Genomic sequencing of SARS-CoV-2

A guide to implementation for maximum impact on public health

8 January 2021

World Health Organization

Genomic sequencing of SARS-CoV-2

A guide to implementation for maximum impact on public health

8 January 2021

Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health

ISBN 978-92-4-001844-0 (electronic version)

ISBN 978-92-4-001845-7 (print version)

Suggested citation. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Contents

Preface

The year 2020 was a turning point in history and in global health. The COVID-19 pandemic has highlighted the potential for deadly epidemic-prone diseases to overwhelm our globalized world. We have learned a hard lesson about the intrinsic vulnerability of our societies to a single pathogen.

Although COVID-19 has brought untold tragedy, it has also shown how science can respond when challenged by a massive global emergency. In short, the pandemic has opened great scientific opportunities and capitalized on them. A technological revolution, building over the past decade, provided several new capacities for a pandemic response. Development of vaccines at lightning speed is one of them. Genomic sequencing is another.

Sequencing enabled the world to rapidly identify SARS-CoV-2; and knowing the genome sequence allowed rapid development of diagnostic tests and other tools for the response. Continued genome sequencing supports the monitoring of the disease's spread and activity and evolution of the virus.

The COVID-19 pandemic is still ongoing, and new viral variants are emerging. The global response will have to continue for the foreseeable future. The progress made since the start of the pandemic with the use of genome sequencing can be consolidated and further expanded to new settings and new uses.

As more countries move to implement sequencing programmes, there will be further opportunities to better understand the world of emerging pathogens and their interactions with humans and animals in a variety of climates, ecosystems, cultures, lifestyles and biomes. This knowledge will shape a new vision of the world and open new paradigms in epidemic and pandemic prevention and control.

Increased urbanization and human mobility are providing the conditions for future epidemics and pandemics. The accelerated integration of genome sequencing into the practices of the global health community is a must if we want to be better prepared for the future threats. We hope this guidance will help pave the way for that preparedness.

Sylvie Briand
Director
Global Infectious Hazard Preparedness
World Health Emergencies Programme
World Health Organization

Acknowledgements

This implementation guide was developed in consultation with experts with experience in the various fields of genome sequencing from the Global Laboratory Alliance of High Threat Pathogens (GLAD-HP), WHO reference laboratories providing confirmatory testing for COVID-19 and the Global Outbreak Alert and Response Network (GOARN). Following initial discussions by a technical writing group led by a temporary adviser and members of the WHO COVID-19 Laboratory Team, contributions were sought from other experts within and outside WHO, and two online meetings were held to resolve outstanding questions. Suggestions for improvement and corrections that could be incorporated into a second edition of this guide should be directed to WHElab@who.int.

Lead drafting and editing group

Sarah C. Hill, Royal Veterinary College, London and University of Oxford, Oxford, United Kingdom
Mark Perkins, Emerging Diseases and Zoonoses, Health Emergencies Programme, WHO, Geneva, Switzerland
Karin J. von Eije, Emerging Diseases and Zoonoses, Health Emergencies Programme, WHO, Geneva, Switzerland

Drafting group

Kim Benschop, Netherlands National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands
Nuno R. Faria, Imperial College, London and University of Oxford, Oxford, United Kingdom
Tanya Golubchik, University of Oxford, Oxford, United Kingdom
Edward Holmes, University of Sydney, Sydney, Australia
Liana Kafetzopoulou, KU Leuven – University of Leuven, Belgium
Philippe Lemey, KU Leuven – University of Leuven, Belgium
Tze Minn Mak, National Centre for Infectious Diseases, Singapore
Meng Ling Moi, Nagasaki University, Nagasaki, Japan
Bas Oude Munnink, Erasmus MC, Rotterdam, Netherlands
Leo Poon, Hong Kong University, Hong Kong Special Administrative Region (SAR), China
James Shepherd, University of Glasgow, Glasgow, United Kingdom
Timothy Vaughan, Eidgenössische Technische Hochschule Zurich (ETH Zurich), Zurich, Switzerland
Erik Volz, Imperial College, London, United Kingdom

Reviewers

Kristian Andersen, Scripps Research, La Jolla, CA, USA
Julio Croda, Ministry of Health, Rio de Janeiro, Brazil
Simon Dellicour, Free University of Brussels, Brussels, Belgium
Túlio de Oliveira, University of KwaZulu-Natal, Durban, South Africa
Nathan Grubaugh, Yale University, New Haven, CT, USA
Marion Koopmans, Erasmus MC, Rotterdam, Netherlands

Tommy Lam, University of Hong Kong , Hong Kong SAR, China
Marcio Roberto Nunes, Evandro Chagas Institute, Ananindeua, Pará, Brazil
Gustavo Palacios, United States Agency for International Development, Washington, DC, USA
Steven Pullan, Public Health England, London, United Kingdom
Josh Quick, University of Birmingham, Birmingham, United Kingdom
Andrew Rambaut, University of Edinburgh, Edinburgh, United Kingdom
Chantal Reusken, Netherlands National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands
Etienne Simon-Loriere, Institut Pasteur, Paris, France
Tanja Stadler, Eidgenössische Technische Hochschule Zurich (ETH Zurich), Switzerland
Marc Suchard, University of California at Los Angeles, Los Angeles, CA, USA
Huaiyu Tian, Beijing Normal University, Beijing, China
Lia van der Hoek, Amsterdam Medical Centre, Amsterdam, Netherlands
Jantina de Vries, Associate Professor in Bioethics, Department of Medicine, University of Cape Town, South Africa

Other contributors

Kazunobu Kojima, Biosecurity and Health Security Interface, Health Emergencies Programme, WHO, Geneva, Switzerland
Lina Moses, Emergency Operations, Health Emergencies Programme, WHO, Geneva, Switzerland
Lane Warmbrod, Epidemiology Team, Health Emergencies Programme, WHO, Geneva, Switzerland
Vasee Sathyamoorthy, Research for Health, Science Division, WHO, Geneva, Switzerland
Katherine Littler, Health Ethics & Governance, WHO, Geneva, Switzerland
Maria van Kerkhove, Emerging Diseases and Zoonoses, Health Emergencies Programme, WHO, Geneva, Switzerland

Abbreviations

ACE	angiotensin-converting enzyme
BDSKY	Birth-Death Skyline Model package
bp	base pair
CDC	Centers for Disease Control and Prevention (USA)
CoV	coronavirus
Ct	cycle threshold
DDBJ	DNA Data Bank of Japan
E	envelope
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
HIV	human immunodeficiency virus
INSDC	International Nucleotide Sequence Data Collaboration
M	membrane
MERS	Middle East respiratory syndrome
MRCA	most recent common ancestor
N	nucleocapsid
NAAT	nucleic acid amplification test
NCBI	National Center for Biotechnology Information (USA)
NGS	next-generation sequencing
nt	nucleotide
ORF	open reading frame
PCR	polymerase chain reaction
R_0	reproduction number
RACE	rapid amplification of cDNA ends
RBD	receptor binding domain
RNA	ribonucleic acid
S	spike
SARS	severe acute respiratory syndrome
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SRA	Sequence Read Archive
TMRCA	time to most recent common ancestor
WHO	World Health Organization

Executive summary

Recent advances have allowed the genomes of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) – the causative agent of COVID-19 – to be sequenced within hours or days of a case being identified. As a result, for the first time, genomic sequencing in real time has been able to inform the public health response to a pandemic. Metagenomic sequencing was fundamental to the detection and characterization of the novel pathogen. Early sharing of SARS-CoV-2 genome sequences allowed molecular diagnostic assays to be developed rapidly, which improved global preparedness, and contributed to the design of countermeasures. Rapid, large-scale virus genome sequencing is contributing to understanding the dynamics of viral epidemics and to evaluating the efficacy of control measures.

Increased recognition that viral genome sequencing can contribute to improving public health is driving more laboratories to invest in this area. However, the cost and work involved in gene sequencing are substantial, and laboratories need to have a clear idea of the expected public health returns on this investment. This document provides guidance for laboratories on maximizing the impact of SARS-CoV-2 sequencing activities now and in the future.

Intended goals of sequencing

Before starting a sequencing programme, it is important to have a clear understanding of the objectives of sequencing, a strategy for analysis, and a plan for how findings will be used to inform public health responses. Each phase in the COVID-19 pandemic will raise different questions that are central to public health, some of which require distinct genome sampling strategies. SARS-CoV-2 gene sequencing can be used in many different areas, including improved diagnostics, development of countermeasures, and investigation of disease epidemiology. Despite the obvious power of sequencing, it is important that those who set out the goals, conduct genomic analyses and use the resulting data are aware of the limitations and potential sources of bias.

Considerations when implementing a sequencing programme

Decisions about sequencing goals should be made in a multidisciplinary framework that includes senior representatives of all stakeholders. Funding sources should be identified to ensure sustainable support, including the cost of specialist personnel, sequencing devices and consumables, and the computational architecture required to process and store data. Ethical aspects of the project should be carefully evaluated. Laboratories should conduct biosafety and biosecurity risk assessments for every step in their chosen protocol.

The goals of sequencing should inform technical considerations about the methods to be used for sequencing and the selection of samples. Several devices are available for sequencing SARS-CoV-2 genomes, and each may be more or less appropriate in particular circumstances, as a result of differences in per-read accuracy, amount of data generated, and turnaround time. For the majority of goals, both virus sequence data and sample metadata are required. Acquiring and translating such data into the correct format for analysis may require extensive resources, but will help to maximize the potential impact of the sequencing. Many analyses rely on the ability

to compare locally acquired virus sequences with the global virus genomic diversity. It is therefore crucial that virus genomic sequences are appropriately shared. Such sharing is taking place at an impressive rate via repositories such as GISAID and GenBank.

Which samples should be sequenced will depend on the question to be answered and the context. Consideration should also be given to sample logistics, such as how material is best transported, and how RNA extraction and sequencing can best be conducted without risking RNA integrity. When multiple organizations carry out sequencing and analysis, a practical and shared sample identification system should be devised.

Once a sample has been sequenced and appropriate metadata collected, bioinformatic analysis is required. The bioinformatic pipeline will depend on the pre-sequencing laboratory stages, sequencing platform, and reagents used. Sequence alignment and phylogenetic analysis will require high-performance computational power, with can be expensive. Analysis and interpretation of the data will require highly trained staff. Results and conclusions should be shared with the relevant stakeholders in a clear and consistent manner to avoid misinterpretation.

Maximizing public health impact

No matter how many SARS-CoV-2 genome sequences are generated, they will have a positive impact on public health only if strategies are defined for subsequently producing and communicating useable and timely results. Programmes should always consider how the results of SARS-CoV-2 sequence analysis can extend, complement, or replace other existing approaches, and decide whether sequencing is the most appropriate or resource-effective method to achieve the desired goals. Results should be communicated in a timely and clear manner to stakeholders who can use the information directly for public health benefit. This may be most efficiently achieved if genomic sequencing and analysis laboratories are closely integrated with existing diagnostic and epidemiological public health programmes.

Building a strong and resilient global sequencing network can maximize the public health impact of sequencing, not only for SARS-CoV-2 but also for future emerging pathogens. Various pathogen-specific laboratory networks have invested in sequencing capacity as part of their surveillance activities. As the costs of sequencing are substantial and many parts of the sequencing workstream can be used for various pathogens and sequencing objectives, national collaboration is encouraged, to ensure optimal use of existing capacity. Long-term investment is required to strengthen capacity for bioinformatic and phylogenetic analysis, as this now lags behind molecular laboratory capacity in many settings. Capacity-building programmes should focus on a stepwise approach to build up competencies. The focus of capacity building will depend on the context: some countries may need to build their wet laboratory capacity, while others may decide to outsource the actual sequencing and focus on the bioinformatics, data management and interpretation. Collaboration between sequencing groups will be facilitated by shared sequencing protocols, standardization of database structure and metadata formats, joint meetings and training, and access to audits and proficiency testing using reference standards.

1. Introduction

Genomic sequences from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) – the virus that causes COVID-19 – are being generated and shared at an unprecedented rate. Recent technological advances have allowed SARS-CoV-2 genomes to be sequenced within hours or days of a case being identified. The use of these genomes to inform public health policy during an ongoing outbreak signifies a revolution in virus genomic investigations. For the first time, genomic sequencing can help to guide the public health response to a pandemic in near-real time.

Virus genome sequencing has already proved fundamental in identifying SARS-CoV-2 as the causative agent of COVID-19 and in investigating its global spread. Moreover, virus genome sequences can be used to investigate outbreak dynamics, including changes in the size of the epidemic over time, spatiotemporal spread and transmission routes. In addition, genomic sequences can help in the design of diagnostic assays, drugs and vaccines, and in monitoring whether hypothetical changes in their efficacy over time might be attributable to changes in the virus genome. Analysis of SARS-CoV-2 virus genomes can therefore complement, augment and support strategies to reduce the burden of COVID-19.

Increased understanding of the potential of genomic sequencing to improve public health is leading more laboratories to invest in this process. However, the potentially high cost and work involved necessitates clarity about the expected returns from this investment, how genomic sequence data can best be used and, the pathways by which a beneficial impact on public health and policy can be achieved.

This guide aims to help public health technical officers and laboratories responsible for, or considering the establishment of, genome sequencing programmes for SARS-CoV-2. It provides information on the considerations to be taken into account when planning or conducting a SARS-CoV-2 sequencing programme, to ensure that best use is made of the results in improving public health. In addition, it raises practical questions, details the possible applications and limitations of genomic analyses, and provides brief guidance on technical strategies for sequencing and analysis.

2. Background

2.1 Growth in virus genomic sequencing

The first two decades of the 21st century have brought a transformational shift in the use of virus genomics in disease outbreaks, from the lengthy protocols and retrospective analyses of the past, to a new ability to investigate genomic epidemiology in near-real time. The widespread application of sequencing has been facilitated by rapid decreases in per-base cost and sample-to-result turnaround time, increases in the volume of data generated and the computational capacity required to process it, and the development of easily deployable, cost-effective benchtop sequencing equipment *(1)*. Sequencing has consequently become a critical tool in clinical microbiology for detecting and characterizing viral pathogens in clinical samples *(2)*, supporting infection control, informing epidemiological investigations and characterizing evolutionary viral responses to vaccines and treatments *(3, 4)*.

The increased importance of virus genomic sequencing for clinical and epidemiological investigations is exemplified by the differences in speed and scale between the genomic responses during the 2002–2003 epidemic of severe acute respiratory syndrome (SARS) and those in the current COVID-19 pandemic. During the SARS epidemic, only three virus genomes were publicly shared in the first month following identification of a coronavirus as the causative pathogen, and only 31 were available within 3 months. Genomics was used to design molecular assays that could establish an association between the disease and the new coronavirus concerned *(5–7)*, but was not sufficiently developed to allow virus epidemiology to be studied in real-time on a large scale. In contrast, during the COVID-19 pandemic, metagenomic sequencing was used to identify the causative pathogen of unexplained pneumonia within a week of the disease being reported *(8, 9)*. The pathogen was announced as a novel coronavirus (SARS-CoV-2, previously known as 2019-nCoV) in the beginning of January 2020 *(9)*. Six genomes were shared publicly before mid-January, allowing the rapid development of diagnostic assays and strategies for extensive virus genomic sequencing. Sequencing efforts have continued as the virus has spread across the world, resulting in a constantly growing data set of more than 60 000 near-complete viral genomes within the 6 months following the identification of SARS-CoV-2. Frequently, genomes have been generated within days of case identification, and used to understand virus spread during the pandemic.

2.2 Growth in virus genomic applications

In recent years, public health emergencies caused by epidemics have fuelled developments in virus genomic sequencing and molecular epidemiology. Viral genomic sequences have allowed us to identify pathogens and to understand their origin, transmission, genetic diversity and outbreak dynamics (Box 1). This understanding has informed the development of diagnostic approaches, provided important background information for vaccine development and drug design, and helped in disease mitigation *(33, 41, 42)*. Genomic analyses are capable of estimating aspects of the epidemiological dynamics of viral disease that are unrecoverable using epidemiological data alone

(3, 41, 43), because they allow insights into periods of an outbreak when cases were unobserved. Powerful insights can be achieved even with relatively sparse genomic data.

SARS-CoV-2 has therefore emerged in a scientific context in which genome sequences can be generated more rapidly and more easily, and can be used to answer a broader range of public health questions, than ever before.

Box 1. The contribution of virus genomics to epidemiological understanding in public health emergencies since the SARS epidemic[1]

Influenza A(H1N1)pdm09 was the first pandemic in which many epidemiological questions could be investigated through genetic analyses. Assessment of virus transmissibility from gene sequences provided early estimates of the basic reproduction number, R_0, that were similar to those produced by epidemiological analysis *(10)*. Retrospective genomic analysis confirmed that the pandemic had begun at least 2 months before the first sampled case, and inferred population growth rates and epidemic doubling times similar to those found in early analyses *(11)*. However, efforts to understand the origins of the A(H1N1)pdm09 epidemic were hindered by a lack of systematic influenza surveillance in swine *(12)*. A retrospective study in 2016 demonstrated extensive diversity among influenza viruses in Mexico, and suggested that swine in Mexico were the most likely source of the virus that gave rise to the 2009 pandemic *(13)*.

Since 2012, several outbreaks of Middle East respiratory syndrome (MERS) caused by the coronavirus MERS-CoV have been reported, raising questions about the origins of the virus and its mode of transmission. Following preliminary serological and epidemiological evidence that supported the involvement of dromedaries (Arabian camels, *Camelus dromedarius*) in these outbreaks *(14)*, genome sequencing was used to identify the presence of the virus in camels *(15, 16)* and to demonstrate multiple independent virus transmission events from camels to humans *(15, 17,18)*. Subsequent sequencing analyses further showed that MERS-CoV is endemic in camels from Eastern Mediterranean and African countries *(19)*. In 2018 a comprehensive genomic study confirmed that the virus is maintained in camels and that humans are terminal hosts *(20)*. Mean R_0 values estimated via virus genomic sequences were less than 0.90, suggesting that MERS-CoV was unlikely to become endemic in humans. This confirmed that focusing on ongoing control efforts among camels was appropriate, while highlighting a continued need to monitor the possible emergence of strains that are more easily transmissible among humans *(20)*.

The 2013–2016 Ebola virus disease epidemic marked the beginning of large-scale genomic epidemiological investigation in an ongoing outbreak. Genomic analyses allowed viral epidemiological surveillance during the unfolding epidemic and assisted understanding of the origin, epidemiology and evolution of the virus. Molecular clock dating techniques estimated that the common ancestor of all sequenced Ebola virus genomes occurred very early in 2014, consistent with epidemiological investigations that placed the first case around late December 2013 *(21–24)*. Evolutionary analyses demonstrated that spread was maintained by human-to-human transmission rather than by multiple separate introductions from an animal reservoir *(21–28)*. Phylodynamic insights into the early spread of the epidemic allowed for R_0 to be estimated and superspreading events in the population to be investigated *(29, 30)*. Molecular genetic

investigations supported the possibility of sexual transmission of Ebola virus, resulting in WHO recommendations to improve safe-sex counselling and testing of Ebola survivors *(31, 32)*. Towards the end of the outbreak, there was a shift towards rapid in-country sequencing that helped to resolve viral transmission chains and community spread *(4, 33–36)*.

On 1 February 2016, WHO declared Zika virus infection a public health emergency of international concern following autochthonous circulation of the virus in 33 countries and strong suspicions that infection during pregnancy was linked to fetal microcephaly and other developmental abnormalities *(37)*. Reconstructing the spread of the virus from epidemiological data alone was challenging because symptoms were often mild or absent, and overlapped with those caused by other co-circulating arboviruses (e.g. dengue, chikungunya), and also because Zika virus molecular diagnostic surveillance was often established long after local transmission had begun *(38)*. Collaborative efforts were initiated to sequence retrospective and new cases in order to gain insights into the origin, transmission routes and genetic diversity of the virus *(38)*. Preliminary phylogenetic and molecular clock analysis showed that the epidemic in the Americas was caused by a single introduction event of an Asian genotype lineage, which was estimated to have occurred a year prior to detection of the disease in May 2015 in Brazil *(37)*. Genomic epidemiological studies have subsequently documented the spread of Zika virus in considerable detail *(37–40)*. For example, widespread sampling of genomic sequences from infected patients and mosquitoes during the sustained 2016 Zika virus outbreak in Florida, USA, allowed R_0 to be estimated as less than 1. This led to the conclusion that multiple introductions of the virus were required for such extensive local transmission *(40,41)*.

[1] See Annex 1 for the sampling strategies employed in the studies cited in this box.

2.3 Phylogenetic and phylodynamic analyses

Many important applications of virus genomics in informing public health responses have been built on phylogenetic or phylodynamic analyses. Phylogenetics is used in almost every branch of biology to investigate evolutionary relationships between different organisms using their genetic sequences. Phylogenetic trees (for example, see Fig. 1) are useful visualizations of such relationships. The branching patterns and the length of the branches can be used to represent evolutionary relatedness. Any two organisms, represented by external or "leaf" nodes (tips), will have a common ancestor where the branches that lead to them intersect (internal nodes). Given homologous genetic sequence data from multiple organisms and a genetic substitution model of how different sites in those sequences change over time, it is possible to assess a large number of trees to determine which is most likely to represent the true relatedness between those organisms.

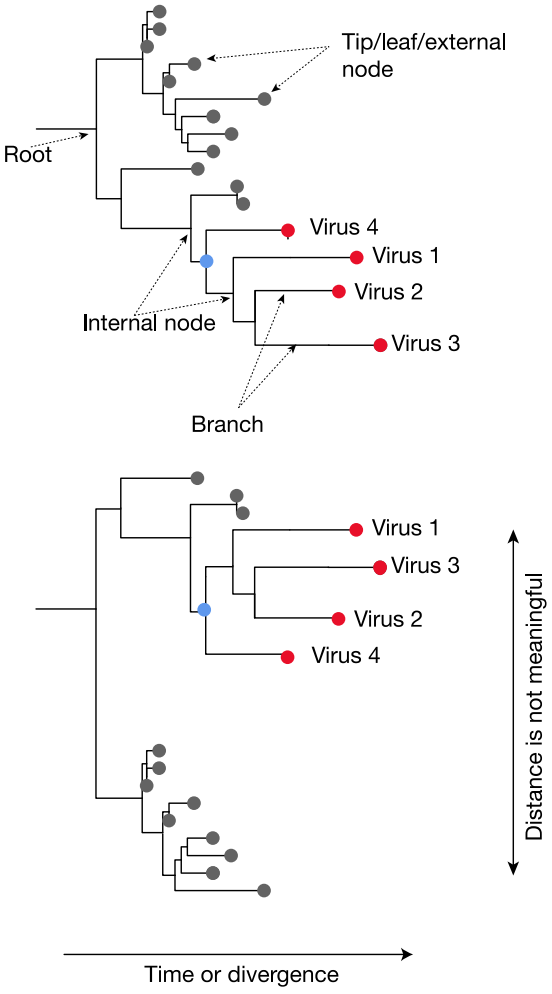

Fig. 1. Phylogenetic trees with key features marked. The distance along the x-axis in the phylogenies displayed as above in a "rectangular" format usually represents either the time or the amount of genetic change that has accrued. The most recent common ancestor of viruses 1–4 is highlighted by the blue node. The distance along the y-axis is not meaningful. Specifically, the clades descending from any node can be rotated around that node without altering the phylogenetic interpretation of the tree. The two trees pictured above are therefore phylogenetically identical.

