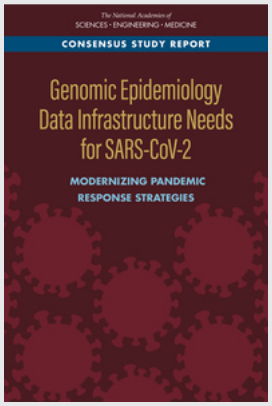


The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2

MODERNIZING PANDEMIC RESPONSE STRATEGIES

# Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies (2020)

## DETAILS

110 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-68091-2 | DOI 10.17226/25879

## CONTRIBUTORS

Committee on Data Needs to Monitor the Evolution of SARS-CoV-2; Board on Health Sciences Policy; Health and Medicine Division; Board on Life Sciences; Division on Earth and Life Studies; National Academies of Sciences, Engineering, and Medicine

## SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2020. *Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25879.

# Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2

## MODERNIZING PANDEMIC RESPONSE STRATEGIES

Committee on Data Needs to Monitor the Evolution of SARS-CoV-2

Board on Health Sciences Policy

Health and Medicine Division

Board on Life Sciences

Division on Earth and Life Studies

A Consensus Study Report of

*The National Academies of*
SCIENCES · ENGINEERING · MEDICINE

THE NATIONAL ACADEMIES PRESS
*Washington, DC*
**www.nap.edu**

**THE NATIONAL ACADEMIES PRESS** **500 Fifth Street, NW** **Washington, DC 20001**

This activity was supported by a contract between the National Academy of Sciences and the U.S. Department of Health and Human Services' Office of the Assistant Secretary for Preparedness and Response (75A50120G00002). Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-68091-2
International Standard Book Number-10: 0-309-68091-3
Digital Object Identifier: https://doi.org/10.17226/25879

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2020. *Genomic epidemiology data infrastructure needs for SARS-CoV-2: Modernizing pandemic response strategies*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25879.

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

**Consensus Study Reports** published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

**Proceedings** published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

## COMMITTEE ON DATA NEEDS TO MONITOR THE EVOLUTION OF SARS-CoV-2

**DIANE GRIFFIN** (*Chair*), Distinguished University Service Professor, W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health

**RALPH BARIC,** William R. Kenan, Jr. Distinguished Professor, University of North Carolina at Chapel Hill

**KENT KESTER,** Vice President and Head, Translational Sciences and Biomarkers, Sanofi Pasteur

**DEVEN McGRAW,** Chief Regulatory Officer, Ciitizen Corporation

**ALEXANDRA PHELAN,** Assistant Professor, Center for Global Health Science and Security, Georgetown University

**SASKIA POPESCU,** Senior Infection Preventionist, HonorHealth; Affiliate Faculty, George Mason University; Adjunct Professor, University of Arizona

**STUART RAY,** Professor of Medicine and Vice Chair of Medicine for Data Integrity and Analytics, Johns Hopkins University School of Medicine

**DAVID RELMAN,** Thomas C. and Joan M. Merigan Professor of Medicine and Professor of Microbiology and Immunology; Co-Director, Center for International Security and Cooperation, Stanford University; Chief of Infectious Diseases, Veterans Affairs Palo Alto Health Care System

**JULIE SEGRE,** Chief and Senior Investigator, Translational and Functional Genomics Branch, National Human Genome Research Institute, National Institutes of Health

**MARK SMOLINSKI,** President, Ending Pandemics

**PAUL TURNER,** Rachel Carson Professor of Ecology and Evolutionary Biology, Yale University

**DEBORAH ZARIN,** Program Director, Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard

*Liaison to the Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats*

**HARVEY FINEBERG,** President, Gordon and Betty Moore Foundation; Chair, Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats

*Study Staff*

**LISA BROWN,** Study Director
**EMMA FINE,** Associate Program Officer
**BENJAMIN KAHN,** Associate Program Officer
**STEVEN MOSS,** Associate Program Officer
**ANDREW M. POPE,** Senior Director, Board on Health Sciences Policy

*Science Writer*

**ANNA NICHOLSON**

# Reviewers

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

**ARAYINDA CHAKRAVARTI,** New York University
**MARK R. DENISON,** Vanderbilt University Medical Center
**KATHLEEN M. NEUZIL,** University of Maryland School of Medicine
**MARK A. ROTHSTEIN,** University of Louisville
**JOSHUA M. SHARFSTEIN,** Johns Hopkins Bloomberg School of Public Health

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by **SUSAN J. CURRY,** The University of Iowa, and **BOBBIE BERKOWITZ,** University of Washington. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the

National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

# Contents

**APPENDIXES**

# Boxes, Figures, and Tables

## BOXES

## FIGURES

## TABLES

# Acronyms and Abbreviations

| | |
|---|---|
| ARMoR | Antimicrobial Resistance Monitoring and Research [Program] |
| CARB | Combating Antibiotic-Resistant Bacteria National Action Plan |
| CDC | U.S. Centers for Disease Control and Prevention |
| COG-UK | COVID-19 Genomics UK |
| CoV | coronavirus |
| COVID-19 | coronavirus disease 2019 |
| DoD | U.S. Department of Defense |
| DPH | department of public health |
| GISAID | Global Initiative on Sharing All Influenza Data |
| HHS | U.S. Department of Health and Human Services |
| HIE | health information exchange |
| HIPAA | Health Insurance Portability and Accountability Act |
| ICU | intensive care unit |
| IHR | International Health Regulations |
| ILI | influenza-like illness |
| ILINet | Influenza-like Illness Surveillance Network |
| IRB | Institutional Review Board |
| MERS | Middle East respiratory syndrome |

| | |
|---|---|
| MIS-C | multisystem inflammatory syndrome in children |
| N3C | National COVID Cohort Collaborative |
| NCBI | National Center for Biotechnology Information |
| NIH | National Institutes of Health |
| NYC | New York City |
| OCR | Office for Civil Rights |
| PHI | protected health information |
| RdRp | RNA-dependent polymerase |
| RT-PCR | reverse transcription polymerase chain reaction |
| SARS | severe acute respiratory syndrome |
| SARS-CoV | severe acute respiratory syndrome coronavirus |
| SARS-CoV-2 | severe acute respiratory syndrome coronavirus 2 |
| SPHERES | Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance |
| WGS | whole genome sequencing |
| WHO | World Health Organization |

# Summary[1]

The 21st century has already seen the emergence of four pandemic viruses (chikungunya virus, Zika virus, 2009 H1N1 influenza virus, and severe acute respiratory syndrome coronavirus 2 [SARS-CoV-2]), several viral epidemics (e.g., 2003 SARS, 2012 Middle East respiratory syndrome [MERS-CoV], 2014 Ebola virus in West Africa, and 2018 Ebola virus in the Democratic Republic of the Congo), and intermittent sporadic outbreaks of other viruses such as H7N9 influenza. At the time of this writing, SARS-CoV-2 had spread worldwide, infecting at least 10 million people with an estimated 500,000 deaths within 6 months. Multiple outbreaks suggest that preparedness and response strategies need modernization. New advances in metagenomics, epidemiology, and big data analyses provide new paradigms for tracing symptomatic and asymptomatic transmission networks, thereby enabling our capacity to break or delay virus transmission to reduce morbidity and mortality. Recognizing this need, the U.S. Department of Health and Human Services' Office of the Assistant Secretary for Preparedness and Response and Office of Science and Technology Policy requested that the National Academies of Sciences, Engineering, and Medicine convene an ad hoc committee to lay out a framework to define and describe the data needs for a system to track and correlate viral genome sequences with clinical and epidemiological data. Such a system would help ensure the integration of data on viral evolution with detection, diagnostic, and countermeasure efforts.

