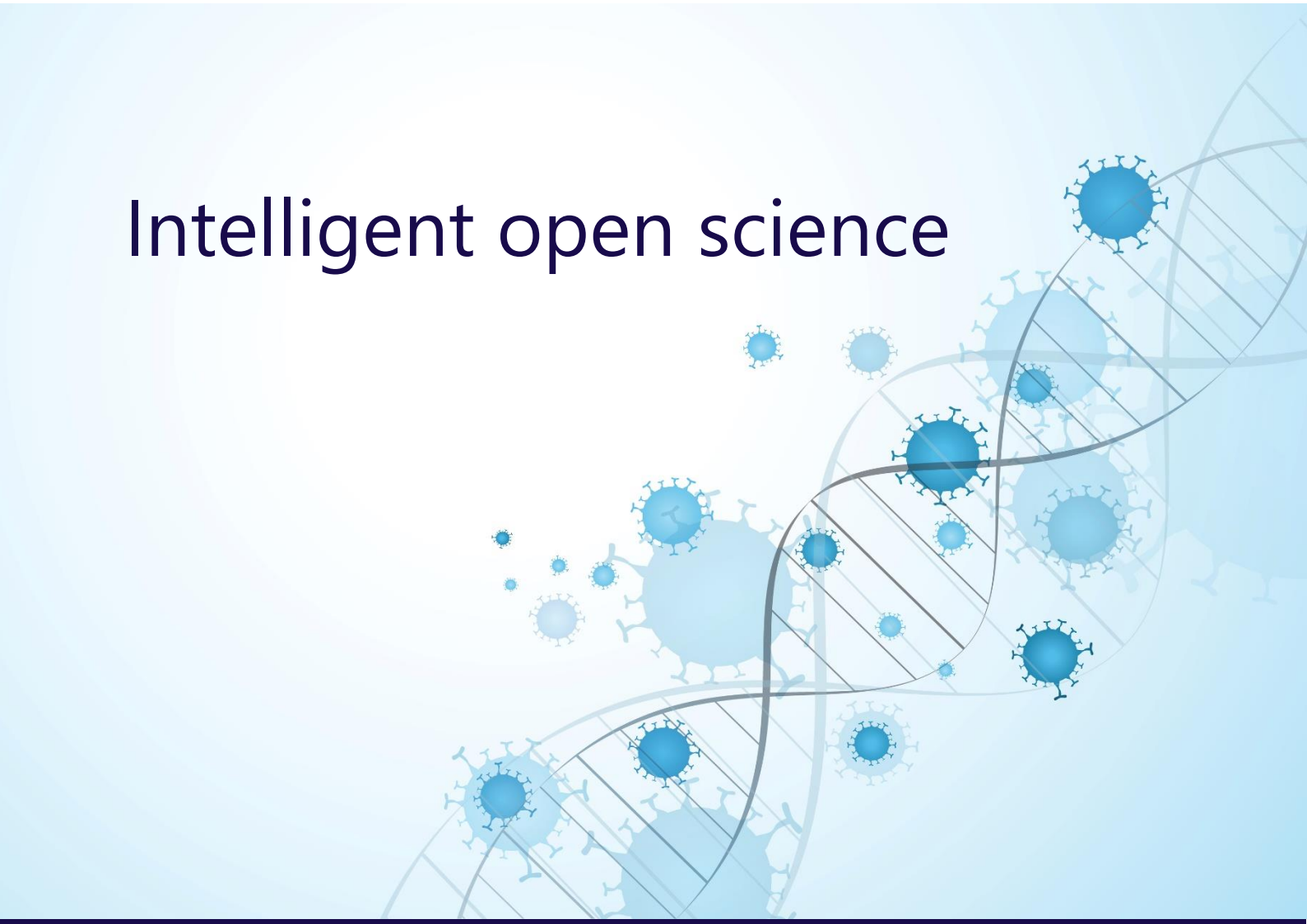
Research Consulting

Intelligent open science



A case study of viral genomic data sharing during the COVID-19 pandemic

Prepared on behalf of The UK Government Department of Business, Energy and Industrial Strategy

November 2022

Intelligent open science

Report commissioned by:

The UK Government Department of Business, Energy and Industrial Strategy

https://www.gov.uk/government/organisations/department-for-business-energy-and-industrial-strategy

Contact:

Katherine Barnes

Head of Multilateral Research & Innovation

Katherine.Barnes@beis.gov.uk

Report authors:

Rob Johnson, Lucia Loffreda, Christine Ferguson, Eleanor Cox, Neil Beagrie, Cephas Avoka, Julian Hiscox.

www.research-consulting.com

Contact:

rob.johnson@research-consulting.com

Report dated: October 2022

Executive Summary

Introduction

Intelligent open science means striking a balance between open and controlled data sharing

Open data sharing can bring enormous benefits to scientists, citizens, governments, and businesses. Fully open data (without any restriction on the end user) can remove the frictions in discovery, access and use that impede rapid development but an equitable data access and re-use ecosystem must also include legitimate boundaries and appropriate incentives. Controls may be needed, for example, to incentivise investment by private actors or to protect the privacy of individuals, public safety and security, or indigenous and other disadvantaged communities. Striking an intelligent balance between fully open and controlled data sharing lies at the heart of the commitment "to promote the efficient processing and sharing of research data as openly as possible and as securely as necessary" in the G7 Research Compact (2021).[1] The appropriate balance for viral genomic data sharing during and after a pandemic remains under active discussion and is explored in this study.

Study aims

Using open science to respond to future global crises

The COVID-19 pandemic has been a significant stress test of global genomic sequencing capacity and the open sharing of sequencing data and associated metadata. As nations look to emerge from the crisis, lessons can be learnt for the future. The UK Government's Department for Business, Energy & Industrial Strategy (BEIS) commissioned this study following a commitment made during the UK's G7 Presidency and published in the G7 Research Compact. The study aims to add depth and precision to existing recommendations on data sharing across borders, and related research practice and cultural issues. It draws on evidence gathered via interviews and focus groups with expert policymakers, infrastructure providers and researchers to inform the use of open science in responding to future global crises. It was not within the scope of the study to make specific recommendations on the creation and sharing of sequencing information.

The role of genomic viral data sharing during COVID-19

Pathogen genomic data was often shared too late to contribute effectively to the emergency response, surveillance efforts and preparedness

Pathogen genomic data is relatively difficult to create, and its interpretation requires both accompanying comprehensive metadata and distinct expertise. However, it has been deployed worldwide to characterise virus outbreaks, track the mutation and spread of the virus and develop public health responses to the COVID-19 pandemic. Genomics has a highly developed culture of data sharing, underpinned by globally recognised databases, and data sharing was strongly encouraged by funders, publishers and policymakers during the COVID-19 pandemic. Despite these strengths, however, efforts to share genomic data during the pandemic yielded mixed results. While sequences were shared more quickly and widely than ever before, in many cases they were shared too late, in too partial a form or with insufficient metadata to contribute effectively to the emergency response. Variations in data quality, formats, associated metadata standards, and arrangements for access and re-use continue to present barriers to the effective sharing and use of genomic data at scale.

Preconditions for successful sequencing

Gaps in underlying sequencing capability led to 'dark spots' in SARS-CoV-2 datasets

Genomics data can only be shared where the incentives and infrastructure exist for it to be generated in the first place. Variable access to, and investment in, sequencing capability, particularly in low- and middle-income countries, led to significant gaps in our understanding of how SARS-CoV-2 mutates and spreads worldwide. Pooled investment mechanisms are being developed to strengthen global sequencing capability and promote data sharing, but these investments in technical infrastructure must be accompanied by equitable access to chemical reagents, recruitment and development of skilled staff and increased availability of open code and software.

The geopolitics of genomics

Geopolitical considerations heavily influence attitudes to data sharing

The willingness and ability of different actors to share data during emergencies is heavily influenced by pre-existing geopolitical considerations. These lead to the pursuit of self-interest over shared interests, resulting in suboptimal outcomes for all.[2] During the COVID-19 pandemic, pre-existing disparities in sequencing capacity and capability have been compounded by unresolved concerns over access and benefit-sharing. Digital sequence information, which includes viral genomic data, is not covered by the Nagoya Protocol[3] to the Convention on Biological Diversity and there are divergent international perspectives on whether it should be considered a common good. Effective data sharing during future emergencies will depend on underpinning work being undertaken in 'peacetime' to better understand the motivations of different stakeholders and reach mutually agreed approaches to access and benefit-sharing through relevant international fora.

From personal choice to community norm

Rapid sharing depends on close collaboration between relevant stakeholders

The reluctance of some actors to immediately share sequencing data in the COVID-19 pandemic is due to a combination of factors. In academic communities, the longstanding practice of withholding data until the point of publication runs counter to the needs of emergency response. Academics who work regularly on pandemics understand the requirements for public health actions and share their data rapidly. However, the adoption of prosocial approaches to data-sharing needs to become a wider community norm. Meanwhile, in hospitals and other public health settings, data governance concerns and the need to comply with data protection legislation frequently inhibit sharing. While these barriers are common worldwide, some nations were able to share sequencing data much more quickly than others, indicating that they are not insurmountable. Progress depends on normalising rapid data sharing in an emergency context, developing strong linkages between research and public health actors, and establishing legislative mechanisms that enable sharing of personal data in an emergency context. Effectively leveraging viral genomic sequence data also relies on the capture of high-quality metadata and the development of secure environments and legal frameworks that allow for the analysis of sequence data in conjunction with sensitive clinical data. Developing secure environments, and the technology and tools that enable them, remains one of the biggest challenges faced and impacts on the effectiveness of other interventions.

Balancing the needs of data generators and users

Researchers favour repositories that allow them to assert or retain rights to, and receive credit for, the reuse of submitted data

Public, open access databases within the International Nucleotide Sequence Database Collaboration (INSDC) are maintained by partner organisations in Japan, Europe and the United States for the benefit of all types of community worldwide. However, the pandemic saw many data generators opt to deposit sequences in a controlled-access repository, GISAID. GISAID was also able to harvest data from open access repositories to supplement direct deposits, thereby making it the most comprehensive available source of global sequencing data. Furthermore, it appealed to many depositors, particularly those with fewer resources to permit immediate analysis of the data, as it allowed them to retain rights over their data and receive credit for its subsequent re-use. The limitations of the GISAID model arise from the requirements imposed on re-use of the data, which meant scientists and public health agencies needed permission to aggregate and re-analyse the data alongside other datasets. Some contributors to this study perceived a lack of transparency and accountability in GISAID's operations in comparison with the INSDC databases, while others noted that the INSDC databases themselves lack global accountability, being operated by a small number of high-income countries. Perspectives on the most appropriate solution vary widely between individuals in research, public health and industry contexts, high-, medium- and low-resource environments and emergency preparedness, response and/or recovery settings.

Lessons learned for open science

Five lessons learnt

The lessons learnt for open science policymakers can be summarised as follows:

1. Invest for the long term

An effective emergency response relies on long-term investment in open data infrastructure, standards and skills

The long-term investment made in developing international standards and infrastructures for data sharing in genomics was repaid many times over when this data became central to the pandemic response. Critical data infrastructures need open and transparent governance mechanisms, sustainable funding, and common standards that enable interoperability and scalability. These must be accompanied by skilled individuals who are able to create, analyse, share and re-use relevant data. All of this relies on a long-term commitment by governments and funders to invest in science, research, and public health infrastructure, as well as a recognition of the critical importance of open, scalable data infrastructure, software, standards and skills.

2. Take a global perspective

Global challenges like COVID-19 require a global and inclusive approach to data sharing

Effectively tackling global crises like the COVID-19 pandemic requires representative data from all parts of the world. Not all countries and regions have sufficient data-generating capacity or trained human resources to collect, disseminate, and analyse these data. The willingness and ability of different actors to share data during emergencies is also heavily influenced by pre-existing geopolitical considerations and the risk of adverse political and economic consequences. Interventions designed to enhance the availability of relevant data must ensure they identify and tackle the root of the problem. In many cases this will be a lack of underlying research capacity and public health infrastructure, or political tensions rather than inadequate uptake of open sharing practices. Open international infrastructures must also be cognisant of the needs of a diverse community of users, with standardisation of data and metadata formats accompanied by a flexible approach to access.

3. Create incentives for equitable data-sharing

Reformed incentives are needed to promote data-sharing across boundaries

If rapid and open data sharing is to be encouraged, the contributions of data generators need to be recognised and rewarded. Informal data access arrangements based on pre-existing knowledge of trusted individuals is not sufficient to enable equitable sharing and re-use of data at scale. Equitable data sharing in an emergency situation – in this context, a disease outbreak with political, social and public health impacts – therefore depends on prior work to understand the motivations of different stakeholders and reach agreement through relevant international fora on prosocial arrangements for access and benefit sharing. There is a compelling need to continue efforts to reform incentives for all data generators to reward the sharing of reusable high-quality data, code and other research objects alongside accompanying metadata. Similarly, there is a need to clarify expectations of speed, quality and transparency for data generators in differing contexts such as routine surveillance in public health. The pandemic has highlighted the crucial importance of cross-boundary collaboration at international, national and local levels, and exposed a need to improve the interface between research and public health in order to maximise the combination and re-use of scientific and clinical data.

4. Adapt to changing circumstances

Established norms for data-sharing must evolve in light of the COVID-19 pandemic

Established norms around the timing and extent of data sharing were in many cases set aside in the COVID-19 crisis, with multiple actors recognising that the immediate availability of data to a broad set of users was paramount. Yet the pandemic also provides an opportunity to re-assess these established norms, whose deficiencies were in some cases sharply exposed. Ongoing efforts to reform academic incentives must be accompanied by corresponding work to incentivise sharing by public health actors, with strengthened expectations for data-sharing by all parties in an emergency context. Public policymakers, research and development funders, institutions in academic and public health, and publishers all have a role to play in setting expectations for open and rapid sharing of all relevant data and information in these circumstances. Open infrastructure providers must be able to identify and respond rapidly to emerging requirements, while new approaches should make provision for sensitive datasets to be used for research purposes in emergency scenarios.

5. Move beyond current sharing paradigms

New sharing paradigms are needed to address competing interests

This study has exposed divergent perspectives within and between the research and public health communities on the merits of open and controlled models of access to genomic viral data. Fully open-access infrastructures for data sharing offer demonstrably greater benefits than controlled access repositories in terms of data re-use and integration, but these benefits cannot be realised in practice unless these infrastructures are accompanied by a transparent and globalised approach to funding, governance and benefits sharing. Proponents of open-access infrastructures must also give greater consideration to mechanisms for incentivising and crediting data deposits and enabling the creation of high-quality metadata. An intelligent approach to open science means moving beyond our existing data-sharing paradigms to be better prepared for future emergencies. Progress is closely tied to the reform of existing incentive structures and must be understood as a long-term endeavour. But it holds out the prospect of data being shared and re-used as widely as possible for the benefit of all populations, while acknowledging and rewarding the efforts of data generators and custodians.

Research Consulting

Contents

Research Consulting

1. Background and methodology

The COVID-19 pandemic has been a significant stress test of global genomic sequencing capacity and the open sharing of sequencing data and associated metadata. As nations look to emerge from the crisis, lessons can be learnt for the future.

This report draws on evidence gathered via desk research, interviews and focus groups with expert policymakers, infrastructure providers and researchers to review approaches to the sharing of viral genomic data during the pandemic. It aims to add depth and precision to existing recommendations on data sharing across borders, and related research practice and cultural issues.

Research Consulting

1. Background and methodology

1.1. Introduction

Background

The UK Government's Department for Business, Energy & Industrial Strategy (BEIS) has commissioned this study to add depth and precision to existing recommendations on data sharing across borders, and related research practice and cultural issues (e.g. legal barriers and required agreements, incentives, and rewards). It was commissioned to fulfil the commitment made in the G7 Research Compact, under the UK's G7 Presidency, to deliver a specific case study focussed on data sharing in an emergency.[1]

Rationale for this case study

This report examines the role of data sharing and open science practices during the pandemic, with a specific focus on genomic viral data and its use in combination with other datasets. Genomic sequencing, a series of laboratory methods that are used to determine the genetic makeup of particular organisms (in this case, SARS-CoV-2), has been critical to the public health response to the pandemic. Yet the COVID-19 outbreak has also been a significant stress test of worldwide genomic sequencing capacity, with a global pandemic declared just three months after the first cases were identified in December 2019.

The literature reviewed in this study, as well as the insights shared by expert policymakers, infrastructure providers and researchers, have highlighted both successes and failures in approaches to viral genomic data sharing in the COVID-19 pandemic. As in other aspects of the pandemic response, there is significant room for future improvement, particularly in the speed of public health responses.[4] Genomics is a scientific field with well-established infrastructures and disciplinary norms for data-sharing, and sequencing data has played a crucial role in the response to the COVID-19 pandemic. It therefore represents an appropriate case study through which the broader role of data sharing and open science in responding to public health emergencies can be assessed.

Overview of research questions

Drawing on experiences and lessons learnt from the COVID-19 pandemic, this study sought to answer the following research questions:

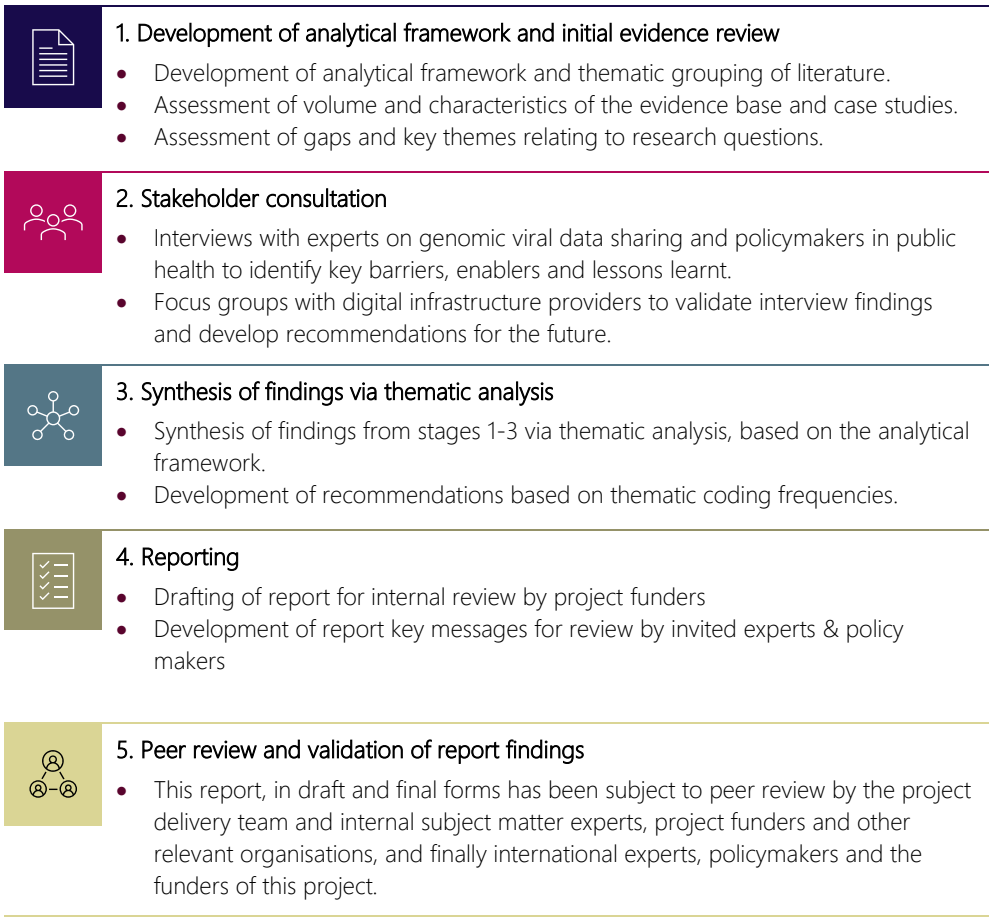
- what worked well to enable sharing and reuse of genomic viral data in the context of public health emergencies;
- what were the key barriers that limited the effective sharing of genomic viral data during the pandemic;
- what lessons have been learnt for open science within the context of COVID-19; and
- what role can open science play in responding effectively to future global emergencies?