When discussing virus evolution, it is extremely important to distinguish between the mutation rate and the evolutionary rate (or substitution rate). The mutation rate is a biochemical measure that considers the number of errors that occur in the copying of RNA from a parent virus to its progeny and is typically measured in mutations per genome per replication. The mutation rate can be estimated experimentally in several ways, such as by sequencing whole virus populations to measure genetic diversity before and after a known number of replications in a laboratory setting. Most mutations are deleterious *(44)*, and individual virions containing such mutations will often fail to replicate.

Only those mutations that increase in frequency and become fixed within a lineage, following genetic drift or the action of natural selection on a virus population, contribute to the evolutionary rate. The evolutionary rate is typically depicted as the number of nucleotide substitutions per site, per year (often abbreviated to subs/site/year). Different virus lineages can have different evolutionary rates. The evolutionary rate can often be inferred directly from virus genomic sequence data obtained from different patients on different dates. The range of sample collection dates (over months or years) needed to allow robust inference of the evolutionary rate will vary for different viruses and outbreaks, because it depends on the substitution rate, the age of the viral lineage under investigation, and the genomic sequence length under investigation. For SARS-CoV-2, inclusion of genomic data collected over intervals of two months appears minimally sufficient *(45)*, although more robust estimates are achieved using data collected over a longer period.

RNA viruses typically have a high evolutionary rate, with many gaining a genetic change every few days or weeks *(46)*. Some RNA viruses therefore acquire genetic substitutions at close to the same time scale as transmission between hosts. In the case of SARS-CoV-2, the rate of transmission events between humans is higher on average than the rate at which transmitting viral lineages acquire genetic substitutions. SARS-CoV-2 lineages accrue genetic diversity over weeks or months rather than days, so that directly neighbouring patients in a transmission chain may be infected by viruses with identical genomes. Analysis of patterns of accumulation of virus genomic diversity during an outbreak can be used to make inferences about epidemiological processes. This is the focus of a body of phylogenetic techniques that come under the umbrella term phylodynamics, coined by Grenfell et al. *(47)*.

Phylodynamic methods are useful in outbreak investigations, as they can complement and augment other epidemiological analyses based on identified confirmed cases. First, several phylodynamic approaches may be less affected – or differently affected – by biases in diagnostic surveillance, such as changes in surveillance effort over time or patchy detection of cases. Secondly, phylodynamics can reveal features of the epidemic that occur outside of the sampling time window (for example, before the first case is identified). Thirdly, phylodynamic analyses provide a direct means of learning about the population dynamics of specific different virus lineages.

Phylodynamic methods use probabilistic models to tie the phylogenetic tree of sampled genomes to epidemiological parameters of interest. As such, they require inference of a dated phylogenetic tree that contains information not only about which sequences cluster together, but also when the unsampled most recent common ancestors (MRCAs) of sampled virus genomes existed. While

sampling dates are known for viruses from sequenced samples (i.e. tree tips, see Fig. 1), the MRCAs (i.e. internal nodes) are phylogenetically inferred and their time of existence must be estimated. Estimation of these dates requires the use of a molecular clock model parameterized by a clock rate – the average rate of genetic substitution along branches of the phylogeny.

There are several distinct families of phylodynamic models: coalescent, birth–death, and simulation-based models. Reviews of these different models are available elsewhere *(48, 49)*.

2.4 SARS-CoV-2 genomic and evolutionary features important for genomic applications

Several fundamental features of any virus determine the possible approaches for the generation and use of virus genomic data to inform public health authorities. These features include its genetic material (RNA or DNA), genomic length, genome structure and composition, and evolutionary rate.

SARS-CoV-2 is classified within the genus *Betacoronavirus* (subgenus *Sarbecovirus*) in the family *Coronaviridae* (subfamily *Orthocoronavirinae*), a family of single-strand positive-sense RNA viruses (50). The International Committee on the Taxonomy of Viruses (ICTV) currently considers SARS-CoV-2 as belonging to the species *Severe acute respiratory syndrome-related coronavirus*, along with SARS-CoV and closely related viruses sampled from non-human species *(51)*. The reference strain of SARS-CoV-2, Wuhan-Hu-1 (GenBank accession MN908947), was sampled from a patient in Wuhan, China, on 26 December 2019 *(52)*. That genome is 29 903 nucleotides (nt) in length and comprises a gene order of similar structure to that seen in other coronaviruses: 5'-replicase ORF1ab-S-E-M-N-3'. The predicted replicase ORF1ab gene of Wuhan-Hu-1 is 21 291 nt in length. The ORF1ab polyprotein is predicted to be cleaved into 16 nonstructural proteins. ORF1ab is followed by a number of downstream open reading frames (ORFs). These include the predicted S (spike), ORF3a, E (envelope), M (membrane) and N (nucleocapsid) genes of lengths 3822, 828, 228, 669 and 1260 nt, respectively (52). Like SARS-CoV, Wuhan-Hu-1 also contains a predicted ORF8 gene (366 nt in length) located between the M and N genes. Finally, the 5' and 3' terminal sequences of Wuhan-Hu-1 are also typical of betacoronaviruses and have lengths of 265 nt and 229 nt, respectively.

Preliminary estimates of the evolutionary rate of SARS-CoV-2 are close to a mean of 1×10^{-3} substitutions per site per year *(45, 53)*, which is similar to the mean evolutionary rate observed in other RNA virus genomes *(46)*.

At the time of writing, there is no accurate estimate of the rate of mutation per genome replication for SARS-CoV-2 (mutation rate). However, it is expected to be similar to those of other coronaviruses. The mutation rate of coronaviruses and other members of the *Nidovirales* order is lower than that of other RNA viruses because they have an intrinsic proof-reading ability to correct replicative mistakes that is absent in other RNA viruses *(50)*.

3. Practical considerations when implementing a virus genomic sequencing programme

Many public health laboratories now recognize the potential impact that virus genomic sequences could have on public health decisions during the current COVID-19 pandemic or future outbreaks (see also section 5).

3.1 Planning a sequencing programme

Laboratories should have clear plans in place. A checklist to assist planning is given in Annex 2. Key questions to be considered before initiating a sequencing programme include the following.

(1) What are the expected outputs of the sequencing programme?
(2) Which samples should be sequenced to achieve the expected outputs identified in step 1? Which metadata or additional data sources are critical?
(3) Who are the key stakeholders and what are their responsibilities? How can they be effectively engaged?
(4) How can samples and information be transferred rapidly and appropriately between stakeholders, as required?
(5) Is the project designed in accordance with local, national and international laws, and ethical guidelines?
(6) Are adequate funding, equipment and human resources available to deliver all stages of specimen retrieval, wet-laboratory sequencing, bioinformatic, phylodynamic and other analyses, data-sharing, and communication of timely results to appropriate stakeholders?
(7) How can goals be achieved without disrupting other areas of laboratory work, such as clinical diagnostics, and avoiding duplication of effort?
(8) How will the programme be evaluated for cost-effectiveness and impact?

3.2 Ethical considerations

When a sequencing programme is being designed, it is important to review all the ethical implications. Possible risks for 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, taking into account the social value and scientific validity of the investigation, selection of participants, risk-benefit ratio, informed consent, and respect for participants *(54, 55)*. Where researchers have little experience in identifying possible ethical issues related to the sequencing of pathogens, international collaboration and engagement of appropriate expertise are strongly encouraged. Collaboration among researchers around the world will help ensure equitable and mutually beneficial research partnerships. Local researchers are more likely to understand their health care and research systems and to be able to translate results into policy, and therefore often best suitable to take leading and active roles throughout

the research process *(54,55)*. Ethical considerations related to data-sharing are discussed in more detail in Chapter 4.

3.3 Identifying expected outputs and necessary data

Before embarking on any sequencing programme, deliverable goals should be set. Possible goals are discussed extensively in section 5; the goals defined will affect the design of the sequencing workstream.

Once goals have been identified, an achievable sampling strategy has to be designed to collect the appropriate genomic sequences and metadata; genomic sequences that lack appropriate metadata are not useful for most applications. Different public health questions will demand different sampling strategies and data. It is therefore vitally important to ensure that there is discussion among the different stakeholders who (a) conduct diagnostic sampling, (b) choose samples for sequencing, (c) choose the sequencing strategy, (d) choose analytical strategies, and (e) use generated information for public health, to ensure that genomic sampling strategies and metadata collection are correctly targeted for the analyses for which they are intended.

3.4 Identifying and liaising with stakeholders

Key stakeholders should be identified, consulted and involved at an early stage (Box 2). Their identity and level of involvement will vary depending on local circumstances and the goals of the programme, but it is reasonable to consider stakeholders involved in all steps of the process, from case identification to use of the findings. It may be relevant to provide educational resources to stakeholders, including the general public, to demonstrate the potential usefulness of a sequencing programme and to explain how sequences will be used and why specific patient metadata are necessary. Close collaboration and communication among relevant stakeholders are critical if sequencing activities are to resolve questions of public health importance.

Box 2. Stakeholders to be engaged when developing sequencing programmes

This list is not exhaustive and additional stakeholders should be considered, depending on the local circumstances.

- **Public health bodies.** Local or national public health bodies, such as ministries of health, will often commission or help deliver SARS-CoV-2 sequencing programmes. Their involvement will ensure that goals respond to key policy questions. In addition, public health bodies can often help secure widespread collection of particular diagnostic samples and metadata.
- **Diagnostic laboratories** should ideally be partners in any sequencing programme for SARS-CoV-2. They typically have the best access to SARS-CoV-2 samples and can often provide residual positive samples and metadata directly to sequencing facilities. In some settings, clinical diagnostic laboratories may be tasked with implementing an in-house sequencing programme, while in others the sequencing may be done by external research or national public health laboratories.

- **Sequencing facilities** may be public or private; some sequencing facilities will have the bioinformatic capacity to generate consensus virus genomes, whereas others will provide raw data that must be further processed elsewhere to generate genomes. Not all bioinformaticians will have the expertise to handle data produced by all the possible wet-laboratory sequencing techniques and platforms. In such cases, support from an expert who can handle the intended data type is strongly recommended.
- **Analytical groups** that will conduct planned phylogenetic, phylodynamic or other genomic analyses must be closely involved in determining which samples should be sequenced, so that genomic sequences are appropriate for the analytical methods to be used. It should not be automatically assumed that expertise to conduct such analyses is present in the molecular genetic wet-laboratories that conduct sequencing. Where relevant, close integration of analysts and those involved in surveillance and response (e.g. public health teams investigating local outbreaks) will increase the potential impact of analyses.
- **Infection prevention and control teams** (e.g. in hospital, elderly homes and public health) can support the identification of emerging disease clusters and are well placed to identify cases that would be useful for sequencing. They can also act on the subsequent findings regarding transmission clusters.
- **Occupational health services** in work related settings they can help to identify potential transmission clusters or transmission routes that can be investigated using virus genomic studies, and to implement infection prevention and control activities emerging from the results of these studies.
- **Patients** should be engaged to ensure that they understand how sequences and metadata are being used and shared, and benefit from results. A properly designed and resourced community engagement programme can help identify and address potential obstacles to research, relating for instance to stigma, and ensure that the programme design is cognisant of and responsive to the sociocultural environment in which the programme will be implemented.

Once key stakeholders have been identified, appropriate channels of communication need to be established between the various groups. As a minimum, programme aims should be defined in a multidisciplinary framework involving senior representatives of all stakeholders. Communication between stakeholders should ideally be maintained throughout the project, and may require daily or weekly meetings between representatives from some or all bodies involved, to ensure appropriate reactions to changing situations during the epidemic (e.g. investigation of transmission clusters as they arise). Epidemiologically focused activities that integrate genomic data analysts directly in public health investigation and response teams are likely to have a greater immediate impact than those in which virus genomic analysis is considered as a separate or secondary activity.

How, when, and with whom any data are shared – with the scientific community or between stakeholders – should be agreed at the outset. Stakeholder responsibilities, including provision of funding if appropriate, should also be agreed. If data or publications will be generated, it is often

helpful to agree in advance on how those involved will be fairly credited for their contribution to data production or analysis.

The results of sequencing analysis should be rapidly communicated to stakeholders in a standardized and easily interpretable written report, and opportunities for discussion should be arranged. The practical take-home message of results and analytical limitations should be conveyed in everyday language, avoiding technical jargon. Where a multidisciplinary approach has been followed in dealing with public health questions (e.g. involving analysis from phylogenetics and mathematical modelling), sequencing results should ideally be discussed alongside results from other fields.

3.5 Project execution: acquisition of data, logistics and human resources

Technical considerations regarding legal and ethical adherence, sample selection, detailed resource evaluation and technical guidance are given in section 6.

3.6 Project evaluation

Regular structured feedback should be sought from stakeholders to identify and deal with any difficulties that may arise.

The potential of virus genomic sequencing continues to grow, and the scientific and public health community are rapidly developing new strategies to maximize its impact in future disease outbreaks. All sequencing efforts should therefore include clear opportunities for regular evaluation by all stakeholders of what was useful, what was missing and what impact sequencing achieved. Identifying and communicating these findings to researchers and the bodies that fund them is important to help guide development of new tools.

4. Data-sharing

4.1 WHO recommendations on data-sharing

The rapid sharing of pathogen genome sequence data, together with the relevant anonymized epidemiological and clinical metadata will maximise the impact of genomic sequencing in the public health response. Such data, generated during an outbreak, should be shared with the global community as rapidly as possible, to ensure maximum usefulness in improving public health. In April 2016, WHO issued a policy statement on data-sharing in the context of public health emergencies: "WHO will advocate that pathogen genome sequences be made publicly available as rapidly as possible through relevant databases and that benefits arising out of the utilization of those sequences be shared equitably with the country from where the pathogen genome sequence originates" *(56)*. One of the critical factors to assure continued sharing of genetic data is due acknowledgement to those who collect clinical samples and generate virus genome sequences. Data sources should be acknowledged where publicly available data are used, and related publications and pre-print articles be cited where available. Also, funders, journal editors and peer-reviewers should encourage sustained data-sharing.

4.2 Sharing of appropriate metadata

Anonymized sample metadata should be shared along with SARS-CoV-2 genomic data to maximize the usefulness of the genomic sequence. Shared metadata should always include at least the date and location of sample collection, but additional metadata will greatly increase the potential applications of the sequence. Where possible, therefore, metadata should include data pertaining to the sample type, how the sequence was obtained, links to other sequenced viruses, patient travel history, and demographic or clinical information. For a detailed description of metadata see section 6, Table 2. When any information is shared, it is important that patient anonymity is protected.

4.3 Sharing of consensus sequences, partial consensus sequences and raw sequence data

As SARS-CoV-2 has only recently emerged in humans, virus genetic diversity remains relatively limited and full-length sequences are therefore important to capture as many phylogenetically informative sites as possible. Where full-length sequencing is unsuccessful, partial sequences may be generated. SARS-CoV-2 genomes that have partial coverage are still valuable and should be shared. While the required genome coverage (proportion of sites without ambiguous bases, i.e. Ns) will vary for different applications and for different viruses, partial genomes often represent important sources of data. For example, Zika virus genomes with as little as 40% coverage (i.e. 60% of sites with Ns) were found to be phylogenetically informative of clade structure *(57)*.

As for full-length genomes, the quality of the partial genome should be checked to ensure that sites with insufficient support are masked before the genome is made publicly available. Partial

genomes in which coverage or sequencing depth is generally very low, but in which a few short regions have very high sequencing depth, may be indicative of contamination with amplicons produced through the polymerase chain reaction (PCR) and should be carefully evaluated prior to sharing.

Sharing of raw sequencing reads (i.e. all individual sequenced fragments of a virus genome before they are assembled into one consensus genome) is important because it allows the effect of different bioinformatic approaches for consensus genome generation to be directly compared and facilitates correction of errors where necessary. Depending on the sequencing strategy adopted and the depth of sequencing coverage, read-level data can also be used for analyses of intra-host variation in virus genomes. Read-level data sets of SARS-CoV-2 should therefore be made available where possible. Given that the data size of sequenced libraries can reach hundreds of gigabytes, sharing read-level data may be more challenging in settings that have limited internet upload speeds or intermittent connections. Raw data containing human reads must be filtered to retain only non-human (i.e. viral) genetic sequence data prior to sharing, in order to ensure patient anonymity (see section 6.7.1).

4.4 Platforms for sharing

Sharing of sequences via commonly used, searchable platforms increases the accessibility of the data. Platforms vary in the type of data that they host, the use-conditions they place on data and the ease with which metadata can be uploaded. Some platforms (e.g. the European Nucleotide Archive) offer spreadsheet templates for sequence data that can be filled in offline and then uploaded in batches.

Sharing mechanisms used for genomic sequence data include public-domain and public-access databases. Public-domain databases provide access to data without requiring the identity of those accessing and using data. In public-access databases, users must identify themselves to ensure transparent use of the data and permit effective oversight, to protect the rights of the data contributors, make best efforts to collaborate with data providers, and to acknowledge their contribution in published results. SARS-CoV-2 genetic sequences with appropriate metadata are frequently shared through multiple platforms. Public-domain databases for sharing consensus genomes include the National Centre for Biotechnology Information (NCBI), the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), and the DNA Data Bank of Japan (DDBJ). Raw read data with appropriate metadata are sharable via repositories of the International Nucleotide Sequence Data Collaboration (INSDC), which includes the NCBI Sequence Read Archive (SRA), the EMBL-EBI ENA and the DDBJ Sequence Read Archive. A public-access database for consensus genomes is for example GISAID EpiCoV™. The COVID-19 Data Portal attempts to facilitate sharing and access to all biomedical data sources that are of relevance to COVID-19 *(58)*.

Laboratories should contact sequence-sharing platforms to update previously submitted partial sequences if an error is identified and corrected.

Preliminary analyses of genetic sequence data are frequently shared on forums and preprint servers, such as medRxiv or bioRxiv. This allows data producers to provide additional

information on initial findings to the wider scientific community. Forums including Virological have proved useful for informal sharing and discussion of initial results with the molecular genetics community, and posts can be continually updated as analyses progress. Preprint servers are often used to share articles at the point of submission to a peer-reviewed journal, and clearly communicate intentions to publish. WHO strongly encourages the sharing of genetic and metadata as soon as possible after data quality checks, with no withholding until after preprint deposition.

Unreviewed preliminary analyses are being used more extensively by the public and media in the current pandemic than ever before. Scientists should therefore be mindful of how analyses might be interpreted or presented in the media, and should provide clear interpretations of their findings so that results cannot be easily misconstrued.

5. Applications of genomics to SARS-CoV-2

This section reviews how SARS-CoV-2 genome sequencing has been used at different phases of the COVID-19 pandemic and suggests possible future applications. It also provides brief guidance on the common limitations of current approaches, to assist in setting realistic goals. For some of the applications considered, virus genomic sequencing represents only a small component of a larger investigation, which may include substantial essential laboratory or clinical investigations.

5.1 Understanding the emergence of SARS-CoV-2

5.1.1 Identifying the causative agent of COVID-19

SARS-CoV-2 was independently identified and sequenced in early 2020 by Wu et al., Lu et al. and Zhou et al. *(52, 59,60)*. Several different metagenomic next-generation sequencing (mNGS) approaches were used to identify the causative pathogen of COVID-19. Metagenomic sequencing permits untargeted sequencing of nucleic acid in a sample, and can therefore identify viral RNA or DNA if present at high enough copy numbers relative to DNA or RNA from other sources (see also section 6.5.1). Completion of the full-length virus genome sequences, including the genome termini, generally involved Sanger sequencing and a 5'/3' rapid amplification of cDNA ends (RACE) method. This method is cost-efficient for sequencing short regions of a genome that may be missed with metagenomic methods, but relies on previous knowledge of the sequence information relatively close to the missing region.

5.1.2 Determining times of origin and early diversification

It was particularly important to determine when SARS-CoV-2 first emerged in humans, since this could provide an indication of whether there was a long period of undetected transmission before the first clinical cases were seen (and hence possibly many undetected cases). SARS-CoV-2 genomes from Wuhan and surrounding areas of Hubei province provided a number of key insights.

All the sequences were extremely closely related, differing by only a few nucleotide variants. Several early molecular clock dating exercises using these sequences gave estimated times for the appearance of the most recent common ancestor of all the sequenced SARS-CoV-2 viruses as the period from November to December 2019. These initial estimates have been confirmed as more sequences have become available. The latest possible date of emergence of SARS-CoV-2 in humans is therefore November–December 2019. This is close to the first identification of the initial cluster of pneumonia cases in Wuhan in mid-December *(59–61)*.

Where only one introduction to humans has occurred, the earliest possible timing of emergence of a zoonotic virus in humans is phylogenetically represented by the time to the most recent common ancestor (TMRCA) of the human zoonotic virus and the non-human animal virus from which it emerged. Inadequate sampling of non-human animal viruses that are closely related to SARS-CoV-2 means that the possible interval in which SARS-CoV-2 could have emerged in

humans is relatively wide when phylogenetic data alone are considered. It is therefore difficult to distinguish phylogenetically between two possible scenarios of SARS-CoV-2 emergence. In the first, SARS-CoV-2 could have emerged in humans in late 2019, close to the time of disease identification. Alternatively, a progenitor of SARS-CoV-2 could have emerged and circulated in humans before acquiring genomic changes that allowed it to cause large numbers of severe cases and initiate the current pandemic *(62)*. However, no samples collected from humans prior to late 2019 have yet been found to be positive for SARS-CoV-2; the second possible scenario is, therefore, currently unsupported by other lines of evidence.

Although the Wuhan sequences exhibit limited genetic diversity, two phylogenetically distinct lineages are apparent, indicating a separation event early in the emergence of the virus. Note that phylogenetic distinction of lineages does not imply phenotypic differences in transmissibility or pathogenicity between lineages, because such distinctions will usually emerge through stochastic processes. These lineages have recently been classified as lineages A and B *(61)* (more rarely referred to as lineages S and L) (see section 6.8.7 for further discussion of SARS-CoV-2 lineage nomenclature). Notably, although lineage B viruses were identified and sequenced first *(52, 59, 60)*, it is likely that lineage A viruses are ancestral because they share two nucleotides with the most closely related coronaviruses in other animals that are not shared in lineage B viruses. Despite the strong quarantine measures adopted in Hubei province, both lineages were exported to the rest of China and have seeded multiple epidemics in other countries.

