SARS-CoV-2 genomic sequencing for public health goals

Interim guidance

8 January 2021

World Health Organization

Key messages:

- Global surveillance of SARS-CoV-2 genetic sequences and related metadata contributes to the COVID-19 outbreak response. This contribution includes tracking the spread of SARS-CoV-2 geographically over time and ensuring that mutations that could potentially influence pathogenicity, transmission or countermeasures (such as vaccines, therapeutics and diagnostics) are detected and assessed in a timely manner.
- While the cost and complexity of genetic sequencing have dropped significantly over time, effective sequencing programmes still require substantial investment in terms of staff, equipment, reagents and bioinformatic infrastructure. Additionally, effective collaboration is needed to ensure that generated data are of good quality and are used in a meaningful way.
- Countries are encouraged to rapidly deposit SARS-CoV-2 sequences in a public database in order to share them with the scientific community for public health purposes. Investments in a tiered global sequencing network for SARS-CoV-2 will contribute to the development of resilient, high-quality global sequencing programmes for the detection and management of other outbreak pathogens in the future.

Background

Over the last decade, genetic sequence data (GSD) of pathogens have come to play a pivotal role in the detection and management of infectious disease outbreaks, supporting the development of diagnostics, drugs and vaccines, and informing the outbreak response *(1–11)*. With the emergence of the novel coronavirus, later named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the importance of GSD has been further underlined. More than 280 000 full genome sequences have been shared via publicly accessible databases within a year of the initial identification of SARS-CoV-2 *(12)*. Near real-time analysis of data has directly impacted the public health response *(12–16)*. The public health objectives of SARS-CoV-2 genomic sequencing are listed in Table 1.

The growing understanding of how sequence information can contribute to improved public health is driving global investments in sequencing facilities and programmes. The falling cost and complexity of generating GSD provides opportunities for expanding sequencing capacity; however, challenges to widespread implementation remain, and sequencing capacity and data are not evenly distributed around the world, with an overrepresentation of SARS-CoV-2 GSD from high-income countries.

Table 1. Public health objectives of SARS-CoV-2 genomic sequencing

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 longer period of time**	
- Identify SARS-CoV-2 as the causative agent of 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 reinfection.	**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).	**Monitor viral movement and activity:** - Investigate geographic spread and reintroductions 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.

Purpose of this document

This document provides national-level policy-makers and stakeholders with guidance on how to maximize the public health benefit of SARS-CoV-2 genomic sequencing activities in the short and long term as the pandemic continues to unfold. Practical considerations for the implementation of a virus genomic sequencing programme and an overview of the public health objectives of genomic sequencing are covered. This guidance focuses on SARS-CoV-2 but is applicable to other pathogens of public health concern. It is recommended that countries wishing to build sequencing capacity for SARS-CoV-2 do so as part of a broader plan to build capacity to detect and monitor other pathogens of public health concern.

Additional WHO guidance

WHO has developed the implementation guide Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health in collaboration with sequencing experts around the world. This guide provides a more complete background on SARS-CoV-2 sequencing and is intended for those who are actively involved in implementing sequencing programmes *(17)*. It provides an in-depth review of the various uses of sequencing and gives technical advice on pathogen sequencing in the context of SARS-CoV-2. Beyond reading these and other published documents, laboratories with limited sequencing experience should actively look for opportunities to collaborate with experienced laboratories and/or join or form laboratory networks with sequencing expertise.

1. Introduction to SARS-CoV-2

SARS-CoV-2 is classified within the genus *Betacoronavirus* (subgenus *Sarbecovirus*) of the family *Coronaviridae (18)*. It is an enveloped, positive-sense, single-stranded ribonucleic acid (RNA) virus with an approximately 30kb genome *(19)*. Genetic sequencing enables the reading of viral genomes. As each pathogen has a unique genomic sequence, this method can be used to identify novel pathogens (as in the case of SARS-CoV-2) *(20)*. The SARS-CoV-2 genome encodes for non-structural proteins, four structural proteins (spike [S], envelope [E], membrane [M], nucleocapsid [N]) and several putative accessory proteins *(21–23)*. SARS-CoV-2 host cell entry requires binding of the viral S protein to the host cell angiotensin-converting enzyme 2 (ACE-2) receptor *(24–27)*. The SARS-CoV-2 spike protein, especially the receptor-binding domain (RBD), is a critical target for natural and vaccine-induced immunity *(28–32)*. Therefore, diversification of the gene encoding the spike protein could potentially impact vaccine efficacy, natural immunity and (monoclonal) antibody therapies *(33)*.