[1] This Summary does not include references. Citations for the discussion presented in the Summary appear in the subsequent report chapters.

Previous efforts to integrate genomic, clinical, and epidemiological data have led to new insights around the transmission and pathogenesis of disease, including for previous outbreaks of SARS-CoV, Ebola virus, Zika virus, seasonal influenza, mumps, foodborne illnesses, and antibiotic-resistant bacteria. The most successful approaches to date have involved multipronged approaches and the timely collaboration of public and private stakeholders.

## CURRENT GENOMIC EPIDEMIOLOGY EFFORTS FOR SARS-CoV-2

Several ongoing efforts are leveraging the power of genomic epidemiology in response to the coronavirus disease 2019 (COVID-19) pandemic. In the United States, the U.S. Centers for Disease Control and Prevention's SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance consortium is working to coordinate a nationwide genomic sequencing effort. The National Institutes of Health supports the National COVID Cohort Collaborative (N3C), a secure portal for patient-level COVID-19 clinical data, and the National Center for Biotechnology Information's reference sequence database. Several regional initiatives have emerged as well, integrating data sharing through existing global efforts like the Global Initiative on Sharing All Influenza Data and Nextstrain. Even as new efforts are being established, the committee found that several limitations blunt their effectiveness, such as insufficient funding, poor coordination, limited capacity for data integration, unrepresentative data, and lack of an adequately trained workforce with the multifaceted expertise needed to conduct this work. Fundamental governance and collaboration issues extending from the top down have led to the fragmentation of approaches and varying capacities at local and national levels.

> *Conclusion: Current sources of SARS-CoV-2 genome sequence data, and current efforts to integrate these data with relevant epidemiological and clinical data, are patchy, typically passive, reactive, uncoordinated, and underfunded in the United States. As a result, currently available data are unrepresentative of many important population features, biased, and inadequate to answer many of the pressing questions about the evolution and transmission of the virus, and the relationships of genome sequence variants with virulence, pathogenesis, clinical outcomes, and the effectiveness of countermeasures. Thus, the viral sequence data and associated data needed are not being collected.*

**RECOMMENDATION 1. The U.S. Department of Health and Human Services should ensure the generation of representative, high-quality full genome sequences of SARS-CoV-2 across the United States, and in the future, from emerging epidemic or pandemic pathogens, in order that these data can be used to meet key needs for genomic surveillance.**

- **Pathogen samples must be obtained from individuals who represent a broad diversity of factors such as race and ethnicity, gender, age, geography, and other demographic features such as housing type, clinical manifestations and outcomes, and transmissibility.**
- **Capacity for genomic sequencing should be developed and supported at many geographically distributed sites performing testing, including public health laboratories and academic and medical centers.**
- **Representative SARS-CoV-2 clinical samples from across the United States should be collected and sequenced on an ongoing basis to provide baseline data and facilitate near-real-time transmission tracking.**
- **Genome sequences should be shared openly on publicly accessible databases, such as the National Center for Biotechnology Information linked to the Global Initiative on Sharing All Influenza Data.**

## BUILDING A FRAMEWORK TO TRACK AND CORRELATE VIRAL GENOME SEQUENCES WITH CLINICAL AND EPIDEMIOLOGICAL DATA

To understand the evolution of SARS-CoV-2 and the implications for transmission and clinical manifestations, the interpretation of genomic data (see Recommendation 1) is reliant on linked clinical and epidemiological data. Table S-1 briefly outlines how viral genome sequence data, when combined with other types of data, can be used to inform questions related to transmission, evolution, and clinical disease.

In order to answer the questions outlined in Table S-1, development of data integration will be crucial. Currently, no central repository exists for the collection and curation of infectious disease outbreak data from multiple sources such as federal, state, and local public health agencies; health care networks; and public health and clinical laboratories. In order to create a more integrated data system, insights can be gleaned from existing efforts to integrate data. Leveraging and expanding existing infrastructure and planning—through programs such as N3C—will be crucial to addressing the data infrastructure challenge in a way that is strategic, innovative, and iterative.

**TABLE S-1** Summary Table of Considerations for Transmission, Evolution, and Clinical Disease

| Goal | Question | Viral Genomic Sequence Data Needs | Clinical and/or Epidemiological Data Needs[a] |
|---|---|---|---|
| Transmission patterns | Is outbreak due to multiple introductions? Where is the virus coming from? | Pathogen samples from individuals who represent broad diversity from outbreaks and many regions/countries | Time and place of virus isolation and travel history of cases |
| | Is outbreak due to local spread? How and/or where is the virus being transmitted? | Sequences from local groups/areas with increased incidence rates | Local population-based information on sites of exposure, gatherings, isolated communities, and congregate living (long-term care facilities, hospitals, prisons) |
| | Is there evidence of super-spreading events and how important are they? | Sequences of virus from groups of people infected in the same setting | Information on sites of exposure, gatherings |
| Evolution/influence of selective pressures | Is the virus changing in transmissibility? | Changes in viral genome sequence associated with increased spread | Calculations of $R_0$ (contact tracing data–number of people infected) |
| | Is resistance to antiviral drugs or other treatments changing? | Changes in viral genome associated with failure to respond to treatment | Hospital or health care center data on patients who do not respond to therapy or show failure of treatment |
| | Is there altered escape from the host immune response/within host evolution? | Changes in viral genome associated with persistence | Hospital data on patients who show prolonged shedding |
| | Is there changed protection from vaccine-induced immunity? | Changes in virus that affect epitopes important for protective immunity and sequences of viruses associated with vaccine failure | Vaccine trial databases and post-marketing vaccine failures |