5.1.3 Identifying the zoonotic origin

SARS-CoV-2 genome sequences and related virus genomes from other animals have been analysed phylogenetically in an attempt to determine the zoonotic reservoir from which SARS-CoV-2 emerged. To date, there has been relatively limited sampling with the aim of identifying the animals involved in the genesis of SARS-CoV-2 and determining when, where and how the virus emerged in humans. Although environmental samples were taken at the Huanan wholesale seafood market in Wuhan at the time of its closure in early January 2020 *(63)* and tested positive, it is currently unclear whether these samples were only from surfaces or were from animals present in the market. If the former, these could simply reflect human contamination. In addition, not all early cases could be linked to this market *(61)*. Identifying the animal source from which SARS-CoV-2 emerged could help combat the spread of conspiracy theories related to the emergence.

Research prior to the COVID-19 pandemic showed that betacoronaviruses are present in a number of mammalian species and exhibit a particularly high phylogenetic diversity in bats *(64–66)*. That bats are likely to have played a role in the evolutionary history of SARS-CoV-2 was confirmed by the identification of a close relative of SARS-CoV-2 (denoted RaTG13) in a species of horseshoe bat (*Rhinolophus affinis*) sampled in Yunnan province, China, in 2013 *(60)*. RaTG13 and SARS-CoV-2 have approximately 96% sequence similarity across the genome as a whole, although this does not rule out decades of evolutionary divergence between them *(67)*. Another coronavirus, RmYN02, was identified in a different horseshoe bat species, *Rhinolophus malayanus*, in Yunnan province in 2019 *(68)*. Although the genome of RmYN02 has experienced a complex set of recombination events, it is the closest relative of SARS-CoV-2, sharing a 97% nucleotide sequence similarity in the ORF1ab gene.

Close relatives of SARS-CoV-2 have also been found in Malayan pangolins (*Manis javanica*) recovered in anti-smuggling activities in Guangdong and Guangxi provinces in southern China. The pangolin coronaviruses are more distantly related to SARS-CoV-2 than RaTG13 and RmYN02 across their genomes as a whole, yet share strong sequence similarity with SARS-CoV-2 in the key receptor binding domain (RBD) of the spike (S) gene (97.4% at the amino acid level) *(69)*.

While it is clear that betacoronaviruses experience frequent and complex recombination events, and that this process has occurred in viruses that are closely related to SARS-CoV-2, there is no evidence at present that recombination played a direct role in the emergence of this virus *(67)*.

Limitations. Although SARS-CoV-2 undoubtedly has animal origins, as did SARS-CoV and MERS-CoV *(64),* the source species will only be resolved with additional sampling of a wide range of non-human animals. It is possible that its origins will never be fully resolved.

5.2 Understanding the biology of SARS-CoV-2

5.2.1 Host receptor usage

Since viruses can replicate only inside the living cells of a host organism, determining the host cellular receptor used by SARS-CoV-2 is essential to understanding its basic biology. Receptor binding is mediated by the S protein of the virus. Genetic similarities in the S protein receptor-binding motif between SARS-CoV-2 and other, previously investigated coronaviruses have helped to identify the cellular receptor to which SARS-CoV-2 binds, and hence the cell types that it might infect. Initial studies indicated that SARS-CoV-2 was likely to use the same angiotensin-converting enzyme 2 (ACE2) cell receptor as the 2002–2003 SARS-CoV, and was likely to bind to this receptor with high affinity *(70, 71)*. Most amino acid residues that are known to be essential for ACE2 binding by SARS-CoV are conserved in SARS-CoV-2 *(70)*. *In vitro* assays confirm the strong specificity for ACE2 suggested by direct structural studies *(72)*.

Limitations. In vitro or *in vivo* experiments were required for full confirmation of genetic sequence findings and are always required to investigate any proposed change in binding affinity.

5.2.2 SARS-CoV-2 evolution: identifying candidate genomic sites that may confer phenotypic changes

All viruses acquire genetic changes as they evolve, and most acquired genetic changes do not substantially affect virulence or transmissibility. Variants between virus genomes sampled from different locations cannot be assumed to cause observed epidemiological differences in COVID-19 between those locations and are instead likely to be stochastic. Despite this, it is possible that a genetic change may occur that causes a corresponding phenotypic change in SARS-CoV-2 of public health importance.

Properly conducted clinical genomic studies could be used to propose candidate variants that might confer clinically observed virus phenotypic changes, but *in vitro* or *in vivo* studies would need to be conducted subsequently to evaluate candidate variants. Virus genomic sequencing

before and after such experimental studies would also be necessary to exclude the possibility that the inferred phenotypic difference is not driven by stochastic virus adaptations to replication within cell culture. Phenotypes observed in cell culture and animal models may not translate to alterations in human disease.

When viruses associated with different phenotypes have several sites that differ between the genomes, it can be difficult to determine which, if any, of those genetic variants cause the observed phenotypic difference. Identified genomic variants could be investigated by reverse genetics to gain a full understanding of their phenotypic characteristics. Reverse genetics can involve systematic synthetic induction of a genetic change in a viral gene and investigation of the phenotypic effect that it causes following production of that protein. Such experiments should only be undertaken under strict compliance with local and (inter)national biosafety and biosecurity laws and regulations.

If a genetic change with a phenotypic effect can be confirmed through these methods, epidemiological phylodynamic studies (section 5.4) can be used to track their global or local spread.

Limitations. It is extremely challenging to identify and provide evidence for genomic changes that may confer phenotypic changes. Virus genomic sequencing is a necessary part of such studies, which should be carefully designed and controlled in order to validate any hypothesized effects. Subsequent *in vitro* and *in vivo* studies with mutant viruses can, in some instances, further support the evaluations of these hypotheses.

5.3 Improving diagnostics and therapeutics

5.3.1 Improving molecular diagnostics

While SARS-CoV-2 was first identified in patients through metagenomic sequencing (section 5.1), this approach is too time-consuming and costly to be used routinely to diagnose viral infection. The development of rapid, inexpensive and sensitive nucleic acid amplification tests (NAATs) for routine molecular detection of SARS-CoV-2 was therefore prioritized early in the outbreak.

The rapid public release of SARS-CoV-2 genomes was important for the design of NAATs. Specifically, these genomes were necessary for the design of primers and probes that would bind effectively to SARS-CoV-2 nucleic acid (through exact or near-exact complementary sequences) but would not bind to other commonly circulating viruses, such as coronaviruses that cause common colds *(73).* Multiple SARS-CoV-2 NAATs were designed and validated by different groups within days of the first genome release (e.g. *74–76*).

As SARS-CoV-2 continues to acquire genetic changes over time during this pandemic, continued generation and sharing of virus genomes will be vital for monitoring the expected sensitivity of the various diagnostic assays in different locations. Mismatches between primers or probes and corresponding binding sites within SARS-CoV-2 genomes could reduce NAAT sensitivity or result in false negatives. Monitoring will be especially important if a variant site is detected in viruses that are phylogenetically closely related. Using multiple targets for SARS-CoV-2 detection, such as a multiplex PCR targeted at two or more regions of the virus genome,

is a cost-effective approach to reducing the chance of false negatives as a result of virus evolution. Consistent failure to detect one target in several clinical samples, or emergence of differences in the sensitivity of assays targeting different regions that were not observed previously and occur in clinical samples but not the established positive control, could be followed up by sequencing of the virus genome or target gene to identify the possible cause.

Several existing platforms allow monitoring of mismatches between user-submitted or publicly available SARS-CoV-2 sequences and the primer/probe binding sites of commonly used NAATs. A number of tools have been developed to monitor such mismatches with common primers and probes, as described elsewhere *(77)*.

Limitations. Genetic sequencing of primer/probe binding regions only is sufficient to investigate the emergence of mismatches. However, whole genome sequencing allows a broader genomic investigation of the spatiotemporal spread of viruses containing mismatches (e.g. to determine when the mismatch variant may have arisen) or the number of times the variant may have independently emerged.

5.3.2 Supporting the design and sensitivity monitoring of serological assays

Virus genomic sequence data can be important in helping to identify virus proteins that are likely to be strongly antigenic, and to indicate how these antigens can be produced for serological assays. Peptide screening has indicated that the four SARS-CoV-2 structural proteins, S, E, M and N, are likely to be the most strongly antigenic *(78, 79)*. SARS-CoV-2 antigens can be synthetically produced for use in commercial assays. Specifically, synthetic coronavirus genes encoding the four proteins can be inserted into expression vector systems *(80, 81)*, where the proteins are produced. This process relies on understanding the genomic sequence and structure of SARS-CoV-2 proteins.

As SARS-CoV-2 acquires genomic substitutions, it is possible that a lineage may emerge with altered antigenic properties (section 5.2.2). This could mean that serological assays fail to detect that an individual has been infected, because the antigen used in the assay is different from that to which the individual was exposed. Also, the antigen detection assays can be impacted by viral change as the capture antibodies might not recognise the adapted viral protein that it aims to detect. Continual assessment of genomic diversity, including in antigenically important sites that may be under selection, could help identify plausible candidate sites that might affect the efficacy of serological assays.

Limitations. In silico predictions of antigenic change from genomic sequence data are inadequate, and the possible sensitivity of serological assays in the detection of genetically diverse infections should always be investigated through laboratory serological validation.

5.3.3 Supporting vaccine design

Several candidate vaccines against SARS-CoV-2 have been designed, and a number have been evaluated clinically *(82)*. SARS-CoV-2 genome sequences have been used in the design of candidate vaccines that rely on inoculation with antigens or mRNA/DNA to stimulate, directly or indirectly, antibody production and cellular responses. Several early candidate mRNA vaccines were designed exclusively on the basis of publicly available SARS-CoV-2 genomes.

Alternatively, synthetic coronavirus genomes can be inserted into expression vector systems *(80, 81)* to produce antigens for vaccines (section 5.3.2).

Limitations. While genomic sequences can assist in the design of candidate vaccines, *in vivo* studies and clinical trials remain critical for evaluating vaccine efficacy.

5.3.4 Supporting design of antiviral therapy

Developing novel antiviral drugs can be time-consuming. Repurposing of existing drugs for SARS-CoV-2 treatment could significantly shorten the time required to obtain approval for clinical use. Genetic and structural information can reveal similarities in proteolytic and replication pathways *(78, 79)* between SARS-CoV-2 and other viruses for which antiviral therapy is already available, and therefore help to determine which existing antivirals might be repurposed. Several candidate drugs that target viral proteins similar to those of SARS-CoV-2 have already been identified *(83)* and are currently undergoing preclinical and clinical studies.

5.3.5 Identifying antiviral resistance or vaccine escape mutations

Once vaccines are implemented and/or antivirals become available, genomic sequencing could be used to support surveillance for variants that may confer antiviral resistance or allow vaccine escape. In-depth genomic or genetic sequencing may be useful in exploring the impact of intra-host diversity on antiviral resistance and vaccine escape (if these occur) or pathogenesis. Genetic sequencing of specific regions of interest, such as the spike gene, may be sufficient to assess the prevalence of specific known variants in pre-identified regions.

Limitations. Such studies are extremely complex and will require targeted and detailed genomic and computational investigation of viruses from patients with a known vaccination history and clinical outcomes. While sequence data from viruses cultured under drug selection pressure may reveal possible antiviral resistance markers, these markers should always be validated by reverse genetics to determine their phenotypic characteristics.

5.4 Investigating virus transmission and spread

5.4.1 Supporting or rejecting evidence for transmission routes or clusters

The placement of sequences within a phylogenetic tree can be used to investigate hypotheses of transmission routes. Phylogenetic clustering of sequences from patients exposed to the same hypothetical source of exposure would be consistent with (although not strong evidence for) that exposure. Sequencing a proportion of cases from outside a hypothesised cluster, and including global reference sequences that are genetically closest to the cluster sequences (to represent the background of genomic diversity), can help evaluate the likelihood that sequences from an identified phylogenetic cluster with a hypothesized epidemiological link are grouped together by chance. The higher the proportion of viruses that are sequenced from the same time and place as the viruses of interest, but which are not identified as probably part of that cluster, the lower the chance that those virus sequences will fall into a cluster by chance. In contrast, considerable phylogenetic separation of virus sequences from two patients (e.g. placement within different

well-supported lineages) would indicate that the two patients had acquired infections from different sources .
Phylogenetic clustering has been used extensively to investigate sources of transmission and exposure events for SARS-CoV-2. In one early study, it was suggested that the clustering of sequences from patients on the Grand Princess cruise ship was consistent with a single introduction of the virus onto that ship, followed by transmission between passengers *(84)*. The observation of monophyletic clades of viruses sampled from members of the same family is consistent with direct transmission between family members, or infection from the same (unsampled) source. Analyses of transmission clusters can guide decisions on whether additional control measures are required to prevent future transmission in identified settings.

Limitations. Phylogenetic information cannot be used to confirm direct virus transmission between two patients or transmission from a single source to several patients, because the involvement of other individuals or sources of exposure that have not been sampled cannot be ruled out. The evolutionary rate of SARS-CoV-2 means that substitutions occur at a slower rate on average than transmissions between patients, and so this remains true even if the sampled genomic sequences are identical.

5.4.2 Identifying and quantifying periods of transmission

Once there is sufficient genetic diversity within a virus lineage, the rate of evolutionary change (substitution rate) can be estimated (section 2.4). If the substitution rate can be estimated, genetic diversity between two sampled viruses with known sampling dates can be used to estimate the TMRCA. The TMRCA of a group of viruses provides a lower-limit estimate of the duration of its circulation within the sampled population. Crucially, the estimated duration of circulation can pre-date the first clinical identification of a case by weeks or months. Molecular clock phylogenetic approaches are particularly useful in identifying where undetected (or cryptic) circulation may have occurred, and in estimating the possible dates of unobserved events.

Initial analyses suggest that SARS-CoV-2 has now acquired sufficient genetic diversity to allow such molecular clock approaches to be applied *(45, 59, 85)*. Accordingly, they have been used to estimate that the pandemic lineage of SARS-CoV-2 emerged in humans by November–December 2019 at the latest *(53, 59, 85)* (section 5.1.2). Important applications of these approaches for the control of COVID-19 include the identification of undetected local transmission in different locations. Identification of largely clinically undetected, local transmission of long duration in an area may suggest that specific locations or populations should be targeted with more extensive or adapted diagnostic surveillance programmes.

Limitations. The time resolution of events that can be investigated is limited by the ratio of the evolutionary rate and the transmission rate. Current estimates of the evolutionary rate of SARS-CoV-2 are that, on average, one substitution occurs approximately every 2 weeks. This means that transmission events between individuals will often not be genomically resolvable, and epidemiologically relevant events that occur on a finer timescale cannot be investigated using these techniques. Early in the outbreak, it was difficult to estimate the duration of cryptic transmission because SARS-CoV-2 had not yet accumulated sufficient genomic diversity. Thus, it was difficult to determine whether a particular genome was the result of local transmission or a new introduction from a location with similar circulating diversity. Studies suggested that SARS-

CoV-2 may have circulated undetected for weeks in Seattle (USA) and in Italy prior to clinical detection of the first community-acquired cases *(84, 86)*. However, a subsequent study argued that the duration of cryptic transmission may have been overestimated by several weeks *(87)*.

Errors in sequencing or consensus-generation can obscure phylogenetic signals when true diversity is low. Sequencing errors can also affect estimates of evolutionary rate variation between lineages and estimated divergence times.

The minimum duration of virus transmission can be estimated even where very few (two or more) cases from a single transmission chain are sequenced. However, incorporating additional samples from a wide geographical area and time frame will reduce the risk that sampled cases cluster closely within a phylogeny by chance, so that the estimated minimum duration is likely to be closer to the true duration.

5.4.3 Identifying importation events and local circulation

If metadata on sampling location are available, sequencing of SARS-CoV-2 genomes can help to determine whether infections have resulted from local transmission or have been imported. Such transmission dynamics may be interpreted cautiously and informally via sequence positioning within a phylogeny (Fig. 2) or investigated through more formal phylogeographical or discrete trait analyses, in which the location at each internal node in the phylogeny is statistically estimated. Incorporation of known sampling times allows the spatiotemporal movement of the outbreak to be reconstructed.

Formal phylogeographical inference includes both discrete and continuous approaches. In the former, virus lineages are considered to be moving between a fixed number of distinct locations *(88)*. The exact areas are defined by the user, and may represent countries, administrative units, cities, etc., depending on the specific questions posed. In the continuous approach, virus lineage movement is modelled on the basis of random walk and diffusion processes between geographical coordinates *(89)*. Both discrete and continuous phylogeographical investigations can be conducted under a number of statistical frameworks, which have different advantages and challenges; these have been extensively reviewed elsewhere *(90, 91)*.

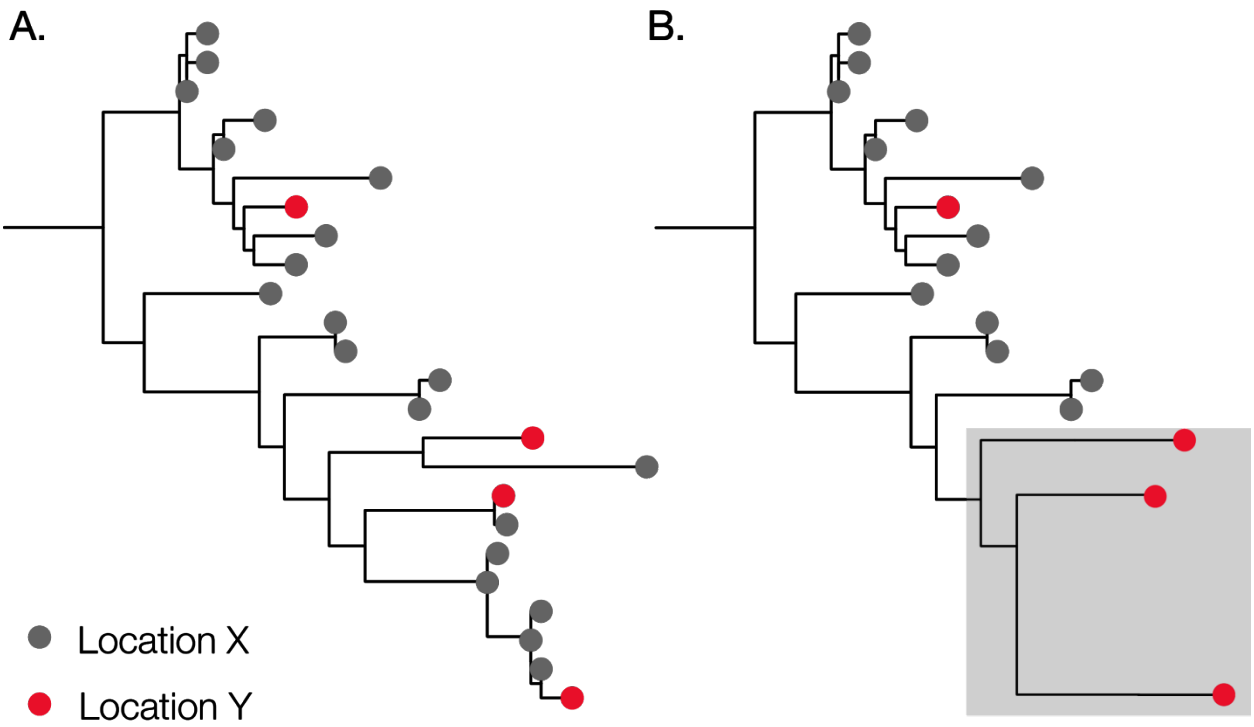

Fig. 2. Potential effect of genomic sampling. Tips in grey represent genomic sequences sampled from location X, and those in red represent sequences sampled from location Y. Panel A: The "true" tree that could be reconstructed in the case of four separate introduction events from location X to location Y. Panel B: Insufficient sampling in location X in the highlighted clade of the tree means that this clade might be (incorrectly) inferred as local transmission in location Y. In this scenario, only two introductions from location X to location Y might be inferred from the phylogeny in the absence of additional travel information

Because the genomic diversity of SARS-CoV-2 was low during the first months of the pandemic, the use of genome sequencing to track its spread was largely limited to national and regional introductions, rather than transmission in communities. Informal, visual interpretation of phylogenetic structures was used extensively in the early literature to infer international or regional movement. For example, genomic epidemiology was used to show that many sequenced cases in Connecticut (USA) were probably imported via domestic travel from other parts of the USA rather than from other countries *(92)*.

Phylogeographical assessment of genomic diversity could be used to assess whether for example stricter quarantine of patients who have visited specific locations are effectively preventing introduction or exportation of SARS-CoV-2 to other regions. For example, in Brazil, continuous phylogeographical analyses showed that spread of SARS-CoV-2 within and between Brazilian states decreased after the implementation of non-pharmaceutical interventions *(53)*.

Phylogeographical approaches that also incorporate sampling times allow estimation of both where and when virus lineage movement events may have occurred. The duration of virus persistence, the number of introductions and relative outbreak size can be determined for each location and can therefore be used to identify specific locations where control measures need to be strengthened.

It may be useful to study viruses in returning travellers to help reconstruct SARS-CoV-2 epidemiology in the country in which the infection was acquired *(93).* New approaches allow patient travel history and sequences from unsampled locations to be incorporated into discrete phylogeographical analyses, thereby allowing more realistic phylogeographical patterns to be revealed and the effect of biased global sampling to be evaluated *(94).*

Limitations. Phylogeographical reconstructions are often computationally demanding. Carefully considered subsampling strategies can help to reduce this computational burden (section 6.8.1). Dispersal of human pathogens is not always well captured by these processes. However, where geographical scales and long-distance travel are restricted, random walks may appropriately capture SARS-CoV-2 movement. Careful consideration should be given to the appropriateness of a continuous process for SARS-CoV-2, as the use of inappropriate diffusion models may lead to incorrect conclusions.

The way in which virus genome sequences are sampled can strongly bias the conclusions of phylogeographical analyses. It is therefore extremely important to conduct analyses and interpret results cautiously, ideally involving experts who are experienced in these methods. There are multiple ways in which phylogeographical analyses can fail to capture the "true patterns" of spread, including the following.

- Undersampling of virus genomes can lead to underestimates of the number of introductions (and hence overestimates of the extent of community transmission) (Fig. 2). This has been clearly highlighted by Lu et al. *(59)*, who showed that a single cluster of closely related sequences sampled from patients in Guangdong actually represented multiple independent introductions via travel. The travel history of patients is important information that should be used to support phylogenetic findings where possible (appropriate data-sharing and protection of patient anonymity is discussed in section 4).
- For discrete phylogeographical analyses, locations of ancestral viruses can only be reliably inferred from the set of locations in which sampled viruses were observed *(89, 95)*. Consequently, with genomic data alone, it is usually impossible to distinguish between direct transmission between two locations, and indirect transmission via an intermediary location in which no genomes have been produced. Distinguishing between these scenarios is only possible in rare situations in which travel information is known *(94)*.
- For certain discrete phylogeographical analyses (particularly those based on discrete trait analysis rather than coalescent or birth–death models), locations that have a higher number of genomic sequences associated with them are more likely to be reconstructed as donor locations from which a virus subsequently spreads *(96)*. Down-sampling of available genomic sequence data from over-represented locations can be useful to investigate whether conclusions are likely to be relatively robust in this regard *(97)*. Use of adjusted Bayes factor support statistics *(98)* may provide additional help in determining whether transition events are supported because of geographically biased sampling.
- Sampling only certain areas of an outbreak can result in inaccurate reconstructions of dispersal history and estimates of dispersal velocity within the continuous phylogeographical framework. Ways of reducing the impact of biased sampling are currently being evaluated *(99)*.