Scope of work

This report identifies a range of ongoing challenges and initiatives relevant to the creation and sharing of sequencing data and draws on these to identify lessons learnt and recommendations on the use of open science to inform responses to future emergencies. It was not within the scope of this work to make specific recommendations on the creation and sharing of sequencing information.

1.2. Methodology

Figure 1. Overview of project methodology



Development of analytical framework and initial evidence review

This study has been informed by an **analytical framework**. The framework was developed to address each of the research questions outlined above, focusing broadly on four key themes: contextual information on genomic viral data creation and sharing, barriers to genomic viral data sharing, enablers of genomic viral data sharing and recommendations arising from the COVID-19 pandemic. The analytical framework is based on an adaptation of the PESTLE framework (i.e., Political, Economic, Social, Technological, Legal and Environmental factors).[5] Limited environmental factors were identified as part of this review and are therefore not presented in our analysis.

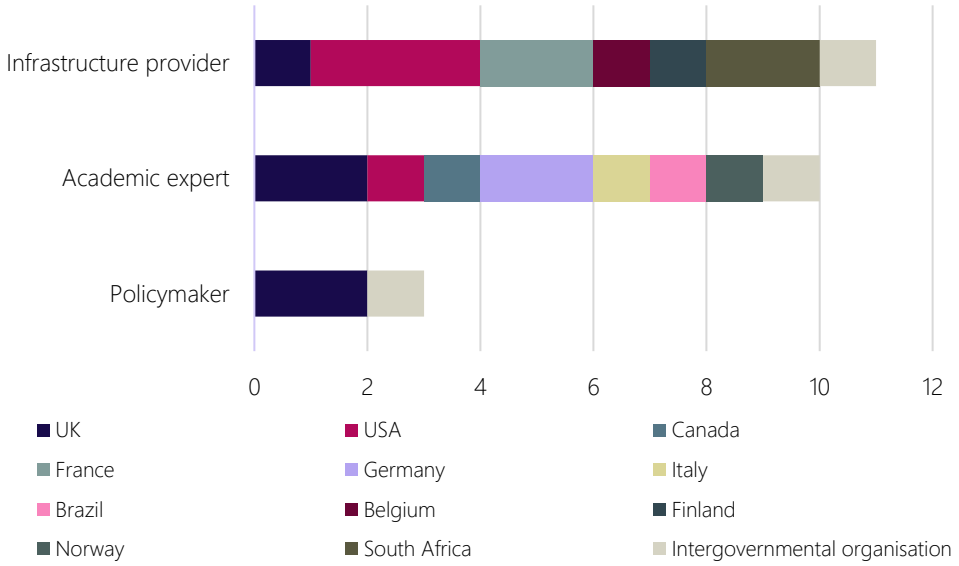
The **evidence review** that informs this study was structured based on the themes identified in the analytical framework. A total of 295 sources were identified via a mix of structured Google searches and snowball sampling (a process where the bibliographies of relevant documents were used to identify additional sources). An additional body of literature (120 sources), initially gathered for a similar study on open research practices during the pandemic[6] was reviewed and sources were filtered based on their relevance to the research questions of this study. The 106 sources selected for inclusion in this report, based on their relevance to our research questions, constitute a mixture of academic articles, reports, blog posts, and relevant websites. These sources have been summarised and compared, with key themes being extracted for discussion in the present document.

Stakeholder consultation

Interviewees and focus group attendees were identified via a mixture of approaches including desk-based research, snowball sampling (where initial participants identified other individuals to be consulted) and a text processing algorithm. In brief, the text processing algorithm approach was used to identify (a) mentions of SARS-CoV-2 genomic sequencing data deposits in peer-reviewed articles and preprints (referred to as "publications") and (b) the researchers that have led these data deposits, using corresponding authors in the associated publications. Full details of the text processing algorithm are included in Appendix C of this report.

The **stakeholder consultation** phase of this work comprised a series of 60-minute one-to-one interviews with expert researchers familiar with genomic viral data sharing, and with policymakers with expertise in public health. Interviews focused on identifying key barriers, enablers and lessons learnt from genomic viral data sharing during the pandemic. Focus groups with digital infrastructure providers were then conducted to validate the lessons learnt highlighted in interviews and to identify areas where future recommendations could be introduced. with A total of 24 stakeholders were engaged across interviews and focus groups (Figure 2). Only three of these stakeholders were from low- and middle-income countries, and therefore the views of these countries are likely to be under-represented in our findings.

Figure 2. Interviewees and focus group attendees involved in the study by country



Synthesis of findings via thematic analysis

Interview and focus group transcripts were generated following the stakeholder consultation phase and subjected to a process of thematic coding using NVivo, a qualitative analysis software package, to identify common themes and concerns. This report presents **a synthesis of findings identified via thematic analysis**, and the prioritisation of issues in this report is based on the frequency of findings in the dataset, their relevance to the project's objectives and the professional judgement of the research team, including peer reviewers.

Reporting and peer review

Evidence gathered from the evidence review and stakeholder consultations phases was synthesised for **reporting**. A report outline of key messages was reviewed by project

funders and members of the G7 Open Science Working Group. Report drafts were then subjected to **peer review** in three phases:

- Internal peer review of report draft by the project delivery team, including review from internal subject matter experts.
- Initial round of peer review by two independent subject matter experts.
- Final round of peer review by international experts and members of the G7 Open Science Working Group. In total, 18 independent experts reviewed this report and provided feedback prior to its publication (Appendix B).

We wish to highlight that **participation in this study (whether as an interviewee, focus group attendee or reviewer) does not imply endorsement of all the report's findings**.

Report structure

Following this introduction, this report is divided as follows:

- Section 2. The role of genomic viral data sharing during COVID-19
- Section 3: Preconditions for successful sequencing
- Section 4: The geopolitics of genomics
- Section 5: From personal choice to community norm
- Section 6: Balancing the needs of data creators and users
- Section 7: Conclusions and lessons learnt

Limitations

The present report is subject to the following limitations:

Evidence review

- Our literature review was designed to provide an informed conclusion on the volume and characteristics of the evidence base and a synthesis of what that evidence indicates in relation to the research question. It did not include a critical appraisal of that evidence.
- Our literature review was primarily conducted in the United Kingdom, with supplementary searches conducted in Ghana. However, we acknowledge the evidence supporting this study is likely to retain some bias towards high-income countries.

Stakeholder consultation

- 24 participants (Appendix A) were recruited to join interviews and focus groups via convenience sampling, that is, we consulted with individuals who were both available and willing to communicate. Therefore, the viewpoints expressed in this report may not be representative of the wider communities relevant to this study.
- We note the balance of contributors to this study was weighted towards interviewees from academic in higher income countries, particularly members of the G7. The perspectives of low- and middle-income countries, intergovernmental organisations, public health organisations and industry may be under-represented as a result.

Synthesis of findings via thematic analysis

- Our analysis of qualitative data (literature, interview and focus group transcripts) is underpinned by thematic coding, which relies on an extent of subjective interpretation.

Acknowledgements

This report was commissioned by The UK Government's Department for Business, Energy & Industrial Strategy. We gratefully acknowledge:

- the guidance and support received from Katherine Barnes (Department for Business, Energy & Industrial Strategy) and Rachel Bruce (UK Research and Innovation);
- the experiences and insights shared by interviewees and focus group attendees listed in Appendix A;
- the peer review of this report by international experts listed in Appendix B; and
- the development of a stakeholder identification strategy and text processing algorithm by Etienne Vignola-Gagné (Science-Metrix, an Elsevier company, Canada) (see Appendix C).

2. The role of genomic viral data sharing during COVID-19

During the COVID-19 pandemic, the genomes of the causative agent, SARS-CoV-2, were sequenced at a hitherto unimaginable rate. This section provides a brief introduction into genomic sequencing methodologies, explaining how sequencing data is shared and how it is used to inform public health responses. It contrasts the complex processes involved in creating pathogen genomic data with the ease with which it can be shared and re-used. In the case of COVID-19, genomic viral data sharing has been critical to characterising virus outbreaks, tracking the mutation and spread of the virus across and within borders, and developing public health responses.

2. The role of genomic viral data sharing during COVID-19

2.1. Sequencing of genomic viral data

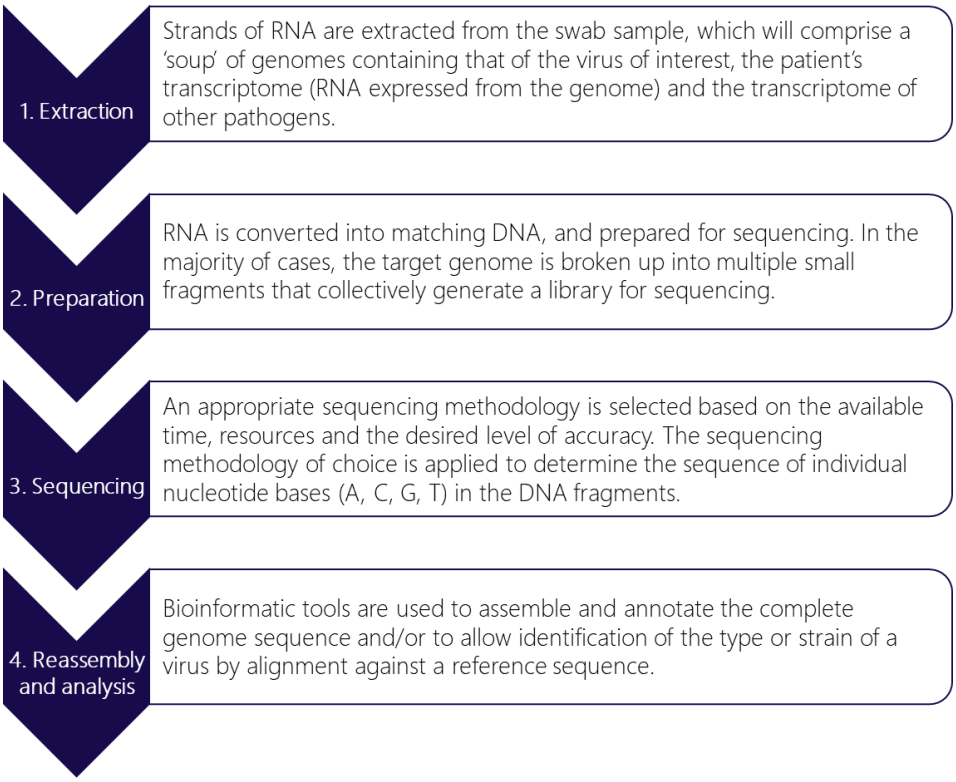
SARS-CoV-2 is a novel coronavirus containing a genome comprised of RNA

At the beginning of the COVID-19 outbreak, little to nothing was known about the novel coronavirus. To begin to understand, and subsequently develop responses to, SARS-CoV-2, a wide range of data types have been gathered. One critical data type is genomic viral data which is used to characterise a virus. In this report, genomic viral data refers to the genome – or genetic material – that makes up SARS-CoV-2, the causative agent of COVID-19. Like all other coronaviruses, SARS-CoV-2 is an RNA virus. The genetic composition of its RNA genome is determined via genomic sequencing.

Viral genomes are obtained via a multi-stage process: extraction, preparation, sequencing, reassembly, and analysis

Raw viral genomes are typically obtained from swab samples from infected individuals, often via health facilities. While the exact sequencing methodologies vary, there are four generalisable steps required to obtain the genomic sequence of a virus from a clinical sample, that can then be uploaded and shared in a data repository (Figure 3).

Figure 3: Overview of the process to obtain the genomic sequence of a virus from a clinical sample



Time and accuracy constraints determine which genomic viral

The selection of a sequencing strategy depends on what one wants to sequence, the available time, resources, and the desired level of accuracy. The vast majority of global publicly deposited genomic data for SARS-CoV-2 were sequenced using platforms from

sequencing strategy is used

Illumina,[7] a US biotechnology company, and Oxford Nanopore, which is based in the UK.[8] Other platforms used, albeit at much lower volumes, included those from Ion Torrent, MGI and Sanger.[9] Table 1 outlines the two main sequencing methodologies applied during the pandemic. Where the term 'sequencing' is used in the remainder of this report, it can be assumed this refers to amplicon-based sequencing, unless otherwise stated.

Table 1. Sequencing methodologies commonly applied during the COVID-19 pandemic

Methodology	Features	Use during COVID-19 pandemic
Metagenomic/ Metatranscriptomic / 'Shotgun' sequencing	• Used to **understand what is present in a sample** (from an infected human or can be from the environment) and in what proportions. • Used to **understand the host response to infection** and **identifying new pathogens or variations**. • Aims to be provide an unbiased sampling of all genomes/transcripts in a sample. Sequenced without prior knowledge of which microbes are present.	• Used to **identify the virus for the first time**. • Otherwise had **limited use during the COVID-19 pandemic** due to low sequence read depth. • This method is **expensive due to its non-targeted methodology and it is too time-consuming for routine sequencing** as many more sequence reads per sample are required for accuracy than for amplicon sequencing.
Amplicon-based sequencing[1]	• Used to **determine the presence of coronavirus** in a sample or to **study coronavirus genetic variation**. • **Some prior knowledge of the sequence is required to use this technique**, because it targets specific regions of interest.	• Used for the **majority of sequences produced during the pandemic**. • **Cheaper due to targeted methodology** compared to shotgun sequencing. Only small sections of viral genomic DNA are amplified for sequencing.

There are five core uses of viral genomic data in public health

All viruses rely on hosts to replicate, spread, and ultimately survive. As viruses mutate, new variants emerge that may be better adapted to evade immune systems, vaccines, and other threats to their survival. Enabling researchers to decipher the genetic material in a virus is now an essential part of understanding the spread of viruses and eventually controlling them. The five core uses of genomic viral sequencing within the context of public health are shown in Figure 4. While all of these uses were deployed in the COVID-19 pandemic, the benefits of viral genomic sequencing extend well beyond public health emergencies and include ongoing surveillance and investigation of prioritised diseases.

[1] An amplicon is a piece of DNA that is the source and/or product of amplification, usually by Polymerase Chain Reaction (PCR), which increase the amount of DNA available for different sequencing techniques.

Figure 4: Core uses of viral genomic sequencing in public health (Centers for Disease Control and Prevention, 2021)[10]

1. Identify, detect, investigate, monitor, and control the virus
2. Understand how people become exposed to the virus – i.e. disease surveillance
3. Learn how the virus is geographically distributed
4. Help trace the source and transmission of outbreaks
5. Learn about virus evolution such as the development of variants of concern

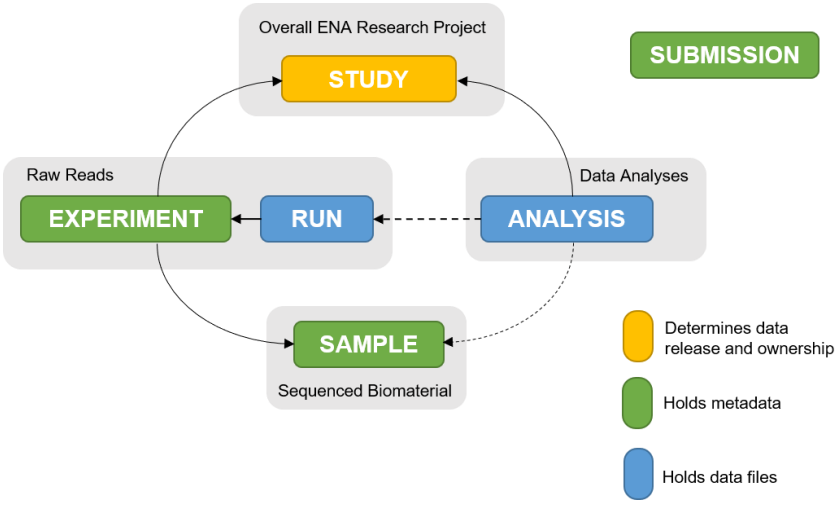
2.2. Sharing genomic viral data

Sequences are shared in a standardised data format, and can then be accessed via genome browsers and APIs

Once a genome sequence has been assembled and annotated the information needs to be stored in a database so that it can be shared. Sequence data is typically uploaded to a database in FASTA format, a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes.[11] A sequence in FASTA format begins with a single-line description, followed by lines of sequence data, and will usually be accompanied by metadata providing contextual information on the sequence itself. GISAID uses a FASTA-like simple format.

An extension of the FASTA format is FASTQ format, which stores both sequence data, associated quality values, and other metadata. Data can be retrieved from the European Nucleotide Archive (ENA) in FASTQ format. Figure 5 shows the metadata model used by the European Nucleotide Archive. Once uploaded, sequences can then be accessed by individual users via a genome browser, a graphical interface for users to browse, search, retrieve and analyse genomic sequence and annotation data, or via application program interfaces (APIs) which allow sequences to be searched, linked and downloaded programmatically.

Figure 5: The European Nucleotide Archive (ENA) Metadata Model (ENA, 2017-2022)[12]



Both raw and assembled SARS-CoV-2 sequences can be shared

While the focus of this report is on open sharing of data across international boundaries, databases of sequencing data exist at multiple levels in order to meet varying local, regional, national and international needs. For the purposes of global pandemic response, however, local and national databases alone are insufficient. Both virus sequence data and sample metadata must be accessible internationally as analyses rely on the ability to compare locally acquired virus sequences with the global virus genomic diversity. In ideal circumstances, both raw sequencing reads (i.e. all individual sequenced fragments of a virus genome before they are assembled into one consensus genome) and full-length genomes would be shared alongside relevant metadata, including the date and location of sample collection as a minimum. However, there are differing opinions on the value of raw read sharing, in addition to the fact that read-level datasets can reach hundreds of gigabytes in size, and so sharing raw sequence data may not be feasible in settings that have limited internet upload speeds or intermittent connections.[13] This is particularly challenging in low- and middle-income countries, where frequent interruption of power supplies and poor internet bandwidth create challenges to upload and download, use or analyse genomic data.[14,15] Taking limited internet connections into account, the COVID-19 data portal developed by the European Molecular Biology Laboratory's bioinformatics institute split sequenced COVID-19 datasets into chunks for download that could easily be recombined into the complete dataset.[16]

Case study: The FAIR and CARE principles in genomics

The FAIR principles aim to maximise the reuse of data, ensuring that it is findable, accessible, interoperable, and reusable, and they are a broadly recognised core principle for open science practices worldwide. To make full use of the increasing amount of genomic sequencing data shared worldwide, the Research Data Alliance has recommended that public health institutions encourage and adopt guidelines for data collection, annotation, storage, and reuse in line with the FAIR data principles.[17]

In disciplines such as genomics, transcriptomics and proteomics data pose unique challenges when it comes to production and reuse. However, making data FAIR can enable this. For example, significant benefits can also be achieved through the deposit of data in globally accessible genomic data repositories (see below) with standardised metadata schema.