**TABLE S-1** Continued

| Goal | Question | Viral Genomic Sequence Data Needs | Clinical and/or Epidemiological Data Needs[a] |
|---|---|---|---|
| Clinical disease | Are there strains/ mutations associated with changes in disease severity? | Sequences of viruses from patients with different disease severity | Severity of symptoms, ICU, ventilation, mortality, length of hospitalization, co-infections |
| | Are there strains/ mutations that affect virus loads or clearance? | Sequences of viruses from patients with viral load data | RT-PCR data to measure viral load of respiratory secretions, blood, and feces over time |
| | Are there strains/ mutations that affect response to different treatments? | Sequences of viruses from before and after treatment | Treatment type, duration, and outcome |
| | Are there strains/ mutations that are associated with response to different treatments? | Sequences of viruses from different body sites and patients with and without specific complications | Clinical data on complications related to different organ systems (e.g., kidney, liver, nervous system) |
| | Are there strains/ mutations that predispose to MIS-C? | Sequences of viruses from children in the same community/family with and without MIS-C | Clinical data over time on immune response, viral load, treatment, and response |

NOTE: ICU = intensive care unit; MIS-C = multisystem inflammatory syndrome in children; $R_0$ = basic reproduction number; RT-PCR = reverse transcription polymerase chain reaction.

[a] The committee recognizes that clinical and epidemiological data often come from very different data collection sources and efforts, but for the purposes of this table these data needs have been incorporated into one column.

**RECOMMENDATION 2. The U.S. Department of Health and Human Services should develop and invest in a national data infrastructure system that constructively builds on existing programmatic infrastructure with the ability to accurately, efficiently, and safely link genomic data, clinical data, epidemiological data, and other relevant data across multiple sources critical to a public health response such as the current SARS-CoV-2 outbreak. Such a system should:**

- **Allow for the linkage of genomic data, clinical data, epidemiological data, and other relevant data in a way that is not overly burdensome to laboratories that collect data regularly.**

- **Create and foster safe data-sharing practices to ensure that individuals' personal identifying information remains unexposed when data are being used and shared across the system.**
- **Be grounded in the pursuit of standardization, interoperability, flexibility, and the practical linkage of data, including consideration of a potential national patient identifier.**
- **Consider not only the data required to create such a system, but also investment in mechanisms supporting the collection and analysis of such data, including promoting formal education in "data wrangling" at the intersection of data science and infectious disease epidemiology.**
- **Conduct regular annual reviews—including scenario-based simulations—to identify capacity gaps, promote process improvement (based on existing U.S. infrastructure to assess the annual risk of seasonal influenza, work could improve usability and coverage of health information exchanges, and other initiatives), and ensure inclusion of entities with supporting functions across scales—including private health care systems that provide data or state and local public health laboratories that collect data—in ongoing system development and evaluation.**

## GOVERNANCE AND LEADERSHIP

In the United States, federal or state laws do not protect or mandate sharing of samples of viral sequence data. As such, any sharing of such data and samples is done voluntarily and generally without concerns about possible regulatory barriers. Conversely, federal and state laws protect clinical and epidemiological data, including through the Health Insurance Portability and Accountability Act and the Common Rule at the federal level. The sharing of viral sequence data and associated information should be guided by national-level leadership to create supportive legal or strategic frameworks that instill principles of good governance. These data-sharing and reporting processes should be clearly established and resourced as an urgent matter, and prior to an emergency. Without a clear and urgent public health rationale, changing reporting processes during an emergency should be avoided, and emergencies should not justify not complying with principles of good governance, including data transparency. Principles and elements of good governance include accountability processes that clarify authorities and responsibilities, as well as maintenance of transparency, equity, participation, and clear and certain legal protections for public health agencies, researchers, and individuals' rights.

**RECOMMENDATION 3. The U.S. Department of Health and Human Services should establish an effective and sustainable science-driven leadership and governance structure for the use of SARS-CoV-2 genome sequences in addressing critical national public health and basic science issues, develop a national strategy, and ensure the funding needed for successful execution of the strategy.**

- **Leaders of this effort must have sufficient authorities and responsibilities to ensure that key issues are identified and prioritized, representative data are generated, and barriers to data sharing are diminished.**
- **A national strategy for SARS-CoV-2 genome sequences linked to clinical and epidemiological data should be developed that articulates goals, priorities, and a path for achieving them.**
- **A board with diverse relevant expertise should be established with broad authority to oversee and advise the national strategy for SARS-CoV-2 genome sequences linked to clinical and epidemiological data, and the delivery of actionable data for related investigations.**

# 1

# Introduction

In December 2019, new cases of severe pneumonia were first detected in Wuhan, China, and the cause was determined to be a novel beta coronavirus related to the severe acute respiratory syndrome (SARS) coronavirus that emerged from a bat reservoir in 2002 (Wu et al., 2020). Within 6 months, this new virus—SARS coronavirus 2 (SARS-CoV-2)—has spread worldwide, infecting at least 10 million people with an estimated 500,000 deaths. Coronavirus disease 2019 (COVID-19), the disease caused by SARS-CoV-2, was declared by the World Health Organization a public health emergency of international concern on January 30, 2020, and a pandemic on March 11, 2020 (WHO, 2020b). To date, there is no approved effective treatment or vaccine for COVID-19, and it continues to spread in many countries. COVID-19 has caused unprecedented global economic and social disruption. In the United States alone, 25 million people have become unemployed and the real gross domestic product contracted 4.8 percent at an annual rate during the first quarter of 2020, with projected losses increasing moving forward (CBO, 2020). Surging numbers of severely ill patients have strained health systems, and population lockdowns to curtail virus transmission have disrupted social interactions, education, and businesses small and large.

Clearly, multiple outbreaks suggest that preparedness and response strategies need modernization. Modern advances in DNA sequencing, genomics, epidemiology, and big data analyses provide new paradigms for tracing symptomatic and asymptomatic transmission networks and identifying sites of spread and at-risk populations, thereby enabling the capacity to break or delay virus transmission to mitigate social and economic disruption and reduce morbidity and mortality. Doing so also allows limited resources to be targeted to key sites of disease expansion, such as long-term care facilities or specific places of work. Another advantage of these 21st-century pathogen disease-tracing methods is that they provide critical time for the implementation of public health intervention strategies, medical countermeasure development, and disease control. As the 21st century has already seen the emergence of four pandemic viruses (chikungunya virus, Zika virus, 2009 H1N1 influenza, and SARS-CoV-2), several viral epidemics (2003 SARS, 2012 Middle East respiratory syndrome [MERS]-CoV, 2014 Ebola virus in West Africa, and 2018 Ebola virus in the Democratic Republic of the Congo), and intermittent sporadic outbreaks of other viruses such as H7N9 influenza (WHO, 2020a), global health dictates a critical need for modernization and integration of public, private, and federal public health response efforts designed for rapid deployment to protect the health of populations and the economy.