- Information on patient location is often limited to the administrative subunit, for example, municipality. It is often appropriate to consider the uncertainty associated with a sampling location when using the continuous phylogeographical approach. For example, rather than using the geographical coordinates of the nearest city, the entire polygon corresponding to the municipality can be used to define an area from which coordinates for that sample can be randomly selected. Repeating this random sampling during the analysis also helps *(100, 101)*.

5.4.4 Evaluation of transmission drivers

The methods described in the previous section can also be used to investigate the factors that have driven virus dispersal *(97)*. In discrete phylogeographical models (including those implemented as structured coalescent and multi-type birth–death models), information about pairs of defined areas is used as a predictor of the virus lineage migration rate between those areas. The information could include human mobility, population characteristics, such as density, and geographical proximity. Dispersal events inferred by continuous phylogeographical reconstruction can also be analysed to determine whether they are influenced by the "landscape" of environmental or human factors through which they occur.

At the time of writing, such analyses have not yet been applied to SARS-CoV-2. Identifying the drivers of transmission may help to shape new strategies for preventing spread. For example, for Ebola virus this method was used to establish that the virus was more likely to spread between countries that share land borders *(102)*) and this method was subsequently used to evaluate the effect of the taken measures *(103)*.

Limitations. These approaches are computationally demanding, involving large data sets of thousands of genomes, and requiring days or weeks to complete. Use of a pre-estimated distribution of empirical trees may reduce the computational time required and is particularly appropriate for preliminary data exploration. Subsampling of specific clades or random subsampling can also reduce computational burden (section 6.8.1). There is also a computational limit to the number of defined areas that can be included in the model.

Some models have relatively limited flexibility in identifying factors that may be driving SARS-CoV-2 transmission at different times and in different places. Epoch models implemented within the discrete phylogeographical framework *(104)* may be appropriate to investigate time-varying effects of different factors where meaningful time-periods can be predefined. However, control measures are changing rapidly in many countries at different times. This might limit the ability to define epidemiologically useful epochs when applying these techniques above the national or regional scale. Substantial expertise is required to specify these models appropriately and to interpret the resulting estimates.

Techniques to evaluate the effects of interventions are likely to be applied retrospectively, perhaps months after the intervention. Analyses of the effect of interventions that have been successful in reducing cases might help guide future strategies in countries where the outbreak is progressing.

Biased sampling can affect results (section 5.4.3).

5.4.5 Discerning involvement of other species

A number of non-human animal species can become naturally infected with SARS-CoV-2, including cats, dogs and mink *(105–107)*. Where epidemiologically linked pairs of an infected human and an infected animal are observed, it is not feasible to determine the directionality of infection between them. Where multiple animals are infected, phylogenetic investigations of clustering can be used to demonstrate that the animals became infected through different routes, as was done for mink on two farms in the Netherlands *(106)*. Strong support (high bootstrap or posterior support) for the placement of a SARS-CoV-2 genome sequence sampled from a human within a cluster of multiple sequences sampled from mink would be consistent with humans becoming infected from animals. If branch order is not strongly supported, the directionality cannot be robustly inferred. More extensive methodologies employing formal discrete ancestral trait reconstruction could also be performed (section 5.4.3).

5.4.6 Discerning transmission chains between patients using intra-host viral diversity

As mentioned previously, because nucleotide substitutions appear to occur approximately every 2 weeks for SARS-CoV-2, answering epidemiological questions on a finer time scale will be challenging. For other viruses, intra-host genetic variation between virions has been used to increase the resolution at which transmission can be phylogenetically inferred. Intra-host virus minority variants (variants that occur at low frequency within an individual) that are transmitted between patients provide information that is obscured with the consensus genome. Analysis of these variants has been used to improve understanding of the pathways of transmission for many different viruses *(108, 109)*.

Intra-host variation exists for coronaviruses that are closely related to SARS-CoV-2, such as MERS-CoV *(110)*. While the (limited) current data support the existence of intra-host genetic variation in SARS-CoV-2, to date there are very few data sets of within-host variation from known epidemiological clusters that could be used to determine whether this variation is transmitted between patients *(111)*. If it is not, the use of these techniques would not be possible.

Specialist bioinformatic and phylogenetic analyses are required to analyse intra-host virus variation. Given the current lack of understanding of the magnitude of SARS-CoV-2 intra-host variation or its transmissibility, these specialist analyses are not covered here.

Limitations. Many virus genomic sequence data sets will not be appropriate for these analyses. Sanger sequencing or next-generation sequencing using devices that have high per-read sequencing error rates without replication *(112)*, will not provide sufficient information about intra-host variations. Noise caused by cross-sample contamination and sequencing errors can also obscure true signals.

5.5 Inferring epidemiological parameters

5.5.1 Reproduction number

The reproduction number, R_0, can be estimated using population genetic modelling, such as coalescent, structured coalescent and sampling birth–death models. These phylodynamic approaches are all based on the concept that epidemic parameters, such as R_0, affect the shape of time-resolved phylogenies. The various approaches are based on different assumptions, have slightly different data requirements and are susceptible to different forms of bias. They are also appropriate at different points of the epidemic, depending on the extent of geographical spread and the population being studied.

In the very early stages of the SARS-CoV-2 pandemic, geographical population structure could be largely ignored; estimates of R_0 drew on sequence data sampled globally under the approximation that all cases were only a few generations distant from the original epidemic in Hubei, China *(113).* Under these sampling conditions, birth–death models *(114)* and coalescent models that are premised on a single panmictic (random mixing) population can be applied.

As SARS-CoV-2 dispersed globally, it became possible and appropriate to estimate R_0 in different countries, regions and cities. Once substantial geographical clade structuring indicative of the predominance of intra-regional transmission was phylogenetically apparent (section 5.4.3), sampling birth–death and coalescent models premised on a panmictic population became invalid. Methods were then applied at the level of individual identified phylogenetic clusters that represent a lineage circulating in the community. This requires a priori definition of phylogenetic clades.

It is possible to use more complex population genetic models that account for multiple importations of SARS-CoV-2 lineages and community transmission; such models do not require a priori definition of clusters. These analyses are possible using structured coalescent or multi-type birth–death models *(85, 113),* which potentially make use of more clinical and demographic metadata that influence transmission rates or transmission patterns. Their development and implementation require considerable expertise and a good grasp of epidemiological modelling. Computational requirements are much higher than for many other phylogenetic or phylodynamic applications.

Limitations. Practitioners should be aware of the robustness of different methods in relation to different forms of bias. All methods are fallible in the presence of biased sampling, such as occurs when sequencing from transmission chains identified through contact-tracing or small clusters identified epidemiologically. Model mis-specification is a source of bias for all methods. This is ameliorated with more complex structured coalescent methods but these require more computational effort. Individual methods are differently affected by different potential sources of bias.

- Coalescent models based on deterministic relationships between R_0 and the demographic model can provide a biased estimate of R_0 when the epidemic size is small or R_0 is close to 1 *(115)* and stochastic effects predominate.

- Birth–death sampling models require an appropriate parameterization of the variation of the sampling rate over time *(116)*. Since many countries have been actively testing for SARS-CoV-2 since before the start of their outbreaks, it may be sensible to assume that the sampling proportion is greater than zero for the entire duration spanned by the analysis. However, if testing strategies changed at some point during this span, the sampling proportion will need to vary similarly. Coalescent models may give more precise estimates than sampling birth–death models if the sampling rate varies over time.
- Analyses based on a priori identified clusters cannot be considered to be representative of the community as a whole because they neglect small chains of transmission that are either not sampled or are below the size threshold required for analysis. Thus, the clades observed are those that have grown the most successfully. Cluster reproduction numbers are likely to be larger in these clades than in the community as a whole.
- When setting up any analysis that assumes the absence of a structured population, it is critical to ensure that there is likely to be only one R_0 parameter within the time period spanned by the tree relating the samples. If quarantine or other measures have been introduced during the period of study, it will be necessary either to exclude sequences collected after the instigation of these measures, or to include all sequences but allow the R_0 parameter to change over time.

Many approaches, including the birth–death models that are implemented in the Birth Death Skyline Model software package (BDSKY) *(114)*, require the explicit incorporation of prior information to fix certain parameters to known values and, therefore, to improve computational tractability. Typically, it is common to fix a parameter that can be verified from clinical data, such as the rate at which infected individuals become non-infectious. Prior parameter specification should be conducted carefully to avoid potential sources of bias. Conducting analyses using alternative prior specifications may help to determine how sensitive phylodynamic results are to the specified prior parameter.

5.5.2 Scale of outbreak over time and infection-to-case reporting ratio

In traditional population genetics, effective population size (the number of individuals in a population that successfully contribute progeny to the next generation) is estimated rather than absolute virus population size (total number of virions) or number of infected individuals (epidemic size). Effective population size can be used to identify relative changes in epidemic size over time if certain conditions are met. Estimating absolute epidemic size from genetic data has been attempted only recently and is an active area of phylodynamic methodological development. A variety of experimental methods have been applied in the current COVID-19 epidemic. In general, any method for reconstructing epidemic size should account for the major factors that influence genetic diversity within the sampling frame, including: geographical structure, variance in transmission rates, exponential growth and nonlinear population dynamics, and the generation time distribution *(117)*.

Three different approaches and their limitations are highlighted below.

- In some situations, effective population size estimated with coalescent models can be translated into epidemic size. For example, Koelle & Rasmussen derived a formula for doing

so that makes use of independent estimates of R_0 and variance in transmission rates under epidemic equilibrium *(118).* This was subsequently extended to a scenario with exponential growth by Li, Grassly and Fraser *(117)*.

Limitations. The latter approach is limited to the early epidemic period with exponential growth and both approaches may be inappropriate where there is substantial geographical or demographic structuring in virus transmission. For example, where virus transmission is occurring separately in two different locations without substantial transmission between the locations, two different R_0 values may be needed.

- Under a birth–death framework such as BDSKY, the sampling proportion can be inferred, and can be combined with the number of sequences to yield a crude estimate of the cumulative number of cases.

  Limitations. While perhaps a useful means of obtaining a quick estimate, this approach is limited, particularly for small sample sizes, as it ignores the effect of stochasticity in the sampling procedure. It is applicable to an unstructured/panmictic population, such as in a single phylogenetic cluster or early in the epidemic. These approaches do not account for high variance in transmission rates. Less limited approaches exist, such as the use of particle filtering to sample the absolute prevalence curve directly as part of the birth–death inference *(119).*

- The structured coalescent models that are implemented in the PhyDyn package *(120)* for the phylogenetics software BEAST2 have been developed to estimate epidemic size by accounting for variables such as geographical structure, nonlinear dynamics, and high variance in transmission rates.

Limitations. These methods require epidemiological modelling expertise and have high computational requirements. Factors such as natural selection, unmodelled geographical structure, or genomic recombination can still confound estimates.

6. Practical guidance on technical aspects of genomic sequencing and analysis of SARS-CoV-2

Broad considerations for implementing a sequencing programme were discussed in section 3. This section focuses on the different technical aspects of genomic sequencing projects for COVID-19.

6.1 Genome sampling strategies and study design

Genome sampling strategies will depend on the answers being sought. For example, the investigation of nosocomial transmission or evaluation of the findings of contact-tracing (section 5.4.1) may require extensive genomic sampling of most identified patients in the epidemiological cluster of interest, as well as samples that are not part of the cluster being investigated. Samples from outside the cluster are important to support the hypothesis that cluster samples are epidemiologically linked more closely to each other than to other community infections. Conversely, phylodynamic approaches (sections 5.4.2–5.5 and Table 1) are easily biased by non-random sampling of all confirmed cases but, typically, tolerate relatively sparse sampling of a low proportion of all cases. In particular, phylodynamic models assume that sequences are collected uniformly at random from each compartment in the underlying model. This assumption can easily be violated if, for instance, samples are collected as a result of contact-tracing.

For phylodynamic approaches, virus genomes should therefore ideally be sequenced in proportion to true case incidence. How this can best be approximated in practice may vary. Where diagnostic coverage is good across a whole region, a random subset of positive, residual diagnostic samples could be sequenced. However, in many settings clinical diagnostics are conducted non-randomly, including where extensive contact tracing is used to identify cases. The proportion of cases from which clinical samples are available may change over time as different sampling regimes are implemented. In some countries, positive samples will not reflect the true distribution of infections because of disparities in resources or accessibility between locations (e.g. disproportionately fewer samples from rural areas because of challenges in transporting samples for centralized testing). In such countries, it may be more appropriate to deliberately select a set of samples for sequencing that compensates for known biases in sampling. For example, if reporting of suspected cases is known to be more representative than reporting of confirmed cases, it could be appropriate to select samples from different times and locations in proportion to number of suspected cases rather than to number of confirmed cases.
It is not possible to give universally appropriate recommendations for SARS-CoV-2 sequencing, as decisions will depend on the outbreak context and questions to be answered. Key requirements are listed in Table 1. In addition, Annex 1 highlights the types of sampling strategies that have been used in other viral outbreaks for the specific phylodynamic applications considered in Box 1. However, required sample numbers for SARS-CoV-2 will differ from those presented because of differences in baseline viral diversity, genome length, substitution rate and transmission dynamics.

Table 1. Genomic sampling and data considerations for selected applications

Application	Minimal sequence metadata	Ideal additional metadata	Sampling considerations of sequences within targeted location	Other necessary data
Investigating transmission clusters (section 5.4.1)	• Hypothesized cluster definitions (e.g. family cluster) • Date and location of sampling		Dense sampling of all or most individuals within the anticipated transmission cluster **and** dense sampling of control individuals from same place and time.	Control virus sequences from unlinked individuals in similar place and time. Appropriate sequences likely to be locally generated, rather than available via sequence sharing repositories.
Duration of transmission (section 5.4.2)	Date and location of sampling.	Travel history during past 14 days	Virus genomes should ideally be sequenced proportionally to the number of COVID-19 cases. Can be informative with very few sequences (>1) from the location of interest. Estimates usually more accurate as genomic sampling density and diversity increase.	Sequences from other locations outside of the location under investigation; can sometimes be obtained via sequence-sharing repositories (section 4)
Importation events and local transmission (section 5.4.3)				
Evaluation of transmission drivers (section 5.4.4)			• Virus genomes should ideally be sequenced proportionally to the number of COVID-19 cases. • Typically, hundreds of sequences are required spanning several months.	• Sequences from other locations outside of the location under investigation; can sometimes be obtained via sequence-sharing repositories (section 4). • Additional epidemiological, population and/or environmental data sources are necessary and are often available in the public or private sector.
Inference of R_0 (section 5.5.1)				• Sequences from other locations outside of the location under investigation are necessary to check for geographical structure; can sometimes be obtained via sequence-sharing repositories (section 4). • Generation times or serial intervals • Knowledge about measures that might have changed R_0 substantially in time frame, e.g. timing of quarantines.

6.2 Appropriate metadata

To ensure that SARS-CoV-2 genomic data are as useful as possible, they should be accompanied by appropriate metadata. Curating metadata and sharing them locally or publicly can be time-consuming, but both are an integral part of any sequencing pipeline. The required resources should be allocated when the study is being designed.

Metadata should include as an absolute minimum the date and location of sample collection. However, release of additional metadata greatly increases the potential applications of a genomic sequence. Where possible, therefore, information on specimen type and how the sequence was obtained in the laboratory should be included (Table 2). Duplicate samples from the same individual or duplicate sequences from the same sample should be clearly identified. Demographic and clinical information, such as age, sex, presence of co-morbidities, disease severity and outcome, and links to other sequences in the database, are encouraged where such information does not risk identifying the patient.

A global consensus on specific formats for metadata (such as date) would allow genomic sequence data from many different laboratories to be rapidly compiled into larger data sets and reduce ambiguity. Some consensus genome repositories, including GISAID, already place format restrictions on certain fields. If data repositories do not already impose formats, the format restrictions for SARS-CoV-2 shown in Table 2 are suggested. Table 2 also highlights examples of analyses that require provision of specific metadata.

WHO strongly encourages rapid public sharing of sequences and metadata (section 4). However, it is vital to protect patient anonymity. Laboratories should carefully consider whether patients could be identified if all available metadata are shared together. Where few COVID-19 cases have been observed, there is a greater risk of patient anonymity being compromised and therefore fewer data can typically be shared. Where it is judged inappropriate to share detailed metadata via publicly available repositories, it may nevertheless be appropriate to grant access to a small number of users via secure locally developed platforms.

Where it is not possible to share all metadata without risking patient confidentiality, the data that are most useful for global studies should be preferentially shared. For example, sampling location, date and travel history are more useful for phylodynamic studies than patient age or sex (Table 2).

Some laboratories choose to add jitters (noise) to provided dates to decrease the chance that patients can be identified. This can be achieved by a number of methods, for example, by choosing a false date within 5 days either side of the date of sample collection or by using the sequencing date as the sample date. Such practices negatively affect molecular clock based phylogenetic inference and should ideally be avoided. If, nevertheless, this practice is followed, information on exactly how the new date was selected should be provided as a note.

Table 2. Metadata format and use[a]

Metadata type	Recommended format if applicable	Analyses for which the metadata are required
Sample-specific metadata		
Date of sample collection	*YYYY-MM-DD* If the date of sampling is unavailable, date received by testing laboratory could be adopted as an alternative, but this should be clearly indicated.	Molecular clock phylogenies (including any models implemented in BEAST or BEAST2) These can provide estimates of dates of introduction, changes in outbreak size over time and evolutionary rate
Location	*Continent/country/region/city* For discrete phylogeographical analyses (section 5.4.3), location resolution can be low (e.g. country level information for consideration of movement between countries) but higher resolution data is preferable to allow finer-scale analyses. Continuous phylogeographical approaches typically require relatively high-resolution data (e.g. city or municipality).	Any phylogenetic interpretation of global or regional virus spread (including models in BEAST or BEAST2)
Host	For example, *human* or *Mustela lutreola*	Host range and virus evolution
Patient age	For humans, give age in years (e.g. *65*) or age with unit if under 1 year (e.g. *1 month*, *7 weeks*). For non-human animals, *juvenile* or *adult*.	Descriptive epidemiology or as a possible trait for discrete phylodynamic inference
Sex	*Male, female* or *unknown*	Descriptive epidemiology
Additional host information	No standard format For animals, this may include context, such as "*domestic - farm*", "*domestic - household*", "*wild*", etc.	Disease surveillance in human or animal hosts

Travel history	No standard format Travel history in the 14 days preceding symptom onset should be obtained from patients where possible Deliberate release of travel history only to a low resolution (e.g. country) may be important to protect patient confidentiality	Phylogeographical or phylodynamic analyses directed at estimating transmission rates or routes between regions
Cluster or isolate name	No standard format Appropriate formats may include “Same epidemiological cluster as sample X”, “Same patient as sample X”, or “Sample from patient XYZ” (where XYZ is an anonymized identifier that cannot be traced back to the patient or used to access other patient data that might compromise confidentiality)	Phylogenetic down-sampling to ensure appropriateness of phylodynamic models Cluster investigation
Date of symptom onset	*YYYY-MM-DD*	Specialist phylodynamic applications that investigate transmission clusters
Symptoms	No standard format Appropriate degree of symptoms; may include "severe”, “mild” and “out of norm”	Descriptive epidemiology
Clinical outcome if known	No standard format Appropriate formats may include “recovered”, “death” and “unknown”	Descriptive epidemiology
Comments	No standard format Appropriate comments may include how samples were selected (e.g. “cluster investigation”, “randomly”), or the storage location of other data files, such as raw read data	Interpretation of data quality or utility
Sequence and sample-specific metadata: extensive data should be shared as patient anonymity is typically unaffected		
Specimen source, sample type	No standard format Examples: “sputum”, “blood”, “serum”, “saliva”, “stool”, “nasopharyngeal swab”	Effect of cell tropism

Passage details, history	No standard format It is important to indicate that cell culture was conducted (e.g. "Cultured"); ideally, this information should include the type of cells used and the number of passages	Removal of cell-cultured viruses (which may have induced genetic changes)
Sequencing technology	No standard format Ideally, this should include the laboratory approach and sequencing platform (e.g. "Metagenomics on Illumina HiSeq 2500" or "ARTIC PCR primer scheme on ONT MinION")	Sequencing artefacts
Assembly method, consensus generation method	No standard format	Sequencing artefacts
Minimum sequencing depth required to call sites during consensus sequence generation	e.g. 20x	Sequencing artefacts

[a] Sharing all of the information listed in this table might compromise patient anonymity. An ethical review should be conducted to determine which metadata can be safely shared. It may be appropriate to share fewer data on public databases than on databases that are held and analysed locally.

6.3 Logistic considerations

6.3.1 Location

The decision on where to base a sequencing laboratory should be carefully considered. Sequencing should generally be conducted by institutions with the necessary experience and infrastructure for next-generation sequencing. If such infrastructure is not available, the decision of where to host the sequencing laboratory should take into account the impact on other work carried out by the laboratory. For example, integrating sequencing in an existing diagnostic laboratory may allow shorter turnaround time, but this potential gain should be balanced against the risk of disrupting other operations in the laboratory, which may already be in the process of scaling up its diagnostic capacity for SARS-CoV-2. Careful consideration should also be given to the availability of space and equipment.

Where handling of PCR amplicons is necessary for sequencing (e.g. methods described in section 6.5.4), it is important to reduce the potential for amplicon contamination through appropriate laboratory management. Physical separation of areas that will be used for pre- and post-PCR handling of SARS-CoV-2 material, and a one-way flow of personnel and materials from pre- to post-PCR areas, are strongly advised. Where separate areas are not already available, laboratories could adopt strategies, such as purchase and use of separate gloveboxes or for pre- or post-PCR activities. Equipment should ideally be designated for use only with either pre- or post-PCR material, and required reagents should ideally be stored separately (e.g. in different freezers or different laboratories) to reduce the risk of contamination. As for all sequencing, negative controls are valuable to detect contamination.

6.3.2 Biosafety and biosecurity

Risk assessments should always be conducted to assess biosafety and biosecurity. The results of such risk assessments should be communicated to workers involved in the relevant processes.

Individual laboratories should always conduct local risk assessments for every step in their SARS-CoV-2 protocol. International, national and local legislation should be consulted to ensure the safe handling of SARS-CoV-2 material. WHO has issued broad biosafety guidelines *(121)*. Samples should be inactivated at the earliest possible stage (usually prior to RNA extraction) using chemical methods that preserve RNA quality. Methods used to extract RNA prior to diagnostic NAATs are generally appropriate for sequencing. As for most NAATs, heat inactivation prior to sample extraction is not recommended because it risks damaging RNA integrity.

6.3.3 Ethical considerations

Ethical reviews should be conducted to ensure that patients have given appropriate consent for samples to be collected and sequenced, and to consider the subsequent use, storage and publication of data.