The CARE Principles (Collective benefit, Authority to control, Responsibility, and Ethics) are concerned with protecting indigenous rights. Although different in scope, the CARE principles can complement the FAIR Principles and critically help to inform the inclusion of Indigenous Peoples in data processes.[18] The CARE Principles aim to strengthen Indigenous control for improved discovery, access, use, reuse, and attribution in contemporary data landscapes. Globally to date, the CARE principles are considered to be 'ahead of practice' and thus are not yet widely adopted.[18,19]

A renewed impetus for early data sharing during the coronavirus pandemic drove database submissions…

The field of genomics has a long-established culture of data-sharing. With a viral (SARS-CoV-2) genomic sequence in hand, the accepted research norms are to upload this to a sequence database, such as those operated by the members of the International Nucleotide Sequence Database Collaboration (INSDC), conduct further comparative analysis to related sequences, or to consider adding it to a publication.[20] In some cases, there is scope for newly generated data to be deposited under embargo. Pre-pandemic cautiousness about data sharing prior to publication has been challenged by a number of initiatives from funders and publishers advocating for the early sharing of data sets

to provide a scientific basis to tackle the pandemic. For example, in January 2020 Wellcome reissued a Joint Statement, signed by 160 prestigious organisations across the international research landscape, which called on researchers, journal publishers, and funders to "ensure that research findings and data relevant to this outbreak are shared rapidly and openly to inform the public health response and help save lives".[21] The Joint Statement was informed by the approach taken during the Zika and Ebola outbreaks, where similar statements were released, and the benefits of open data sharing were realised. A recent study of the Joint Statement's impact shows that it helped to align the efforts of key stakeholders, influenced their policy requirements and organisational targets and shaped the long-term vision for open research.[6] On 28 May 2020, the G7 Science and Technology Ministers' Declaration on COVID-19 was issued, calling for government-sponsored COVID-19 epidemiological and related research results, data, and information to be accessible to the public to the greatest extent possible.[22] Supporting this, the G7 Research Compact also highlights the commitment to open sharing practices, particularly the open sharing of research data, in the context of public health emergencies.[1]

...But subsequent analysis shows poor efficacy of sharing genomic viral data

Researchers generating sequence data were thus encouraged and enabled to deposit and share this before publication. However, an analysis conducted more than two years into the pandemic assessed the efficacy of sharing genomic viral data as 'poor... [with] an urgent need to increase timely and full sharing of sequences, standardisation of metadata files and support for countries with limited sequencing and bioinformatics capacity'.[9] The remainder of this report explores the barriers to and enablers of sharing of viral genomic data in the context of the COVID-19 pandemic, and considers the lessons for open science.

VIRAL DATA SHARING DURING COVID-19 - LESSONS LEARNT FOR OPEN SCIENCE

- The long-term investment made in developing international standards and infrastructures for data sharing in genomics was repaid many times over when this data became central to the pandemic response.
- Open international infrastructures must be cognisant of the needs of a global community of users, with standardisation of data and metadata formats accompanied by a flexible approach to access that takes account of technology and bandwidth limitations in certain regions.
- Calls from leading funders, publishers and policymakers were important in setting expectations for open and rapid sharing of research results, data and information, but were not sufficient to deliver a wholesale shift in data-sharing practices.

3. Preconditions for successful sequencing

The successful creation, sharing and analysis of sequencing data depends on a pipeline of activities, requiring investment at each stage. Adequate funding, technical infrastructure, chemical reagents, and skilled staff are essential pre-conditions for the generation of sequencing data.

Inconsistent access to each of these elements has led to variability in the availability and quality of sequencing data. In turn, this contributes to gaps in our understanding of how SARS-CoV-2 mutates and spreads worldwide.

Research Consulting

3. Preconditions for successful sequencing

3.1. Creating a healthy data pipeline

Sequencing capacity is like a pipeline, requiring investments at each stage of the process

Increasing genomic sequencing capacity can be thought of as a pipeline, requiring investments across national borders, economic sectors, and stages of the pandemic response. The key stages of the capacity strengthening pipeline for genomic sequencing can be considered as: having the necessary samples and funds for sequencing, ensuring sufficient supplies of chemical reagents, availability of and access to high quality technologies and sequencing machines, having and retaining trained staff to carry out sequencing and analyse the results, and having the supporting systems and principles needed to share data quickly and transparently.[23,24]

Global sequencing datasets are heavily skewed towards the developed world

Uneven availability of these pipeline elements leads to significant disparities in the availability of sequencing data across the world (Figure 6). Sequencing data cannot be shared if there are no samples being sequenced in the first place, and it cannot be used effectively for analysis and interpretation if the relevant metadata is not captured at the point of sequencing. Current datasets are therefore weighted towards both high-income countries and urban centres, where sequencing capability, and the infrastructure to support it, is most likely to be found.

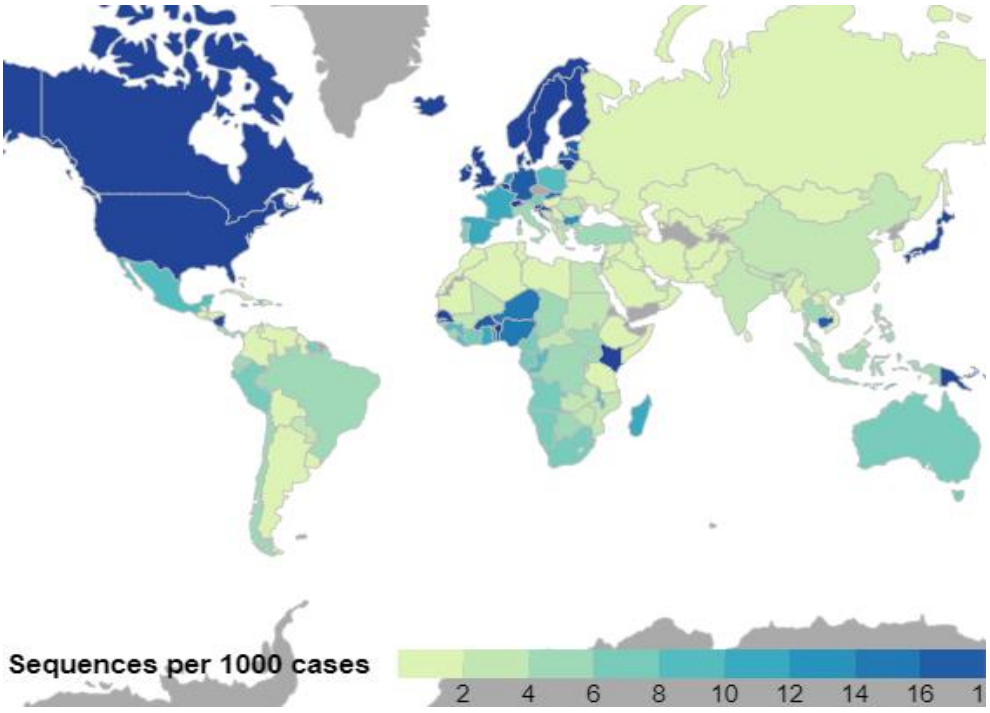


Figure 6. Number of genomic sequences shared via the GISAID Initiative per 1,000 cases (Source: covidcg.org September 2022)[2]

[2] We gratefully acknowledge all the Authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this Figure is based (Shu and McCauley, 2017).[23]

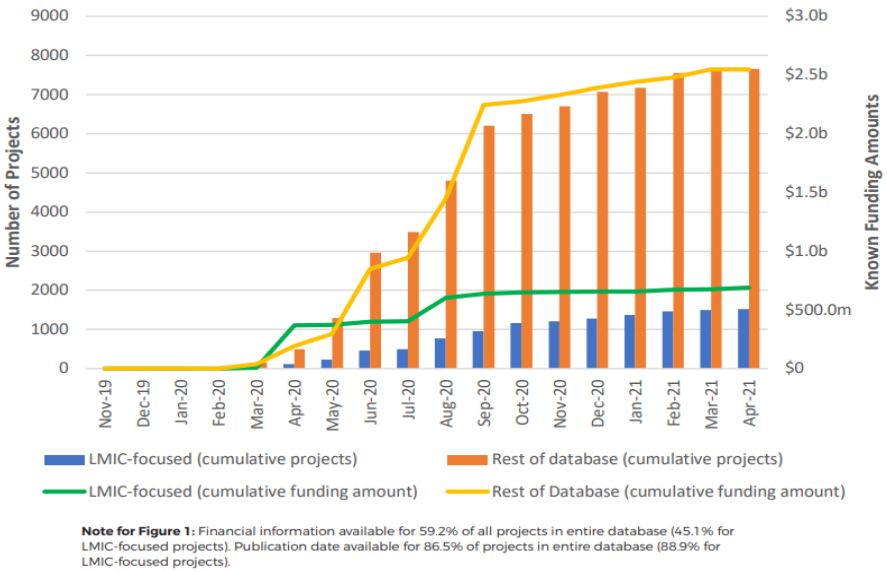
3.2. The need for rapid mobilisation of funding

Funding for sequencing was slow to be mobilised and unevenly distributed

Following the outbreak of the pandemic, it took some time for funds to be allocated in a rapidly emerging landscape of research projects (Figure 7). Access to funding to scale-up sequencing capability was a global challenge, but laboratories in low- and middle-income countries were particularly constrained, a problem exacerbated by a lack of sustained investment in health research predating the pandemic. In the context of COVID-19, this meant that rapidly scaling up sequencing capacity in response to the outbreak was next to impossible in many parts of the world. For those countries with limited capacity WHO identified SARS-CoV-2 international reference laboratories where samples could be sent for sequencing.[25]

> ***"We have a very low budget to react to the challenge that we had... because of the bureaucracy and all this stuff... the money took a long time, almost a year to come to our laboratories to initiate the process"***
>
> *Academic expert*

Figure 7. Cumulative number of projects and known funding amounts by publication date of award information on projects in the UKCDR & GloPID-R COVID-19 Project Tracker (Source: UKCDR and GloPID-R 2021)[26]



Note for Figure 1: Financial information available for 59.2% of all projects in entire database (45.1% for LMIC-focused projects). Publication date available for 86.5% of projects in entire database (88.9% for LMIC-focused projects).

Genomic surveillance efforts remain fragmented, despite longstanding WHO initiatives

The need to decisively improve implementation, coordination and cooperation in the area of collaborative surveillance has been acknowledged in the G7 Pact for Pandemic Readiness.[27] Globally networked surveillance and research to prevent and detect emerging or escalating infectious diseases remains uneven, despite longstanding WHO-led initiatives such as GOARN (the Global Outbreak Alert and Response Network)[28] and GISRS (the Global Influenza Surveillance and Response System).[29] The newly-inaugurated WHO Hub for Pandemic and Epidemic Intelligence,[30] based in Berlin, the Rockefeller

Foundation's Pandemic Prevention Institute (PPI),[31] the Africa CDC Institute of Pathogen Genomics[29] and the ACT-A diagnostic working group lead by FIND are among several initiatives that have emerged during the pandemic with the aim of enhancing genomic surveillance capabilities.[33] The G7 Pact sets out the intention to strengthen a global network approach that will break down barriers to rapid sharing of information, data and samples across borders and sectors on a voluntary basis, following a logic of scientific collaboration and excluding any commercial or industrial benefit. Also significant will be WHO global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022-2032, which provides a high-level unifying framework to leverage existing capacities, address barriers and strengthen the use of genomic surveillance in the detection, monitoring and response to public health threats.

"It will be very good for the world if, even in developing countries, we could have a continuous source of funding for studies on infectious diseases and emergence... This will have connections with the sharing of data as one of the requirements could be that you have to share your data. I think that many people are working on this in funding agencies around the world."

Academic expert

3.3. Securing access to technology and reagents

Sequencing technology has developed rapidly in recent years, but 'dark spots' remain

The increased affordability and widespread availability of sequencing technologies meant that many nations were equipped with the tools needed to contribute to large-scale genomic sequencing efforts. Technology for genomic sequencing has also improved over time, with modern next-generation sequencing (NGS) technologies being applied at increasing scales since the Ebola and Zika outbreaks.[34] Infrastructure providers consulted during this study noted that bolstering sequencing and data-sharing capacity in business-as-usual scenarios, for example through maintaining access to relevant sequencing technologies and analytical tools, is vital to ensuring sequencing efforts can be scaled up rapidly in emergencies. In the long term, increasing available sequencing capacity will also ensure that 'dark spots' in surveillance efforts are limited.

"A lot of infrastructures and databases are run on piecemeal approaches and run through, very often, the goodwill of people. We have to think about the value of maintaining really good repositories and really good publicly available analysis tools because they've proved to be absolutely vital."

Academic expert

Uneven access to reagents for genomic sequencing

Access to reagents for sample preparation is essential to the timely sequencing of SARS-CoV-2 genomes, and the subsequent sharing of sequencing data. At the extraction phase

limits global capacity to identify and track viruses

of genomic sequencing, reagents are needed to ensure the isolation of RNA from cells and tissues collected in test samples. However, the pressure of the pandemic highlighted the global shortage of diagnostics, reagents, and consumables, with the limited access to relevant materials being particularly acute in LMICs. The cost of sequencing reagents for African nations, relative to those in Europe, similarly limited their sampling and sequencing capacity. Additionally, the closure of land borders and airspaces during the pandemic made real-time contributions to global sequencing from LMICs unrealistic, as samples had to be shipped to laboratories elsewhere for sequencing.[35] This in turn contributed to significant disparities in sequencing volumes across the world (see Figure 6, above). As a result, capacity to track the spread and mutation of the coronavirus worldwide has been limited, creating opportunities for variants to emerge and spread undetected.

"The logistics and the infrastructure will affect the sequencing, which will affect the data sharing. At one point during the outbreak, it was getting fairly down to the bone on getting the reagents to be able to sequence, because everybody was sequencing."

Academic expert

3.4. Training and retaining skilled individuals

Having trained staff to carry out and interpret sequencing is crucial

Retaining a skilled research base has been a longstanding challenge in LMICs, exacerbated by a lack of political will to sustainably finance health research systems.[36] During COVID-19 the limited research capacity in LMICs was exposed, resulting in an increased reliance on research collaborations and external funding, often from the 'Global North'.[35] Weaknesses across multiple stages of the sequencing pipeline have been highlighted, including frontline tasks such as dispatching samples, entering metadata alongside samples and datasets, and interpreting sequencing data in public health contexts. To address limitations in research capacity in the short-term, centralised genomics hubs at strategic locations were used to collate and analyse samples from multiple African countries.[35] In the longer term, researcher training and development at the individual level must be linked with public health capacity building (including surveillance).[26]

Case study: Canadian Bioinformatics Workshops

Founded in 1999, The Canadian Bioinformatics Workshops series began offering one- and two-week short courses in bioinformatics, genomics and proteomics in response to an identified need for a skilled bioinformatics workforce in Canada.[37] Over the following 20-year period, thousands of individuals nationwide across public health, academia, government and industry received training in the handling of and analysis of genomic data.

Andrew McArthur, Associate Professor at McMaster University Canada, runs a lab dedicated to genomic surveillance of infectious pathogens. As he explains, the benefit of long-term, cross-sector investment in genomics skills became immediately apparent as the pandemic unfolded:

'It meant that when we called regional hospitals and said: "We need to do surveillance for variants", there was usually someone there who had taken one of those courses and knew what we were talking about. So this was a wonderful thing.'

Bioinformatics skills are a key gap in LMICs

The interviewees consulted in this study drew particular attention to a gap in bioinformatics training – a subdiscipline of genomics research using computational methods to analyse and disseminate biological information, as well as to support the troubleshooting of sequencing failures. During the pandemic, the issue was especially prominent in LMICs where there has been a significant disconnect between researchers capable of conducting genomic sequencing, and those capable of interpreting the data within a public health context. A growing interest in bioinformatics training, beyond the end of the COVID-19 pandemic, is beginning to be met by research and development funders. For example, Wellcome have designed distributed bioinformatics training sessions, based on data from the COVID-19 pandemic, and GISAID, EMBL-EBI and others have similarly continued to provide bioinformatics training workshops.[38,39] Further investment in this area will be crucial in enhancing global preparedness for future pandemics.

"One of the things that remains a challenge... is how to incorporate bioinformatics in the analysis of all sequences, and from the sequence to epidemiological data. This is very well done by Nordic countries, by the UK and sometimes in the US, but this is not commonplace for many countries."

Academic expert

PRECONDITIONS FOR SUCCESSUL SEQUENCING - LESSONS LEARNT FOR OPEN SCIENCE

- For some global crises, it is essential to obtain representative data from all parts of the world, but not all countries and regions have sufficient data-generating capacity or trained human resources to collect, disseminate, and analyse data for global surveillance efforts.
- Interventions designed to enhance the availability of relevant data must ensure they identify and tackle the root of the problem, which in many cases will be a lack of underlying research capacity rather than inadequate uptake of open sharing practices.
- The training and development of skilled individuals in fields such as bioinformatics is critical if the benefits of open data are to be realised in practice.

Research Consulting

4. The geopolitics of genomics

Disparities in sequencing capacity and capability are compounded by unresolved concerns over access and benefit-sharing. Digital sequence information, which includes viral genomic data, is not covered by the Nagoya Protocol to the Convention on Biological Diversity and there are divergent international perspectives on whether it should be considered a common good. Even where sequencing data exists, governments and other actors may be reluctant to share it on the grounds they will receive no benefit in return and may even suffer adverse political and economic consequences.