## CORONAVIRUS EVOLUTION AND SARS-CoV-2

Coronaviruses demonstrate the capacity for continuous emergence to cause significant and potentially pandemic disease in humans and animals. In the 21st century, three new human coronaviruses have emerged to cause epidemic or pandemic disease outbreaks including SARS-CoV in 2003, MERS-CoV in 2012, and SARS-CoV-2 in 2019 (WHO, 2020a). Concomitantly, three novel coronaviruses have emerged in the 21st century to cause major pandemics in swine, including porcine epidemic diarrhea virus, porcine delta coronavirus, and severe acute diarrhea disease virus in China (Vlasova et al., 2020). The continual emergence of coronavirus epidemics and pandemics underscores the critical importance of developing robust metrics of genome evolution, which could be used to inform the medical and public health communities of high-impact mutations and changes in evolutionary trajectories that might impact spread to human or domesticated animals.

Coronaviruses have large (28–32 kb), message-sense RNA genomes. The replicative machinery of the virus is encoded in the first 20 kb of the genome as two large open reading frames. Downstream of this machinery, all coronaviruses encode essential structural proteins, membrane (M), envelope (E), spike (S), and nucleocapsid (N). For SARS-CoV-2, the best characterized human epithelial cell receptor is angiotensin-converting enzyme 2, which is bound by the virus's spike protein. This S protein contains the receptor-binding domain that is the target for neutralizing antibody and the focus of vaccine development efforts. RNA viruses depend on an RNA-dependent polymerase (RdRp) to replicate the viral genome. Much of the reason for the high mutation rates in RNA viruses is the error-prone nature of RdRp enzymes. To compensate for their large genome size, coronaviruses have adapted to mutate less frequently by encoding a novel proofreading 3′-to-5′ exoribonuclease that associates with the polymerase complex to improve genome replication fidelity (Smith et al., 2015).

Although coronavirus mutation rates are slower than for many other RNA viruses, recombination, insertion, and deletion events also produce changes in the viral genome. Genetic recombination drives the creation of viral diversity in many positive-strand RNA viruses by the formation of novel chimeric genomes. During controlled mixed infections in vitro, rates of coronavirus genome RNA recombinations approach 25 percent or more and can be accompanied by deletions and insertions. Intergenic and intragenic recombination allow for rapid acquisition of novel functions and modular exchange of functional components between viruses. Under natural conditions, recombination-based processes have resulted in viruses with altered host range as well as altered immunogenicity and virulence, and thus provide a rapid approach to escape antibody neutralization (Ballesteros et al., 1997; Gallagher et al., 1990; Sánchez et al., 1992).

These processes provide extensive opportunities to overcome reductions due to population bottlenecks caused by antiviral drugs, host immunity, or human-to-human and animal-to-human transmission events. Monitoring RNA virus genomic change is an important step for anticipating viral emergence, predicting disease severity, evaluating drug and vaccine performance, and tracing symptomatic and asymptomatic transmission networks throughout a host population. An important consideration for any genetic epidemiology study is an understanding of the basic factors and selective pressures influencing viral evolution. Monitoring the evolution of genetic diversity in SARS-CoV-2 has the potential to inform targets for countermeasures, such as antiviral drugs and vaccines, and to improve diagnostics (van Dorp et al., 2020).

Phylogenetic studies estimate that SARS-CoV-2 spilled over to humans in late 2019 (Dawood, 2020; van Dorp et al., 2020; Wu et al., 2020). One study of more than 7,000 sequences found 198 sites at which the SARS-CoV-2 genome appeared to have already undergone recurrent independent mutations (van Dorp et al., 2020). Analysis of genomic sequences from the Global Initiative on Sharing All Influenza Data database found eight novel recurrent mutations and potentially coexistent European, North American, and Asian strains characterized by different mutation patterns (Pachetti et al., 2020).

## THE POWER OF GENOMICS IN UNDERSTANDING SARS-CoV-2

Viral genome sequence data are an increasingly important tool for detecting and understanding the spread of infectious diseases in real time and for mounting effective responses. Advances in the speed, granularity, affordability, and portability of genomic sequencing technologies have created transformative potential for widespread rapid genomic surveillance during infectious disease outbreaks, particularly when data from genomic sequencing are integrated with and analyzed alongside patient-based clinical and population-based epidemiological data (Ladner et al., 2019). The rapidly advancing field of phylodynamics uses Bayesian statistical frameworks to obtain both epidemiological and evolutionary information from pathogens to trace their history, infer transmission dynamics, and construct phylogenetic trees (Baele et al., 2016; Grenfell et al., 2004). Prior to the advent of viral genome sequencing and phylodynamics, estimates of critical epidemic parameters to inform the public health response to an outbreak, such as the basic reproductive number ($R_0$), relied exclusively on epidemiological incidence data (Grubaugh et al., 2019). Today, the ability to collectively harness genomic, epidemiological, and clinical data contributes to enhanced, multidimensional understanding of an outbreak and enables molecularly precise and targeted responses that were not previously possible. Sequencing of viral genomes can help to answer—and in some cases, may provide the only way to answer—questions that are foundational to understanding, mitigating, and controlling a virus outbreak: the identity and novelty of the causal virus; its origin in terms of reservoir host and geography; its introduction in humans; its linkages with other outbreaks; and its potential to evolve and locally adapt (Grubaugh et al., 2019).

When combined with epidemiological information, genomic sequencing can be particularly useful for investigating outbreaks of RNA viruses, such as SARS-CoV-2. Calls have already been made to enhance the response to the COVID-19 pandemic by integrating genomic, epidemiological, and clinical data (Koks et al., 2020). Rapidly developing and deploying precise and targeted interventions and countermeasures will require a more granular understanding of exposure and infection by virus variants within and across populations, as well as the genetic, comorbidity, social, and environmental cofactors that modulate disease severity.

Virus genome sequencing is a cornerstone of the emerging field of "genomic epidemiology," which leverages phylodynamic approaches to clarify pathogen transmission patterns and events in greater detail than is possible with traditional epidemic investigations (Gardy and Loman, 2018; Grubaugh et al., 2019) (see Box 1-1).

Increasing evidence supports the value of viral genome sequence data across all stages of an infectious disease outbreak. During the initial stages of an outbreak, unbiased DNA sequencing of infected tissue can help to genetically identify the causative novel pathogen from which rapid screening tests can be developed. The data can also contribute to identifying the reservoir host and geographic location of the virus's origin. Compared to traditional approaches, such as interview-based contact tracing, approaches

**BOX 1-1**
**Defining Genomic Epidemiology**

Genomic epidemiology is defined as the use of pathogen genome sequencing to understand infectious disease transmission and epidemiology.