When viruses replicate, especially RNA viruses such as SARS-CoV-2, changes (mutations) occur in the genome. If an acquired mutation does not have an evolutionary disadvantage, it may become fixed in SARS-CoV-2 populations. The rate of evolutionary change in SARS-CoV-2 is currently estimated to be 1×10^{-3} substitutions per site per year at the nucleotide level *(34)*. This translates into approximately one substitution in the genome every two weeks *(35)*. This relatively low rate of evolution limits the time resolution of individual transmission events *(35)*. Studying SARS-CoV-2 evolution and rapidly identifying substitutions, insertions or deletions that could impact viral properties (phenotypic change) is an important tool for epidemic monitoring. Among the most obvious yields of such work is the detection of mutations that are associated with changes in the transmissibility and/or pathogenicity of the virus, or that could reduce the utility of medical countermeasures (diagnostics, vaccines and therapeutics). Following virus mutations over time and space can also help to track the spread of the pathogen and support an enhanced understanding of potential transmission routes and dynamics. Reconstructing the evolutionary history of pathogens can be achieved through phylogenetic analysis. Phylogenetic and phylodynamic (i.e., how virus phylogenies are shaped by epidemiological and evolutionary processes) analyses can provide extensive information to support outbreak response.

2. Establishing an optimal SARS-CoV-2 sequencing approach in the local context

2.1 Context-specific prioritization of sequencing objectives and approach

Although the cost of gene sequencing has fallen significantly over the past decades, sequencing still requires substantial investment in resources (financial, infrastructure and human). Before initiating a sequencing project, the critical first step is to determine whether sequencing is truly valuable for reaching a specific objective or whether there are other more time-effective or cost-effective approaches available. This decision may involve considering whether virus sequencing alone is sufficient to reach the defined objective or whether it should be included as a smaller component within a multidisciplinary approach. Epidemiologically focused activities that integrate genomic data analysts directly into public health investigation and response teams are likely to have a greater immediate impact than those in which virus genomic analysis exists as a separate or secondary activity.

Where resources to support sequencing are limited, it may be necessary to limit objectives of a sequencing programme to those activities with high clinical and/or public health potential, which can be sustained. Such a programme may prioritize the sequencing of SARS-CoV-2 i) from individuals vaccinated for SARS-CoV-2 but who later become infected with SARS-CoV-2 despite exhibiting an appropriate immune response to the vaccine; ii) in risk settings, such as where there is close human–animal interaction with a large number of animals that are susceptible to SARS-CoV-2 infection, or where there are immunocompromised patients with prolonged shedding, especially when receiving antibody therapy against SARS-CoV-2; iii) when there is an unexpected increase or change in SARS-CoV-2 transmissibility and/or virulence; iv) when there is suspicion of a change in the performance of diagnostic (antibody, antigen, molecular assays) methods or therapies; and v) during cluster investigations when sequencing can support understanding of transmission events and/or evaluate the efficacy of infection control procedures.

Fig. 1 provides an overview of the basic pillars required for sequencing. If there is no or limited capacity available in all three pillars, it will likely be necessary to build partnerships with other groups in order to achieve sequencing goals. Conversely, if there is adequate capacity and resources for one or more pillars, the laboratory could consider supporting other partners with nascent sequencing programmes. Variable demand on capacity will occur throughout different phases of an outbreak and might require laboratories to shift from one strategy to another.

