal pathogen Candida auris. Emerg Microbes Infect. 2018. 7(1): p. 43

238. Timeline: community work. Oxford Nanopore Technologies. 2021; Available from: https://nanoporetech.com/covid-19/community-timeline

239. ARTIC Network - Real-Time Molecular Epidemiology for Outbreak Response. 2020; Available from: https://artic.network/

240. Instruments for Sanger Sequencing and Fragment Analysis by Capillary Electrophoresis. Thermo Fisher Scientific. 2020; Available from: https://www.thermofisher.com/uk/en/home/life-science/sequencing/sanger-sequencing/sanger-sequencing-technology-accessories.html.

241. Resende, P. C., Motta, F. C., Roy, S., et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. bioRxiv. 2020. p. 10.1101/2020.04.30.069039

242. Gohl, D. M., Garbe, J., Grady, P., et al. A Rapid, Cost-Effective Tailed Amplicon Method for Sequencing SARS-CoV-2. bioRxiv. 2020. p. 10.1101/2020.05.11.088724

243. Wang, M., Fu, A., Hu, B., et al. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. medRxiv. 2020. p. 10.1101/2020.03.04.20029538

244. Sequencing of SARS-CoV-2 - first update. European Centre for Disease Prevention and Control. 2021; Available from: https://www.ecdc.europa.eu/en/publications-data/sequencing-sars-cov-2

245. Pastorino, B., Touret, F., Gilles, M., et al. Evaluation of heating and chemical protocols for inactivating SARS-CoV-2. Viruses. 2020. 12(6): p. 624

246. The Crick COVID-19 Consortium. The Crick Institute. 2020; Available from: https://www.crick.ac.uk/research/covid-19/covid19-consortium

247. DNA/RNA Shield reagent. Cambridge Bioscience. 2020; Available from: https://www.bioscience.co.uk/cpl/dna-rna-shield-reagent

248. DNA/RNA Shield. Zymo Research. 2020; Available from: https://www.zymoresearch.com/collections/dna-rna-shield/products/dna-rna-shield.

249. Batéjat, C., Grassin, Q., Manuguerra, J.-C., et al. Heat inactivation of the Severe Acute Respiratory Syndrome Coronavirus 2. J Biosaf Biosec. 2021. 3(1): pp. 1-3

250. Kampf, G., Voss, A., Scheithauer, S. Inactivation of coronaviruses by heat. J Hosp Infect. 2020. 105(2): pp. 348-349

251. Wang, T., Lien, C., Liu, S., et al. Effective Heat Inactivation of SARS-CoV-2. medRxiv. 2020. p. 10.1101/2020.04.29.20085498

252. QIAamp Viral RNA Mini Kit. Qiagen. 2020; Available from: https://www.qiagen.com/ie/products/diagnostics-and-clinical-research/sample-processing/qiaamp-viral-rna-mini-kit

253. High-throughput SARS-CoV-2 RNA extraction. Cambridge Bioscience. 2020; Available from: https://www.bioscience.co.uk/cpl/viral-rna-extraction-covid-19-detection

254. COVID-19 Viral RNA Extraction Kits & Related Products. NBS Biologicals. 2020; Available from: https://www.nbsbio.co.uk/products/covid-19

255. innuPREP Virus TS RNA Kit. Analytik Jena. 2020; Available from: https://www.analytik-jena.com/products/kits-assays-reagents/kits-for-dnarna-extraction/innuprep-virus-ts-rna-kit/.