Some sequencing approaches, such as metagenomics, will generate human genomic data. Any human genomic sequences should be removed from the viral data set via an automatic analysis pipeline at the earliest possible stage, without manual operation by staff (see section 6.7.1), unless ethical approval and explicit patient consent to process human genetic data have been obtained. If personal or human data have to be stored, proper encryption of all such files is highly recommended.

Ethical reviews should determine the maximum possible relevant metadata that can be shared without risking patient confidentiality.

6.3.4 Human resources

It is important to ensure that there are sufficient staff to support all aspects of the sequencing programme, from clinical sampling to communication of results and sharing of sequences and metadata. The costing of a sequencing programme should include personnel costs, as well as the costs of personal protective equipment, consumables, purchase and maintenance of other equipment and computational architecture. If several laboratories or institutes are involved in collaborative investigations, it may be valuable to obtain written agreement on each laboratory's responsibilities (e.g. in relation to funding, staff that can be committed and work to be performed) and the expected benefits before the project begins. The content of such agreements will vary; existing institutional collaboration agreements or material transfer agreements may provide appropriate templates.

The human resource implications of any planned sequencing programme should be considered with reference to the expected working patterns. In general, a normal working pattern should be encouraged so as to avoid staff burnout. The probability of staff sickness and unavailability in the context of the COVID-19 pandemic should also be accounted for. Attempts to build extra capacity into the workflow should be considered early, while appreciating that generating pathogen genomes from clinical samples requires a multidisciplinary team with highly specific skill sets. Workload intensity and predictability will depend on the goals of the project (Table 3).

Diagnostic laboratories are often central to the identification of positive cases and the safe processing and storing of patient samples. If a large-scale sequencing project is planned it is recommended that a representative from the diagnostic laboratory is designated to liaise directly with the sequencing team to ensure efficient retrieval of samples and relevant metadata for downstream applications.

Table 3. Expected workloads for specific goals of the sequencing programme

Goal	Typical speed of sequencing required for impact	Work intensity	Workload
Contribution to global phylodynamics	Low (often retrospective)	Variable	Predictable, though may change in response to changing size of outbreak
Identification of importation events and local circulation	Low (often retrospective)	Variable	Predictable, though may change in response to changing size of outbreak
Investigation of diagnostic assay specificity	Moderate	Low	Unpredictable if in response to observed change in assay specificity; predictable if part of continuous monitoring
Supporting or rejecting evidence for transmission routes or clusters	High	High	Unpredictable, in response to clinical need

6.4. Choosing appropriate material for sequencing

6.4.1 Material for sequencing

The acquisition of sufficient, high quality SARS-CoV-2 RNA helps to maximize sequencing yield and the ultimate quality of genome sequence data. The quantity and quality of an RNA sample are affected by: choice of clinical sample; handling of clinical sample; method of viral RNA isolation; and the technical proficiency of personnel.

Where several different sample types are available, it is beneficial to select one that has a high viral load and low levels of human or bacterial genetic material contaminants (Table 4). Such samples can be sequenced using both metagenomic and SARS-CoV-2 targeted techniques (section 6.5). Some materials, such as faeces, may require centrifugation and filtration prior to viral RNA extraction, to deplete human or bacterial cellular material that may reduce the sensitivity of sequencing.

Table 4. Direct sequencing of clinical specimen and cell culture

Starting material	Quantity of viral RNA	Content of non-viral material	Reference
Serum, blood	Detection very infrequent	High in whole blood, low in serum	*(77, 122–126)*
Respiratory samples (naso-oropharyngeal swabs, sputum, bronchial lavage fluid)	Detection frequent at high levels	High but can be reduced through filtration and centrifugation	*(122, 127–135)*
Oral fluids and gargling, mouth washes	Detection highly variable depending on collection and handling process; can be frequent	High but can be reduced through filtration and centrifugation	*(133, 136–143)*

Faecal and anal swabs, faeces	Detection variable, but when detected this can be at high levels.	High, but can be reduced through filtration and centrifugation	*(122, 144–147)*
Autopsy, tissue samples	Detection possible, although samples are rarely accessible	Very high, challenging to reduce through filtration and centrifugation	*(148–155)*
Viral isolate from clinical sample (cell culture, animal model) (biosafety level 3 facility required)	Levels high following culture, but culture may induce artificial variants	Moderate/high but can sometimes be reduced via filtration and centrifugation depending on exact sample type	*(8, 156, 157)*

In many settings, the only samples routinely available for virus genome sequencing will be residual diagnostic samples. Samples collected for NAAT diagnostics are typically also appropriate for sequencing *(77)*. Nasal swabs, throat swabs and saliva have been found to have high viral loads shortly after symptom onset and for up to 25 days afterwards *(140, 158, 159).* SARS-CoV-2 viral load and viral RNA abundance in samples is normally highest in the first week following disease onset *(158, 160)*.

If feasible, isolates for sequencing should be selected from positive samples that have already been processed by a molecular diagnostics laboratory (Fig. 3). Sharing resources in this way prevents duplication of work in sample processing and nucleic acid extraction and can therefore save human and other resources, and cost. Some commercial molecular diagnostic kits use viral lysates as inputs, and do not allow storing of extracted RNA. In such cases where the components of the commercial lysis buffer are not disclosed, it can be extremely challenging to reuse prepared lysates with other commercial extraction kits and it may be necessary to perform fresh inactivation and extraction directly from the original clinical sample. Disclosure of the components of commercial lysis buffers would assist researchers in developing strategies to reuse already inactivated lysates for use in subsequent sequencing activities.

A practical and effective system of sample identification should be used if samples move between laboratories; ideally, the same sample identification should be used in all handling laboratories.

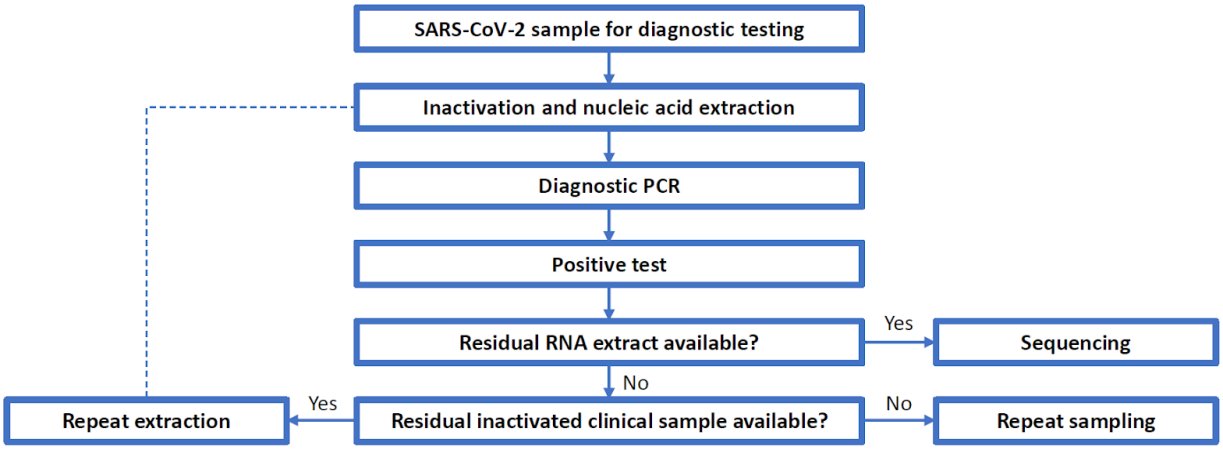

Fig. 3. Example workflow for specimen retrieval from a diagnostic laboratory.

Preserving viral RNA is important for the production of high-quality sequence data. This can be achieved by maintaining a cold-chain between sample collection and analysis, reducing the number of times that RNA or samples are frozen and thawed, and minimizing the time between sample collection and sequencing. RNA that is stored or shipped at 4 ℃ for longer than a few days is unlikely to be of sufficiently high quality for sequencing unless it was first preserved in an RNA-stabilization solution. Quality will be substantially higher if RNA can be stored at –20 ℃ or preferably –80 ℃. Viral lysates typically cannot be stored at 4 ℃ for as long as extracted RNA. Many sequencing protocols include steps that improve the storage capability of a sample, including reverse transcribing RNA to cDNA, or second-strand synthesis/generation of double-stranded DNA PCR amplicons. PCR amplicons can be stored at 4 ℃ for many months without reduction in sequencing quality. In some contexts, it may therefore be appropriate to perform these steps rapidly after the diagnostic PCRs, so that material can be stored or shipped with fewer temperature constraints prior to library preparation.

6.4.2 Control samples

Negative control samples, such as buffer or water, should always be included in any sequencing run that contains multiple samples. They should be included at the earliest stage possible and should proceed with samples through all stages of the sequencing pipeline. This is extremely important to rule out contamination during a sequencing run that occurs in the laboratory or during bioinformatic processing.

Positive control samples with known genetic sequences can be useful to validate newly adopted or adapted bioinformatic pipelines for consensus calling, but do not need to be included in every sequencing run.

6.5 Enriching SARS-CoV-2 genetic material prior to library preparation

Sequencing strategies for SARS-CoV-2 include metagenomic approaches, which do not require prior knowledge of the genomic sequence, and targeted approaches, which rely on knowledge of the genome. Both approaches typically attempt to enrich SARS-CoV-2 genetic material relative to other RNA/DNA prior to sequencing. If sufficient residual RNA is available and has been stored appropriately (section 6.4.1), most approaches can be performed using RNA extracted for diagnostic assays. Many different protocols have already been shared for SARS-CoV-2 sequencing. Some of these are highlighted below; others have been collated by the United States Centers for Disease Control and Prevention (CDC) *(161)*.

6.5.1 Metagenomic analyses of uncultured clinical samples

Metagenomic protocols permit untargeted sequencing of nucleic acid in a sample, including viral genomic material if present *(162).* These protocols offer a hypothesis-free approach to pathogen discovery, as they require little prior knowledge of the pathogen of interest *(163).*

Depletion of host or other non-SARS-CoV-2 genetic material in a sample leads to a higher proportion of SARS-CoV-2 reads in generated sequence data and therefore a higher chance of recovering a full genome. SARS-CoV-2 metagenomic approaches therefore typically include steps to remove host and bacterial cells, through either centrifugation or filtration prior to RNA extraction, or chemical or enzymatic removal of unwanted DNA/RNA. This is easier for liquid samples, from which cells can be more easily separated, such as bronchoalveolar lavage (Table 4). Ribosomal RNA (rRNA) and DNA content are also commonly depleted during library preparation for virus RNA sequencing, and carrier RNA is often omitted from extractions or replaced with linear polyacrylamide. Despite such measures, samples may still contain high quantities of off-target host DNA/RNA that may also be sequenced. Metagenomic approaches therefore generally benefit from input of samples with high virus loads (such that a reasonable proportion of the genetic material in the sample is virus). Alternatively, a large number of reads usually needs to be generated; in this way, even if SARS-CoV-2 genetic material represents only a small proportion of the reads, it will still be possible to obtain the entire virus genome.

Metagenomic sequencing typically produces high numbers of off-target, non-virus reads. It is also often (though not always, depending on the sequencing platform and multiplexing) more costly than targeted capture-based or amplicon-based sequencing approaches, because more data have to be produced to generate one SARS-CoV-2 genome. Moreover, pretreatment steps that are particularly beneficial for metagenomics, such as centrifugation, are not typically performed for molecular diagnostic assays so new extractions that incorporate pretreatment steps may have to be performed for metagenomic sequencing. Targeted sequencing approaches (sections 6.5.3 and 6.5.4) are often more cost-effective and require fewer resources; they may therefore be more appropriate where the benefits of metagenomic approaches (e.g. pathogen discovery, detection of co-infections) are not required. The success of metagenomic approaches varies between methods. Several studies have shown a rapid reduction in the success of several metagenomic sequencing analyses in samples with real-time PCR (qPCR) cycle thresholds (Cts) of over approximately 25–30. For such samples, multiplex and capture-based PCR methods achieve

consistently higher coverage across the genome than metagenomic sequencing *(57, 164)*. The number of sequencing reads per sample that must be generated to obtain the full genome will depend on sample type, pretreatment procedures to remove host material and level of viraemia.

6.5.2 Metagenomic approaches following cell culture

For samples with a low viral load, the proportion of viral genetic material can theoretically be increased by allowing the virus to replicate in cell culture. However, the biosafety risks associated with virus culture are significantly higher than those associated with uncultured clinical samples. Biosafety level 3 facilities are required, with extensive additional procedures to ensure safe handling and storage. In addition, passage in cell culture can result in artificial mutations in the sequences, which were not present in the original clinical sample. This can have major implications for subsequent analyses. Using cell culture solely for the purpose of amplifying virus genetic material for SARS-CoV-2 sequencing should therefore be avoided, especially now that other bait-capture and amplicon-based approaches are available to improve sequencing sensitivity.

6.5.3 Targeted capture-based approaches

Following preparation of a metagenomic sequencing library, capture-based approaches that enrich for SARS-CoV-2 genetic material can be performed before sequencing. Such approaches rely on hybridizing DNA that has been reverse transcribed from viral RNA, to DNA or RNA baits. These baits are designed to be complementary to regions of the SARS-CoV-2 genome. Off-target library material that has not successfully bound to a bait (e.g. host DNA) can be removed using enzymatic or physical approaches. This reduces the chance of detecting other co-infections but increases the expected number of sequencing reads that will map to the SARS-CoV-2 genome, allowing more samples to be effectively sequenced together in multiplexed runs.

One advantage of using a capture-based approach over a PCR amplicon-based approach (section 6.5.4) is that capture-based approaches can tolerate sequence differences from the probe sequences of 10–20%. This is higher than the mismatch tolerated by PCR, where such a divergence from the primer sequences would result in a high risk of amplicon failure. Capture-based approaches can therefore be used to enrich successfully for relatively divergent SARS-CoV-2 sequences. Capture-based approaches are typically more complex to establish and more expensive than PCR amplicon-based approaches.

Several specific SARS-CoV-2 capture panels that are commercially available or can be designed to order can result in a 100–10 000-fold increase in sensitivity. When multiple samples are to be sequenced together in a single pool, it is most cost-effective to perform capture on an entire pool of up to 96 multiplexed samples after sample barcoding. Several published protocols have been validated for capture-based SARS-CoV-2 sequencing (e.g. based on *(165)*).

6.5.4. Targeted amplicon-based approaches

PCRs that generate amplicons tiling the whole SARS-CoV-2 genome can be used to amplify virus material prior to sequencing library preparation. Unlike capture-based approaches, amplicon-based approaches do not tolerate substantial mismatch between the targeted sequence and the primers that are used. The targeted genomic diversity must therefore be relatively low, and/or the

target sequence sufficiently known to allow primers to be designed to target more conserved genomic regions. Given that SARS-CoV-2 has only recently emerged in humans and therefore shows relatively low global genomic diversity, PCR-based approaches are currently highly appropriate for SARS-CoV-2 sequencing. However, the occurrence of amplicon failures needs to be monitored and primers replaced where failure occurs as a result of substitutions in primer binding sites.

Optimized PCR-based approaches are highly specific and sensitive and allow whole SARS-CoV-2 virus genomes to be routinely generated from samples with PCR Ct values of up to 30. Partial genomes can be routinely generated from samples with Ct values of 30–35. However, these values are an approximation; Ct is not a perfect predictor of amplification success as it can vary between different diagnostic methods *(166)*, and use of different types and quality of sample will affect sensitivity. In addition, genomic regions targeted in PCR diagnostic assays are typically far shorter than those used in common amplicon-based sequencing approaches, so RNA degradation will typically affect PCR-based sequencing more than PCR diagnostics. Where targeted genomic diversity is low, PCR-based approaches are a cheap, rapid and convenient way of increasing the amount of virus genetic material available in a sample prior to sequencing.

Several different primer sets for amplicon-based full genome sequencing have been described. These target amplicons of different lengths, typically 400–2000 base pairs (bp). Longer amplicons require fewer PCR primers to scaffold the whole genome, but may result in larger gaps in the consensus genome in the event of an amplification failure of one primer pair. Longer amplicons are suitable for long-read platforms but require fragmentation for short-read sequencing tools. The most widely used scheme is currently the tiling amplicon-based approach designed by the ARTIC Network *(167)*. While the ARTIC protocol focuses largely on nanopore sequencing from Oxford Nanopore Technologies, several laboratories have validated the ARTIC approach on other sequencing platforms *(112, 168)*.

It is vital to adopt strategies to prevent amplicon contamination of other diagnostic testing or further sequencing (section 6.3.1).

6.6 Selecting sequencing technology

After initial sample preparation to enrich for SARS-CoV-2 genetic material, libraries can typically be prepared using standard sequencing protocols that are appropriate for any virus. The protocol will depend on the instrument used. Before investing in sequencing capacity for the first time, or adopting an alternative technology, consideration should be given to run-time, costs, ease of use, subsequent data processing, throughput (rate of data production) and sequencing accuracy of the various technologies (Table 5) (see also section 6.7).

Conventional sequencing (Sanger sequencing) can be used to sequence individual fragments (up to 1000 bp) in separate reactions. Whole genome sequencing of SARS-CoV-2 would require at least 30 individual amplicons to be separately sequenced per patient sample. Sanger sequencing is therefore likely to be most useful for sequencing short fragments of genomes, for example, to fill gaps in assemblies following next-generation sequencing or for investigating virus diversity in short regions, such as primer binding sites, following the failure of a diagnostic assay.

Next-generation sequencing platforms are more appropriate for routine, whole genome sequencing. Sequencing platforms that have been commonly used to date for SARS-CoV-2 include those from Illumina, IonTorrent and Oxford Nanopore Technologies. Unlike Sanger sequencing, in which all DNA molecules in a sample must have the same or highly similar sequences (e.g. following PCR of a single amplicon), these technologies allow concurrent sequencing of multiple fragments of the SARS-CoV-2 genome. All next-generation sequencing platforms allow multiple samples to be sequenced together in a single run. The key advantages and limitations of each technology are summarized in Table 5. While all platforms are appropriate for generating consensus genomes of SARS-CoV-2, some may be better suited to meet specific sequencing programme goals. For example, a fast turnaround time may be important for clinical applications, while read-level accuracy may be more important for investigation of intra-host diversity.

Table 5. Commonly used platforms for sequence analysis of SARS-CoV-2 and their characteristics[a]

Instrument	Advantages	Limitations	Instrument run-time	Sequencing throughput	Relative availability and cost
Sanger sequencing	Widely accessible Easy to use Cost-effective sequencing if few targets required	Very low throughput Amplicons (often no more than 1000 bp) must be individually amplified and sequenced Expensive for full genomes Inappropriate for metagenomics	Typically a few hours	100 kB-2 Mb per single run	Widely available Relatively low cost for a few targets
Illumina (e.g. iSeq, MiniSeq, MiSeq, NextSeq, HiSeq, NovaSeq)	Very high sequencing yields possible Very high accuracy iSeq is portable Methods for handling data are well established	With the exception of Illumina iSeq, expensive to purchase and maintain compared with some other platforms Maximum read length 2 x 300 bp.	10–55 h, depending on the instrument	1.2–6000 Gb, depending on instrument	High maintenance and start-up costs Moderate running costs
Oxford Nanopore Technologies (Flongle, MinION, GridION, PromethION)	Portable, direct sequencing Real-time data Low start-up and maintenance costs Can stop sequencing as soon as sufficient data are achieved Very long read lengths achievable (exceeding the full length of the SARS-CoV-2 genome)	Challenges with homopolymers Error rate per read is ~5% (R9.4 flowcells) so use of appropriate pipelines is critical to obtain high-accuracy consensus sequences Currently unsuitable for determining intra-host variation unless replicate sequencing is used *(112)*	Reads available immediately Can be monitored and run for up to several days as required	Ranging from < 2 Gb for Flongle flow cell to 220 Gb for PromethION flow cell Up to 48 flow cells can be used on PromethION	No maintenance and low startup costs Moderate running costs.
Ion Torrent	Fast turnaround once sequencing starts	Challenges with homopolymers Expensive to purchase Maximum typical read lengths around 400 bp.	2 h–1 day, depending on chip and device	30 Mb–50Gb depending on device and chips	Moderate costs.

[a] This listing of the various instruments is to provide an overview of most commonly used tools for SARS-CoV-2 genomic sequencing and does not imply WHO endorsement of these products.

6.7. Bioinformatic protocols

The selection of an appropriate bioinformatic protocol that can process raw read data into whole genome consensus sequences is usually as important as that of the sequencing platform. The use of an inappropriate bioinformatic protocol could produce erroneous results that can severely affect downstream analyses.

6.7.1 Overview of typical bioinformatic steps

Archiving of raw read data

Sequencing generates large volumes of data (Table 5). The costs of the computational architecture required to store and handle these data should be considered when a sequencing pipeline is being developed. The volume of raw data produced, usually stored as FASTQ files (which store genetic sequences along with the quality score of each base in the sequence), will depend on the number of samples processed. Short-read data that has been enriched for viral sequences, either by bait capture or by PCR amplification, may often comprise 1–2 million reads per sample, and require up to 1 Gb of disk space, depending on read length. Unenriched samples that have been sequenced metagenomically will typically require 100-fold greater read numbers to obtain good genomic coverage of SARS-CoV-2, since the proportion of viral reads in such samples can be less than 1% of total reads *(164)*.

If storage capacity is limited, permanent storage of raw data may not be feasible. While it is preferable to store raw read data locally for as long as possible, it is not always critical if such storage becomes a barrier to additional sequencing. An exception is the storage of raw data from metagenomic or metatranscriptomic sequencing, which may contain information about co-infection with other viruses or bacteria. Such samples represent a valuable asset and efforts should be made to preserve the information even if raw reads cannot usually be stored in other circumstances.

A best-practice alternative to permanent local archiving of raw read data is to upload data to a repository, such as SRA (NCBI), DDBJ or ENA.

Unless ethical review has approved the investigation and sharing of human genomic sequences, and all participants have given explicit informed consent to this, data submitted to public repositories should first be stripped of reads of human origin. For SARS-CoV-2 targeted sequencing approaches, all sequencing reads can be mapped to the SARS-CoV-2 genome and mapped reads extracted. The extracted reads that are subsequently shown not to map to the human genome can typically be submitted to repositories. Existing software can facilitate this for different platforms, for example, nanostripper for data produced using Oxford Nanopore Technologies devices *(169)*. For metagenomic projects in which one of the aims is to identify co-infections, strategies to remove human reads are more complex. Some repositories, such as SRA, can remove human genetic reads from metagenomic datasets if contacted directly. Pipelines to remove human reads can also be established using taxonomic classification software, such as Kraken2 or CLARK *(170, 171)*, or software for removal of reads mapping to human genomes, such as GSNAP *(172)*. Processes to remove human reads should always be evaluated as part of

the ethical review of any project and should be extensively tested to ensure their efficacy. Other data-sharing approaches and ethical considerations are covered more extensively in section 4.