Research Consulting

4. The geopolitics of genomics

4.1. Gaps in sequencing are compounded by variable approaches to data sharing

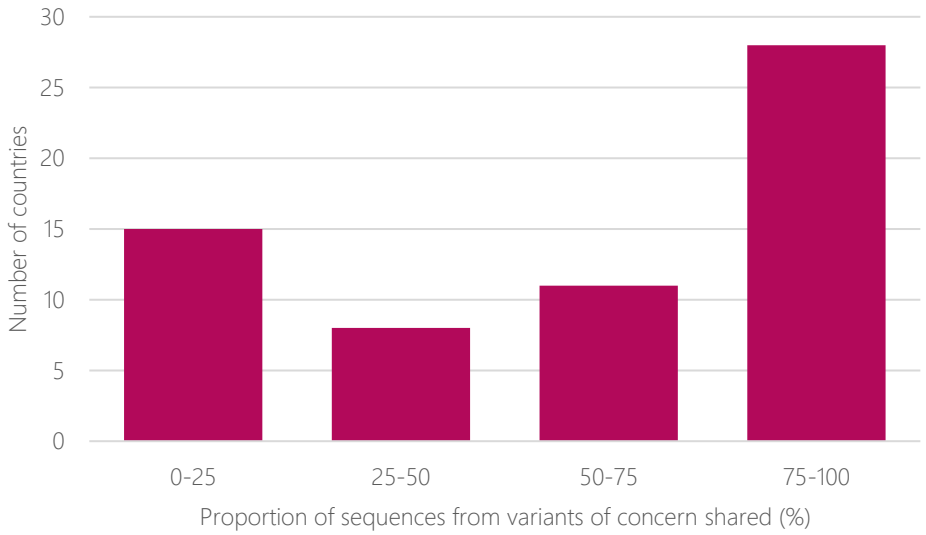
Even where sequencing is undertaken, the results are not always being shared

Disparities in sequencing capacity are compounded by two factors. Firstly, not all sequence data are shared, and secondly there are significantly variations in the quality of the data that are made available. During COVID-19, these inconsistencies in data generation, sharing and quality globally have contributed to racial and social biases in datasets, potentially skewing the subsequent public health outcomes that they inform.[40]

There are striking variations between G7 countries in their approaches to sharing

As Sachs et al have noted, a successful pandemic response relies on, 'an ethical framework of prosociality - the orientation of individuals and government regulations to the needs of society as a whole, rather than to narrow individual interests'.[2] In the field of genomics, a prosocial approach means making relevant data available globally, rather than retaining it for local and/or national use alone. A recent study by Chen et al found that, of 62 countries that report sequencing volumes, more than a third had uploaded fewer than 50% of their identified sequences to public repositories (Figure 8).[9] Within the G7, the countries with the highest levels of sharing were the United States (100%, with some accompanying caveats over data completeness) and Japan (95%). Canada had the lowest level of public availability (26%), with the UK, Germany, Italy and France all sharing at least 50% of their reported sequences. These data should be treated with caution, as there will be a number of different factors at play, including variations in how a sequence is defined and counted and differences in the quality control thresholds used for submission. Nevertheless, they indicate that a significant proportion of available sequencing data is not being shared internationally, typically due to variations in capacity (human and technical infrastructure) rather than by design.

Figure 8: Proportion of sequences from variants of concern shared by 62 countries (Source: Mallapaty, 2022)[41]



Global variations in levels of sharing are even greater

About 27% of high-income countries uploaded less than 50% of their total variant of concern sequences, while 56% of low-middle-income countries uploaded less than 50% of their total variant of concern sequences. In Thailand, for example, the publicly available proportion of Alpha, Beta and Delta variants was 13.6%, 15.4% and 9.8%, respectively. As such, it has been suggested that more than 80% of variant-related sequences are not uploaded to public databases on a timely basis. Similarly, high income countries seen to be uploading less than 50% of their sequences include Austria, Cyprus, Greece, Hungary and Panama.[9]

4.2. Political tensions underlie global variations in sharing

Political and economic concerns may lead governments to withhold sequencing data

Public health laboratories undertook a significant amount of sequencing during the pandemic, particularly in high-income countries. Academic norms alone cannot, therefore, explain the size of the gap between reported and deposited sequences. Instead, many observers point to the political ramifications of sharing sequence information, including the repercussions of being the first country to report a new variant of concern. In some countries, governments need to review and approve sequences before they are uploaded. For nations that are highly dependent on tourism, this can be problematic. Firstly, from a public health perspective, the mandatory review and approval processes can take time, allowing variants to spread between populations. Secondly, tourism-dependent nations have a strong economic incentive to withhold adverse data in the context of a pandemic or smaller-scale outbreak.[41]

Case study: The consequences of viral genomic data sharing in South Africa

In November 2021, genomic surveillance teams in South Africa and Botswana detected a new SARS-CoV-2 variant associated with a rapid resurgence of infections in Gauteng province, South Africa. Within three days of the first genome being uploaded, it was designated a variant of concern (Omicron, B.1.1.529) by the World Health Organization and, within three weeks, had been identified in 87 countries.[42]

The speed with which the Omicron variant was identified and designated as a variant of concern can be attributed in large part to the rapid sharing of genomic data and transparent reporting of its implications. Within just a few hours of the sequences being shared, international scientists were able to confirm the variant's potentially worrying mutations. It was quickly dubbed B.1.1.529, prioritised for further study and evidence submitted to the WHO.[43]

The day after the new variant was announced, dozens of countries announced travel restrictions on countries in Southern Africa. 56,000 hotel bookings were cancelled in Cape Town alone, and tourism officials estimated daily losses at R200,000,000 (circa $12.6 million) in that city alone.[44] A statement by the South African foreign ministry strongly criticised the travel bans, which it considered:

> ***"...akin to punishing South Africa for its advanced genomic sequencing and the ability to detect new variants quicker"***

(Department of International Relations and Cooperation, 2021).[45]

Professor Tulio de Oliveira (2021), one of several scientists at South Africa's University of KwaZulu-Natal who first identified the Omicron variant, put it more succinctly:

> *"If the world keeps punishing Africa for the discovery of Omicron and 'global health scientists' keep taking the data, who will share early data again?".*[46]

4.3. The status of digital sequencing information remains contested

Mechanisms for access and benefit sharing of genetic material exist, but are not universally adopted

The 1992 Convention on Biological Diversity (CBD) established that genetic resources are under national sovereignty and the 2014 Nagoya Protocol to the CBD set out a detailed mechanism for access and benefit sharing (ABS) of "genetic material of actual or potential value" where genetic material is "of biological origin containing functional units of heredity" It also requires countries to pay due regard to present or imminent emergencies that threaten or damage human, animal or plant health, as determined nationally or internationally (Article 8(b)).[3] While there are 133 parties to the Protocol, and a number of these have developed extensive guidance on its implementation, non-signatories include Canada, China, Russia and the United States, and negotiations on a global multilateral benefit-sharing mechanism for genetic resources in transboundary situations remain ongoing.[47]

The relationship of digital sequence information to the Convention on Biological Diversity and the Nagoya Protocol is unresolved

The term 'digital sequence information' (DSI) is used to refer to a wide range of digitised genetic information, including pathogen genomic data such as the sequences of SARS-CoV-2 considered in this report. However, there is no accepted definition of DSI, and it is unclear whether DSI should be considered a genetic resource, the utilisation of a genetic resource or its application.[48] Additional complexities emerge in cases where samples are transferred beyond national boundaries for sequencing, for example in reference laboratories in other countries. A science and policy-based process, which includes the convening of an Ad Hoc Technical Expert Group (AHTEG), has been established by the CBD to address the issue of digital sequence information on genetic resources.[49] This has allowed six policy options for access and benefit sharing to be identified as shown in Table 2, together with a set of criteria against which to access them.[50] However, the parties to the CBD and the Nagoya protocol have divergent views on the most appropriate policy options, and it is unclear whether agreement on a preferred option can be reached.

Table 2: Policy options for access and benefit sharing of digital sequence information (source: Co-leads' report on the work of the informal co-chairs' advisory group on digital sequence information on genetic resources, 2021)[51]

Policy option	Description
0 Status quo	Parties have not agreed on how to address access and benefit sharing for digital sequence information of genetic resources
1 DSI fully integrated	DSI is fully integrated into the approach of the Convention on Biological Diversity and Nagoya Protocol with usage subject to prior informed consent (PIC) and mutually agreed terms (MAT) (i.e. DSI is treated in the same way as the underlying genetic resources). A tracking and tracing system would be required to not only determine the country of origin of each DSI record

	uploaded to relevant databases but also how the DSI was being utilized and by whom so researchers could comply with that country's ABS obligations.
2 Standard mutually agreed terms	There is benefit-sharing from the use of DSI, but it is decoupled from access, i.e. there are mutually agreed terms but no prior informed consent. This alternative requires downstream monitoring of DSI use for implementation or enforcement, and monitoring
3 No prior informed consent, No mutually agreed terms	A payment or contribution for access goes into a multilateral fund. It avoids the need for tracing the origin of the genetic resource from which the DSI was extracted, or the need to monitor the downstream utilization of the product or service derived from DSI. This option includes various possible forms of payments and contributions, with one sub-option being linked to the DSI itself, and the other being separate from the information itself.
4 Enhanced technical and scientific capacity and cooperation	Under this option, systematic and mandated technical and scientific cooperation and capacity development related to DSI are promoted. There is enhanced capacity support for developing countries aiming to ensure that each country has improved/expanded capacity and opportunity to generate, access and use DSI to its full potential.
5 No benefit-sharing from digital sequence information on genetic resources	This option entails that the international community decides that no explicit benefit-sharing is necessary from the use of DSI from genetic resources and, thus, no additional mechanisms are proposed for benefit-sharing to be implemented.
6 1 per cent levy on retail sales of genetic resources	A multilateral fund would be established and financed through a 1 per cent levy on all retail sales of goods in developed countries arising from the utilization of genetic resources. Funds would be distributed through a competitive project-based approach for conservation and sustainable use by indigenous peoples and local communities and others.

"I think there is a problem with the Nagoya Protocol and the property of sequences. I personally think that the pathogen should [not be included] in the Nagoya Protocol and that there should be no property owner of a pathogen."

Academic expert

4.4. There is no international agreement on access and benefit-sharing for DSI

In the absence of agreement, DSI is perceived as a loophole that inhibits fair and equitable benefit-sharing

DSI has historically tended to be shared in online open-access databases such as INSDC, where use is disconnected from physical access and accompanying permits. Some biodiverse nations, many of which are low- and middle-income countries (LMICs), claim their sovereign rights have been undermined because any potential gains from DSI through commercialization or development of diagnostics and vaccines are not shared with them, as they would be with a genetic resource. Thus DSI is perceived as a loophole that inhibits fair and equitable benefit-sharing.[52] In the case of COVID-19, some of our interviewees questioned why researchers in LMICs should be expected to share genomic data openly if the vaccines and other public health benefits that it informs are not also shared on an equitable basis.

> *"For the first time, I've started seeing pure researchers saying "why should we provide this information... because we're not getting access to the technologies, the knowhow or the vaccines, and a lot of our people are dying because of the delays. So why should we actually share our information in the first place?""*
>
> *Policymaker*

Case study: Proposals for a multilateral benefit-sharing framework for digital sequence information (DSI)

In a recent paper in Nature Communications, 41 researchers from 17 countries proposed a benefit-sharing framework for DSI that is multilateral in nature and addresses five fundamental objectives:

'1. Open access. Any future benefit-sharing system must guarantee open data access, which is required to be able to use and understand DSI. Only open access enables efficient and broad scale knowledge generation and capacity building. The existing core DSI infrastructure already fulfils these requirements.

2. Simplicity. The DSI data ecosystem is highly complex, even for expert users. For benefit-sharing to happen, the policy framework must be simple. If a complex regulatory layer is added on top of a complex technical system, it is doomed to fail.

3. Harmonize. The DSI dilemma is an opportunity to learn from current inefficiencies in ABS and minimize transaction costs. Furthermore, as DSI is being discussed in multiple international fora, DSI users need a harmonized framework to address benefit-sharing.

4. Biodiversity. Any mechanism needs to effectively support biodiversity conservation and sustainable use (the first two objectives of the CBD). The framework should incentivize and reward biodiversity knowledge generation, and fill in the blank spots on the world map of biodiversity.

5. Fairness. The framework should treat all users and providers fairly and create a level playing field by facilitating both access and compliance evenly across the globe. A multilateral system could nevertheless include opportunities for country-specific recognition and differentiated distribution of funds.'[52]

Referencing the INSDC, which contains over 200 million annotated sequences, the authors argue that 'a bilateral system, modelled on the principles of the Nagoya Protocol, that required permission between the end-user and the country of origin for every sequence and user transaction, would be prohibitively complex, affect data interoperability, and be ill-suited for generating knowledge' (ibid.). Instead, the authors propose to de-couple access to DSI from benefit-sharing, with new monetary mechanisms established upstream of DSI generation, and benefit-sharing based on the entire global DSI dataset and not on individual sequences.

4.5. Political considerations constrain sharing within as well as between nations

Disparities in access and benefit-sharing arise at multiple levels

As Cochrane *et al* have observed, 'data sharing for public health purposes is most effective at global level and with full openness', but the lack of an agreed mechanism for access and benefit-sharing weakens the case for pathogen genomic data sharing by LMICs.[53] To these concerns can be added the lack of protection afforded Indigenous Peoples within the open data and open science movements, and local or regional disparities in access to data and benefits that mirror those seen at the global level and create similar disincentives to sharing.[54]

"Even within wealthy countries where there's disparities, like the United States, the benefits of gathering and sharing the data aren't fair and equitable to the communities from whom the samples might come. That threatens the timeliness of genomic sequencing data, and it threatens your ability to get sequence data from mild cases."

Academic expert

A diverse and inclusive approach is essential

Interviewees noted that political considerations can also inhibit the gathering and sharing of viral genomic data within individual nations. Barriers to participation in sequencing include a lack of diversity within project teams and the genomics profession; an underestimation of time and resources required to recruit diverse participants; recruitment and information resources that fail to visually represent certain communities; and contemporary and historical experiences of discrimination.[55] Political tensions emerged in many countries over where limited resources for sequencing should be directed, and how the resulting data should be shared between federal and state authorities, or between academic laboratories and public health agencies.

> ***"I think the largest lesson that I've learnt from this situation is that you need to pull in a diverse set of voices... We have to think about how the least among us are going to benefit from [sequencing] and why they might care about it. And if we fail at that task, then we will fail to get representative data every time."***
>
> *Academic expert*

Case study: The GLOPID-R Principles of Data Sharing in Public Health Emergencies

The GLOPID-R Principles of Data Sharing in Public Health Emergencies (2018) provide a framework for timely data sharing during an outbreak.[56] They can be used to support data sharing during such emergencies to inform pandemic preparedness, public health responses and the development of vaccines, diagnostics and therapeutics.

These data sharing practices were assessed during past outbreaks (Ebola outbreak in West Africa; Middle East Respiratory Syndrome (MERS), Cholera and Yellow Fever outbreaks in Africa and China, and Zika in Latin America).[57] GLOPID-R presents a framework for the timely sharing of data in emergency scenarios, considering the following aspects: accessibility, transparency, equality, fairness, data quality, and ethics.

Key barriers to data sharing and a roadmap of five recommendations and priorities for GLOPID-R funders were published in 2019.[58] The barriers to data sharing, addressed by these recommendations for regional outbreaks, are relevant to the COVID-19 pandemic and living systematic reviews such as this show that outbreaks are exacerbated where resources are lacking.[59,60]

THE GEOPOLITICS OF GENOMICS - LESSONS LEARNT FOR OPEN SCIENCE

- The willingness and ability of different actors to share data during emergencies is heavily influenced by pre-existing geopolitical considerations. In many cases, these led to the pursuit of self-interest over shared interests during the COVID-19 pandemic, resulting in suboptimal outcomes for all.
- Effective data sharing in an emergency context depends on prior work to understand the motivations of different stakeholders and reach agreement through relevant international fora on mechanisms to promote prosocial behaviours.
- A diverse and inclusive approach is needed at global, national and local levels in order to maximise the availability and representativeness of scientific data.

Research Consulting

5. From personal choice to community norm

In academia, the longstanding practice of withholding data until the point of publication runs counter to the needs of emergency response. Meanwhile, in public health, data governance concerns and the need to comply with data protection legislation frequently inhibit sharing. While the pandemic highlighted the value of pre-existing collaborations, it also exposed the different sharing cultures within research and public health, and the need to improve the interfaces between these two communities.

Research Consulting

5. From personal choice to community norm

5.1. Sharing of pathogen genomic data often occurs too late

Practices for sharing pathogen genomic are well-established…

The field of genomics has often been cited as the branch of biology that has led the way in data sharing.[61] Sequencing data is easy to share via established international databases and pathogen genomic data (in isolation) is unencumbered by the privacy and data protection considerations that often constrain data sharing in human genomics. Funders and leading journals in the field of genomics also have long-standing policies for data sharing. For example the NIH Genomic Data Sharing Policy become effective in 2015,[62] and the Japan Agency for Medical Research and Development's Policy was established in 2016.[63] In April 2016, WHO issued a clear policy statement on data-sharing in the context of public health emergencies stating: *"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"*.[64]

… but current sharing paradigms fail to meet the needs of emergency response

Despite these advantages, and the vast numbers of sequences that have been made available over the course of the pandemic, especially in GISAID (see section 6), there is broad consensus that existing sharing paradigms are not well-adapted to an emergency context in which near real-time sharing is the desired goal. In the COVID-19 pandemic, the sharing of sequencing data continued to be delayed in a significant number of cases, often until the point of publication. However, the collection to submission lag was reduced as the pandemic progressed, suggesting that as time went on more-streamlined approaches were developed in local and national contexts.