256. VolTRAX. Oxford Nanopore Technologies. 2020; Available from: https://nanoporetech.com/products/voltrax

257. SalivaDirect: RNA extraction-free SARS-CoV-2 diagnostics V.1. 2020; Available from: https://www.protocols.io/view/salivadirect-rna-extraction-free-sars-cov-2-diagno-bhfej3je

258. Brown, J. R., Atkinson, L., Shah, D., et al. Validation of an extraction-free RT-PCR protocol for detection of SARS-CoV2 RNA. medRxiv. 2020. p. 10.1101/2020.04.29.20085910

259. Beltrán-Pavez, C., Márquez, C. L., Muñoz, G., et al. SARS-CoV-2 detection from nasopharyngeal swab samples without RNA extraction. bioRxiv. 2020. p. 10.1101/2020.03.28.013508

260. CoV-GLUE. University of Glasgow. 2021; Available from: http://cov-glue.cvr.gla.ac.uk/#/home

261. Pan African Bioinformatics Network for H3Africa. H3ABioNet. 2021; Available from: https://www.h3abionet.org/

262. Baichoo, S., Souilmi, Y., Panji, S., et al. Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. BMC Bioinformatics. 2018. 19(1): p. 457

263. Galaxy Community Hub. Galaxy Project. 2020; Available from: https://galaxyproject.org/.

264. ELIXIR. 2020; Available from: https://elixir-europe.org/

265. MacCannell, D. SARS-CoV-2 Sequencing Resources. GitHub. 2020; Available from: https://github.com/CDCgov/SARS-CoV-2_Sequencing

266. SARS-CoV-2 sequencing protocols. COG-UK. 2021; Available from: https://www.cogconsortium.uk/tools-analysis/public-data-analysis/

267. Starting with COVID-19: whole genome sequencing. Oxford Nanopore Technologies 2020; Available from: https://nanoporetech.com/covid-19/workflows

268. CLC workflow for CDC NCIRD/DVD ONT & Illumina NEBNext Sequencing Protocol 2020; Available from: https://github.com/CDCgov/SARS-CoV-2_Sequencing/issues/11

269. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol V.4. Wellcome Sanger Institute. 2020; Available from: https://www.protocols.io/view/covid-19-artic-v3-illumina-library-construction-an-bgxjjxkn/abstract

270. Protocol for use with NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (E7645, E7103). New England Biolabs. 2020; Available from: https://international.neb.com/protocols/2015/09/16/protocol-for-use-with-nebext-ultra-ii-dna-library-prep-kit-for-illumina-and-with-sample-purification-beads-e7645-e7103

271. Pillay, S., Giandhari, J., Tegally, H., et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation During a Pandemic. Genes (Basel). 2020. 11(8): p. 949

272. Enrichment workflow for detecting coronavirus using Illumina NGS systems. Illumina 2020; Available from: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/ngs-enrichment-coronavirus-app-note-1270-2020-002.pdf

273. Illumina DNA Prep with Enrichment. Illumina 2020; Available from: https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-flex-enrichment.html

274. Comprehensive workflow for detecting coronavirus using Illumina benchtop systems. Illumina. 2020; Available from: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/ngs-coronavirus-app-note-1270-2020-001.pdf

275. Direct RNA Sequencing Protocol for the MinION™ using SQK-RNA001. Oxford Nanopore Technologies 2020; Available from: https://store.nanoporetech.com/media/wysiwyg/Example_Direct_RNA_Sequencing_Protocol_v4.pdf

276. Viehweger, A., Krautwurst, S., Lamkiewicz, K., et al. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 2019. 29(9): pp. 1545-1554

277. Epitranscriptomics and RNA Dynamics Lab (Novoa Lab) and the Bioinformatics Core Facility (BioCore) at the CRG: Uniform analysis of coronavirus and SARS-COV-2 nanopore direct RNA sequencing datasets using MasterOfPores. 2020; Available from: https://biocorecrg.github.io/covid/

278. Illumina COVIDSeq Test Instructions for Use. Illumina. 2020; Available from: https://emea.support.illumina.com/clinical_support/clinical_kits/illumina-covidseq-test-ivd/documentation.html

279. COVID-19 Strategic Preparedness and Response Plan (SPRP 2021). World Health Organization. 2021; Available from: https://www.who.int/publications/m/item/covid-19-strategic-preparedness-and-response-plan