Genome assembly from raw data

A number of freely available software pipelines tuned for SARS-CoV-2 sequencing have been developed. Many require minimal local set-up and have clear instructions for use. A useful (non-exhaustive) repository of links to sequencing pipelines, including bioinformatics where established, is maintained by CDC *(161)*. Further packages for virus sequencing are available and would be appropriate following extensive customization to SARS-CoV-2.

The bioinformatic pipeline will depend on the pre-sequencing laboratory stages (e.g. PCR amplification requires bioinformatic trimming of primer sites) and the sequencing platform and reagents used. Bioinformatic pipelines will often include steps similar to those shown in Table 6.

Table 6. Common steps in bioinformatic consensus building for the two most commonly used next-generation sequencing platforms.[a]

<table>
<tr><th>Stage</th><th>Illumina[b]</th><th>Oxford Nanopore Technologies (ONT)[b]</th></tr>
<tr><td>Base calling of raw read signal into FASTQ format data</td><td rowspan="2">Bcl2Fastq (Illumina)
Sequencing facilities will often conduct these stages before sending to data users</td><td>Guppy (ONT)</td></tr>
<tr><td>Demultiplexing of reads into those from different samples</td><td rowspan="2">Porechop (173) for demultiplexing and adaptor trimming</td></tr>
<tr><td>Removal of sequencing artefacts, including sequencing adaptors</td><td>Cutadapt for adaptor trimming (174)</td></tr>
<tr><td>Trimming of low-quality base pairs</td><td>Trimmomatic (175)</td><td>Reads that are substantially longer or shorter than the expected read-length may be removed for multiplex PCR schemes</td></tr>
<tr><td>Removal of optical duplicates for short-read data from protocols that include enrichment or amplification</td><td>Picard Mark Duplicates</td><td>N/A</td></tr>
</table>

Alignment of on-target reads to a canonical reference genome, such as the genome NCBI reference sequence NC_045512 *(176)*	Bowtie2 *(177)*	Minimap2 *(178)* or BWA *(179)*
Removal of sequencing artefacts from on-target reads, including primers for multiplex schemes (optional, depending on sequencing method)	Pipelines such as iVar for primer trimming *(112)*	Pipelines such as by the ARTIC network *(167)*
Identification of variants from the reference sequence, with appropriate quality thresholds to distinguish true variants from sequencing errors. Variant calling methodology is strongly dependent on the library protocol and sequencing technology, and in most cases requires substantial tuning of parameters to distinguish true variants from false positive calls. The simplest protocols for variant filtering follow steps to remove positions with low read depth or those supported by reads with insufficient quality, and require that a significant proportion of the base calls supports a variant from the reference	Samtools mpileup followed by BCFtools filter and call *(179, 180)*. Variants could be kept in the following cases, for example: – a minimum depth of 5 reads at each position, or greater for PCR-amplified samples – a minimum average base quality of 15 – at least 75% of reads at the position supporting the call – of reads spanning the position, at least one in the forward orientation and at least 1 in reverse (for paired-end Illumina sequencing)	Use of Nanopolish *(181)* or Medaka (ONT) to improve consensus sequences It is important to use established pipelines that have been fully validated. Pipelines can include various conditions, such as: – a minimum depth of 20 reads for Oxford Nanopore data to account for error rates – thresholds at which sites are not resolved, but are marked as ambiguous

[a] Based on the SARS-CoV-2 sequences submitted to GISAID during the first three months of the pandemic. The software listed is for illustrative purposes only; other appropriate software is available at each stage. For Ion Torrent, similar software as for the Illumina platform can be used.

[b] The mention of specific instruments and software does not imply WHO endorsement of the products.

Regardless of the pipeline, nucleotide variants should not be called if the number of unique supporting reads at the site is lower than the required depth for confidence. Instead, such sites should be called as ambiguous bases (N) in the final consensus genome. Depending on the accuracy of raw reads in the chosen methods, any sites with fewer than 5–20 supporting unique reads cannot be accurately called. The minimum expected contamination level can be determined from the number of SARS-CoV-2 reads observed in the negative control, and sites should only be called if depth greatly exceeds this level.

Metagenomic and capture-methods are quantitative, meaning that the read depth of the samples will approximately reflect the number of viral genome copies in the starting library. For samples with a low viral load, variant calling should be performed with caution, as even a small number of contaminating reads can interfere with the signal from the sample. Negative controls should also be sequenced to allow the likelihood of contamination to be assessed.

Variants in samples with high Cts that probably have few start RNA copy numbers should be evaluated cautiously, because stochastic presence of certain variants among the few copies present may lead to artefactual errors. Variants should also be considered very cautiously if the enzymes used during reverse transcription and/or PCR frequently induce errors. High-fidelity enzymes should be used where feasible to protect against such errors.

6.7.2 Dealing with multiplexed data

It is cost-effective to sequence multiple viral samples in a single sequencing run. This is generally accomplished by the addition of unique adapters or barcodes to the sequencing reads. When raw data are generated, it can be de-multiplexed by allocating reads to samples with matching barcodes. Multiplexing introduces new complexity to the process of quality control of bioinformatic outputs, since it is possible for barcodes to be incorrectly determined, due to a process known as index hopping or index misassignment. These artefacts particularly affect samples with a low viral load, as a small number of contaminating reads can have a disproportionate effect on the genome consensus. To guard against this, it is recommended that multiplexed pools contain at least one negative (buffer) control and, if feasible, one non-SARS-CoV-2 control, and that the number of misassigned reads in the run is determined on the basis of observations of control reads in samples and the negative control. Unique systems with dual indexing (e.g. Illumina applications), or double-end barcoding (e.g. Oxford Nanopore Technologies applications and some Ion Torrent preparations), should be used where feasible, and there should be stringent controls on sample demultiplexing. Demultiplexing should be conducted using stringent settings (for example, depending on the technology, requiring barcodes to be present on both ends of a sequencing read, with few or no mismatches to that barcode).

6.8 Analysis tools

6.8.1 Subsampling data prior to analysis

As of mid-November 2020, 180 000 full genomes with good coverage were publicly available, and the number was rising exponentially. Many of these genomes are likely to be almost identical. If a complete genome of thousands of near-identical sequences is not required, down-sampling strategies can be employed to reduce the computational demands of alignment and subsequent analyses. Down-sampling strategies must be carefully considered, as they can severely affect downstream analyses.

One possible procedure is to run a clustering tool, such as cd-hit-est *(182)*, at a high clustering threshold (> 99% sequence similarity) and to construct an alignment using the representative genomes from this analysis. This is computationally lightweight and auditable, as a clustering report is produced indicating which sequences were selected for each cluster and listing the full cluster membership.

An alternative may be to select clades of interest from a larger previously computed tree. This may be a useful strategy, particularly where a geographical region or other feature is of primary importance to the analysis, and the full global diversity of the viral genomes is less relevant. Nextstrain *(183)* allows clades to be selected from a global tree, and the metadata of sequences in those clades to be subsequently extracted and used to help subsample large available data sets.

For phylogeographical inference in which researchers are interested in capturing virus lineage movements between locations, but not within locations, it may be appropriate to perform subsampling based on phylogenetic criteria. Here, monophyletic clades of sequences from the same location could be subsampled to a single sequence from that clade, as additional sequences within the clade may not add further information of interest regarding inter-location viral lineage movements *(103, 184)*.

6.8.2 Sequence alignments

Alignment of thousands of SARS-CoV-2 genome sequences, many of which include regions of ambiguity due to partially determined genomes, is computationally challenging. Very few existing tools can cope with alignments of this length, and it is worth noting that each time a new sequence is generated, it has the potential to modify the previously determined alignment. It is possible to use alignment software, such as MAFFT, to add a small number of new sequences to a small existing alignment with relatively little computational overhead *(185)*. Alignments of up to several hundred sequences can also be curated with the help of experts, and the authors of MAFFT *(186)* offer this service for SARS-CoV-2 alignments. However, for larger sample sets a different strategy may be required. The shiver pipeline *(187)* produces a version of each assembled genome that is aligned to maintain coordinate placement. In this way, every genome processed can simply be added into a growing alignment without needing to re-align all sequences every time a sequence is added, although care must be taken to ensure that novel insertions are not missed.

It is often appropriate to trim non-coding regions, including the 5' and 3' ends of an alignment, prior to further analyses. It can be challenging to analyse such regions phylogenetically because they incur insertions, deletions and multiple substitutions at the same site more frequently than coding regions that may be under more intense selection.

6.8.3 Quality control

Generated sequences should always be subject to quality control before being used in any analysis. Quality control procedures should be conducted at different stages, to determine multiple features that may be associated with poor quality sequences.

Removing sequences with ambiguous bases, indels or frame-shifts based on unaligned/aligned sequences

The majority of phylogenetic tree-building software tools, including all maximum likelihood methods, are vulnerable to large numbers of ambiguous bases within sequenced genomes. More extensive analyses are needed to evaluate the effect of partial sequences on phylogenies, but removing sequences with > 10% Ns in regions of interest may be appropriate in the first instance.

Sequences with suspected underlying sequencing errors (for example, induced by misassemblies) should be investigated, and usually removed. Sequencing errors can manifest as high divergence compared with other sequences or as high numbers of substitutions in short regions that may indicate local misassemblies. High numbers of non-ACGTN bases may be indicative of mixed viral populations as a result of contamination.

Several useful tools are available to help detect ambiguous bases, indels (insertions or deletions of bases) and frame-shifts, including the Nextclade Quality Control Metric feature within Nextstrain *(183),* CoV-GLUE *(188)* and Pangolin *(189).*

Removing sequences that form long phylogenetic branches

Sequences that form suspiciously long branches on a phylogenetic tree (that suggest unusually high evolutionary divergence) should be curated very carefully. Such branches may reflect real effects, such as large indels or recombination events, but in the case of highly conserved genomes, including SARS-CoV-2, they more commonly indicate a substantial error rate in the underlying sequence or misalignment (Fig. 4).

Removing sequences in which the divergence is substantially greater or less than expected

Suspect sequences can also be identified using a phylogenetic tree and tools such as TempEst *(190)* or TreeTime *(191)*. Specifically, if a sequence is substantially more or less divergent than expected given the time at which it was sampled, it should be carefully checked for potential errors and possibly removed. Sequences that are more or less divergent than expected may arise from bioinformatic problems (e.g. poor variant calling or inappropriate trimming) or metadata misattribution, i.e. an incorrect sampling date. Exactly what constitutes "too divergent" is not formally defined, but clear outliers should be investigated. Manual inspection to identify features that may indicate errors in virus genomic assembly is often useful for smaller data sets. Such features may include insertions, substitutions or deletions that lead to stop codons within

expected coding sequences, or short strings of bases that are highly divergent from all other sequences in the alignment, in particular when neighbouring sites have ambiguous base calls.

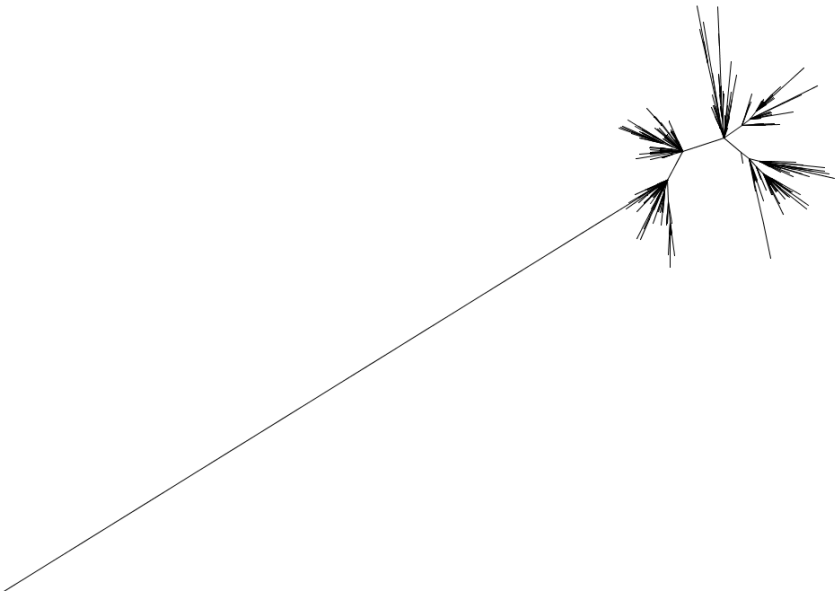

Fig. 4. Spurious long branch shown on an unrooted maximum likelihood phylogeny constructed from an alignment of complete and partial SARS-CoV-2 genomes, where a single genome was misaligned relative to the rest. This is an extreme example. Small misassemblies or short regions of misalignment may result in a terminal branch that is longer than average but not so extreme.

6.8.4 Removing recombinant sequences

While there is no evidence to date of recombination within SARS-CoV-2, coronaviruses are known to recombine, and sequences should be checked for recombinant forms as the pandemic expands. Recombinant viruses cannot be appropriately placed within a phylogenetic tree from a single analysis of the whole genome, as genome sections from each ancestral virus would have different histories and would each therefore be placed at different phylogenetic positions. Inclusion of recombinant sequences can lead to incorrect estimates of evolutionary rate and phylogenetic positioning. If recombinant sequences are detected, they can be removed or several phylogenetic trees can be estimated from the subsections of the alignment that fall either side of recombinant break-points.

Detecting recombination is challenging for many SARS-CoV-2 data sets because existing tools are not designed for use on such extremely large data sets, with thousands of sequences that also have relatively low genetic diversity. Detection of multiple homoplasies (where a substitution has arisen independently in separate phylogenetic lineages) may indicate the possibility of a recombination but should be carefully investigated as homoplasies can also be caused by mutation. The software RDP4 can be used to examine up to 2500 aligned sequences using various tests of recombination *(192)*, although its sensitivity for accurate detection of recombination within SARS-CoV-2 lineages has not yet been determined. Improved or

benchmarked strategies for recombination detection within SARS-CoV-2 data sets would be beneficial.

6.8.5 Phylogenetic tools

With high-quality genome alignment, it is possible to reconstruct the corresponding phylogenetic tree. Neighbour-joining phylogenetic methods are rapid and can be useful for initial exploration of large genetic data sets. However, they only consider a single possible tree and should not be used to make inferences about phylogenetic relatedness. FastTree is also rapid and produces an approximate maximum likelihood phylogenetic estimation, which can be an appropriate alternative to neighbour-joining methods for data exploration *(193)*.

Many maximum likelihood and Bayesian phylogenetic and phylodynamic programs are appropriate for phylogenetic inference. Each requires specification of a model of site evolution. This can be chosen on the basis of information contained within the alignment, using software such as ModelTest-NG *(195)*. Commonly used software for maximum likelihood tree inference includes PhyML *(195)*, RAxML *(196)* and IQ-TREE *(197, 198)*. RAxML is specifically designed for speed of execution where the alignment contains thousands of sequences, while PhyML and IQ-TREE are slower but have been consistently demonstrated to be highly accurate. IQ-TREE has the added functionality of performing a model test first to identify the most appropriate choice of substitution model from the data, and also provides an ultrafast bootstrapping method for estimating branch support. IQ-TREE also performs a complexity check on input data and rejects sequences that contain too many ambiguities or other artefacts that are expected to interfere with phylogeny reconstruction. Branch support statistics (e.g. support from 100 bootstraps, in which 100 trees are re-estimated based on fictional alignments generated from random resampling with replacement of sites in the true alignment sites) should always be calculated to assess the robustness of clustering patterns. Such phylogenetic approaches are useful for investigating evolutionary relatedness, but cannot be used to perform phylodynamic inferences (sections 5.4 and 5.5).

For small data sets (ideally no more than 500–1000 to avoid extremely slow run completion and convergence issues, although the exact number depends on the availability of high performance computing and the dataset in question), it may be possible to use probabilistic methods such as those implemented in BEAST *(199)* or BEAST2 *(200)*. These methods can be used to estimate time of emergence of particular clades of interest (e.g. local outbreaks), the geographical spread of an outbreak, and demographic parameters, including the population size of the virus over time (sections 5.4 and 5.5). For analyses focused exclusively on estimating the time since divergence for a group of viral genomes, especially when these data sets are large, it may be sufficient and more computationally tractable to use less complex methods that combine sampling dates with pre-computed maximum likelihood trees, such as the least-squares dating (LSD) *(201)* or TreeTime *(191)* software programs. All these methods require a sufficient "temporal signal" within the data set, such that virus lineages can be seen to evolve in a clock-like manner with substitutions occurring at a relatively predictable rate. Exactly how to draw the line between insufficient and sufficient temporal signalling with respect to SARS-CoV-2 was the focus of much of the early phylodynamic work *(45)*. There have now been multiple examples of time-scaled phylogenetic and phylodynamic analyses of SARS-CoV-2 *(59, 85)*. While the phylodynamic threshold (the point in time at which sufficient molecular evolutionary change has

accumulated in available genome samples to obtain robust phylodynamic estimates) has been reached for some analyses, subsets of the available sequence data corresponding to local clusters within specific geographical areas should be treated with care and reassessed prior to use, in order to determine the applicability of phylodynamic methods.

While network-based methods (e.g. haplotype joining methods, median-joining networks) are rapid and simple to perform and are present in the published SARS-CoV-2 literature, networks lack appropriate phylogenetic rooting that is important for the understanding of evolutionary histories. They also lack an appropriate model of site evolution, being based instead on similarity of genome sequences alone, and do not assess or capture the robustness of displayed connectivity patterns. Construction of a phylogenetic tree will therefore usually be as appropriate, or more appropriate, than construction of a network to analyse SARS-CoV-2 viral genome sequences *(202)*.

6.8.6 Visualization

Phylogenetic trees can be visualized locally using a wide variety of freely available (e.g. FigTree and MEGA *(203)*) and commercial software.

The web application Microreact provides an interactive display of a user-entered phylogenetic tree, allowing phylogenetic structuring by location (longitude and latitude), category (e.g. country) and time to be visualized *(204)*. Mapping of phylogenetic tip locations relative to tree position can be useful for exploring geographical structuring of SARS-CoV-2 diversity, and for rapid confirmation, where relevant, that any data has been geocoded properly. Microreact requires an input file containing metadata, such as sampling date and location, and a phylogenetic tree. Uploaded projects can be shared publicly or kept private, and updated by the user as required. Currently available publicly accessible projects include a global distribution of SARS-CoV-2 lineages that is being updated by the COVID-19 Genomics UK Consortium. Phylogenetic tree files are not shown with branch support statistics, and therefore trees from publicly available data sets have to be downloaded for local inspection if required. Additional information on the methods used to construct phylogenies are provided at the project author's discretion; this is useful to allow adequate consideration of these phylogenies.

Nextstrain *(183)* provides an interactive display of the evolution and geographical diversity of SARS-CoV-2 and other pathogens. Contributors and developers curate global and regional online phylogenetic visualizations that have been frequently accessed during the COVID-19 pandemic. Users can set up their own local Augur phylogeny and map visualizations to analyse data based on input files of sequences, phylogenies and metadata. Nextstrain is a powerful and rapid tool for exploring broad-scale patterns of geographical structuring. Nevertheless, any phylogeny should always be interpreted with caution, taking into consideration the confidence intervals in divergence dates provided and the uncertainty in the geographical location displayed, and within the context of explanatory "Nextstrain narratives" where available. Phylogenetic trees are not displayed with branch support statistics, so the branching order shown should not be assumed to be exact or used to inform policy decisions without further investigation to confirm the finding. The geographical locations and timing of divergence of phylogenetic branches are inferred using less complex but more rapid methods than those commonly employed in BEAST or BEAST2. Analyses comparing the extent of agreement between the different methods for

SARS-CoV-2 would be valuable: disparities between different methods are not uncommon *(205)*.

As described in section 5, non-random sampling of sequences can bias phylogenetic and phylogeographical interpretations and conclusions. It is important to be mindful of these possible biases when interpreting any phylogenetic visualizations.

6.8.7 Lineage classification

There is currently no universally accepted formal naming system for SARS-CoV-2 evolutionary lineages. Several proposed nomenclatures use the same names (e.g. "A1") to refer to different lineages, and it is therefore important to state which nomenclature is being used in any description. Global adoption of a single nomenclature system would facilitate scientific communication about specific lineages and avoid the confusion generated by the use of multiple systems.

There are currently three commonly used nomenclature systems for SARS-CoV-2 clades/lineages. Both GISAID EpiCoV™ and Nextstrain aim to provide a broad categorization of globally circulating diversity through naming of different phylogenetic clades. Rambaut et al. *(189)* proposed a dynamic nomenclature for SARS-CoV-2 lineages that focuses on actively circulating virus lineages, and those that spread to new locations. Software to allow users to assign their own sequences to these lineages is available, including via Pangolin, Nextstrain and CoV-Glue *(183, 188, 189)*.

Given that there is currently no universally accepted nomenclature, the best approach when reporting lineages is to state the nomenclature of particular clades in all three of the commonly used systems, or at least to state explicitly which nomenclature is being used.

6.8.8 Phylogenetic rooting

Regardless of the phylogenetic software and method used, the choice of one or more outgroups is important, and will have an effect on how the root of the tree is determined. This in turn will affect estimates of time since divergence. An outgroup is a sequence selected to be as closely related as possible to the sequences of interest but known not to be part of the same clade. In practice, the earliest available SARS-CoV-2 reference sequence is often used as an outgroup when constructing a phylogeny of genomes from a variety of geographical sources. For investigation of local clusters, it may be appropriate to choose a more closely related genome from outside the data set to be analysed.

7. Conclusions and future needs

Rapid sequencing of virus genomes is now achievable in varied settings, and analyses of SARS-CoV-2 genomic sequences have a huge potential for informing public health efforts surrounding COVID-19. The rapid generation and global sharing of virus genomic sequences provides information that will contribute to the understanding of transmission and the design of clinical and epidemiological mitigation strategies.

Dialogue between public health bodies, data generators and analysts is critical to ensure that the data are generated and used appropriately for maximum public health benefit. Careful prior consideration of why sequencing is being conducted is required, as this will affect the choice of samples, the collation of metadata and subsequent analyses. Sequencing should be conducted with due consideration of available resources and capacities, and should not draw capacity away from other equally vital areas. Clear communication channels should be established for sharing results, samples and data with appropriate stakeholders, so that information can be used to improve public health as rapidly as possible.

Translating SARS-CoV-2 genome sequences into informative results is complex, and often requires substantial specialist training to ensure that violations of model assumptions do not lead to incorrect understanding of virus epidemiology. A clear understanding of the benefits and limitations of genomic analyses will allow a confident assessment of where genomic tools can augment or support existing approaches and where, conversely, epidemiological modelling or laboratory experimentation may be more robust. Partnership between experts with different skill sets is valuable, as not all laboratories will have existing local expertise in all areas.
Despite recent advances in the ease with which virus sequences can be generated, challenges remain. In many settings, the need for rapid importation of temperature-sensitive reagents was a significant barrier to the adoption of within-country portable sequencing approaches early during the during COVID-19. Solutions must be found if countries are to develop their capacity to conduct sequencing activities in future public health emergencies as well as during the current pandemic. Funding that supports activities to validate and compare different published sequencing and analysis strategies would also be beneficial to ensure appropriate informed selection.