SOURCE: Adapted from Gardy and Loman, 2018.

that integrate genomics offer quicker and more comprehensive methods to build an understanding of a virus's transmission chain and dynamics (e.g., human-to-human or zoonotic) (Grubaugh et al., 2019; Houldcroft et al., 2017). When integrated with other sets of contextual metadata, genomic epidemiology has transformed the ability to map the spatiotemporal patterns, social drivers, and transmission chains through which cases are emerging as an outbreak continues to unfold (Grubaugh et al., 2019). For example, genomic data and location can serve as proxies for estimating epidemiological connections (Wohl et al., 2020). Understanding the spatiotemporal characteristics of virus transmission—within and across different populations—as well as the virus's genetic changes over time can inform the design of more effective, targeted interventions and countermeasures (Ladner et al., 2019). During periods between outbreaks, genomic epidemiology also contributes to tracking the evolution and transmission dynamics of viruses in both humans and reservoir species (Grubaugh et al., 2019).

The use of genomic data has substantial practical implications for public health practice around infectious disease control by improving the capacities for ongoing surveillance, rapid diagnosis, and real-time disease tracking (Gardy and Loman, 2018). In situations where there is prior knowledge of mutations that affect specific characteristics of the virus such as virulence, drug susceptibility and antigenicity, whole genome sequencing (WGS) of pathogens—which can now be conducted in (near) real time directly from clinical samples—can provide this information during an outbreak (Koks et al., 2020; Ladner et al., 2019). This information can also enable point-of-care molecular diagnostics and inform individualized treatment regimens akin to the use of human genetic data in precision medicine (Gardy and Loman, 2018; Houldcroft et al., 2017) (see Figure 1-1). At the population level, WGS combined with epidemiological data can use pathogen mutations as markers of transmission events to "reveal patterns of epidemic transmission at a fine-scale resolution" to inform more precise and targeted large-scale public health interventions than traditional approaches (Ladner et al., 2019). WGS of pathogens fits within the broader

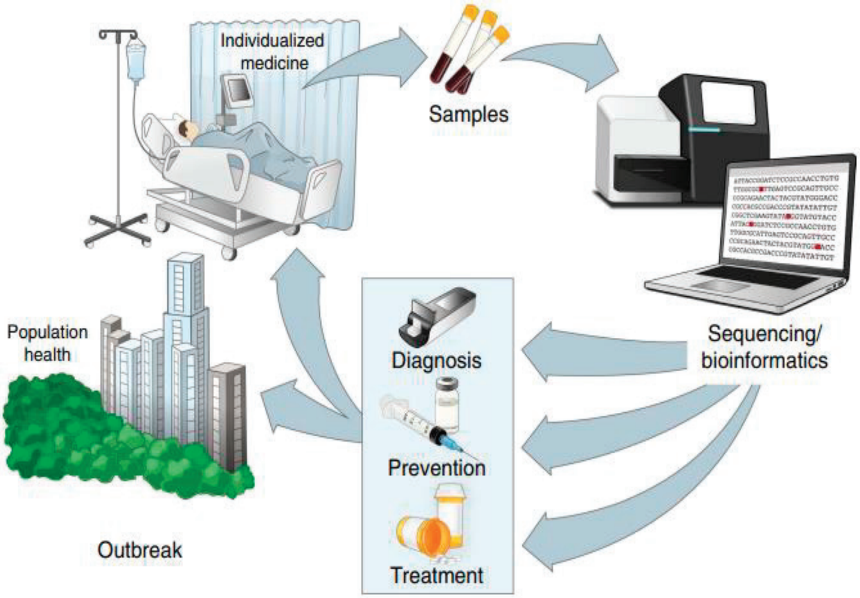


**FIGURE 1-1** Pathogen sequencing during infectious disease outbreaks can inform precise interventions.
SOURCES: Reprinted by permission from Springer Nature: Springer Nature, *Nature Medicine*, Precision epidemiology for infectious disease control; Ladner et al., 2019.

paradigm of the One Health approach, which considers human, animal, and environmental health as a whole. Given that most emerging infectious diseases have zoonotic origins and they often spillover to humans in settings of high biological diversity, the application of genomic epidemiology across all three domains could bolster the One Health approach to surveillance, prevention, and control of those diseases (Gardy and Loman, 2018).

## STUDY CHARGE

After a rapid telephonic consultation on May 7, 2020,[1] with the U.S. Department of Health and Human Services' Office of the Assistant Secretary for Preparedness and Response and Office of Science and Technology Policy, the National Academies of Sciences, Engineering, and Medicine convened an ad hoc committee to lay out a framework to define and

[1] See https://www.nationalacademies.org/event/05-07-2020/standing-committee-on-emerging-infectious-diseases-and-21st-century-health-threats-expert-call-on-genotypic-drift-and-potential-phenotypic-manifestations-of-sars-cov-2 (accessed June 25, 2020).

describe the data needs for a system to track and correlate viral genome sequences with clinical and epidemiological data. Such a system would help ensure the integration of data on viral evolution with detection, diagnostic, and countermeasure efforts. The full charge to the committee is presented in Box 1-2. The committee comprises experts in the fields of infectious disease and epidemiology; clinical care; immunology, evolutionary biology, microbiology, and molecular genetics; data sharing and genomic surveillance; legal and regulatory issues; and therapeutic and diagnostic development. The biographies of the committee members are presented in Appendix A.

### BOX 1-2
### Statement of Task

The National Academies of Sciences, Engineering, and Medicine will establish an ad hoc committee to lay out a framework to define and describe the data needs for a system to track and correlate viral genome sequences with clinical and epidemiological data. This system would help ensure the integration of data on viral evolution with detection, diagnostic, and countermeasure efforts. Issues to be considered include

- Data collection mechanisms to ensure a representative global sample set of all relevant extant sequences, with respect to geography, time since beginning of pandemic, patient subpopulations, and sample type, among other potential sources of sequence diversity.
- Challenges and opportunities for coordination across existing domestic, global, and regional data sources.
- Rates and mechanisms of genome evolution as a function of geography, time since introduction into human population, and host features such as immune status.
- Correlation of viral genotype with clinical outcomes.
- Correlation of variant viral sequence features with viral escape from host immune recognition or vaccine-induced immunity, therapeutic beneficial effects, or detection by currently deployed methods.
- Identification of features of viral genome evolution most relevant to the design and development of therapeutics and vaccines.
- Potential for viral evolvability in response to selection pressure.

The ad hoc committee will produce a short consensus report with recommendations to address these issues.