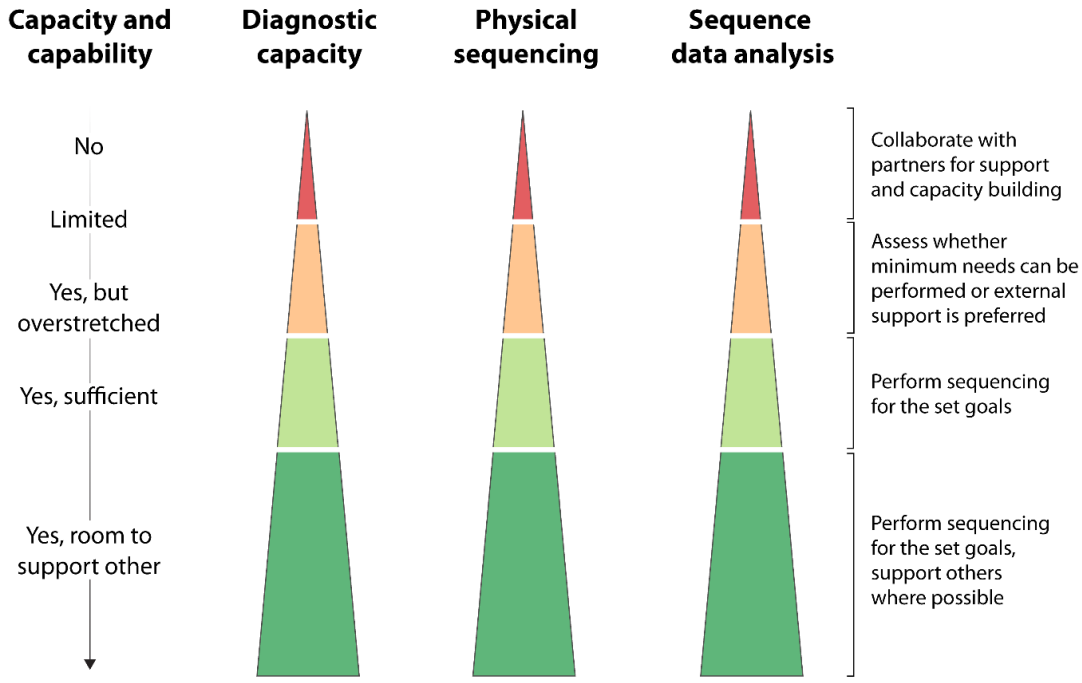

Fig. 1. Establishing capacity in the pillars required for sequencing in order to select the best approach for the setup of a sequencing programme

2.2 Investing in global sustainable sequencing capacity for SARS-CoV-2 and other (emerging/re-emerging) pathogens of public health concern

Tiered WHO laboratory networks have proven functionality and enable global collaboration, with regional adaptation of networks to specific national and regional needs *(36–39)*. Building such a strong and resilient global sequencing network can maximize the public health impact of sequencing for SARS-CoV-2 and emerging/re-emerging pathogens. Currently, the WHO reference laboratories providing confirmatory testing for COVID-19 are supporting some of these sequencing and analysis needs *(40)*. Several regions have, or are in the process of developing, sequencing capacities that will be able to join the global network of laboratories/sequencing groups. To determine the contribution laboratories in the network can make, a global estimation of capacity for each pillar listed in Fig. 1 can be undertaken. Various pathogen-specific laboratory networks (such as those working on antimicrobial resistance, MERS-CoV, influenza, measles, rubella, poliovirus and tuberculosis) have invested in sequencing capacity as part of their surveillance activities *(8, 9, 41–43)*. As the costs of sequencing are still substantial and many parts of the sequencing workstream can be used for various pathogens or sequencing objectives, national collaboration to ensure the optimal use of existing capacity is encouraged. Capacity-building programmes should focus on a stepwise approach to build competencies. The priorities of capacity building should be context-dependent. For some countries, building wet-laboratory capacity would make sense, whereas in other settings, outsourcing the actual sequencing and focusing on bioinformatics, data management and interpretation would have a bigger impact. For effective collaboration, data sharing, standardized protocols for sequencing, joint meetings and training, audits, proficiency testing (sequencing and analysis) and the development of reference standards for the evaluation of different procedures will support the further development of high-quality sequencing programmes for SARS-CoV-2 and for the detection of and response to future emerging pathogens. Where samples are shared in a network, appropriate mechanisms to ship samples under adequate requirements should also be in place.

3. Practical considerations for the implementation of a virus genomic sequencing programme

Here, we provide a general overview of the technical requirements to establish a sequencing programme. For detailed information regarding SARS-CoV-2 sequencing, refer to the full SARS-CoV-2 sequencing implementation guide *(17)*.