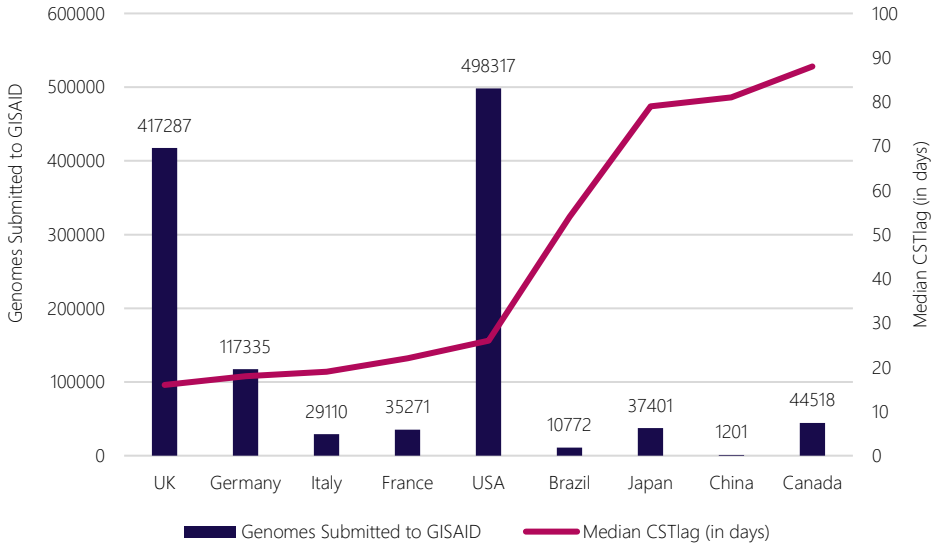
> ***"In pathogen genomic data very specifically it's clear to me that the predominant paradigm is inadequate. We need the data to be available way before any associated publications."***
>
> *Policymaker*

The speed of sequence submission varies significantly between countries

An analysis of the collection to submission time (CST) lag for SARS-CoV-2 sequences to GISAID, prepared in mid-2021, showed wide variation between countries in the speed with which sequences were submitted.[65] At 16 days, the median CST lag from the United Kingdom, which submitted 417,000 genomes, was almost a week faster than the lags of 25 and 26 days for the rest of Europe and the United States, which submitted 590,000 and 489,000 genomes, respectively. The median lag for Japan's 37,000 genomes was 79 days, while for Canada the lag was over five times as long as the UK's, at 88 days for 44,000 genomes (Figure 9). As the pandemic progressed, the turnaround time shortened in all regions, but wide variations between countries remain.[9] This suggests

that speed of sharing to GISAID is determined by a range of national and regional factors, at least some of which can be influenced by governments and policymakers. Data may of course have been shared in repositories other than GISAID at an earlier date, and analyses undertaken at a national-level may have been shared directly with entities like the WHO. Nevertheless, as Kalia et al have observed, those countries with the shortest median CST lags, such as the UK and Denmark, have strong public health systems allowing efficient sample and metadata collection and high levels of coordination between sample collection centres, RNA isolation laboratories and sequencing labs.[61]

Figure 9: Collection to submission time lags for selected countries (based on submissions to GISAID as of 27 May 2021) (Kalia et al, 2021)[65]

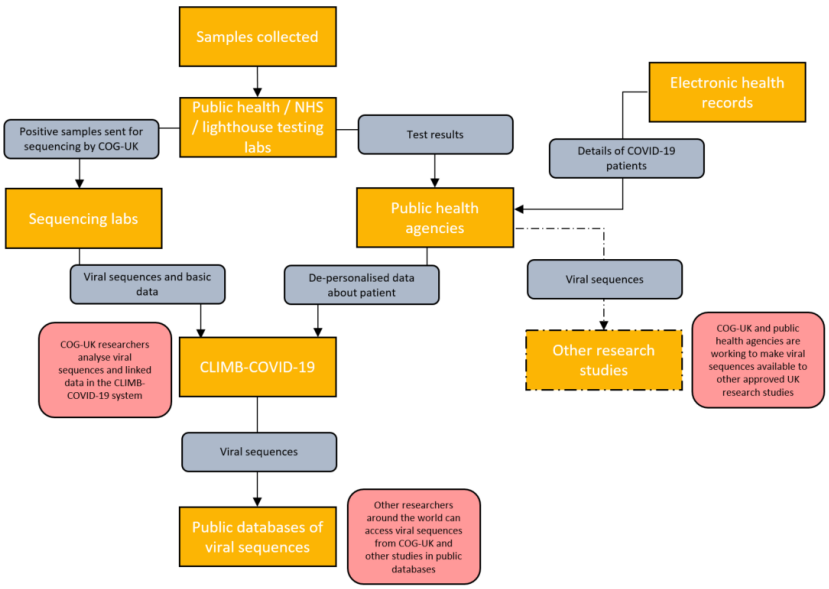


Case study: the COVID-19 Genomics UK consortium (COG-UK)

In the UK, the COVID-19 Genomics UK consortium (COG-UK) was established to coordinate testing and sequencing strategies. Since it was established in 2020, COG-UK has played a critical role in managing the collection of samples, coordinating the sharing of samples and test results between sequencing labs and public health agencies, and depositing data in public databases of viral sequences. The COG-UK data flow below shows the coordination needed between public health agencies and sequencing labs.[66]

As a consortium of four UK public health agencies and 16 academic institutions, high levels of coordination were needed between groups to ensure CST is kept to a minimum. One of our interviewees highlighted how beneficial the tight links between labs and public health agencies had been during COVID-19:

"I think the beauty of it was that there was a COG-UK lab, and our campus, next door to the testing lab. So we could get positives to them, and they could create a genome sequence within 24 hours. And so we were able to spot local transmission clusters and pass that information over to local public health immediately."



Thanks to sustained investments in academic research and microbial genomics, the UK was in a strong position to develop and support initiatives such as COG-UK:

"The reason I think we were so phenomenally successful from the genomics perspective, is because of how much investment has gone into academic research and microbial genomics. The UK is the *place in the world for microbial genomics and viral genomics, not just with the Sanger Institute, but with the amount of world leading research groups that do microbial genomics in the United Kingdom… Undoubtedly, that investment we've had in genomics research in the UK is why we did so well."*

By September 2022, over 3 million SARS-CoV-2 viral genomes have been sequenced according to the CLIMB Genome Counter– a public database of viral sequences supported by COG-UK infrastructure – allowing other researchers to access and build on openly available data. From a public health perspective, COG-UK's role has been critical. Throughout the pandemic, the organisation has provided sequencing data in real time to UK public health agencies and advisory groups, such as the UK's Scientific Advisory Group for Emergencies (SAGE), continuously informing the national pandemic response with the latest available data.

A critical impasse occurs at the stage of submitting sequences and metadata

Successful provisioning of sequences is the result of a number of conditions, as outlined in the preceding sections. However, as Bernasconi *et al* have noted, 'the most critical 'impasse' is met at the stage of submitting sequences and associated metadata… which has become almost a deliberate political act in the current times' (2021, p.672).[23] While the pandemic has shifted practice in favour of greater openness, it has also exposed a number of longstanding social-cultural factors that continue to inhibit rapid sharing of sequence data. Pratt and Bull conceptualised data sharing in epidemics and pandemics as a critical tension between utility norms (such as rapid, real-time sharing for effective response) and equity norms (such as researcher recognition and equitable access).[59]

"At the very beginning, many people worked as we always did in the past. We were very jealous about our sequences, and people were very afraid that 'if I share my sequences…

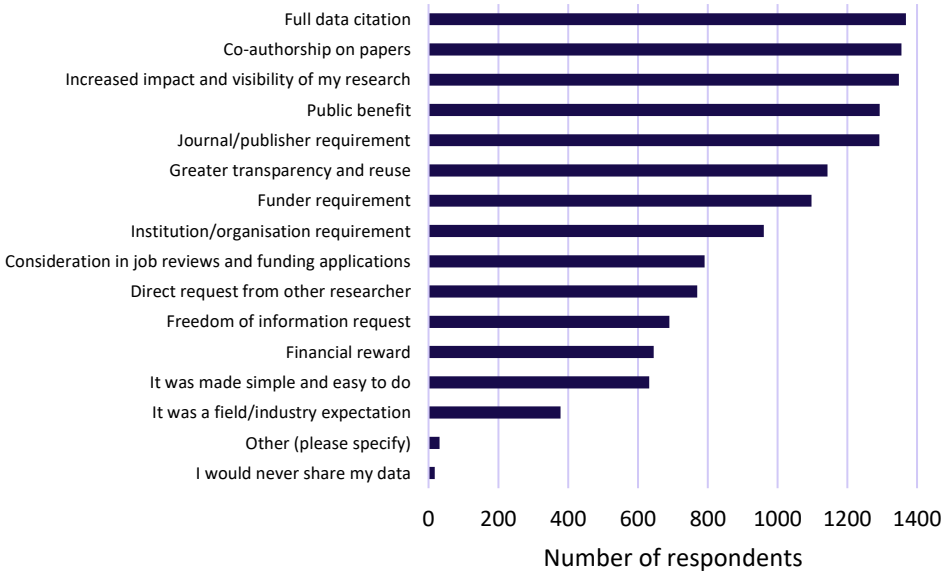
somebody else will use them before me'. This [attitude] was, let's say, a little deprecated through the time of the pandemic."

Academic expert

Genomics suffers from the same barriers to data sharing as other life sciences disciplines

Our consultation with genomics researchers confirmed that, while national context is an important factor, a perceived lack of academic credit remains a barrier to the timely and open sharing of sequencing data. In this respect, genomics mirrors the broader picture in science, where data citation, co-authorship and increased research impact are the primary motivations for data sharing (see Figure 10). While public benefit assumed greater importance as a motivator for sharing during the pandemic, altruistic motives alone were not sufficient to lead to a widespread change in scientific data sharing practices. It should be noted (see Figure 10) that publisher and funder requirements requiring sequence data to be shared are also influential and their impact and effectiveness could be worthy of future detailed study.

Figure 10: Circumstances motivating researchers in biology and medicine to share their data (Digital Science, 2020, n=2,130)[3]



Early sharing must become a community norm, rather than a personal choice

Current incentives within academia mean the early sharing of genomic viral data continues to be largely unrewarded by the institutions and funders that support researchers, rather than being a community norm. Progress relies on developing new

[3] The State of Open Data is the longest-running longitudinal survey and analysis on open data and sharing data. It is run by Digital Science, Figshare and Springer Nature and respondents can be from any discipline. The 2020 survey dataset was selected for further analysis for this study because of its potential to review broader issues of data sharing in Biology/Medicine; for its questions on the COVID-19 pandemic; for its global coverage and number of respondents; and for its licensing under CC-BY. The raw dataset for 2016-2020 was downloaded as an Excel spreadsheet and cleaned to align it with the criteria for usable responses in the published report (i.e. only respondents who have published within the last 5 years are used) and to include only respondents from the 2020 survey. There were 4945 usable responses in total to the 2020 survey. Out of the usable responses, 4563 provided their main area of interest. 47% (2130) described Biology (945) or Medicine (1185) as their main area of interest and these were selected for the additional analysis presented in this study.

norms within the relevant disciplinary communities, as well as robust demonstration and increased awareness of the benefits and impacts of sharing at national and international levels, as outlined in the previous section.

5.2. The role of informal sharing

Genomic sequencing efforts must be supported by close collaboration of relevant stakeholders

Many stakeholders are involved in developing evidence-based public health responses. In the context of the COVID-19 pandemic, connections between key stakeholder groups (i.e. those with access to samples and those with capability to sequence and analyse them) and other networks that shared information and collaborated internationally have been critical (see section 6.1). The World Health Organization has also developed an initial list of stakeholders (Table 3) that are integral to ensuring that genomic sequencing activities are able to solve questions of public health importance.

Table 3. Stakeholders to be engaged when developing sequencing programmes (World Health Organization 2021)[13]

Stakeholder group	Role/capability
Public health bodies	• Commission or deliver SARS-CoV-2 sequencing programmes • Answer key policy questions • Secure widespread collection of particular diagnostic samples and metadata
Diagnostic laboratories	• Have access to SARS-CoV-2 samples for sequencing • Can provide positive samples and metadata directly to sequencing facilities • May be capable of managing in-house sequencing
Sequencing facilities	• May have the bioinformatic capacity to generate consensus virus genomes • May provide raw data that must be further processed elsewhere to generate genomes.
Analytical groups	• Conduct genomic analyses and determine which samples should be sequenced
Infection prevention and control teams	• May be based in hospitals or treatment centres • Can support the identification of emerging disease clusters • Are well placed to identify cases that would be useful for sequencing • Can act on subsequent findings regarding transmission clusters
Occupational health services	• Can help to identify potential transmission clusters or transmission routes that can be investigated using virus genomic studies • Can implement infection prevention and control activities emerging from the results
Patients	• Should be engaged to ensure that they understand how sequences and metadata are being used and shared, and benefit from results

International collaboration is critical to timely sharing

In the absence of open, timely sharing as the default model, rapid sharing of sequencing data around the globe often remained reliant on individual relationships and pre-existing networks. From the sharing of biological samples to the sequences themselves, the role of these informal channels in alerting the world to emerging threats cannot be overstated.

The further development of scientific collaborations in the field of genomics between high- and low-income countries, and between the Americas, Europe and the Asia-Pacific region, constitutes a key part of world's pandemic warning system. These collaborations should also extend to the sharing of cloud computing resources, bioinformatics tools, knowledge and practice, in addition to data (see section 6.1).

> ***"We have enough colleagues that have been in larger consortia – Europe-wide, or across the world - beforehand. And these were our beacons, they helped us to establish the international connections to different countries."***
>
> *Academic expert*

As the pandemic progressed, local and regional networks came into their own

As SARS-CoV-2 took root in individual countries and communities, the focus of collaboration shifted from the international to the national and local level. Sourcing, sequencing and sharing of samples relied on the development of robust linkages between public health, academic and industry stakeholders. As with other aspects of the pandemic response, however, these partnerships were not always as smooth as if collaboration channels had been pre-established (McKinsey, 2021).[67]

> ***"The main takeaway point is really close collaboration with the relevant local actors. It may take a state data protection agency a month or so to respond to your request, but if you talk to your local institutional review board, health authority or city diagnostic labs you can have a response within hours. At the state or national level, things were very slow and very sluggish. But that did not prevent us from implementing stuff at the local level."***
>
> *Academic expert*

5.3. Strengthening the public health-research interface

Culture clashes between research and public health actors were common

While many existing networks and collaborations were mobilised to support the pandemic responses, others had to be created from scratch, or rapidly scaled up. Almost every infectious disease program within the US Centers for Disease Control and Prevention generates and analyses pathogen sequence, while agencies such as Public Health England, the Public Health Agency of Canada and the European Centre for Disease Prevention and Control, also have large sequencing programs. However, in many cases, these agencies relied on academic partners to supplement their sequencing capability, and to assist them in assembling, analysing and interpreting genomic data.[68] These new partnerships often exposed significant cultural differences between different communities' approaches to data sharing and re-use which highlights the importance of developing trusted systems that can meet the needs of all stakeholders.

> ***"A key aspect of ensuring we had good surveillance was a strong link between diagnostic laboratories and phylogenetic laboratories. The phylogenetic laboratories were often in academic institutions, whereas routine diagnostic laboratories are in government institutions. So, to create this kind of partnership was critical because that way we could get the specimens quickly."***
>
> *Policymaker*

Public health and research communities have different attitudes to sharing

When it comes to the speed at which sequencing data was shared, there is agreement that both public health and academic researchers fell short, but for different reasons. Public health systems were used to generate data for specific purposes, such as border control, and were often focussed on sensitive activities such as contact tracing. While many had little reason to withhold sequencing data per se[53], they were often unfamiliar with open sharing practices[69], and approaches for pooling and sharing public health surveillance data were not well-established. Academics, by contrast, were typically motivated to get data in the public domain but our interviews indicate that many continued to see publication as the primary mechanism for achieving this. In each case, there were delays to the open sharing of sequencing data. Although the Bermuda Principles of rapid automatic open sharing of genetic sequences have been in place since 1997, many researchers still choose to delay sharing data, or apply an embargo, out of a fear of their findings being 'scooped'.[70,71]

Incentive changes and cross-silo working are needed in research and public health

As public health and research communities become more interdependent, there is a growing need to support complex flows of data and interpretation between the two.[53] Policymakers seeking to encourage the early sharing of genomic viral data must also be prepared to adapt their approaches to the needs of different communities, taking appropriate account of their differing incentives and priorities.

> ***"We are in this bubble of open science and .... [clinical labs] are in their own bubble. Breaking those silos within science is a tremendous amount of work, and a much bigger issue than I ever anticipated."***
>
> *Infrastructure provider*

5.4. Leveraging genomic viral data for public health

Linking genomic viral data with other data types is a necessary condition to effectively respond to public health emergencies

Genomic sequencing data alone cannot be drawn upon to develop public health responses designed to protect human populations. Instead, viral genomic sequencing data must be linked to clinical, social, economic, and epidemiological data to effectively respond to public health threats. For example, using other data sources such as geographical data from smartphones and patterns of viral concentrations in sewage can

determine if an area has an increased number of cases, and thus can be subjected to public health measures such as quarantines, to reduce the spread of the virus.[72] The integration of sequencing data with each of these other types of data involves a range of complex technical and legal issues, but is crucial to leveraging its potential for public health.

> ***"Sequencing alone is not enough. It is one of the tools, and looking back to the alpha variant, it was actually a trend in epidemiology that made them go and look at the sequence...sequence data is not enough, it has to be linked with clinical data."***
>
> *Academic expert*

The quality and completeness of metadata is crucial

Figure 11, prepared by Bernasconi et al (2020), illustrates the various factors that can influence an individual person's measurable observable traits i.e. their phenotype. As seen in the figure, a patient's phenotype is determined by both their genomic make-up ('Patient Genomics' in the grey box, or genotype) and by environmental factors (highlighted in the green boxes).[23] Each box represents a type of metadata (genomic or clinical) that describes the virus-infected patient, and it is possible to connect each patient to the genomic sequence of the SARS-CoV-2 virus. Pulling this metadata together for populations enables powerful studies that can determine for e.g., the role of host genetic factors in susceptibility and severity of the coronavirus pandemic (The COVID-19 Host Genetics Initiative, 2020).[73] The quality and completeness of metadata is essential for these sorts of studies, however the more metadata requested by a platform, the less complete entries are likely to be, and the more difficult robust anonymisation of individuals becomes.

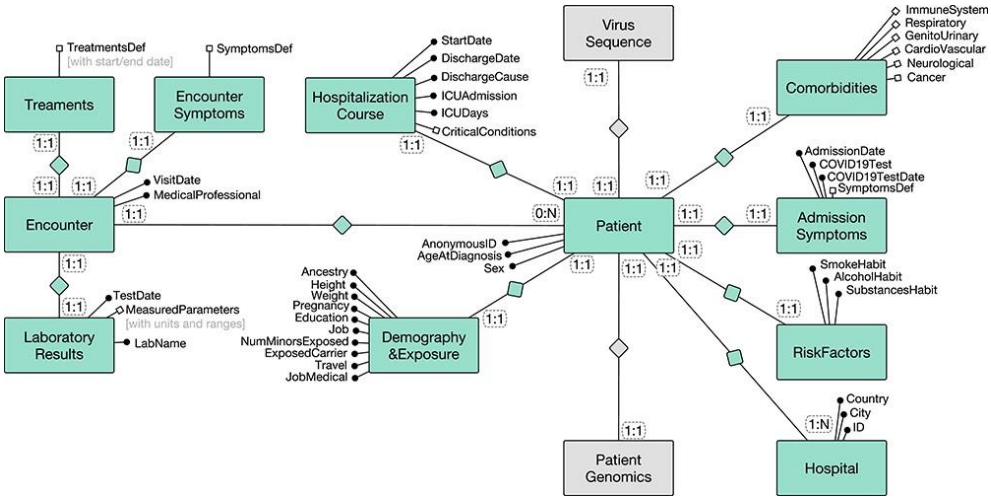


Figure 11: Entity-relationship diagram of the Phenotype Data Dictionary proposed within the COVID-19 Host Genetics Initiative. (source: Bernasconi et al, 2021, PMC Open Access Subset)[23]

Case study: Elevating data analysis: ICODA's efforts to combine data from different sources

The International COVID-19 Data Alliance (ICODA) was convened by Health Data Research UK in 2020 at the outbreak of the pandemic.[74] ICODA was developed when it was recognised that close to 100 high quality data repositories housing COVID-19 data existed in isolation. ICODA aimed to increase the co-ordination across these repositories so that relevant health information from different sources across institutional and geographic boundaries could be easily found and aggregated for analysis but pivoted

instead to fund specific data re-use research projects including ten through the Grand Challenges ICODA pilot programme This demonstrates just how difficult it is to achieve global interoperability of health data.[75–78]

ICODA projects demonstrated that, despite the challenges, COVID-19 health data sharing and re-use was possible, and produced important and clinically relevant findings, with a wide range of research outputs supported by ICODA emerging since 2020.[79] However, a legacy of integrated repositories and standardised and transparent access was not built. A secure trusted research environment was provided through the ICODA workbench only for ICODA funded researchers, and it's not yet clear whether data used in the research projects will be any more accessible in the future than they were before ICODA began.).[81]

The benefits of sharing genomic and clinical data must be weighed against privacy concerns

Genomic and clinical metadata of patients can include personal information such as age, gender, location, comorbidities, etc. Potentially sensitive personal information can also simply be revealed by linking one or more of pathogen genetic sequences, demographic metadata, human genomics, clinical outcomes and data from wearable devices.[40] Sharing this information without anonymisation, then, can cause issues for the person whose information this is. For example, the public disclosure of some comorbid illness can lead to discrimination and workplace harrassment.[82] As such, it is hard to balance the need for personal information, in order to inform transmission and diagnostics, while also ensuring that personal information is not exploited.