## ABOUT THIS REPORT

### Study Approach

As the nation is in the midst of the pandemic, the committee deliberated and developed this report and the recommendations presented herein on a compressed timeline. The committee held three virtual meetings in June 2020, two of which included public information-gathering sessions that allowed the committee to hear from the study sponsors and other experts and stakeholders. At the first meeting, a representative of the U.S. Centers for Disease Control and Prevention's SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance initiative spoke about the program. At the second meeting, the public session included several speaker panels covering scientific principles of viral evolution (including genomic epidemiology and phylodynamics), policy and ethical concerns, and examples from prior and ongoing initiatives. The public meeting agendas can be found in Appendix B. Staff and committee members conducted targeted searches of literature to ensure adequate background knowledge of the issue at the time of this writing. Given the rapidly evolving nature of the work around SARS-CoV-2, the committee closely monitored ongoing initiatives and concurrent, complementary work throughout the study process.

### Organization of the Report

The report is organized into five chapters. Chapter 2 discusses applications of genomic epidemiology in previous infectious disease outbreaks and Chapter 3 highlights current efforts related to SARS-CoV-2. Together, these chapters explore the evolution of genomic epidemiology to present day. Chapter 4 presents a framework to track and correlate viral genome sequences with clinical and epidemiological data and Chapter 5 discusses regulatory and governance considerations.

## REFERENCES

Baele, G., M. A. Suchard, A. Rambaut, and P. Lemey. 2016. Emerging concepts of data integration in pathogen phylodynamics. *Systematic Biology* 66(1):e47–e65.

Ballesteros, M. L., C. M. Sánchez, and L. Enjuanes. 1997. Two amino acid changes at the n-terminus of transmissible gastroenteritis coronavirus spike protein result in the loss of enteric tropism. *Virology* 227(2):378–388.

CBO (Congressional Budget Office). 2020. *Interim economic projections for 2020 and 2021.* https://www.cbo.gov/system/files/2020-05/56351-CBO-interim-projections.pdf (accessed June 25, 2020).

Dawood, A. A. 2020. Mutated COVID-19 may foretell a great risk for mankind in the future. *New Microbes and New Infections* 35:100673.

Forster, P., L. Forster, C. Renfrew, and M. Forster. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences* 117(17):9241–9243.

Gallagher, T. M., S. E. Parker, and M. J. Buchmeier. 1990. Neutralization-resistant variants of a neurotropic coronavirus are generated by deletions within the amino-terminal half of the spike glycoprotein. *Journal of Virology* 64(2):731–741.

Gardy, J. L., and N. J. Loman. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* 19(1):9–20.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656):327–332.

Grubaugh, N. D., J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, and K. G. Andersen. 2019. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology* 4(1):10–19.

Houldcroft, C. J., M. A. Beale, and J. Breuer. 2017. Clinical and biological insights from viral genome sequencing. *Nature Reviews Microbiology* 15(3):183–192.

Koks, S., R. W. Williams, J. Quinn, F. Farzaneh, N. Conran, S. J. Tsai, G. Awandare, and S. R. Goodman. 2020. COVID-19: Time for precision epidemiology. *Experimental Biology and Medicine (Maywood)* 245(8):677–679.

Ladner, J. T., N. D. Grubaugh, O. G. Pybus, and K. G. Andersen. 2019. Precision epidemiology for infectious disease control. *Nature Medicine* 25(2):206–211.

Pachetti, M., B. Marini, F. Benedetti, F. Giudici, E. Mauro, P. Storici, C. Masciovecchio, S. Angeletti, M. Ciccozzi, R. C. Gallo, D. Zella, and R. Ippodrino. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine* 18(1):179.

Sánchez, C. M., F. Gebauer, C. Suñé, A. Mendez, J. Dopazo, and L. Enjuanes. 1992. Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. *Virology* 190(1):9–105.

Smith, E. C., J. B. Case, H. Blanc, O. Isakov, N. Shomron, M. Vignuzzi, and M. R. Denison. 2015. Mutations in coronavirus nonstructural protein 10 decrease virus replication fidelity. *Journal of Virology* 89(12):6418–6426.

van Dorp, L., M. Acman, D. Richard, L. P. Shaw, C. E. Ford, L. Ormond, C. J. Owen, J. Pang, C. C. S. Tan, F. A. T. Boshier, A. T. Ortiz, and F. Balloux. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* 83:104351.

Vlasova, A. N., Q. Wang, K. Jung, S. N. Langel, Y. S. Malik, and L. J. Saif. 2020. Porcine coronaviruses. *Emerging and Transboundary Animal Viruses* 79–110.

WHO (World Health Organization). 2020a. *Disease outbreaks by year*. https://www.who.int/csr/don/archive/year/en (accessed July 2, 2020).

WHO. 2020b. *Timeline of WHO's response to COVID-19*. https://www.who.int/news-room/detail/29-06-2020-covidtimeline (accessed July 7, 2020).

Wohl, S., H. C. Metsky, S. F. Schaffner, A. Piantadosi, M. Burns, J. A. Lewnard, B. Chak, L. A. Krasilnikova, K. J. Siddle, C. B. Matranga, B. Bankamp, S. Hennigan, B. Sabina, E. H. Byrne, R. J. McNall, R. R. Shah, J. Qu, D. J. Park, S. Gharib, S. Fitzgerald, P. Barreira, S. Fleming, S. Lett, P. A. Rota, L. C. Madoff, N. L. Yozwiak, B. L. MacInnis, S. Smole, Y. H. Grad, and P. C. Sabeti. 2020. Combining genomics and epidemiology to track mumps virus transmission in the United States. *PLOS Biology* 18(2):e3000611.

Wu, F., S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579(7798):265–269.

# 2

# Application of Genomic Epidemiology in Previous Infectious Disease Outbreaks

When genomic, clinical, and epidemiological data analyses are integrated, they can provide a real-time picture of an outbreak that is more nuanced, precise, and rich than any of the data types considered in isolation. Genomic epidemiology can inform rapid deployment of targeted interventions to protect the public as an outbreak unfolds. Previous efforts to combine and analyze viral genome sequence data with clinical and epidemiological data have demonstrated the value of this approach and have indicated its potential to transform the response to infectious disease outbreaks in the future.