The analysis and interpretation of virus genomic sequence data are not straightforward. Laboratories planning to adopt sequencing for the first time could benefit from programmes that provide support for the formal validation of their sequencing pipelines. The global genomic data sets generated for SARS-CoV-2 are too large for many current tools; improvements are needed to allow increasingly large data sets to be analysed rapidly during public health emergencies and, where possible, to increase the level of automation. A better academic understanding of what public health agencies need and of how results can best be presented to emphasize the practical implications while nevertheless taking into account analytical uncertainty would also be beneficial.

Public health laboratories generally have more expertise in molecular genetics than in computational phylogenetics and bioinformatics. Strengthened, long-term investment in

phylogenetics and bioinformatics training is necessary to obtain the maximum benefit from the growth in laboratory sequencing possibilities in this and subsequent epidemics.

Repositories such as GISAID have encouraged and facilitated the sharing of data on COVID-19. However, broader discussions are still needed to ensure continued improvements in data-sharing during public health emergencies. Currently, many researchers remain reluctant to share genomic sequence data until a pre-print publication has been prepared. The reasons for this should be sought and solutions proposed. More extensive discussion and agreement on appropriate accreditation for data producers in different circumstances is also necessary to encourage data sharing. There is a need to develop new data accreditation standards or metrics and for journals to commit to uphold fair data-use practices.

More extensive public engagement by scientists is important to reduce the spread of false information during the current and future public health emergencies. Increased support and training for scientists in how scientific messages can be effectively shared with the general public would be beneficial. Ensuring that patients and the public understand the value and limitations of virus genomic sequence data is essential to underpin public consultations on the appropriate use of patient metadata during public health emergencies.

References

1. Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J et al. Next-generation sequencing informatics: challenges and strategies for implementation in a clinical environment. Arch Pathol Lab Med. 2016;140:958-75. doi: 10.5858/arpa.2015-0507-RA.
2. Gu W, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. Annu Rev Pathol. 2019;14:319-38. doi: 10.1146/annurev-pathmechdis-012418-012751.
3. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nat Rev Microbiol. 2017;15:183-92. doi: 10.1038/nrmicro.2016.182.
4. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228-32. doi: 10.1038/nature16996.
5. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W et al. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet. 2003;361:1319-25. doi: 10.1016/s0140-6736(03)13077-2.
6. Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Eng J Med. 2003;348:1967-76. doi: 10.1056/NEJMoa030747.
7. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S et al. A novel coronavirus associated with severe acute respiratory syndrome. N Eng J Med. 2003;348:1953-66. doi: 10.1056/NEJMoa030781.
8. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J et al. A novel coronavirus from patients with pneumonia in China, 2019. N Eng J Med. 2020;382:727-33. doi: 10.1056/NEJMoa2001017.
9. World Health Organization. Novel coronavirus (2019-nCoV): Situation report 1. Geneva; 2020 (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4, accessed 2 November 2020).
10. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD et al. Pandemic potential of a strain of influenza A(H1N1): early findings. Science. 2009;324:1557-61. doi: 10.1126/science.1176062.
11. Rambaut A, Holmes E. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Curr. 2009;1:RRN1003. doi: 10.1371/currents.rrn1003.
12. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459:1122-5. doi: 10.1038/nature08182.
13. Mena I, Nelson MI, Quezada-Monroy F, Dutta J, Cortes-Fernández R, Lara-Puente JH et al. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. eLife. 2016;5:e16777. doi: 10.7554/eLife.16777.
14. WHO MERS-CoV Research Group. State of knowledge and data gaps of Middle East respiratory syndrome coronavirus (MERS-CoV) in humans. PLoS Curr. 2013;5. doi: 10.1371/currents.outbreaks.0bf719e352e7478f8ad85fa30127ddb8.
15. Haagmans BL, Al Dhahiry SHS, Reusken CBEM, Raj VS, Galiano M, Myers R et al. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. Lancet Infect Dis. 2014;14:140-5. doi: 10.1016/S1473-3099(13)70690-X.

16. Sabir JSM, Lam TTY, Ahmed MMM, Li L, Shen Y, Abo-Aba SEM et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science. 2016;351:81-4. doi: 10.1126/science.aac8608.
17. Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM et al. Evidence for camel-to-human transmission of MERS coronavirus. N Eng J Med. 2014;370:2499-505. doi: 10.1056/NEJMoa1401505.
18. Memish ZA, Cotten M, Meyer B, Watson SJ, Alsahafi AJ, Al Rabeeah AA et al. Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013. Emerg Infect Dis. 2014;20:1012-5. doi: 10.3201/eid2006.140402.
19. Chu DKW, Hui KPY, Perera RAPM, Miguel E, Niemeyer D, Zhao J et al. MERS coronaviruses from camels in Africa exhibit region-dependent genetic diversity. Proc Natl Acad Sci USA. 2018;115:3144-9. doi: 10.1073/pnas.1718769115.
20. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. eLife. 2018;7:e31257. doi: 10.7554/eLife.31257.
21. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba NF et al. Emergence of Zaire Ebola virus disease in Guinea. N Eng J Med. 2014;371:1418-25. doi: 10.1056/NEJMoa1404505.
22. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369-72. doi: 10.1126/science.1259657.
23. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature. 2015;524:97-101. doi: 10.1038/nature14594.
24. Dudas G, Rambaut A. Phylogenetic analysis of Guinea 2014 ebov ebolavirus outbreak. PLoS Curr. 2014;6. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.
25. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell. 2015;161:1516-26. doi: 10.1016/j.cell.2015.06.007.
26. Simon-Loriere E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. Nature. 2015;524:102-4. doi: 10.1038/nature14612.
27. Tong Y-G, Shi W-F, Liu D, Qian J, Liang L, Bo X-C et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015;524:93-6. doi: 10.1038/nature14490.
28. Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S et al. Evolution and spread of Ebola virus in Liberia, 2014-2015. Cell Host Microbe. 2015;18:659-69. doi: 10.1016/j.chom.2015.11.008.
29. Volz E, Pond S. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. PLOS Curr. 2014;24 doi:10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e.
30. Stadler T, Kühnert D, Rasmussen DA, Plessis DL. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. PLOS Curr. 2014. doi: 10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.

31. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T et al. Molecular evidence of sexual transmission of Ebola virus. N Eng J Med. 2015;373:2448-54. doi: 10.1056/NEJMoa1509773.
32. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zoology. 1978;27:401-10. doi: 10.2307/2412923.
33. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. Nature. 2016;538:193-200. doi: 10.1038/nature19790.
34. Arias A, Watson SJ, Asogun D, Tobin EA, Lu J, Phan MVT et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. Vir Evol. 2016;2:vew016. doi: 10.1093/ve/vew016.
35. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD et al. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. Emerg Infect Dis. 2016;22:331-4. doi: 10.3201/eid2202.151796.
36. Smits SL, Pas SD, Reusken CB, Haagmans BL, Pertile P, Cancedda C et al. Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015. Euro Surveill. 2015;20. doi: 10.2807/1560-7917.ES.2015.20.40.30035.
38. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406-10. doi: 10.1038/nature22401.
37. Faria NR, Azevedo RdSdS, Kraemer MUG, Souza R, Cunha MS, Hill SC et al. Zika virus in the Americas: early epidemiological and genetic findings. Science. 2016;352:345-9. doi: 10.1126/science.aaf5036.
39. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM et al. Zika virus evolution and spread in the Americas. Nature. 2017;546:411-5. doi: 10.1038/nature22402.
40. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K et al. Genomic epidemiology reveals multiple introductions of zika virus into the United States. Nature. 2017;546:401-5. doi: 10.1038/nature22400.
41. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol. 2019;4:10. doi: 10.1038/s41564-018-0296-2.
42. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19:9-20. doi: 10.1038/nrg.2017.88.
43. Rasmussen AL, Katze MG. Genomic signatures of emerging viruses: a new era of systems epidemiology. Cell Host Microbe. 2016;19:611-8. doi: 10.1016/j.chom.2016.04.016.
44. Loewe L, Hill WG. The population genetics of mutations: Good, bad and indifferent. Philos Trans R Soc Lond B Biol Sci. 2010;365:1153-67. doi: 10.1098/rstb.2009.0317.
45. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. 2020: 19;6(2). doi:10.1093/ve/veaa061.
46. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9:267-76. doi: 10.1038/nrg2323.
47. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303:327-32. doi: 10.1126/science.1090727.

48. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9. doi: 10.1371/journal.pcbi.1002947.
49. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10:540-50. doi: 10.1038/nrg2583.
50. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci. 2016;73:4433-48. doi: 10.1007/s00018-016-2299-6.
51. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus : classifying 2019-nCov and naming it SARS-CoV-2. Nat Microbiol. 2020;5:536-44. doi: 10.1038/s41564-020-0695-z.
52. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579:265-9. doi: 10.1038/s41586-020-2008-3.
53. Candido DDS, Claro IM, Jesus DJG, Souza DWM, Moreira FRR, Dellicour S et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science (New York, NY). 2020;369:1255-60. doi: 10.1101/2020.06.11.20128249.
54. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000; 283(20): 2701-11.
55. WHO guidelines on ethical issues in public health surveillance. Geneva: World Health Organization; 2017 (https://www.who.int/ethics/publications/public-health-surveillance/en/, accessed 15 November 2020).
56. World Health Organization. Policy statement on data sharing by the World Health Organization in the context of public health emergencies. Geneva; 2016.
57. Thézé J, Li T, Plessis dL, Bouquet J, Kraemer MUG, Somasekar S et al. Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico. Cell Host Microbe. 2018;23:855-64.e7. doi: 10.1016/j.chom.2018.04.017.
58. COVID-19 data portal. 2020 (https://www.covid19dataportal.org/sequences, accessed 1 November 2020).
59. Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell. 2020;181:997-1003.e9. doi: 10.1016/j.cell.2020.04.023.
60. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270-3. doi: 10.1038/s41586-020-2012-7.
61. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382:1199-207. doi: 10.1056/NEJMoa2001316.
62. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CovV2. Nature Medicine. 2020;26:450-2. doi: 10.1038/s41591-020-0820-9.
63. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Geneva: World Health Organization; 2020.(https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19), accessed 28 December 2020)
64. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17:181-92. doi: 10.1038/s41579-018-0118-9.

65. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017;13:e1006698. doi: 10.1371/journal.ppat.1006698.
66. Lin X-D, Wang W, Hao Z-Y, Wang Z-X, Guo W-P, Guan X-Q et al. Extensive diversity of coronaviruses in bats from China. Virology. 2017;507:1-10. doi: 10.1016/j.virol.2017.03.019.
67. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry B, Castoe T et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature Microbiology. 2020;5:1408-17. doi: 10.1101/2020.03.30.015008.
68. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y et al. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. Curr Biol. 2020;30:2196-203.e3. doi: 10.1016/j.cub.2020.05.023.
69. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020:1-4. doi: 10.1038/s41586-020-2169-0.
70. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020;181:271-80.e8. doi: 10.1016/j.cell.2020.02.052.
71. 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;5:562-9. doi: 10.1038/s41564-020-0688-y.
72. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O et al. Cryo-em structure of the 2019-nCoV spike in the prefusion conformation. Science (New York, Ny). 2020;367:1260-3. doi: 10.1126/science.abb2507.
73. Colson P, Scola BL, Esteves-Vieira V, Ninove L, Zandotti C, Jimeno M-T et al. Plenty of coronaviruses but no SARS-CoV-2. (Letter to the editor). Euro Surveill. 2020;25:2000171. doi: 10.2807/1560-7917.ES.2020.25.8.2000171.
74. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020;25. doi: 10.2807/1560-7917.ES.2020.25.3.2000045.
75. Information for laboratories about coronavirus (COVID-19). Atlanta: Centers for Disease Control and Prevention; 2020. (https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html, accessed 26 June 2020).
76. University of Hong Kong, School of Public Health. Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR. 2020 (https://www.who.int/docs/default-source/coronaviruse/peiris-protocol-16-1-20.pdf?sfvrsn=af1aac73_4.
77. Diagnostic testing for SARS-CoV-2. Interim guidance. Geneva: World Health Organization; 2020 (https://www.who.int/publications/i/item/diagnostic-testing-for-sars-cov-2, accessed 19 November 2020).
78. Ren Y, Zhou Z, Liu J, Lin L, Li S, Wang H et al. A strategy for searching antigenic regions in the SARS-CoV spike protein. Genomics Proteomics Bioinformatics. 2003;1:207-15. doi: 10.1016/s1672-0229(03)01026-x.
79. Kumar S, Maurya VK, Prasad AK, Bhatt MLB, Saxena SK. Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV). Virusdisease. 2020:1-9. doi: 10.1007/s13337-020-00571-5.

80. Melén K, Kakkola L, He F, Airenne K, Vapalahti O, Karlberg H et al. Production, purification and immunogenicity of recombinant Ebola virus proteins - a comparison of Freund's adjuvant and adjuvant system 03. J Virol Methods. 2017;242:35-45. doi: 10.1016/j.jviromet.2016.12.014.
81. Ziegler T, Matikainen S, Rönkkö E, Österlund P, Sillanpää M, Sirén J et al. Severe acute respiratory syndrome coronavirus fails to activate cytokine-mediated innate immune responses in cultured human monocyte-derived dendritic cells. J Virol. 2005;79:13800-5. doi: 10.1128/JVI.79.21.13800-13805.2005.
82. Draft landscape of COVID-19 candidate vaccines. Geneva: World Health Organization; 2020 (https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines, accessed 26 June 2020).
83. Li G, Clercq ED. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat Rev Drug Discov. 2020;19:149-50. doi: 10.1038/d41573-020-00016-0.
84. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L et al. Cryptic transmission of SARS-CoV-2 in Washington State. Science. 2020;370:571-5. doi: 10.1101/2020.04.02.20051417.
85. Volz E, Fu H, Wang H, Xi X, Chen W, Liu D et al. Genomic epidemiology of a densely sampled COVID19 outbreak in China. medRxiv. 2020:2020.03.09.20033365. doi: 10.1101/2020.03.09.20033365.
86. Zehender G, Lai A, Bergna A, Meroni L, Riva A, Balotta C et al. Genomic characterization and phylogenetic analysis of SARS-CoV-2 in Italy. J Med Virol. 2020;92(9):1637-1640. doi: 10.1002/jmv.25794.
87. Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB et al. The emergence of SARS-CoV-2 in Europe and North America. Science. 2020;370:564-70.
88. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5:e1000520. doi: 10.1371/journal.pcbi.1000520.
89. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol. 2010;27:1877-85. doi: 10.1093/molbev/msq067.
90. Bloomquist EW, Lemey P, Suchard MA. Three roads diverged? Routes to phylogeographic inference. Trends Ecol Evol. 2010;25:626-32. doi: 10.1016/j.tree.2010.08.010.
91. Faria NR, Suchard MA, Rambaut A, Lemey P. Towards a quantitative understanding of viral phylogeography. Curr OpinVirol. 2011;1:423-9. doi: 10.1016/j.coviro.2011.10.003.
92. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG et al. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell. 2020;181:990-6.e5. doi: 10.1016/j.cell.2020.04.021.
93. Eden J-S, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. Virus Evol. 2020;6. doi: 10.1093/ve/veaa027.
94. Lemey P, Hong S, Hill V, Baele G, Poletto C, Colizza V et al. Accommodating individual travel history, global mobility, and unsampled diversity in phylogeography: a SARS-CoV-2 case study. bioRxiv. 2020:2020.06.22.165464. doi: 10.1101/2020.06.22.165464.

95. Ewing G, Rodrigo A. Estimating population parameters using the structured serial coalescent with Bayesian mcmc inference when some demes are hidden. Evol Bioinform Online. 2006;2:117693430600200026. doi: 10.1177/117693430600200026.
96. Maio ND, Wu C-H, O'Reilly KM, Wilson D. New routes to phylogeography: A Bayesian structured coalescent approximation. PLoS Genet. 2015;11:e1005421. doi: 10.1371/journal.pgen.1005421.
97. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10:e1003932. doi: 10.1371/journal.ppat.1003932.
98. Chaillon A, Gianella S, Dellicour S, Rawlings SA, Schlub TE, Oliveira MFD et al. HIV persists throughout deep tissues with repopulation from multiple anatomical sources. The J Clin Invest. 2020;130:1699-712. doi: 10.1172/JCI134815.
99. Kalkauskas A, Perron U, Sun Y, Goldman N, Baele G, Guindon S et al. Sampling bias and model choice in continuous phylogeography: getting lost on a random walk. bioRxiv. 2020:2020.02.18.954057. doi: 10.1101/2020.02.18.954057.
100. Nylinder S, Lemey P, De Bruyn M, Suchard MA, Pfeil BE, Walsh N et al. On the biogeography of centipeda: a species-tree diffusion approach. Syst Biol. 2014;63:178-91. doi: 10.1093/sysbio/syt102.
101. Dellicour S, Lemey P, Artois J, Lam TT, Fusaro A, Monne I et al. Incorporating heterogeneous sampling probabilities in continuous phylogeographic inference — application to H5N1 spread in the Mekong region. Bioinformatics. 2020;36:2098-104. doi: 10.1093/bioinformatics/btz882.
102. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544:309-15. doi: 10.1038/nature22040.
103. Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA et al. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. Nature Communications. 2018;9:1-9. doi: 10.1038/s41467-018-03763-2.
104. Bielejec F, Lemey P, Baele G, Rambaut A, Suchard MA. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. Syst Biol. 2014;63:493-504. doi: 10.1093/sysbio/syu015.
105. Sit THC, Brackman CJ, Ip SM, Tam KWS, Law PYT, To EMW et al. Infection of dogs with SARS-CoV-2. Nature. 2020. doi: 10.1038/s41586-020-2334-5.
106. Oreshkova N, Molenaar RJ, Vreman S, Harders F, Munnink BBO, Honing RWH et al. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. Euro Surveill. 2020;25:2001005. doi: 10.2807/1560-7917.ES.2020.25.23.2001005.
107. Segalés J, Puig M, Rodon J, Avila-Nieto C, Carrillo J, Cantero G et al. Detection of SARS-CoV-2 in a cat owned by a COVID-19-affected patient in Spain. PNAS. 2020;117(40):24790-24793. doi: 10.1073/pnas.2010817117
108. Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D et al. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS Pathog. 2012;8. doi: 10.1371/journal.ppat.1003081.

109. Worby CJ, Lipsitch M, Hanage WP. Shared genomic variants: identification of transmission routes using pathogen deep-sequence data. Am J Epidemiol. 2017;186:1209-16. doi: 10.1093/aje/kwx182.
110. Cotten M, Lam TT, Watson SJ, Palser AL, Petrova V, Grant P et al. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. Emerg Infect Dis. 2013;19:736-42B. doi: 10.3201/eid1905.130057.
111. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L et al. Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. Clin Infect Dis. 2020; 71(15):713-720 doi: 10.1093/cid/ciaa203.
112. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using primalseq and ivar. Genome Biol. 2019;20:8. doi: 10.1186/s13059-018-1618-7.
113. Volz E, Baguelin M, Bhatia S, Boonyasiri A, Cori A, Cucunubá Z et al. Report 5 - phylogenetic analysis of SARS-CoV-2. London: Imperial College; 2020 (http://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-5-phylogenetics-of-sars-cov-2/, accessed 26 June 2020).
114. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci USA. 2013;110:228-33. doi: 10.1073/pnas.1207965110.
115. Boskova V, Bonhoeffer S, Stadler T. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. PLoS Comput Biol. 2014;10:e1003913. doi: 10.1371/journal.pcbi.1003913.
116. Volz EM, Frost SDW. Sampling through time and phylodynamic inference with coalescent and birth–death models. J R Soc Interface. 2014;11. doi: 10.1098/rsif.2014.0945.
117. Li LM, Grassly NC, Fraser C. Quantifying transmission heterogeneity using both pathogen phylogenies and incidence time series. Mol Biol Evol. 2017;34:2982-95. doi: 10.1093/molbev/msx195.
118. Koelle K, Rasmussen DA. Rates of coalescence for common epidemiological models at equilibrium. J R Soc Interface. 2012;9:997-1007. doi: 10.1098/rsif.2011.0495.
119. Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T. Estimating epidemic incidence and prevalence from genomic data. Mol Biol Evol. 2019;36:1804-16. doi: 10.1093/molbev/msz106.
120. Volz EM, Siveroni I. Bayesian phylodynamic inference with complex models. PLoS Comput Biol. 2018;14:e1006546. doi: 10.1371/journal.pcbi.1006546.
121. Laboratory biosafety guidance related to coronavirus disease (COVID-19). Geneva: World Health Organization; 2020 (https://apps.who.int/iris/handle/10665/332076, accessed 21 November 2020).
122. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G et al. Detection of SARS-CoV-2 in different types of clinical specimens. JAMA. 2020;323:1843-1844. doi: 10.1001/jama.2020.3786.
123. Chen W, Lan Y, Yuan X, Deng X, Li Y, Cai X et al. Detectable 2019-nCoV viral RNA in blood is a strong indicator for the further clinical severity. Emerg Microbes Infect. 2020;9:469-73. doi: 10.1080/22221751.2020.1732837.

124. Chen X, Zhao B, Qu Y, Chen Y, Xiong J, Feng Y et al. Detectable serum SARS-CoV-2 viral load (RNAaemia) is closely correlated with drastically elevated interleukin 6 (il-6) level in critically ill COVID-19 patients. Clin Infect Dis. 2020; 71(8):1937-1942. doi: 10.1093/cid/ciaa449.
125. Corman VM, Rabenau HF, Adams O, Oberle D, Funk MB, Keller-Stanislawski B et al. SARS-CoV-2 asymptomatic and symptomatic patients and risk for transfusion transmission. Transfusion. 2020; 60(6):1119-1122 doi: 10.1111/trf.15841.
126. Zhang W, Du RH, Li B, Zheng XS, Yang XL, Hu B et al. Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes. Emerg Microbes Infect. 2020;9:386-9. doi: 10.1080/22221751.2020.1729071.
127. Winichakoon P, Chaiwarith R, Liwsrisakun C, Salee P, Goonna A, Limsukon A et al. Negative nasopharyngeal and oropharyngeal swabs do not rule out COVID-19. J Clin Microbiol. 2020;58. doi: 10.1128/JCM.00297-20.
128. Ek P, Bottiger B, Dahlman D, Hansen KB, Nyman M, Nilsson AC. A combination of naso- and oropharyngeal swabs improves the diagnostic yield of respiratory viruses in adult emergency department patients. Infect Dis (Lond). 2019;51:241-8. doi: 10.1080/23744235.2018.1546055.
129. Hammitt LL, Kazungu S, Welch S, Bett A, Onyango CO, Gunson RN et al. Added value of an oropharyngeal swab in detection of viruses in children hospitalized with lower respiratory tract infection. J Clin Microbiol. 2011;49:2318-20. doi: 10.1128/JCM.02605-10.
130. The 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;26:861-868. doi: 10.1038/s41591-020-0877-5.
131. Sutjipto HL, Yant TJ, Mendis SM, Abdad MY, Marimuthu K, Ng OT et al. The effect of sample site, illness duration and the presence of pneumonia on the detection of SARS-CoV-2 by real-time reverse-transcription pcr. Open Forum Infect Dis. 2020; 7(9):ofaa335. doi: 10.1093/ofid/ofaa335.
132. Zou L, Ruan F, Huang M, Liang L, Huang H, Hong Z et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. N Engl J Med. 2020;382:1177-9. doi: 10.1056/NEJMc2001737.
133. Lai CKC, Chen Z, Lui G, Ling L, Li T, Wong MCS et al. Prospective study comparing deep-throat saliva with other respiratory tract specimens in the diagnosis of novel coronavirus disease (COVID-19). J Infect Dis. 2020; 222(10):1612-1619. doi: 10.1093/infdis/jiaa487.
134. Liu R, Han H, Liu F, Lv Z, Wu K, Liu Y et al. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. Clin Chim Acta. 2020;505:172-5. doi: 10.1016/j.cca.2020.03.009.
135. Huang Y, Chen S, Yang Z, Guan W, Liu D, Lin Z et al. SARS-CoV-2 viral load in clinical samples from critically ill patients. Am J Respir Crit Care Med. 2020;201:1435-8. doi: 10.1164/rccm.202003-0572LE.
136. Williams E, Bond K, Zhang B, Putland M, Williamson DA. Saliva as a non-invasive specimen for detection of SARS-CoV-2. J Clin Microbiol. 2020; 24(5):422-427. doi: 10.1128/JCM.00776-20.