> ***Now what seems to be the issue is really the metadata that comes with the samples. In many cases, you can be blind to the underlying information on the individual…you shouldn't have easy access to that [information], but to do any kind of evaluation of what is going on in different populations, especially people with comorbidities or populations that are at risk, is more difficult***"
>
> *Policymaker*

Data protection legislation in some regions appears to have inhibited data sharing

In the Europe Union, personal data is protected under the General Data Protection Regulation (GDPR), meaning that personal data cannot be shared without the explicit consent of the individual to which it pertains.[83] Several of our European interviewees cited the GDPR as constraining the sharing of metadata with sequences, a trend which is borne out by the literature. For example, concerns relating to the General Data Protection Regulation (GDPR) in Denmark meant that no SARS-CoV-2 sequences could be shared internationally for almost two months in early 2021, until a new law was passed.[84] Similarly, privacy concerns were cited as a reason for holding back data in a review of Canada's COVID-19 virus-sequencing effort.[79] There is also a need for training and guidance on regulations as many researchers feel unconfident about what they are allowed to share.

A combination of technical and legislative solutions are needed to successfully combine relevant datasets

It is standard practice that 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, 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 by the WHO.[13] Therefore,

currently, use of personal data is restricted and highly regulated, but legislative solutions can also be found to overcome some of these constraints in emergency situations. For example, South Korea amended certain privacy laws after its 2015 MERS outbreak to accelerate data sharing in the event of a future infectious disease emergency,[86] while in the United Kingdom existing legislation was invoked to require the sharing of patient information by healthcare organisations and local authorities.[87] Similarly, a clause in the European GDPR explictly allows the processing of personal data for "reasons of public interest in the area of public health".[88]

Working with the Information and Commissioner's Officer and the National Data Guardian was important, making sure we were joined up. The powers we took weren't emergency powers, the Control of Patient Information (COPI) notices were successful in galvanising the sort of changes we needed to see.

UK policymaker

FROM PERSONAL CHOICE TO COMMUNITY NORM - LESSONS LEARNT

- Established norms around the timing and extent of data sharing may not be appropriate for a crisis, in which the immediate availability of data is paramount. Ongoing efforts to reform academic incentives must be accompanied by the development of strengthened expectations for data-sharing in an emergency context.
- The pandemic has highlighted the importance of collaboration at international, national and local levels, and exposed a need to improve the interface between research and public health.
- Effectively leveraging research data for public health relies on the capture of high-quality metadata and the deployment of a range of technical and legislative solutions that enable sensitive datasets to be used for research and public health purposes at scale.

Research Consulting

6. Balancing the needs of data creators and users

The pandemic saw many researchers opt to deposit sequences in a controlled-access repository, GISAID, which allowed them to retain rights over their data and receive credit for its subsequent re-use. The downsides of the GISAID model lie in the same constraints on re-use of the data, which meant national surveillance efforts were unable to utilise the data systematically. This illustrates the importance of striking an appropriate balance between the competing interests of data generators and data users.

Research Consulting

6. Balancing the needs of data creators and users

6.1. Open data infrastructure for genomic viral data is well-established

The landscape for depositing viral genomic sequences is dominated by a few key databases

Public databases such as the European Nucleotide Archive (ENA), maintained by EMBL's European Bioinformatics Institute (EMBL-EBI) in the UK,[39] and Genbank, maintained by the National Center for Biotechnology Information (NCBI) in the United States, have been go-to repositories for sequence generators for years before the pandemic.[89] Together with Japan's DNA Data Bank (DDBJ),[90] these databases are members of the International Nucleotide Sequence Database collaboration (INSDC.org).[20] The INSDC has a policy that enables sequence generators to deposit raw data reads, and allows users to access these and other data records with no restrictions. The facility exists for the users to give appropriate credit to data creators by citing the original data accession numbers. This open approach facilitates alignments, assemblies, functional annotations and access to other contextual information relating to samples and experimental configurations. Unrestricted access of this nature has led to ease of re-use but also to concerns that data providers rights are not being sufficiently protected.[91] INSDC partners share/mirror their data so the collaboration between the relevant infrastructures is very close, and the data exists in multiple locations and jurisdictions for the benefit of all types of community worldwide. Nevertheless, as some contributors to our study observed, the INSDC databases are hosted and operated by a small number of high-income countries and so the extent to which they serve (or are perceived to serve) the needs of global stakeholders equally remains contested.

Case Study: Understanding GISAID

GISAID (the Global Initiative on Sharing Avian Influenza Data) was launched in 2008 in response to the widespread reluctance to share data on avian H5N1 influenza viruses. As with public domain databases such as Genbank and The European Nucleotide Archive, usage of the viral genomic data on GISAID is free. However, in GISAID's case there is a proviso: that users sign a database access agreement that confirms their identity and prohibits republishing the site's data unless permission is granted from the data provider. This allows researchers depositing sequences to assert rights to viral data and commits users to acknowledge data submitters in publications, a feature welcomed by many as promoting equity and sovereignty.

Having registered to submit sequence data or use sequence data, a user can access the GISAID platform. The platform offers the EpiCoV™ database which houses the genomic sequences and associated metadata as well as the CoV Server tool which allows users to identify candidates for phenotypic changes or special epidemiological relevance. It was developed to aid the research community with the identification, analysis and interpretation of amino acid changes in coronavirus genomes. In addition to coronavirus data, GISAID also offers data for two other viruses that cause respiratory disease in humans, influenza virus and Respiratory Syncytial Virus (RSV). GISAID curators ensure improvement of deposits or withdrawal if duplicates sequences are identified; deposited

sequences may not be held under embargo for any reason but are made available after curation with immediate access to registered users.

According to its website, the GISAID Initiative's activities are governed by several organizational bodies that operate independently of each other. These include the Executive Board of Freunde von GISAID, a registered non-profit association providing administrative support, a Scientific Advisory Council and a Database Technical Group.

International, intergovernmental, and national infrastructures for sharing genomic viral DNA exist in G7 countries and beyond

Across the globe a range of initiatives emerged to support the sharing of genomic viral data. Initiatives exist at a range of levels, including international, intergovernmental and national. In Europe, national initiatives such as COVID-19 Genomics UK (COG-UK)[92] are supplemented by intergovernmental organisations such as ELIXIR,[93] which offers pooled resources (databases, software tools, cloud storage and supercomputers) coordinated into a single infrastructure. Genome Canada launched its national Viruseq Data portal[94], which, together with the Canadian COVID-19 Genomics Network (CanCOGeN),[95] allows national genomic sequencing uploads and tracking of variants of concern. In the US, the SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance (SPHERES)[96] collaboration was established to coordinate SARS-CoV-2 sequencing, while the COVID-19 Genomic Surveillance Network in Japan (COG-JP) fulfils a broadly similar function. There are of course other notable databases and initiatives used by researchers outside the G7, including:

- The National Genomics Data Center, formerly known as the BIG data centre is located near Beijing, China, which is a key source cited in COVID-19 related publications by Chinese researchers.[97]
- The Indian SARS-CoV-2 Genomics Consortium (INSACOG) which maintains a national sequencing database at two sites.[98]
- The African Center of Excellence in Genomics of Infectious Diseases (ACEGID).[99]
- The COVID-19 Genomic Surveillance Regional Network which aims to strengthen timely generation of COVID-19 genomic data in the Latin America and Caribbean (LAC) region.[100]

In many cases these initiatives will also upload relevant data to international databases such as INSDC and GISAID, but as noted in section 4.1, it is common practice for some sequencing data to be retained at local and national levels only.

6.2. Many data creators favour approaches that allow retention of rights

GISAID has become the de facto destination for SARS-CoV-2 sequences worldwide

Since January 2020, GISAID's data sharing platform has been the most popular primary source of genomic and associated data from SARS-CoV-2 cases. Sequences have been deposited by more than 200 countries worldwide, and 98% of sequences shared on

globally-recognised databases can be found in GISAID.[4] Because of the vast number of genomic sequences uploaded from around the world, GISAID has also provided the data that powers analytic dashboards such as Nextstrain (an open-source project to harness the scientific and public health potential of pathogen genome data),[101] which also gathers data from INSDC databases.[20]

GISAID benefitted from a first-mover advantage and its protection of the rights of data providers

GISAID appears to have benefited from a first-mover advantage in the early stages of the pandemic, in part because it was already widely used by the influenza community. In many cases, those working on SAS-CoV-2 were from the same public health community, and thus it was the obvious platform for them to use. GISAID's receipt of many early sequences of SARS-CoV-2 then made it more likely that other scientists would choose to deposit sequences in the same location (what is known as a 'network effect', where the value of a service increases when the number of people who use that service increases). GISAID was also able to harvest data from open access repositories to supplement direct deposits, thereby becoming the most comprehensive available resource.[5] Furthermore, its status as a controlled-access repository made it a more appealing option than the open access databases of the INSDC in the eyes of many data creators. Depositors retain rights over their data and are therefore more likely to receive credit for its subsequent re-use. This overcomes the lack of incentives for researchers to deposit data described in section 5, and partially compensates for the absence of international agreement on access and benefit-sharing for digital sequence information, as described in section 4.

> ***"The reason why GISAID is preferred by our scientists is because it constrains the use of the data [and] you have to acknowledge the source... Most of the other genetic data banks just allow complete free use of any data on the data set, which is good for the users, but not good for the suppliers. I think GISAID is a better balance between what's good for users and suppliers."***
>
> *Policymaker*

GISAID's success poses a challenge to open science

In many respects, GISAID represents an imperfect solution to the problem of genomic viral data sharing. It is not a fully open access database, it imposes restrictions on data users, and in the early days of the pandemic it did not accept raw sequence data, only assemblies of viral genomes.[91] Nevertheless, the majority of the researchers and several policymakers we interviewed asserted that GISAID's overall impact on the pandemic response has been positive, and many doubted that fully open access repositories could have achieved a similar rapid level of global uptake. Other policymakers and infrastructure providers, by contrast, argued that GISAID may have simply diverted deposits that would

[4] As of 8 May 2022, RCoV19, which aggregates data from multiple international databases, listed 10,872,843 SARS-CoV-2 sequences, while the GISAID submission tracker listed 10,663,247 (GISAID - Submission Tracker Global, 8/5/22). RCoV19 integrates genomic and proteomic sequences as well as their metadata information from the National GeneBank Database (CNGBdb), GenBank (which itself integrates sequences submitted to the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan), GISAID, Genome Warehose (GWH) and National Microbiome Data Collaborative (NMDC) databases.

[5] By contrast, open access repositories are unable to harvest data from GISAID in return due to its controlled access restrictions.

otherwise have been made available in open access repositories. Perspectives on the most appropriate solution for sharing genomic data vary widely between individuals in research, public health and industry contexts, high-, medium- and low-resource environments and emergency preparedness, response and/or recovery settings.

> ***"Data would not have been shared in a timely manner, giving vaccine manufacturers a head start making variant-specific vaccines, had we not had the GISAID sharing mechanism."***
>
> *Academic expert*

6.3. The limitations of controlled access

Protection of data suppliers' rights arguably constrains the ability of users to make full use of GISAID data

The limitations of the GISAID model lie in the same constraints on retrieval and re-use of the data. Some contributors to this study asserted that national surveillance efforts suffered from an inability to utilise the data systematically as a result of these constraints, while others emphasised that it responded to multiple requests from public health agencies by providing data feeds. These findings echo those of a 2021 Science article which identified a similar divergence of opinion.[102] As of September 2022, more than 800 representatives of renowned scientific organisations had signed an open letter calling for raw and assembled SARS-CoV-2 sequence data to be submitted to the databases of the INSDC. In their words, 'only fast, comprehensive and global SARS-CoV-2 sequencing and a rapid flow of SARS-CoV-2 sequence data into the INSDC databases will ensure the rapid dissemination of data with maximal impact due to their connectivity to the global bioinformatics data infrastructure'.[103] Despite this, the continued volume of deposits to GISAID suggests many data generators continue to prefer controlled over open access models of sharing.

> ***"[GISAID] has, I think, been very useful during the early stages of the pandemic, but later on, [it has been] difficult to get feeds to automatically download the GISAID data for our public health authority projects, and that's been a bit frustrating. So in the future, I think we would want a repository that is more fully committed to open data sharing."***
>
> *Academic expert*

Metadata concerns affect GISAID and other databases alike

Incomplete metadata attached to GISAID sequences has been found to be common globally, with about 63% of sequences missing demographic information (age and sex), 84% missing sampling strategies and more than 95% of sequences missing patient-level clinical data (e.g., symptom history, clinical outcome and vaccination status).[9] However, while GISAID's metadata has received greater scrutiny given its prominence in the COVID-19 response, similar criticisms have been levelled at the INSDC databases in other contexts.[104]

> ***"My main issue with a database like GISAID is that it is consensus level data. There's no associated metadata that tells you ...whether that consensus was generated from 500x, or whether it was generated from 1x...if it's 1x coverage then you know there are more errors than there would be if somebody taking a consensus from 500x for example and so you don't have a sense of that quality."***
>
> *Academic expert*

Others have raised questions over GISAID's governance and sustainability

GISAID's status as an independent, non-profit, public-private partnership (albeit one hosted and supported by the German government) has also led to concerns in some quarters that it lacks transparency. Some contributors to our consultation questioned whether it was adequately structured and resourced to fulfil such a central role in pandemic preparedness and argued that critical resources like GISAID should fall within the jurisdiction of intergovernmental bodies such as the WHO (although this would be a contentious move since GISAID is not a database for exclusively human pathogens). In the absence of formal governance procedures of this nature, some see a risk that access to data could be withdrawn arbitrarily and with limited means for recourse, while others noted that GISAID had resisted the creation of institutional access agreements that could have eased some of the concerns about limited access.

These concerns mirror those expressed in relation to the sharing of influenza data over the preceding decade

The debate over GISAID's merits is not new, and dates back to its genesis as a mechanism for rapid sharing of both published and 'unpublished' influenza data.[105] As Shu and McCauley (2017) observed, traditional public-domain archives such as GenBank, where sharing and use of data takes place anonymously, fulfilled a need for an archive of largely published data.[106] However, conventional methods of data exchange had 'not been successful in encouraging rapid sharing of important data in epidemic or (potential) pandemic situations, such as those caused by Middle East respiratory syndrome coronavirus (MERS-CoV) and Ebola viruses'. While methods of data exchange have evolved since GISAID was created in 2008, a clear lesson of the COVID-19 pandemic is that traditional sharing paradigms continue to inhibit rapid data-sharing in emergency contexts.

6.4. Balancing the needs of data creators and users

Current sharing infrastructures remain sub-optimal

Open data-sharing can bring enormous benefits to scientists, research funders citizens, and businesses. Fully open data (without any restriction on the end-user) can remove the frictions in discovery, access and use that impede rapid development and combination of data. However, open research data without controls is not always an unqualified good. to Controls may be needed, for example, to incentivise investment by private actors or to protect the privacy of individuals, public safety and security, or indigenous and other disadvantaged communities. In the case of GISAID, access controls have also provided a means of promoting attribution of data creators. Striking an intelligent balance between fully open and controlled data sharing lies at the heart of the commitment "to promote

the efficient processing and sharing of research data as openly as possible and as securely as necessary" in the G7 Research Compact (2021).[1]

Globally governed, fully open-access and interoperable infrastructure remains the ideal for many

Fully open-access infrastructures for data sharing offer demonstrably greater benefits than controlled access repositories for data re-use and integration. These fully open infrastructures remain the ideal for many policymakers, infrastructure providers and researchers, but they must be coupled with a transparent and globalised approach to funding, governance and benefits sharing, underpinned by appropriate incentives for data generators to submit to them. As experiences in the COVID-19 pandemic demonstrate, they must also be able to respond rapidly and flexibly to emerging needs in an emergency response situation.

In the COVID-19 pandemic, controlled access repositories represented a compromise between competing interests

However, it must also be recognised that controlled access repositories such as GISAID have played a critical role in enabling increased data sharing within an emergency context and addressing the concerns of many researchers, particularly in LMIC countries, around unrestricted rapid data sharing. Efforts to address misaligned incentives and resolve divergent international views on access and benefit-sharing are crucial but may take many years, or even decades, to yield tangible results. In the short term, and possibly beyond, controlled access repositories are likely to continue to play an important role in enabling the sharing of genomic data alongside fully open access repositories.

"We're not supportive of closed databases with embargo, but to support open science we don't have all the options that we need, so there needs to be something in between. There's only one model in between, which is GISAID. That is why they became the default option."

Policymaker

New models are needed that can maximize access while recognizing rights

Over time, there is a need to move beyond our existing data-sharing paradigms, infrastructure and tools in order to be better prepared for future emergencies. Future models of sharing need to provide appropriate credit to data generators without placing excessive constraints on access. This will require continuing efforts to reform academic incentives and pursuing international agreement on access and benefit-sharing for digital sequence information. Progress relies on working together with all relevant stakeholders to identify the most effective approaches that will allow data to be equitably and ethically shared and re-used as widely as possible.

CONTROLLED VERSUS OPEN ACCESS - LESSONS LEARNT

- Different models of access currently involve trade-offs between the interests of data generators and data users. The appropriate balance between these interests is likely to differ from the norm in an emergency context.
- If early data sharing is to become normalised, data generators need to receive reward and recognition for their contributions without placing undue restrictions on subsequent re-use of the data.
- Critical data infrastructures need open and transparent governance mechanisms that can both ensure their sustainability and demonstrate accountability to the wider international community.

Research Consulting

7. Conclusions and lessons learned

Open science offers clear benefits for emergency response, but this case study shows that the cultural and technical barriers to its adoption worldwide remain significant. Leveraging the potential of open science to respond to future emergencies means investing for the long term, taking a global perspective, incentivising equitable data-sharing, adapting to changing circumstances and identifying new sharing paradigms that can meet broad stakeholder needs.