3.1 Practical considerations when developing a SARS-CoV-2 sequencing programme

The sequencing objectives will determine the design of the sequencing workstream (Table 1). Relevant key questions to aid this process can be found in Annex I. Annex II contains a checklist with considerations for planning a SARS-CoV-2 sequencing programme. Fig. 2 depicts the workstream for SARS-CoV-2 whole genome sequencing. All staff involved in a sequencing programme should receive the appropriate training and instruction to comply with the mandated task. Key stakeholders should be identified, consulted and involved at an early stage. Stakeholders to be engaged when developing sequencing programmes include public health bodies, diagnostic laboratories, sequencing facilities, analytical groups, and, depending on the setting, infection prevention and control teams or occupational health services, patient advocacy groups, and other institutions involved in human–animal interface research where appropriate. Communication channels and pathways that are aligned with the objectives of the programme should be developed and maintained throughout the project in order to ensure that the sequencing data are used to the greatest effect. Regular evaluations of the project's progress and end evaluation are key to ensure that lessons are learned and improvements are made where needed. Successfully achieving the objectives of a sequencing programme focused on emerging pathogens requires the involvement of experts in different fields: (i) wet-laboratory sequencing and safe handling of virus samples; (ii) generation of accurate genomes from raw data; (iii) analysis of genomes to generate meaningful results that are useful for the outbreak response; and (iv) pathogens. Most experts will be skilled in only one or two of these areas. For (ii) and (iii), powerful computer resources are required to achieve fast results. Collaboration between experts with different skillsets and pooling of resources is therefore often key to generate timely, accurate and effective results that can truly impact public health.

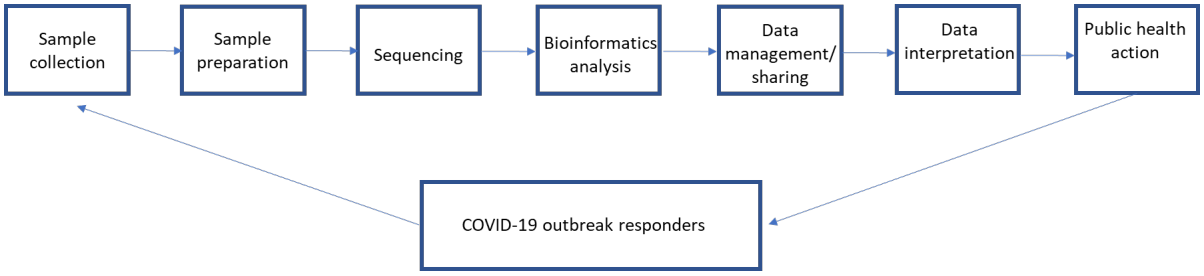

Fig. 2. Workstream for SARS-CoV-2 whole genome sequencing. Note that successfully implementing this workstream will involve bilateral communication between experts involved at different stages; for example, those conducting data interpretation would ideally directly discuss which samples will be chosen from sequencing with those involved in sample selection and preparation.

3.2 Ethical considerations

It is important to review ethical implications when designing a sequencing programme. Possible risks of social harm to research participants should be identified, and mitigation strategies should be defined. Any proposed investigations should be evaluated and approved by an ethical review committee, which takes into account the social value, scientific validity, participant selection, risk–benefit ratio, informed consent, and ongoing respect for participants *(44–46)*. Where researchers are not experienced in identifying possible ethical issues surrounding the sequencing of outbreak pathogens such as SARS-CoV-2, international collaboration and engagement of such expertise is strongly encouraged *(44)*. Collaboration between researchers across the world should ensure equitable and mutually beneficial collaborative research partnerships. Local researchers should be encouraged to take leading and active roles throughout the research process, as they are more likely to understand their health care and research systems and be able to translate results into policy *(44, 45)*. Ethical considerations for genomic sequence and metadata sharing are discussed in section 3.7.