137. Pasomsub E, Watcharananan SP, Boonyawat K, Janchompoo P, Wongtabtim G, Suksuwan W et al. Saliva sample as a non-invasive specimen for the diagnosis of coronavirus disease-2019 (COVID-19): a cross-sectional study. Clin Microbiol Infect. 2020. doi: 10.1016/j.cmi.2020.05.001.
138. Yang JR, Deng DT, Wu N, Yang B, Li HJ, Pan XB. Persistent viral RNA positivity during the recovery period of a patient with SARS-CoV-2 infection. J Med Virol. 2020; 92(9):1681-1683. doi: 10.1002/jmv.25940.
139. Guo WL, Jiang Q, Ye F, Li SQ, Hong C, Chen LY et al. Effect of throat washings on detection of 2019 novel coronavirus. Clin Infect Dis. 2020; 71(8):1980-1981. doi: 10.1093/cid/ciaa416.
140. To KK, Tsang OT, Leung WS, Tam AR, Wu TC, Lung DC et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. Lancet Infect Dis. 2020;20:565-74. doi: 10.1016/S1473-3099(20)30196-1.
141. Azzi L, Carcano G, Gianfagna F, Grossi P, Gasperina D, Genoni A et al. Saliva is a reliable tool to detect SARS-CoV-2. J Infect. 2020;81. doi: 10.1016/j.jinf.2020.04.005.
142. McCormick-Baw C, Morgan K, Gaffney D, Cazares Y, Jaworski K, Byrd A et al. Saliva as an alternate specimen source for detection of SARS.CoV-2 in symptomatic patients using cepheid xpert xpress SARS-CoV-2. J Clin Microbiol. 2020. doi: 10.1128/JCM.01109-20.
143. Wyllie AL, Fournier J, Casanovas-Massana A, Campbell M, Tokuyama M, Vijayakumar P et al. Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2. 383(13):1283-1286. N Engl J Med. 2020. doi: 10.1056/NEJMc2016359.
144. Lescure FX, Bouadma L, Nguyen D, Parisey M, Wicky PH, Behillil S et al. Clinical and virological data of the first cases of COVID-19 in Europe: a case series. Lancet Infect Dis. 2020; 20(6):697-706. doi: 10.1016/S1473-3099(20)30200-0.
145. Xing YH, Ni W, Wu Q, Li WJ, Li GJ, Wang WD et al. Prolonged viral shedding in feces of pediatric patients with coronavirus disease 2019. J Microbiol Immunol Infect. 2020; 53(3):473-480. doi: 10.1016/j.jmii.2020.03.021.
146. Zheng S, Fan J, Yu F, Feng B, Lou B, Zou Q et al. Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January-March 2020: retrospective cohort study. BMJ. 2020;369:1443. doi: 10.1136/bmj.m1443.
147. Wong MC, Huang J, Lai C, Ng R, Chan FKL, Chan PKS. Detection of SARS-CoV-2 RNA in fecal specimens of patients with confirmed COVID-19: a meta-analysis. J Infect. 2020;81:e31-e8. doi: 10.1016/j.jinf.2020.06.012.
148. Tang JW, To KF, Lo AW, Sung JJ, Ng HK, Chan PK. Quantitative temporal-spatial distribution of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) in post-mortem tissues. J Med Virol. 2007;79:1245-53. doi: 10.1002/jmv.20873.
149. Nicholls JM, Poon LL, Lee KC, Ng WF, Lai ST, Leung CY et al. Lung pathology of fatal severe acute respiratory syndrome. Lancet. 2003;361:1773-8. doi: 10.1016/s0140-6736(03)13413-7.
150. Pomara C, Li Volti G, Cappello F. COVID-19 deaths: are we sure it is pneumonia? Please, autopsy, autopsy, autopsy! J Clin Med. 2020;9. doi: 10.3390/jcm9051259.
151. Salerno M, Sessa F, Piscopo A, Montana A, Torrisi M, Patane F et al. No autopsies on COVID-19 deaths: a missed opportunity and the lockdown of science. J Clin Med. 2020;9. doi: 10.3390/jcm9051472.

152. Hanley B, Lucas SB, Youd E, Swift B, Osborn M. Autopsy in suspected COVID-19 cases. J Clin Pathol. 2020;73:239-42. doi: 10.1136/jclinpath-2020-206522.
153. Basso C, Calabrese F, Sbaraglia M, Del Vecchio C, Carretta G, Saieva A et al. Feasibility of postmortem examination in the era of COVID-19 pandemic: the experience of a northeast Italy university hospital. Virchows Arch. 2020 477(3):341-347. doi: 10.1007/s00428-020-02861-1.
154. Tian S, Xiong Y, Liu H, Niu L, Guo J, Liao M et al. Pathological study of the 2019 novel coronavirus disease (COVID-19) through postmortem core biopsies. Mod Pathol. 2020;33:1007-14. doi: 10.1038/s41379-020-0536-x.
155. Sekulic M, Harper H, Nezami BG, Shen DL, Sekulic SP, Koeth AT et al. Molecular detection of SARS-CoV-2 infection in FFPE samples and histopathologic findings in fatal SARS-CoV-2 cases. Am J Clin Pathol. 2020; 154(2):190-200. doi: 10.1093/ajcp/aqaa091.
156. Park WB, Kwon NJ, Choi SJ, Kang CK, Choe PG, Kim JY et al. Virus isolation from the first patient with SARS-CoV-2 in Korea. J Korean Med Sci. 2020;35:e84. doi: 10.3346/jkms.2020.35.e84.
157. Le TQM, Takemura T, Moi ML, Nabeshima T, Nguyen LKH, Hoang VMP et al. Severe acute respiratory syndrome coronavirus 2 shedding by travelers, Vietnam, 2020. Emerg Infect Dis . 2020;26:1624-6. doi: 10.3201/eid2607.200591.
158. Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect Dis. 2020;20:411-2. doi: 10.1016/S1473-3099(20)30113-4.
159. Wyllie AL, Fournier J, Casanovas-Massana A, Campbell M, Tokuyama M, Vijayakumar P et al. Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2. N Eng J Med. 2020;383:1283-6. doi: 10.1101/2020.04.16.20067835.
160. Wolfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Muller MA et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020; 581(7809):465-469. doi: 10.1038/s41586-020-2196-x.
161. MacCannell D. SARS-CoV-2 sequencing. 2020 (https://github.com/CDCgov/SARS-CoV-2_Sequencing, accessed 1 November 2020).
162. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35:833-44. doi: 10.1038/nbt.3935.
163. Bragg L, Tyson GW. Metagenomics using next-generation sequencing. Methods in Mol Biol. 2014;1096:183-201. doi: 10.1007/978-1-62703-712-9_15.
164. Xiao M, Liu X, Ji J, Li M, Li J, Yang L et al. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. Genome Med. 2020;12:57. doi: 10.1186/s13073-020-00751-4.
165. Cesare MD. Probe-based target enrichment of SARS-CoV-2. Protocolsio. 2020; 66(11):1450-1458. doi: 10.17504/protocols.io.bd5di826.
166. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC et al. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT–qPCR primer–probe sets. Nat Microbiol. 2020:1-7. doi: 10.1038/s41564-020-0761-6.
167. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K et al. Multiplex PCR method for Minion and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protoc. 2017;12:1261-76. doi: 10.1038/nprot.2017.066.

168. Matteson N. Primalseq: generation of tiled virus amplicons for miseq sequencing. Protocolsio. 2020. doi: 10.17504/protocols.io.bez7jf9n.
169. Gordon P, Mabon P. Nanostripper2020 (https://github.com/nodrogluap/nanostripper, accessed 15 July 2020).
170. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257. doi: 10.1186/s13059-019-1891-0.
171. Ounit R, Wanamaker S, Close TJ, Lonardi S. Clark. Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236. doi: 10.1186/s12864-015-1419-2.
172. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. Gmap and gsnap for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol Biol. 2016;1418:283-334. doi: 10.1007/978-1-4939-3578-9_15.
173. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex minion sequencing. Microb Genom. 2017;3:e000132. doi: 10.1099/mgen.0.000132.
174. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10-12.
175. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20. doi: 10.1093/bioinformatics/btu170.
176. NCBI. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. 2020 (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2, accessed 1 November 2020).
177. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357-9. doi: 10.1038/nmeth.1923.
178. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094-100. doi: 10.1093/bioinformatics/bty191.
179. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-60. doi: 10.1093/bioinformatics/btp324.
180. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987-93. doi: 10.1093/bioinformatics/btr509.
181. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12:733-5. doi: 10.1038/nmeth.3444.
182. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 2001;17:282-3. doi: 10.1093/bioinformatics/17.3.282.
183. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121-3. doi: 10.1093/bioinformatics/bty407.
184. Hong SL, Dellicour S, Vrancken B, Suchard MA, Pyne MT, Hillyard DR et al. In search of covariates of HIV-1 subtype B spread in the United States—a cautionary tale of large-scale Bayesian phylogeography. Viruses. 2020;12:182. doi: 10.3390/v12020182.
185. Katoh K, Standley DM. Mafft multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772-80. doi: 10.1093/molbev/mst010.

186. Katoh K, Rozewicki J, Yamada KD. Mafft online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20:1160-6. doi: 10.1093/bib/bbx108.
187. Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. Virus Evol. 2018;4. doi: 10.1093/ve/vey007.
188. Singer J, Gifford R, Cotten M, Robertson D. CoV-glue: a web application for tracking SARS-CoV-2 genomic variation. Preprints 2020; 2020060225. doi: 10.20944/preprints202006.0225.v1.
189. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020:1-5. doi: 10.1038/s41564-020-0770-5.
190. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-o-gen). Virus Evol. 2016;2. doi: 10.1093/ve/vew007.
191. Sagulenko P, Puller V, Neher RA. Treetime: maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4. doi: 10.1093/ve/vex042.
192. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. Rdp4: detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015;1. doi: 10.1093/ve/vev003.
193. Price MN, Dehal PS, Arkin AP. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641-50. doi: 10.1093/molbev/msp077.
194. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. Modeltest-ng: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol. 2019; 37(1):291-294. doi: 10.1093/molbev/msz189.
195. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307-21. doi: 10.1093/sysbio/syq010.
196. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35:4453-5. doi: 10.1093/bioinformatics/btz305.
197. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268-74. doi: 10.1093/molbev/msu300.
198. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530-4. doi: 10.1093/molbev/msaa015.
199. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4:vey016. doi: 10.1093/ve/vey016.
200. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15:e1006650. doi: 10.1371/journal.pcbi.1006650.
201. To T-H, Jung M, Lycett S, Gascuel O. Fast dating using least-squares criteria and algorithms. Syst Biol. 2016;65:82-97. doi: 10.1093/sysbio/syv068.

202. Kong S, Sánchez-Pacheco SJ, Murphy RW. On the use of median-joining networks in evolutionary biology. Cladistics. 2016;32:691-9. doi: 10.1111/cla.12147.
203. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. Mega x: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35:1547-9. doi: 10.1093/molbev/msy096.
204. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom. 2016;2. doi: 10.1099/mgen.0.000093.
205. Nadeau SA, Vaughan TG, Sciré J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. 2020 (http://medrxiv.org/lookup/doi/10.1101/2020.06.10.20127738, accessed 17 July 2020).

Annex 1. Examples of sequencing studies for molecular epidemiology

Study (referenced in section 2.2, Box 1)	Type of analysis	No. of sequences	Sampling characteristics
Influenza A(H1N1)pdm09 virus			
Fraser et al. *(1)*	Time-measured phylogenetic analysis, evolutionary rate and R_0 estimation	11	30 March 2009 to 25 April 2009
		23	30 March 2009 to 29 April 2009
Mena et al. *(2)*	Time-measured phylogenetic analysis	58 x 8 segments	2010 to 2014
	Phylogeographical analysis	422	1 March 2009 to 31 May 2009 Swine sampled across 20 countries, and humans sampled globally
Rambaut & Holmes *(3)*	Time-measured phylogenetic analysis, evolutionary rate and growth rate estimation	242	23 countries
Smith et al. *(4)*	Time-measured phylogenetic analysis and evolutionary rate estimation	168	30 March 2009 to 2 May 2009
Coronavirus MERS-CoV			
Azhar et al. *(5)*	Phylogenetic analysis	27 (spike gene) 34 (whole genome)	2012 to 2013
Dudas et al. *(6)*	Time-measured phylogenetic and host-structured coalescent analysis	274	5 February 2013 to 17 August 2015

Haagmans et al. *(7)*	Phylogenetic analysis	20	NA
Memish et al. *(8)*	Time-measured phylogenetic analysis	69	2012 to 2013
Sabir et al. *(9)*	Phylogenetic analysis and time-measured phylogenetic analysis	173	May 2014 to April 2015
Ebola virus			
Arias et al. *(10)*	Phylogenetic analysis and time-measured phylogenetic analysis	1573 1058	2014 to 2015
Baize S et al. *(11)*	Time-measured phylogenetic analysis	51	1976 to 2014 Democratic Republic of the Congo, Gabon and Guinea
Carroll et al. *(12)*	Time-measured phylogenetic analysis and phylogenetic analysis	179 262	27 March 2014 to 31 January 2015 1976 to 2015
Dudas & Rambaut *(13)*	Time-measured phylogenetic analysis	49	1976 to 2014
Gire et al. *(14)*	Time-measured phylogenetic analysis	81	17 March 2014 to 18 June 2014
		123	1976 to 2014
Hoenen et al. *(15)*	Time-measured phylogenetic and evolutionary rate analysis	296	November 2014 to January 2015
Ladner et al. *(16)*	Time-measured phylogenetic analysis	922	March 2014 to February 2015
Park et al. *(17)*	Time-measured phylogenetic analysis	318	17 March 2014 to 12 March 2015
Quick et al. *(18)*	Time-measured phylogenetic analysis and evolutionary rate estimation	728	17 March 2014 to 24 October 2015
Simon-Loriere et al. *(19)*	Time-measured phylogenetic analysis	195	January 2014 to October 2015
Stadler et al. *(20)*	Time-measured phylogenetic	72	May to June 2014

	analysis and phylodynamic analysis		
Tong et al. *(21)*	Time-measured phylogenetic analysis	256	17 March 2014 to 11 November 2014
Volz et al. *(22)*	Time-measured phylogenetic analysis and phylodynamic analysis	78	May 2014 to June 2015
Zika virus			
Faria et al. *(23)*	Time-measured phylogenetic analysis	23	19 February 2013 to 15 December 2015
Faria et al. *(24)*	Time-measured phylogenetic analysis and phylogeographical analysis	254 328	23 February 2015 to 12 October 2015 Brazil and the Americas
Grubaugh et al. *(25)*	Time-measured phylogenetic analysis and evolutionary rate estimation	104	28 November 2013 to 27 April 2016
Metsky et al. *(26)*	Time-measured phylogenetic analysis	174	12 December 2014 to 12 October 2016

NA, not applicable

References

1. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD et al. Pandemic potential of a strain of influenza A(H1N1): early findings. Science. 2009;324:1557-61. doi: 10.1126/science.1176062.
2. Mena I, Nelson MI, Quezada-Monroy F, Dutta J, Cortes-Fernández R, Lara-Puente JH et al. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. eLife. 2016;5:e16777. doi: 10.7554/eLife.16777.
3. Rambaut A, Holmes E. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Curr. 2009;1:RRN1003. doi: 10.1371/currents.rrn1003.
4. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459:1122-5. doi: 10.1038/nature08182.
5. Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM et al. Evidence for camel-to-human transmission of MERS coronavirus. N Eng J Med. 2014;370:2499-505. doi: 10.1056/NEJMoa1401505.
6. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. eLife. 2018;7:e31257. doi: 10.7554/eLife.31257.

7. Haagmans BL, Al Dhahiry SHS, Reusken CBEM, Raj VS, Galiano M, Myers R et al. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. Lancet Infect Dis. 2014;14:140-5. doi: 10.1016/S1473-3099(13)70690-X.
8. Memish ZA, Cotten M, Meyer B, Watson SJ, Alsahafi AJ, Al Rabeeah AA et al. Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013. Emerg Infect Dis. 2014;20:1012-5. doi: 10.3201/eid2006.140402
9. Sabir JSM, Lam TTY, Ahmed MMM, Li L, Shen Y, Abo-Aba SEM et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science. 2016;351:81-4. doi: 10.1126/science.aac8608.
10. Arias A, Watson SJ, Asogun D, Tobin EA, Lu J, Phan MVT et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. Virus Evol. 2016;2:vew016. doi: 10.1093/ve/vew016.
11. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba NF et al. Emergence of Zaire Ebola virus disease in Guinea. N Eng J Med. 2014;371:1418-25. doi: 10.1056/NEJMoa1404505.
12. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature. 2015;524:97-101. doi: 10.1038/nature14594.
13. Dudas G, Rambaut A. Phylogenetic analysis of Guinea 2014 ebov ebolavirus outbreak. PLoS Curr. 2014;6. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.
14. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369-72. doi: 10.1126/science.1259657.
15. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD et al. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. Emerg Infect Dis . 2016;22:331-4. doi: 10.3201/eid2202.151796.
16. Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S et al. Evolution and spread of Ebola virus in Liberia, 2014-2015. Cell Host Microbe. 2015;18:659-69. doi: 10.1016/j.chom.2015.11.008.
17. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell. 2015;161:1516-26. doi: 10.1016/j.cell.2015.06.007.
18. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228-32. doi: 10.1038/nature16996.
19. Simon-Loriere E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. Nature. 2015;524:102-4. doi: 10.1038/nature14612.
20. Stadler T, Kühnert D, Rasmussen DA, Plessis DL. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. PLoS Curr. 2014. doi: 10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.
21. Tong Y-G, Shi W-F, Liu D, Qian J, Liang L, Bo X-C et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015;524:93-6. doi: 10.1038/nature14490.

22. Volz E, Pond S. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. PLoS Curr. 2014;24:ecurrents.outbreaks.6f7025f1271821d4c815385b08f5f80e.
23. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406-10. doi: 10.1038/nature22401.
24. Faria NR, Azevedo RdSdS, Kraemer MUG, Souza R, Cunha MS, Hill SC et al. Zika virus in the Americas: early epidemiological and genetic findings. Science. 2016;352:345-9. doi: 10.1126/science.aaf5036.
25. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K et al. Genomic epidemiology reveals multiple introductions of zika virus into the United States. Nature. 2017;546:401-5. doi: 10.1038/nature22400.
26. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM et al. Zika virus evolution and spread in the Americas. Nature. 2017;546:411-5. doi: 10.1038/nature22402.

Annex 2. Checklist for setting up a sequencing programme

Aims

- ☐ Define the expected aims of the sequencing programme; what information will sequencing be likely to provide that is additional to or more cost-effective than existing approaches?

Stakeholder identification and engagement

- ☐ Identify key stakeholders.
- ☐ Discuss the programme aims with senior representatives of stakeholder groups and define the responsibilities of each group.
- ☐ Consider sharing educational materials about the potential and requirements of SARS-CoV-2 sequencing with stakeholders.
- ☐ Identify the links needed between key stakeholders to allow rapid movement of samples, requests for information and use of results.
- ☐ Ensure that clear, appropriate links between stakeholders are established.

Technical considerations

- ☐ Determine the level of genomic sampling required to achieve the desired goals, in discussion with senior members of case-identification and analytical teams.
- ☐ Identify the metadata required to achieve the desired goals, in discussion with senior members of case-identification and analytical teams.
- ☐ Choose appropriate sample and library preparation protocols.
- ☐ Choose appropriate bioinformatic protocols.
- ☐ Choose appropriate analytical protocols.

Logistical considerations

- ☐ Consider where sequencing and analysis will be conducted (e.g. an existing diagnostic laboratory or external commercial or academic laboratory).
- ☐ Identify appropriate sources of funding that will be adequate to support laboratory sequencing, data storage and data analysis.
- ☐ Ensure that sufficient reagents and computational resources are available and can be sustainably obtained as required.
- ☐ Ensure that there are sufficient and appropriate human resources to deliver the programme at every stage.
- ☐ Ensure that sample integrity can be maintained at all steps throughout the pipeline via cold-chain or other measures.

- Ensure adequate collection and storage of metadata and correct association with biological sample.

- Consider the possible additional pressure sequencing will place on existing arms of the public health response and seek ways to alleviate this.
- For large-scale sequencing programmes, identify how to streamline the sharing of data and samples between participating groups (e.g. the feasibility of using a single sample identification and identical metadata formats).

Ensuring a safe and ethical environment

- Conduct appropriate ethical reviews for the generation, use and storage of sequence data and associated metadata.
- Conduct risk assessments of sequencing activities to ensure appropriate biosafety at all stages.
- Conduct risk assessments of sequencing activities to ensure appropriate biosecurity, if relevant under national and regional law.
- Consider the impact on human resources, including the reallocation of staff or hiring of additional staff to maintain the individual workload at reasonable levels.
- Ensure that staff can commute to work and be in the workplace safely and in accordance with national guidelines on preventing transmission during the COVID-19 outbreak.
- Define strategies for maintaining the sequencing programme if key staff members become ill or must self-isolate.

Data-sharing

- Ensure that all stakeholders are in agreement as to which sequences and metadata will be shared publicly, via which platforms and when.
- Ensure that all stakeholders are in agreement as to whether any metadata are to be restricted to a limited number of local users and devise strategies for securely sharing those data.
- Ensure data-sharing complies with national and international regulatory frameworks.

Evaluation

- Ensure regular opportunities for evaluating the sequencing programme, including successes and continuing challenges.
- Ensure a monitoring and evaluation framework is implemented to assess technical performance of the sequencing programme and confirm that the programme meets the objectives

9789240018440