Research Consulting

7. Conclusions and lessons learnt

Efforts to share genomic data during the pandemic yielded mixed results

This report has explored approaches to data-sharing in an emergency context within pathogen genomics, a scientific field with a highly developed culture of data sharing, underpinned by globally recognised databases. It shows that efforts to share genomic data during the pandemic yielded mixed results. While sequencing data was shared more quickly and widely than ever before, in many cases it was shared too late, in too partial a form or with insufficient metadata to contribute effectively to the emergency response. Variations in data quality, formats, associated metadata standards, and arrangements for access and re-use continue to present barriers to the effective sharing and use of genomic data at scale.

Open science offers clear benefits for emergency response…

Policymakers and infrastructure providers and many researchers from G7 countries have expressed a clear preference for open access models of data-sharing. Fully open-access infrastructures for data sharing offer demonstrably greater benefits than controlled access repositories for data re-use and integration for the purposes of emergency response. These benefits are clearly evident in the way genomic data was shared and used to support pandemic preparedness, public health responses and the development of vaccines, diagnostics and therapeutics.

…but cultural and technical barriers to its adoption worldwide remain significant

However, our consultation highlighted that significant number of researchers and policymakers remain unconvinced of the benefits of open and rapid sharing. These concerns were reflected in the preference for controlled over open access repositories when sharing of genomic data in the COVID-19 pandemic. Technical and legislative barriers also constrained the effective use of genomic data at scale for emergency response in many contexts.

Five lessons learnt

The lessons learnt from this case study for open science policymakers can be summarised under five headings, as follows:

1. Invest for the long term

An effective emergency response relies on long-term investment in open data infrastructure, standards and skills

The long-term investment made in developing international standards and infrastructures for data sharing in genomics was repaid many times over when this data became central to the pandemic response. Critical data infrastructures need open and transparent governance mechanisms, sustainable funding, and common standards that enable interoperability and scalability. These must be accompanied by skilled individuals who are able to create, analyse, share and re-use relevant data. All of this relies on a long-term commitment by governments and funders to invest in science, research, and public health infrastructure, as well as a recognition of the critical importance of open data infrastructure, software, standards and skills.

2. Take a global perspective

Global challenges like COVID-19 require a global and inclusive

Effectively tackling global crises like the COVID-19 pandemic requires representative data from all parts of the world. Not all countries and regions have sufficient data-generating capacity or trained human resources to collect, disseminate, and analyse these data. The willingness and ability of different actors to share data during

approach to data sharing

emergencies is also heavily influenced by pre-existing geopolitical considerations and the risk of adverse political and economic consequences. Interventions designed to enhance the availability of relevant data must ensure they identify and tackle the root of the problem. In many cases this will be a lack of underlying research capacity and public health infrastructure, or political tensions rather than inadequate uptake of open sharing practices. Open international infrastructures must also be cognisant of the needs of a diverse community of users, with standardisation of data and metadata formats accompanied by a flexible approach to access.

3. Create incentives for equitable data-sharing

Reformed incentives are needed to promote data-sharing across boundaries

If rapid and open data sharing is to be encouraged, the contributions of data generators need to be recognised and rewarded. Informal data access arrangements based on pre-existing knowledge of trusted individuals are not sufficient to enable equitable sharing and re-use of data at scale. Equitable data sharing in an emergency situation – in this context, a disease outbreak with political, social and public health impacts – depends on prior work to understand the motivations of different stakeholders and reach agreement through relevant international fora on prosocial arrangements for access and benefit sharing. There is a compelling need to continue efforts to reform incentives for all data generators to reward the sharing of reusable, high-quality data, code and other research objects alongside accompanying metadata. Similarly, there is a need to clarify expectations of speed, quality and transparency for data generators in differing contexts such as routine surveillance in public health. The pandemic has highlighted the crucial importance of cross-boundary collaboration at international, national and local levels, and exposed a need to improve the interface between research and public health in order to maximise the combination and re-use of scientific and clinical data.

4. Adapt to changing circumstances

Established norms for data-sharing must evolve in light of the COVID-19 pandemic

Established norms around the timing and extent of data sharing were in many cases set aside in the COVID-19 crisis, with multiple actors recognising that the immediate availability of data to a broad set of users was paramount. Yet the pandemic also provides an opportunity to re-assess these established norms, whose deficiencies were in some cases sharply exposed. Ongoing efforts to reform academic incentives must be accompanied by corresponding work to incentivise sharing by public health actors, and the development of strengthened expectations for data-sharing by all parties in an emergency context. Public policymakers, research and development funders, institutions in academic and public health, and publishers all have a role to play in setting expectations for open and rapid sharing of all relevant data and information in these circumstances. Open infrastructure providers must be able to identify and respond rapidly to emerging requirements, while new approaches should make provision for sensitive datasets to be used for research purposes in emergency scenarios.

5. Move beyond current sharing paradigms

New sharing paradigms are needed to address competing interests

This study has exposed divergent perspectives within and between the research and public health communities on the merits of open and controlled models of access to genomic viral data. Fully open-access infrastructures for data sharing offer demonstrably greater benefits than controlled access repositories in terms of data re-use and integration at scale, but these benefits cannot be realised in practice unless these

infrastructures are accompanied by a transparent and globalised approach to funding, governance and benefits sharing. Proponents of open-access infrastructures must give greater consideration to mechanisms for incentivising and crediting data deposits and enabling the creation of high-quality metadata.

Towards intelligent open science

For open science policymakers, this case study illustrates the complex and inter-connected nature of the data-sharing landscape, and the impact that wider political and cultural considerations have on the availability of scientific data. It further demonstrates the need to consider diverse perspectives and the risk of unintended consequences when formulating policy interventions. Different approaches can be taken to address the often-competing interests of data generators and data users, and perceptions of the most appropriate solution remain highly context-dependent.

An intelligent approach to open science means moving beyond our existing data-sharing paradigms to be better prepared for future emergencies. This will involve promoting prosocial approaches to the sharing and re-use of data at the level of individuals, institutions and nations, by developing incentive structures and community norms that recognise and reward these behaviours. This process of cultural change must be accompanied by sustained investment in technical infrastructures and skilled individuals and the development of legal frameworks that enable datasets to be aggregated and analysed at scale. All of this will require international collaboration between actors in government, public health, research and the private sector and in high-, middle- and low-resource contexts, in the interest of providing both broad access for data users and broad recognition of data generators and custodians. Progress towards these goals is closely tied to the reform of existing incentive structures and must be understood as a long-term endeavour. But it holds out the prospect of data being shared and re-used as widely as possible for the benefit of all populations, while acknowledging and rewarding the efforts of data generators and custodians.

Research Consulting

References

1. G7 UK. G7 Research Compact. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1001133/G7_2021_Research_Compact__PDF__356KB__2_pages_.pdf (2021).
2. Sachs, J. D. *et al.* The Lancet Commission on lessons for the future from the COVID-19 pandemic. *The Lancet* **0**, (2022).
3. United Nation. Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity. https://www.cbd.int/abs/doc/protocol/nagoya-protocol-en.pdf (2011).
4. Singh, S. *et al.* How an outbreak became a pandemic: a chronological analysis of crucial junctures and international obligations in the early months of the COVID-19 pandemic. *The Lancet* **398**, 2109–2124 (2021).
5. CIPD. PESTLE Analysis | Factsheets. *CIPD* https://www.cipd.co.uk/knowledge/strategy/organisational-development/pestle-analysis-factsheet.
6. Chiarelli, A. *et al. From intent to impact: Investigating the effects of open sharing commitments*. https://zenodo.org/record/6620854 (2022) doi:10.5281/zenodo.6620854.
7. Illumina & Nature Research Custom Media. How next-generation sequencing can help identify and track SARS-CoV-2.
8. Oxford Nanopore. Oxford Nanopore Technologies. https://nanoporetech.com/ (2022).
9. Chen, Z. *et al.* Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat. Genet.* 1–9 (2022) doi:10.1038/s41588-022-01033-y.
10. Centers of Disease Control and Prevention. How it Works | Advanced Molecular Detection (AMD). https://www.cdc.gov/amd/how-it-works/index.html (2021).
11. National Library of Medicine. BLAST Topics. https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp (2020).
12. European Nucleotide Archive. The ENA Metadata Model — ENA Training Modules 1 documentation. https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html.
13. World Health Organization. *Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health*. https://www.who.int/publications-detail-redirect/9789240018440 (2021).
14. van Reisen, M. *et al.* Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. *Adv. Genet. Hoboken Nj* **2**, e10050 (2021).
15. Foreign and Agricultural Organization of the United Nations. *Final Meeting Report: Technical Meeting on the impact of Whole Genome Sequencing (WGS) on food safety management: within a One Health approach : The 9th meeting of the Global Microbial Identifier (GMI9)*. (Foreign and Agricultural Organization of the United Nations, 2016).
16. Harrison, P. W. *et al.* The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res.* **49**, W619–W623 (2021).
17. RDA COVID-19 Working Group. RDA COVID-19 Recommendations and Guidelines on Data Sharing. (2020) doi:10.15497/rda00052.
18. Carroll, S. R. *et al.* The CARE Principles for Indigenous Data Governance. *Data Sci. J.* **19**, 43 (2020).
19. Parry, D. iSchool leads effort to improve stewardship of Indigenous data. https://ischool.uw.edu/news/2022/01/ischool-leads-effort-improve-stewardship-indigenous-data (2022).
20. INSDC. International Nucleotide Sequence Database Collaboration. https://www.insdc.org/ (2022).
21. Wellcome Trust. Coronavirus (COVID-19): sharing research data. *Wellcome* https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak (2020).
22. US Department of State. G7 Science and Technology Ministers' Declaration on COVID-19. *United States Department of State* https://www.state.gov/g7-science-and-technology-ministers-declaration-on-covid-19/ (2020).
23. Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P. & Ceri, S. A review on viral data sources and search systems for perspective mitigation of COVID-19. *Brief. Bioinform.* **22**, 664–675 (2020).
24. Kirka, D. UK virus hunting labs seek to bolster global variant network. https://medicalxpress.com/news/2022-01-uk-virus-labs-bolster-global.html (2022).
25. World Health Organization. WHO reference laboratories providing confirmatory testing for COVID-19. https://www.who.int/publications/m/item/who-reference-laboratories-providing-confirmatory-testing-for-covid-19 (2020).
26. Covid Circle. Funding and undertaking research during the first year of the COVID-19 pandemic. https://www.ukcdr.org.uk/wp-content/uploads/2021/11/Covid-Circle_Lessons-for-funders_Report_2-11-21.pdf (2021).

27. G7. *G7 Pact for Pandemic Readiness*. https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/G/G7/20220520_English_G7_Pact_for_Pandemic_Readiness.pdf (2022).
28. GOARN. Welcome to GOARN. https://extranet.who.int/goarn/ (2022).
29. GISRS & World Health Organization. Global Influenza Surveillance and Response System. https://www.who.int/initiatives/global-influenza-surveillance-and-response-system (2022).
30. World Health Organization. WHO, Germany open Hub for Pandemic and Epidemic Intelligence in Berlin. https://www.who.int/news/item/01-09-2021-who-germany-open-hub-for-pandemic-and-epidemic-intelligence-in-berlin (2021).
31. Rockefeller Foundation. The Rockefeller Foundation Invests USD150 Million to Prevent Future Pandemics and Calls for Greater Collaboration to Build a Global Early Warning System. *The Rockefeller Foundation* https://www.rockefellerfoundation.org/news/the-rockefeller-foundation-invests-150-million-to-preventing-future-pandemics-calls-for-greater-collaboration-to-build-global-early-warning-system/ (2021).
32. Africa CDC. Institute of Pathogen Genomics (IPG). *Africa CDC* https://africacdc.org/institutes/ipg/ (2022).
33. Prime Minister's Office, 10 Downing Street. PM announces plan for 'Global Pandemic Radar'. *GOV.UK* https://www.gov.uk/government/news/pm-announces-plan-for-global-pandemic-radar.
34. Beckett, A. H., Cook, K. F. & Robson, S. C. A pandemic in the age of next-generation sequencing. *The Biochemist* **43**, 10–15 (2021).
35. Oboh, M. A. *et al.* Translation of genomic epidemiology of infectious pathogens: Enhancing African genomics hubs for outbreaks. *Int. J. Infect. Dis.* **99**, 449–451 (2020).
36. Kamp, M., Krause, A. & Ramsay, M. Has translational genomics come of age in Africa? *Hum. Mol. Genet.* **30**, R164–R173 (2021).
37. Canadian Bioinformatics. 2021 Workshops. https://bioinformatics.ca/workshops/ (2021).
38. GISAID. GISAID Bioinformatics Training at Queen's University Belfast. (2022) doi:10.17616/R3Q59F.
39. European Bioinformatics Institute. EMBL-EBI: EMBL's European Bioinformatics Institute. https://ebi.ac.uk/.
40. Kucharski, A. J., Hodcroft, E. B. & Kraemer, M. U. G. Sharing, synthesis and sustainability of data analysis for epidemic preparedness in Europe. *Lancet Reg. Health – Eur.* **9**, (2021).
41. Mallapaty, S. Genome data gaps could stymie search for next COVID variant. *Nature* (2022) doi:10.1038/d41586-022-00894-x.
42. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
43. Barrett, J. Scientists sharing Omicron data were heroic. Let's ensure they don't regret it. *The Guardian* (2021).
44. Seydi, C. O. Southern Africa: Last in line for vaccines, first in line for travel bans. *Bill & Melinda Gates Foundation* https://www.gatesfoundation.org/ideas/articles/omicron-covid-africa-travel (2021).
45. International Relations and Cooperation, Republic of South Africa. South Africa's response to travel restrictions imposed by several countries. http://www.dirco.gov.za/docs/2021/covid-19_1127.htm (2021).
46. Tulio de Oliveira [@tuliodna]. If the world keeps punishing Africa for the discovery of #Omicron and 'Global Health Scientists' keep taking the data, who will share early data again? Just saying thanks and keep hoarding of vaccine, antiviral, diagnostics, PPEs, funding, etc - Thread on unfair practices 1/4. *Twitter* https://twitter.com/tuliodna/status/1467710032821338112 (2021).
47. Biosafety Unit. Parties to the Nagoya Protocol. https://www.cbd.int/abs/nagoya-protocol/signatories/ (2014).
48. Smyth, S. J., Macall, D. M., Phillips, P. W. B. & de Beer, J. Implications of biological information digitization: Access and benefit sharing of plant genetic resources. *J. World Intellect. Prop.* **23**, 267–287 (2020).
49. Biosafety Unit. What has been done? https://www.cbd.int/dsi-gr/whatdone.shtml (2021).
50. Secretariat for the Convention on Biological, Secretariat for the Convention on Biological, & Secretariat for the Convention on Biological Diversity. *Criteria to Consider for Policy Options on Digital Sequence Information on Genetic Resources*. https://www.cbd.int/abs/DSI-webinar/CriteriaSummaryPaper2021.pdf (2021).
51. Covention on Biological Diversity. *Co-leads' report on the work of the informal co-chairs' advisory group on digital sequence information on genetic resources*. https://www.cbd.int/doc/c/079d/1142/339a68fee2d22e95fb2b1c4c/wg2020-03-inf-08-en.pdf (2021).
52. Scholz, A. H. *et al.* Multilateral benefit-sharing from digital sequence information will support both science and biodiversity conservation. *Nat. Commun.* **13**, 1086 (2022).
53. Cochrane, G., Lauer, K., Blomberg, N., Apweiler, R. & Birney, E. *Pathogen genomics data sharing: public health meets research*. https://zenodo.org/record/6368840 (2022) doi:10.5281/zenodo.6368840.
54. Kukutai, T., Russo Carroll, S. & Walter, M. Indigenous Data Sovereignty. https://researchcommons.waikato.ac.nz/bitstream/handle/10289/13633/indigenous%20data.pdf?sequence=9&isAllowed=y (2020).
55. Raza, S. Minding the genomic data gap: COVID-19, genomics and health inequalities. *Ada Lovelace Institute* https://www.adalovelaceinstitute.org/blog/data-gap-covid-19-genomics-health-inequalities/ (2021).

56. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
57. GLOPID-R. Data Sharing in Public Health Emergencies: Learning from past outbreaks. https://www.glopid-r.org/wp-content/uploads/2017/02/data-sharing-in-public-health-emergencies-case-studies-workshop-reportv2.pdf (2019).
58. GLOPID-R. GLOPID-R Roadmap for Data Sharing in Public Health Emergencies. https://www.glopid-r.org/wp-content/uploads/2019/06/glopid-r-roadmap-for-data-sharing.pdf (2020).
59. Pratt, B. & Bull, S. Equitable data sharing in epidemics and pandemics. *BMC Med. Ethics* **22**, 136 (2021).
60. Norton, A. *et al.* A living mapping review for COVID-19 funded research projects: nine-month update. at https://doi.org/10.12688/wellcomeopenres.16259.4 (2021).
61. Nanda, S. & Kowalczuk, M. K. Unpublished genomic data–how to share? *BMC Genomics* **15**, 5 (2014).
62. National Institutes of Health. NIH Genomic Data Sharing. *Office of Science Policy* https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/ (2015).
63. AMED. Data Sharing Policy for the Realization of Genomic Medicine. https://www.amed.go.jp/content/000053433.pdf (2016).
64. World Health Organization = Organisation mondiale de la Santé. Policy statement on data sharing by WHO in the context of public health emergencies (as of 13 April 2016). *Wkly. Epidemiol. Rec. Relevé Épidémiologique Hebd.* **91**, 237–240 (2016).
65. Kalia, K., Saberwal, G. & Sharma, G. The lag in SARS-CoV-2 genome submissions to GISAID. *Nat. Biotechnol.* **39**, 1058–1060 (2021).
66. Health Data Research UK. How do we collect, store and manage data about people's SARS-CoV-2 samples? *HDR UK* https://www.hdruk.org/news/how-do-we-collect-store-and-manage-data-about-peoples-sars-cov-2-samples/ (2021).
67. McKinsey and Company. Preventing pandemics with investments in public health. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/not-the-last-pandemic-investing-now-to-reimagine-public-health-systems (2021).
68. Black, A., MacCannell, D. R., Sibley, T. R. & Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat. Med.* **26**, 832–841 (2020).
69. Gooderham, M. Sequencing the Crisis: How genomics morphed from a COVID-19 research tool to a critical part of the pandemic response. *Public Policy Forum* https://ppforum.ca/publications/sequencing-the-crisis-how-genomics-morphed-from-a-covid-19-research-tool-to-a-critical-part-of-the-pandemic-response/ (2021).
70. National Human Genome Research Institute. 1997: Bermuda Meeting Affirms Principle of Data Release. *Genome.gov* https://www.genome.gov/25520385/online-education-kit-1997-bermuda-meeting-affirms-principle-of-data-release.
71. Pearson, H. Competition in biology: It's a scoop! *Nature* (2003) doi:10.1038/news031124-9.
72. Robishaw, J. D. *et al.* Genomic surveillance to combat COVID-19: challenges and opportunities. *Lancet Microbe* **2**, e481–e484 (2021).
73. The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* **28**, 715–718 (2020).
74. ICODA. Home. *ICODA - A globally coordinated, health data-led research response to tackle the COVID-19 pandemic.* https://icoda-research.org/ (2022).
75. ICODA. The urgent need for better coordination across the global data sharing landscape. *ICODA - A globally coordinated, health data-led research response to tackle the COVID-19 pandemic.* https://icoda-research.org/the-urgent-need-for-better-coordination-across-the-global-data-sharing-landscape/ (2021).
76. Covid-19 Data Portal. COVID-19 Data Portal - accelerating scientific research through data. https://www.covid19dataportal.org/ (2020).
77. Vivli. Vivli - Center for Global Clinical Research Data. *Vivli* https://vivli.org/.
78. Global Alliance for Genomics & Health. Enabling responsible genomic data sharing for the benefit of human health. https://www.ga4gh.org/ (2022).
79. ICODA. Publications. *ICODA - A globally coordinated, health data-led research response to tackle the COVID-19 pandemic.* https://icoda-research.org/research/publications/.
80. ICODA. Partners. *ICODA - A globally coordinated, health data-led research response to tackle the COVID-19 pandemic.* https://icoda-research.org/partners/ (2022).
81. ICODA - Portal. https://portal.covid-19.aridhia.io/.
82. House of Commons: Science and Technology Committee. Oral Evidence: The Right to Privacy: digital data, HC 1000. https://committees.parliament.uk/oralevidence/9979/html/?utm_source=Science+and+Technology+Committee+Weekly+Update&utm_campaign=744c8dcf3c-EMAIL_CAMPAIGN_2020_09_11_12_12_COPY_01&utm_medium=email&utm_term=0_b7e0da2ad0-744c8dcf3c-104200378&mc_cid=744c8dcf3c&mc_eid=4903c69f06 (2022).
83. GDPR.eu. What is GDPR, the EU's new data protection law? *GDPR.eu* https://gdpr.eu/what-is-gdpr/ (2018).