3.3 Considerations for sampling strategy and sample preparation

Once goals have been identified, an appropriate sampling strategy needs to be developed with relevant stakeholders. Details on sampling can be found in the SARS-CoV-2 sequencing implementation guide *(17)*. Ideally, the reasons for the choice of specimens for sequencing should be recorded in the metadata, as the inclusion of non-random subsets of samples can affect the reliability of certain genetic analyses such as phylogenetic and phylodynamic analyses. Practical advice on how to collect clinical samples is covered in the SARS-CoV-2 diagnostic guidance *(47)*. Before sequencing, it is recommended to enrich the sample for SARS-CoV-2 genetic material relative to other genetic material. In this step, take care not to contaminate the sample *(17, 48, 49)*. PCR-based approaches are an inexpensive, rapid and convenient way of increasing the amount of virus genetic material available in a sample prior to sequencing, for example, the approach designed by the ARTIC Network *(51–53)*. For more technical details and how to select the optimal method for different settings, we refer to the SARS-CoV-2 sequencing implementation guide *(17)*. After initial sample preparation to enrich the SARS-CoV-2 genetic material, libraries can typically be prepared using standard sequencing protocols that are appropriate for any virus.

3.4 Laboratory considerations

Sequencing strategies for SARS-CoV-2 include targeted approaches that rely on knowledge of the genome, and metagenomic approaches that do not require prior knowledge of the genomic sequence *(54, 55)*. Annex III summarizes the key advantages and limitations of each commonly used sequencing technology. Before investing in sequencing capacity, consideration should be given to the requirements of the various technologies in terms of human resources, staff competencies, laboratory infrastructure, run-time, costs, ease of use, subsequent data processing, throughput (rate of data production) and sequencing accuracy. The number of samples that need to be analysed will depend on the sequencing objective. When calculating costs, consider not only the procurement of sequencing equipment, but also the recurrent costs for reagents, maintenance and service contracts. This guidance does not cover costs, but an extensive recent overview can be found in *(8)*. Basic infrastructure should be in place to support reliable sequencing, including reliable Internet connection and electricity supply, appropriate environment (e.g., vibration and dust free, temperature and humidity logged and regulated required for some platforms), and logged storage of

samples. Appropriate biosafety and biosecurity measures should be implemented. Assessing the costs and basic infrastructure requirements can help to decide whether the actual sequencing should be done in-house or would be better outsourced. Technology changes rapidly; consequently, certain techniques will become obsolete or manufacturers will shift to different machines and/or reagents. Before making large investments, it is recommended to establish how long the manufacturer will commit to supplying reagents and supporting the maintenance and troubleshooting of the selected platforms of interest. When planning a programme, the availability of ancillary reagents and additional equipment to support the sequencing work should also be taken into account (e.g., extraction methods [either automated or manual], instruments to quantify genetic material, amplification and incubation instruments, sample purification, and sample and reagent storage). Laboratories that conduct genomic sequencing should have high-quality SARS-CoV-2 PCR capacity confirmed by internal and external quality assurance. In addition, for each step in the process, quality indicators should be established and monitored.

3.5 Bioinformatic and computational considerations

Hardware requirements differ depending of the approach taken (for details, we refer to the implementation guide *(17)*). The volume of raw data produced depends on the sequencing method (see Annex III) and the number of samples sequenced *(56)*. The computational power required for data analysis also differs according to the sequencing objective and method. For example, genome phylogenetics and alignment may require high-performance computational power, especially where datasets are large. The costs of the computational architecture required to store and handle these data should be considered when developing a sequencing pipeline. The bioinformatic pipeline will depend on the pre-sequencing laboratory stages, sequencing platform, and reagents used. For a detailed description of bioinformatic pipelines, we refer to the implementation guidance *(17)*.

3.6 Considerations for virus naming and nomenclature

A consistent nomenclature has not yet been established for SARS-CoV-2. In the absence of an agreed upon consistent nomenclature, three main nomenclature strategies are generally used. Lineages or clades can be defined based on viruses that share a phylogenetically determined common ancestor. Both GISAID and Nextstrain aim to provide a broad categorization of globally circulating diversity by naming different phylogenetic clades. Rambaut et al. proposed a dynamic nomenclature for SARS-CoV-2 lineages that focuses on actively circulating virus lineages and those that spread to new locations *(57)*. Software is available to automatically assign new sequences to a lineage and/or clade *(58–60)*. With the increasing diversity in SARS-CoV-2 genomes, the demand for a uniform nomenclature is growing *(57, 61, 62)*. While no consistent nomenclature exists, the best approach would be to list particular lineages and/or clades using all three of the commonly used systems, or at minimum state explicitly which nomenclature is being used.