## PREVIOUS EFFORTS TO INTEGRATE ANALYSES OF GENOMIC, CLINICAL, AND EPIDEMIOLOGICAL DATA

### SARS-CoV

The severe acute respiratory syndrome coronavirus (SARS-CoV) outbreak was among the first epidemics characterized by extensive genome sequencing and was defined by a chronological set of sequence changes, providing a unique opportunity to identify the genetic basis for zoonotic virus transmission and pathogenesis across species during a growing epidemic (Chinese SARS Molecular Epidemiology Consortium, 2004; Kan et al., 2005; Lu et al., 2004; Ruan et al., 2003). First, direct sequencing of clinical samples from the index patients identified a novel coronavirus as the microbial origin of disease, which led to rapid development of diagnostic tests. Sequence analyses and epidemiological data demonstrated that several independent zoonotic strain introductions caused human cases and this predated the expanding outbreak in early fall 2002. Second, hospitals were epicenters for disease expansion events characterized by sequential mutation and super-spreading events, leading to travelers seeding cases and hospitals in Hong Kong, Taiwan, and Vietnam and then globally. Sequence comparisons with SARS-like CoV in open markets rapidly identified civets and raccoon dogs as potential reservoirs for human infections, leading to the rapid closure of open markets and suppression of the expanding outbreak (Guan et al., 2003). In general by comparing the early, middle-, and late-phase human isolates (GZ02, CHUK-W1, and Urbani, respectively) to the civet isolates SZ16 or HZ/SZ/61/03, sequencing revealed 9–12 amino acid changes in ORF1a and ORF1b, 6–17 in the S glycoprotein, 3–4 in ORF3a, 1 in the M glycoprotein, and the ORF8 29 nucleotide deletion. Insufficient patient metadata usually prevented definitive association of any particular mutation with cross-species transmission, person-to-person transmission, and increased pathogenesis and virulence. In the S glycoprotein, changes at K479N and to a lesser extent S487T were shown to be essential for efficient hACE2 recognition and human infection, while the change G5S in the M glycoprotein appears to enhance virus yields per cell (Li et al., 2005; Pacciarini et al., 2008). The functions of the other mutations that evolved during the expanded SARS-CoV 2003 epidemic remain unknown. Thus, genomic sequencing led to identification of the source of human infection and mutations that facilitated human infection.

### Ebola Virus

The power of genomic epidemiology to shape the response to an infectious disease outbreak became increasingly evident during the Ebola virus epidemic in West Africa (Gardy and Loman, 2018; Ladner et al., 2019). In studies published around the outbreak's peak, large-scale whole genome sequencing efforts using portable sequencing platforms and real-time analyses were able to elucidate transmission dynamics and trends. For instance, viral genomic sequencing was used to establish the outbreak's origin in a single spillover event—rather than multiple zoonotic transmissions—followed by sustained human-to-human transmission (Holmes et al., 2016).[1] Crucially, analyses of genomic and epidemiological data later established that there was transmission through breast milk (Holmes, 2017) and that sexual transmission was possible for asymptomatic survivors long after they had recovered (Christie et al., 2015; Diallo et al., 2016; Mate et al., 2015),

[1] In contrast to the transmission dynamics for Ebola virus, genomic epidemiology efforts were able to establish that human cases of Lassa fever, which is endemic in areas of West Africa, have been caused by multiple distinct spillover events from rodents rather than extensive human-to-human transmission (Andersen et al., 2015; Ladner et al., 2019; Siddle et al., 2018).

a previously unknown factor that contributed to both the initial epidemic and subsequent flare-ups (Whitmer et al., 2018). This led to immediate changes in policy and guidance for male survivors to include the recommendation that they have repeated semen tests for the presence of Ebola virus RNA after recovery (Ladner et al., 2019), as well as the implementation of behavior modifications to reduce transmission. Genome sequencing studies during the epidemic were able to identify mutations that were later determined to be likely instances of novel adaptation to human hosts (Diehl et al., 2016; Dietzel et al., 2017; Urbanowicz et al., 2016). More exhaustive post hoc genomic sequencing reconstructed the geographic spread of the virus across the region, suggesting possible missed interventions that could be incorporated for future preparedness planning (Dudas et al., 2017). Ebola virus demonstrated the critical role for genomic–epidemiological analyses to span local, regional, and national levels, integrated as a real-time response during an epidemic that included transmission from asymptomatic carriers (Gardy and Loman, 2018).

## Zika Virus

During the 2015–2016 epidemic of the Zika virus in the Americas, whole genome sequencing, phylogenetic molecular clock analysis, and epidemiological data for this mosquito-borne virus revealed that viral strains in the Americas share a common ancestor with strains in French Polynesia and that the virus was likely circulating in Brazil more than 1 year earlier than initially believed (Faria et al., 2016, 2017; Metsky et al., 2017). Genomic surveillance also revealed a previously unreported 2017 outbreak of Zika in Cuba based on viral sequencing of samples collected in the United States and Europe from travelers to Cuba. Genomics indicated viral introduction from neighboring islands in 2016, 1 year after peak transmission elsewhere in the Caribbean, indicating that aggressive anticipatory mosquito control measures by Cuban public health authorities likely delayed the outbreak (Grubaugh et al., 2019). Sequencing of Zika virus genomes from humans and mosquitoes in Florida suggested that hundreds of introductions of the virus may have been necessary to drive that outbreak (Grubaugh et al., 2017, 2018). Therefore, genomic data can inform outbreak response and mitigation efforts through traveler education and surveillance (Ladner et al., 2019).

## Seasonal Influenza

Existing domestic and international networks for tracking the epidemiology of seasonal influenza have leveraged genomic epidemiology to track strain variants and understand vaccine responses in ways that are directly applicable to SARS-CoV-2. Influenza viruses evolve constantly, making surveillance crucial for vaccine development and modification over time. Because annual seasonal influenza vaccines are developed and produced ahead of flu season based on expert predictions, ongoing data collection is critical for rapidly identifying the major strain(s) circulating in humans during influenza season and assessing the antigenic properties of new strains. The U.S. Centers for Disease Control and Prevention's (CDC's) FluView[2] is a robust surveillance tool for tracking the prevalence of influenza viruses over time and in different population groups. The system collects data from state, local, tribal, and territorial health departments about the location and timing of influenza activity, viruses that are circulating and how they are changing, and clinical outcomes such as outpatient illness, hospitalization, and death (CDC, 2020). The Global Initiative on Sharing All Influenza Data (GISAID) is a network launched by scientists to enable broader sharing of the genetic sequences of influenza viruses (GISAID, 2020). GISAID plays a key role in facilitating data sharing among World Health Organization Collaborating Centers and National Influenza Centers, as well as contributing to vaccine recommendations (GISAID, 2020). In the context of the coronavirus disease 2019 (COVID-19) pandemic, similar collection of viral genetic sequence data, clinical data, epidemiological data, geographic data, and animal virus data could be harnessed to help researchers understand how SARS-CoV-2 is evolving and spreading (GISAID, 2020).

### Mumps

Mumps, which is an infectious disease caused by a paramyxovirus, has been well controlled in the United States with a 99 percent decline in incidence since a vaccine was introduced in 1967 (CDC, 2019). However, from 2016–2017, hundreds of cases were reported to the Massachusetts Department of Public Health by 18 colleges and universities, as well as other close-contact settings (Wohl et al., 2020). Viral genomic sequencing of 158 cases from Massachusetts and 43 cases from other states, collected contemporaneously, revealed that a single viral lineage has been circulating since 2006 mostly within the United States, rather than globally (Wohl et al., 2020). Finer scale genomic analyses demonstrated that multiple introductions of individuals (some with travel history to other regions) onto college campuses started smaller clusters of transmission. Importantly, this study demonstrated how clinical metadata such as "contact links" (e.g., dormitories) or "activity links" (e.g., sports teams or clubs) could be aggregated and combined with genomic data to identify risk factors, untangle, and stem a local–national outbreak of mumps (Wohl et al., 2020).