84. Statens Serum Institur. SSI resumes data sharing of virus sequences. https://en.ssi.dk/news/news/2021/ssi-is-once-again-sharing-virus-sequences (2021).
85. Semeniuk, I. In genetic arms race with COVID-19 variants, Canada's labs fight for better ways to share findings with each other and the world. *The Globe and Mail* (2021).
86. Park, S., Choi, G. J. & Ko, H. Information Technology–Based Tracing Strategy in Response to COVID-19 in South Korea—Privacy Controversies | Global Health | JAMA | JAMA Network. *JAMA* **323**, 2129–2130 (2020).
87. Department of Health and Social Care. Coronavirus (COVID-19): notification to organisations to share information. *GOV.UK* https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information (2020).
88. McLennan, S., Celi, L. A. & Buyx, A. COVID-19: Putting the General Data Protection Regulation to the Test. *JMIR Public Health Surveill.* **6**, e19279 (2020).
89. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.
90. DDBJ. DDBJ. https://www.ddbj.nig.ac.jp/index-e.html (2020).
91. Van Noorden, R. Scientists call for fully open sharing of coronavirus genome data. *Nature* **590**, 195–196 (2021).
92. COG-UK. COVID-19 Genomics UK Consortium. *COVID-19 Genomics UK Consortium | UK-Wide Genomic Sequencing* https://www.cogconsortium.uk/ (2021).
93. ELIXIR. ELIXIR. *ELIXIR* https://elixir-europe.org/.
94. CanCOGeN & VirusSeq. VirusSeq Portal. https://virusseq-dataportal.ca/ (2022).
95. CanCOGeN. *GenomeCanada* https://genomecanada.ca/challenge-areas/cancogen/.
96. Centers of Disease Control and Prevention. Cases, Data, and Surveillance (SPHERES). *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/variants/spheres.html (2020).
97. National Genomics Data Center. Home - National Genomics Data Center. https://ngdc.cncb.ac.cn/ (2020).
98. Press Information Bureau, Government of India. Indian SARS-CoV-2 Genomic Consortia (INSACOG) launched,coordinated by Department of Biotechnology (DBT) along with MoH&FW, ICMR, and CSIR. https://pib.gov.in/pib.gov.in/Pressreleaseshare.aspx?PRID=1684782 (2020).
99. ACEGID. Acegid – Welcome Online. https://acegid.org/.
100. Pan American Health Organization. COVID-19 Genomic Surveillance Regional Network - PAHO/WHO | Pan American Health Organization. https://www.paho.org/en/topics/influenza-and-other-respiratory-viruses/covid-19-genomic-surveillance-regional-network (2020).
101. Nextstrain. Nextstrain / ncov / gisaid / global / 6m. https://nextstrain.org/ncov/gisaid/global/6m (2022).
102. Wadman, M. Critics decry access, transparency issues with key trove of coronavirus sequences. (2021).
103. Covid-19 Data Portal. Open letter: Support data sharing for COVID-19. https://www.covid19dataportal.org/support-data-sharing-covid19 (2020).
104. Crandall, E. D. *et al.* Metadata preservation and stewardship for genomic data is possible, but must happen now. 2022.09.12.507034 at https://doi.org/10.1101/2022.09.12.507034 (2022).
105. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).
106. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

Appendix A. Interviewees and focus group attendees

The following stakeholders participated in interviews or focus groups undertaken as part of this study. We wish to highlight that participation in this study (whether as an interviewee, focus group attendee or reviewer) does not imply endorsement of all the report's findings.

Table A1. Project contributors.

Name	Role	Organisation	Country	Contribution
Miles Carroll	Head of Research & Development Institute	University of Oxford	United Kingdom	Interview
Guy Cochrane	Group Leader	European Nucleotide Archive	United Kingdom	Focus Group
Frederik Coppens	Head of Node for ELIXIR Belgium	Flemish Institute for Biotechnology	Belgium	Focus Group
Antonino Di Caro	Director of Microbiology Laboratory and Infectious Diseases Biorepository	National Institute for Infectious Diseases	Italy	Interview
Alexander Dilthey	Professor of Genomic Microbiology and Immunity	University of Düsseldorf	Germany	Interview
Elodie Ghedin	Chief of the Systems Genomics Sections	National Institute of Allergy and Infectious Diseases	United States	Interview
Jeremy Kamil	Associate Professor of Microbiology and Immunology	LSU Health Shreveport	United States	Interview and Focus Group
Salim Abdool Karim	Unit Diretor (SAMRC) and Professor of Global Health at Columbia University	South African Medical Research Council	South Africa	Interview
Mari Kleemola	Development Manager at Finnish Social Science Data Archive	Tampere University	Finland	Focus Group
Brad Langhorst	Group Leader	New England Biolabs	United States	Focus Group
Katy Lindfield	Deputy Director of Data Policy	Department of Health and Social Care, UK Government	United Kingdom	Interview
Duncan MacCannell	Chief Science Officer, Office of Advanced Molecular Detection	Centre for Disease, Control and Prevention	United States	Focus Group
Andrew McArthur	Associate Professor of Biochemistry and Biomedical Sciences	McMaster University	Canada	Interview

Jo McEntyre	Associate Director	EMBL-EBI	United Kingdom	Interview
Alan McNally	Professor of Microbial Evolutionary Genomics	University of Birmingham	United Kingdom	Interview
Vasee Morthy	Senior Adviser, Research and Development Science Division	WHO	Switzerland	Interview
Nicola Mulder	Professor and Head of Computational Biology	University of Cape Town	South Africa	Focus Group
Howard Needham	Scientific Liaison Officer	European Centre for Disease Prevention and Control	Europe	Interview
David Salgado	NGS Team Manager	Aix-Marseille Université	France	Focus Group
Melanie Saville	Director of Vaccine Research & Development	CEPI	Norway	Interview
Joachim Schultze	Director of Systems Medicine	University of Bonn	Germany	Interview
Fernando Spilki	Professor of Molecular Biology	Universidade Feevale	Brazil	Interview
Jacques Van-Helden	Co-Director of the Institut Français de Bioinformatique	Aix-Marseille Université	France	Focus Group
Joe Watts	Head of Cross Government Data Strategy	Department of Health and Social Care, UK Government	United Kingdom	Interview

Appendix B. Peer reviewers

The following individuals acted as independent reviewers of this study. We wish to highlight that participation in this study (whether as an interviewee, focus group attendee or reviewer) does not imply endorsement of all the report's findings.

Name	Role	Organisation	Country
Michael Arentoft	Head of Open Science Unit	European Commission Directorate-General for Research & Innovation	Intergovernmental organisation
Masanori Arita	Professor	DNA Data Bank of Japan, National Institute of Genetics	Japan
Alistair Darby	Co-Director	Centre for Genomic Research, University of Liverpool	United Kingdom
Ron Fouchier	Deputy Head of Department of Viroscience and Co-Chair	Erasmus University Medical Centre and GISAID	Netherlands
Nina Gadson	UK Health Security Team	Department of Health and Social Care	United Kingdom
Carole Goble	Professor of Computer Science	University of Manchester	United Kingdom
Josie Golding	Head of Epidemics & Epidemiology	Wellcome	United Kingdom
Timothy Hancox	UK Health Security Team	Department of Health and Social Care	United Kingdom
Georgina Humphreys	Independent Consultant	Not applicable	United Kingdom
Toby McMaster	Covid-19 Vaccine Deployment Policy	Department of Health and Social Care	United Kingdom
John McCauley	Co-Chair	WHO Collaborating Centre for Reference and Research on Influenza, The Francis Crick Institute and GISAID	United Kingdom
Vasee Moorthy	Senior Advisor R&D and genomic sequencing lead	World Health Organization	Intergovernmental organisation
Collins M Morang'a	Bioinformatician	West African Centre for Cell Biology of Infectious Pathogens, University of Ghana	Republic of Ghana
Claire Newland	Director of Policy, Ethics and Governance	Medical Research Council	United Kingdom

Kostas Repanas	Policy Officer	European Commission Directorate-General for Research & Innovation	Intergovernmental organisation
Jerry Sheehan	Deputy Director	National Institutes of Health	United States
Fernando Spilki	Professor and Vice-Rector for Research	Feevale University	Brazil
Yochanna Yehudi	Executive Director	University of Manchester	United Kingdom

Research Consulting

Appendix C. Stakeholder identification and text processing methodology

Text processing methodology

A text mining approach was used to identify researchers for interview

As noted in Section 1.2 of this report, a text processing algorithm, developed by Science-Metrix (an Elsevier company), was applied to identify a longlist of researchers for interview in the context of this project.

The algorithm was based on regular expression (regex) queries with two key aims:

- to identify mentions of COVID-19 genomic sequencing data deposits in peer-reviewed articles and preprints (henceforth referred to as "publications"); and
- to identify the researchers that have led these data deposits, using corresponding authors in the associated publications as a reasonable proxy.

To complement the thematic queries outlined above, a two-step identification strategy was deployed by Science-Metrix, aiming to identify relevant repositories noted in the data availability statements (DAS) of publications.

The following sections outlines the steps taken throughout the text processing methodology in more detail, with the overall aim of developing a shortlist of potential interviewees to contribute to this study.

The text mining approach relied on publications data sourced from an existing database

As a first step, publications with a thematic association with COVID-19 research were identified. This exercise was completed as part of another project led by Research Consulting and Science-Metrix, evaluating the impact of joint statement on the sharing of research data and findings relevant to COVID-19 outbreak[6]. The delineation of this thematic publication set also used regex queries for text processing, focusing on isolating Covid-19-associated terms within publications title, abstract and keywords.

A thematic query was applied to both peer-reviewed articles and preprints

The regex query was applied to all identified publications. These comprised preprints uploaded to Medrxiv and Biorxiv, with metadata and full text records downloaded from the servers' Amazon S3 buckets in October and November 2021. In addition, metadata for peer-reviewed articles was accessed through Science-Metrix's custom implementation of the Scopus database. Peer-reviewed article full texts, which were queried in the process of isolating data availability statements (DAS) and data sharing cases (see following sections), were obtained from Scopus indexing records (noting that this source contains full texts for approximately only 75% of articles indexed by Scopus).

Additional regexes were developed to identify mentions of relevant repositories in data availability statements

A two-step identification strategy was deployed by Science-Metrix to refine the results of the above-described queries. This approach aimed to identify relevant repositories noted in the data availability statements (DAS) of publications. This was achieved using regex queries to identify and excerpt text from formal DAS sections and by identifying mentions to selected repositories within the DAS section excerpts.

Formulation of regex queries to identify formal DAS sections were based on adaptations of similar queries

Formulation of regex queries to identify formal DAS sections were informed by similar queries used in the Oddpub R package,[6] adapted to the Databricks and data architecture environment used by Science-Metrix. DAS-identification queries were complicated by certain formulations in publications that discussed matters of data availability as a methodological challenge, rather than as part of a formal DAS. To mitigate (but still not fully eliminating) the capture of these cases by the algorithm, only DAS sections containing a clear section heading were captured by the regex queries. In practice, this approach often entailed isolating only DAS formulations that started a new sentence (DAS formulations needed to be preceded by two non-alphanumeric characters, allowing for instance for a period and a space; or for a period and new line).

This approach was applied to peer-reviewed articles and BioRxiv preprints. For MedRxiv preprints this step was skipped given that their metadata records contain a pre-parsed and specific field with the content of the DAS. MedRxiv mandates the formulation of a DAS as part of its submission process.

Limitations of the text processing approach

It should be noted that the set of publications delineated with the text processing algorithm cannot be used as a basis for statistical analysis. The approach enabled a convenience sampling strategy of genomic sequence data deposition instances, with ultimate precision ensured by manual review of individual cases rather than purely through optimization of the regex queries.

Identifying relevant researchers in genomic viral sequencing to interview for this project

A subset of researchers were identified for inclusion in this review

The final dataset of 61 publications identified by applying the search queries denoted above, included the DOI, title of publication, Journal/preprint server matching the DOI, date of publication/posting, corresponding author(s) and their country of affiliation.

A sample of publications was manually selected and characterised by the following:

- The list comprised 5-10 authors from each of the G7 countries and 3-5 authors where possible from countries outside of the G7 (including upper middle-income countries to LMIC).
- It included approximately 80% from peer reviewed articles; 20 % from preprints (that have subsequently been accepted after journal peer review); most were published in 2020 and 2021
- The publications mentioned sequence depositions in a range of databases: note that GISAID dominated the selection, but it also featured for eg. DDBJ, Big Data centre of the Beijing Institute in China.
- A gender balance of corresponding authors was ensured as far as possible.

Specific queries used

[6] https://github.com/quest-bih/oddpub/tree/v6

Specific queries for assembling the Covid-19 research publication set

The following search terms were employed in the regex queries to delineate the Covid-19 research publication set:

- ncov.?19
- covid.?19
- sars.?cov.?2
- 2019.?ncov
- 2019 novel coronavirus
- novel coronavirus 2019
- coronavirus disease 2019

Specific queries for assembling the DAS excerpt set

The following search terms were employed in the regex queries to identify and extract text excerpts from the DAS sections. Note that all queries were case insensitive.

- [^a-z][^a-z]Data[^a-z]{0,1}Sharing[^a-z]{0,1}(?!platform) [^a-z]
- [^a-z][^a-z]Data[^a-z]{0,1}Availability[^a-z]
- [^a-z]Data[^a-z]{0,1}Deposition[^a-z]
- [^a-z]Data[^a-z]{0,1}Archiving[^a-z]
- [^a-z]Data[^a-z]{0,1}Accessibility[^a-z]
- [^a-z] [^a-z]Availability[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]
- [^a-z]Deposition[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]
- [^a-z]Archiving[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]
- [^a-z]Accessibility[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}Sharing[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}Availability[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}Deposition[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}Archiving[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}Accessibility[^a-z]
- [^a-z]Availability[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]
- [^a-z]Deposition[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]
- [^a-z]Archiving[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]
- [^a-z]Accessibility[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}Sharing[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]{0,1}Availability[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]{0,1}Deposition[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]{0,1}Archiving[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]{0,1}Accessibility[^a-z]
- [^a-z]Availability[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]
- [^a-z]Deposition[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}material[s]? [^a-z]
- [^a-z]Archiving[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]
- [^a-z]Accessibility[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}material[s]? [^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]{0,1}Availability[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]{0,1}Deposition[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]{0,1}Archiving[^a-z]
- [^a-z]Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]{0,1}Accessibility[^a-z]
- [^a-z]Availability[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]

- [^a-z]Deposition[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]
- [^a-z]Archiving[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]
- [^a-z]Accessibility[^a-z]{0,1}of[^a-z]{0,1}Data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]
- [^a-z]Deposited[^a-z]{0,1}data[^a-z]
- [^a-z] (?!must include a data)Availability[^a-z]{0,1}Statement[^a-z]
- [^a-z]Deposition[^a-z]{0,1}statement[^a-z]
- [^a-z]archiving[^a-z]{0,1}statement[^a-z]
- [^a-z]accessibility[^a-z]{0,1}statement[^a-z]
- [^a-z]open[^a-z]{0,1}data[^a-z]{0,1}statement[^a-z]
- [^a-z]open[^a-z]{0,1}data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}statement[^a-z]
- [^a-z]open[^a-z]{0,1}data[^a-z]{0,1}and[^a-z]{0,1}materials[^a-z]{0,1}statement[^a-z]
- [^a-z]Data[^a-z]{0,1}statement[^a-z]
- [^a-z]data[^a-z]{0,1}and[^a-z]{0,1}code[^a-z]{0,1}statement[^a-z]
- [^a-z]data[^a-z]{0,1}and[^a-z]{0,1}material[s]?[^a-z]{0,1}statement[^a-z]
- [^a-z]data[^a-z]{0,1}and[^a-z]{0,1}software[^a-z]{0,1}statement[^a-z]

Specific queries for assembling publications containing DAS mentions of selected repositories

The following search terms were employed within the DAS excerpts to identify deposits of datasets to open repositories from the selected list. Note that all queries were case insensitive.

- [^a-z]ENA[^a-z]
- [^a-z]Genbank[^a-z]
- [^a-z]SRA[^a-z]
- National Center for Biotechnology Information
- Sequence Read Archive
- European Nucleotide Archive
- ArrayExpress
- Array[^a-z]Express
- European Molecular Biology Laboratory
- [^a-z]DDBJ[^a-z]
- DNA Data Bank of Japan
- NCBI Assembly
- [^a-z]ClinicalTrials.gov[^a-z]
- [^a-z]DNA DataBank of Japan[^a-z]
- [^a-z]European Nucleotide Archive[^a-z]
- [^a-z]19[^a-z]data platform[^a-z]
- [^a-z]19[^a-z]data portal[^a-z]
- [^a-z]INSDC[^a-z]
- [^a-z]international nucleotide sequence database[^a-z]
- [^a-z]ncbi[^a-z]virus[^a-z]
- [^a-z]gsa[^a-z]
- [^a-z]genome sequence archive[^a-z]
- [^a-z]gisaid[^a-z]
- [^a-z]genome expression omnibus[^a-z]
- [^a-z]GEO[^a-z]

Intelligent open science

A case study of viral genomic data sharing during the COVID-19 pandemic

Research Consulting

The Ingenuity Centre, University of Nottingham Innovation Park, Nottingham, NG7 2TU, UK

www.research-consulting.com