3.7 Genomic sequence and metadata sharing

The rapid sharing of pathogen GSD, together with the relevant anonymized epidemiological and clinical metadata, will maximize the impact of genomic sequencing in the public health response *(63–65)*. The wide sharing of SARS-CoV-2 sequences, as well as diagnostic protocols, sequencing protocols and samples, has been globally beneficial to achieving worldwide molecular diagnostic capacity *(66–68)*. The scientific/medical community should continue to build on the global collaboration and timely data sharing during SARS-CoV-2 and future emerging outbreaks. There are two distinct choices available for SARS-CoV-2 genomic sequence data sharing: “public-domain” and “public-access” *(69)*. Public-domain databases provide access to data without requiring the identity of those accessing and using the data, for example, the INSDC, operated by DDBJ, EMBL-EBI and NCBI. In public-access databases, such as GISAID, users must identify themselves to ensure transparent use of the data, permit effective oversight, protect the rights of the data contributors, make best efforts to collaborate with data providers, and acknowledge their contribution in published results. The examples mentioned are free of charge and accessible to the public. When pathogen sequencing projects are developed, it is imperative to determine which, if either, of these choices is most appropriate and whether other methods to access and share GSD are necessary *(44)*. One of the critical factors to ensure continued sharing of genetic data is giving due acknowledgement to those who collect clinical samples and generate virus genomic sequences. Data sources should always be acknowledged where publicly available data are used, and related publications and pre-print articles should be cited where available.

Sequence data, including consensus sequences, partial consensus sequences and raw sequence data, can be valuably shared in multiple formats. The quality of the sequence data, including potential contamination with amplicons produced through PCR, should be carefully evaluated prior to sharing. Laboratories should contact sequence-sharing platforms to update previously submitted partial sequences if an error is identified and corrected. Sharing of raw

virus sequencing reads (i.e., all individual sequenced fragments of a virus genome before they are assembled into one consensus genome) is important because it enables the direct comparison of the effect of different bioinformatic approaches for consensus genome generation and facilitates the correction of errors if necessary. Given the large data size of sequenced libraries, the sharing of read-level data may be more challenging in settings that have limited Internet upload speeds or intermittent connections. Any shared data should protect patient anonymity. To ensure patient anonymity, raw data containing human reads must be filtered to retain only non-human (i.e., viral) GSD prior to sharing *(43)*. Sharing of linked metadata, such as date of sample collection or approximate sampling location, is necessary to enable sequence data to be used in many phylogenetic applications. However, which metadata can be reasonably shared without compromising patient anonymity should be carefully considered.

Preliminary analyses of GSD are frequently shared through forums, platforms and preprint servers *(70–72)*. Through their publications, as with all scientific reports, scientists should consider the strengths and weaknesses of their analyses and how analyses might be interpreted or presented by various audiences prior to peer-review. Scientists are encouraged to provide a clear interpretation of their findings so that misunderstandings or misuse of results are minimised.

4. Public health objectives of SARS-CoV-2 genomic sequencing

Summarized below are examples of the key public health objectives of SARS-CoV-2 sequencing; for detailed descriptions, please refer to the SARS-CoV-2 sequencing implementation guide *(17)*.

4.1 Identification and characterization of SARS-CoV-2 and development of countermeasures

The sharing of the complete genetic sequence of the novel virus in early January 2020 was fundamental to characterize SARS-CoV-2, enabling the rapid development of diagnostics and supporting the development of therapies and vaccines *(73–80)*. Genomic sequencing enhances our understanding of the origins and transmission of novel viruses. By studying the initial SARS-CoV-2 genomes available from Wuhan, People's Republic of China and surrounding areas, it was possible to determine the latest possible date of emergence in humans as November–December 2019 *(74, 75, 81, 82)*. Sampling of a wide range of animals is supporting research around the identification of the initial animal source and/or potential intermediate hosts *(81, 83, 84)*.

4.2 Monitoring transmission and geographic spread

Phylogenetics is a method for investigating evolutionary relationships between different organisms using their genetic sequences. It is used in almost every branch of biology and has many important applications in informing public health responses *(17, 85–87)*. The availability of epidemiological or clinical data related to the sampling of the virus genomic sequence (often referred to as metadata, e.g., date of sampling, location of patient, clinical parameters) enhances the interpretation of phylogenetic analyses. Which metadata are required differs according to the objective of the genomic sequencing. The technical aspects of phylogenetic and phylodynamic analyses, metadata and common risks of misinterpretation can be found in the SARS-CoV-2 sequencing implementation guide *(17)*.

4.2.1 Investigating geographical spread and reintroductions between populations

Phylogeographic analyses that use virus genomic sequences and information on sampling location are being used to track SARS-CoV-2 circulation globally *(13, 47, 88–90)*. Phylogeographic reconstructions are often computationally demanding, and subsampling strategies can help to reduce this computational burden. Inferring virus movement or country of origin of specific clades/lineages can be valuable, but should be done with caution because several factors can bias phylogeographic reconstruction. For example, the lack of available SARS-CoV-2 genomes from certain areas can make it less likely that those areas will be reconstructed as the geographic origin of a lineage/clade. Genomic sequences may be associated in some databases with the location of virus sampling, instead of with the suspected location of infection of a patient. Where these locations differ because a patient travelled between the times of infection and virus sampling, phylogeographic analysis can result in an inaccurate reconstruction of the origin of specific clades/lineages *(91)*. These results should be interpreted cautiously and not under the assumption that phylogeographic results represent the true patterns of spatio-temporal (time and space) viral spread.

Methods to infer spatio-temporal spread of an outbreak can also be used to investigate factors that have driven virus dispersal *(92)*. Identifying drivers of transmission may help to shape new strategies for preventing spread. This approach has been used, for example, in the Ebola virus disease outbreaks in West Africa *(93, 94)*. For SARS-CoV-2, several countries have used genomic sequencing to establish the contribution of local transmission compared to

imported cases, and used this information to help make policy decisions *(89, 90, 95–100)*. The phylodynamic identification of factors that are important for understanding transmission is often computationally demanding and requires the curation of extensive data on potential explanatory factors (e.g., human population density, human mobility). Analyses are therefore often completed weeks or months after virus genomic sequencing. However, even retrospective analyses are useful to guide interventions for SARS-CoV-2 or potential emerging pathogens.

4.2.2 Evaluating evidence on transmission routes or clusters

Phylogenetic clustering has been used to support cluster and outbreak investigations for SARS-CoV-2. Analyses of transmission clusters can guide local decisions on whether control measures are needed to prevent future transmission in identified outbreak settings *(101)*. Given the relatively slow evolutionary rate of SARS-CoV-2 (i.e., one nucleotide substitution every two weeks), it is expected that many individual transmission events will not be traceable based on genomic sequence data *(35)*. Phylogenetic clustering of sequences from patients with the same hypothesised source of exposure would be consistent with (although not strong evidence of) that exposure. By contrast, phylogenetic separation of virus sequences from patients with the same hypothesised source of exposure would strongly indicate that the common source of infection has been incorrectly identified.

4.2.3 Quantifying periods of transmission and following the reproduction number over time

Molecular clock phylogenetic approaches can help to estimate the upper and lower limits of time of circulation of the sampled virus's genetic lineages in a given population *(74, 90, 102–106)*. This approach can provide more accurate information on the period of viral transmission than the clinical identification of cases, particularly in the early or late phases of an outbreak when surveillance is limited. Studying the variation in the genomic sequences detected can determine whether there is clinically undetected local transmission. In these settings, improved diagnostic surveillance programmes would need to be implemented where undetected circulation is suspected.

Genomic sequence analysis can also estimate how many individuals are infected by one individual in a given population (the reproduction number [R_0]) and support the assessment of relative changes in outbreak size over time. This information could be used to evaluate the impact of specific control measures.

4.2.4 Environmental surveillance in wastewater and sludge

For pathogens such as poliovirus, wastewater monitoring is an important tool for tracing the silent circulation of viruses in a community. This approach provides opportunities to detect circulation (before the initial patients have been clinically detected), estimate prevalence, and understand the genetic linkage and diversity *(107, 108)*. Several countries have demonstrated molecular detection of SARS-CoV-2 RNA in wastewater *(109–115)*. Consequently, environmental surveillance is a promising approach, especially in low prevalence settings, to identify unrecognized carriers and serve as an "early warning" system for SARS-CoV-2 introduction or changes in prevalence *(109, 116, 117)*.

4.2.5 Investigating potential reinfections

Seasonal coronaviruses can reinfect humans *(118)*. For SARS-CoV-2, cases of reinfection have been documented