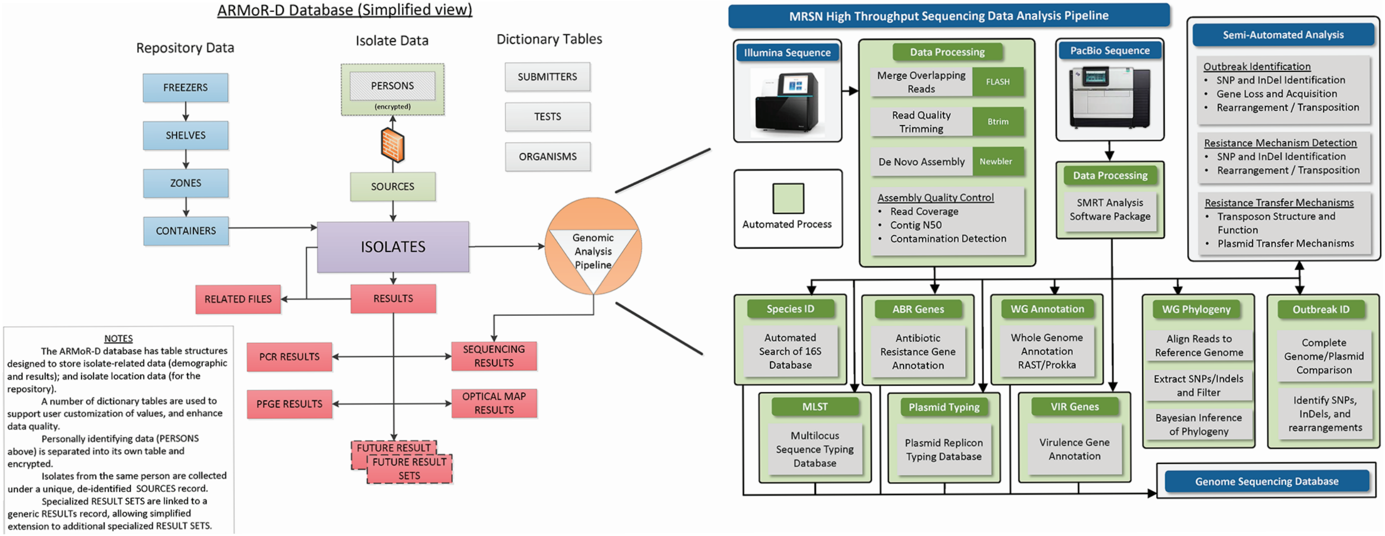
[2] See https://www.cdc.gov/flu/weekly/fluviewinteractive.htm (accessed June 25, 2020).

### Antibiotic-Resistant Bacteria

Genomic epidemiology studies of multidrug-resistant bacteria have demonstrated the capacity for sequencing within hospitals and public health labs to generate clinically actionable data with which to optimize strategies to prevent transmission within hospitals. As the evolution of resistance continues to outpace the development of new antibiotics, preventing transmission of multidrug-resistant bacteria is critical to the mitigation of this threat. Controlling the spread of multidrug-resistant bacteria has required regions and localities to develop systems to integrate patient movement and genomic data. For example, as carbapenem-resistant *Klebsiella pneumoniae* was expanding its hold in U.S. hospitals, one center undertook genomic sequencing combined with epidemiology to reconstruct its earliest transmission events and articulate specific surveillance and clinical practices, which helped to stem an outbreak (Snitkin et al., 2012). Similarly, expanded genomic–epidemiological analyses demonstrated how the transmission of carbapenem-resistant *Klebsiella pneumoniae* across a four-county area of Illinois was linked by a patient-sharing network of hospitals, nursing homes, and long-term acute care facilities (Snitkin et al., 2012, 2017). As part of the National Action Plan for Combating Antibiotic-Resistant Bacteria, a national strategy was established that included the Antibiotic Resistance Laboratory Network to perform and coordinate genomic–epidemiological studies.[3] Genomic sequence data have made it easier for facilities to communicate about possible undetected transmissions.

Also in response to the escalating threat from multidrug-resistant bacteria, the U.S. Department of Defense (DoD) launched in 2009 the Antimicrobial Resistance Monitoring and Research (ARMoR) Program, its own global network to track and characterize multidrug-resistant bacteria (Lesho et al., 2014, 2016). The ARMoR Program is an enterprise-wide initiative, consisting of epidemiologists, bioinformaticists, microbiology researchers, policy makers, hospital-based infection preventionists, and health care providers, that implemented next-generation sequencing across the DoD health care system and surveillance network to gain insight into the molecular epidemiology of carbapenemase-producing bacteria to elicit more accurate and actionable data for infection control (see Figure 2-1). Through this program, outbreaks and emerging pathogens were detected earlier, adjustments were made in clinical standard operating procedures and patient care policies, and advances were made in diagnostic assay and software development.

[3] See https://aspe.hhs.gov/pdf-report/national-action-plan-combating-antibiotic-resistant-bacteria-progress-report-year-3 (accessed June 25, 2020).


ARMoR-D Database (Simplified view)
Repository Data
FREEZERS
SHELVES
ZONES
CONTAINERS
Isolate Data
PERSONS
(encrypted)
SOURCES
ISOLATES
Dictionary Tables
SUBMITTERS
TESTS
ORGANISMS
Genomic Analysis Pipeline
RELATED FILES
RESULTS
PCR RESULTS
SEQUENCING RESULTS
PFGE RESULTS
OPTICAL MAP RESULTS
FUTURE RESULT
FUTURE RESULT SETS
MRSN High Throughput Sequencing Data Analysis Pipeline
Illumina Sequence
Automated Process
Data Processing
Merge Overlapping Reads
FLASH
Read Quality Trimming
Btrim
De Novo Assembly
Newbler
Assembly Quality Control
Read Coverage
Contig N50
Contamination Detection
PacBio Sequence
Data Processing
SMRT Analysis Software Package
Semi-Automated Analysis
Outbreak Identification
SNP and InDel Identification
Gene Loss and Acquisition
Rearrangement / Transposition
Resistance Mechanism Detection
SNP and InDel Identification
Rearrangement / Transposition
Resistance Transfer Mechanisms
Transposon Structure and Function
Plasmid Transfer Mechanisms
Species ID
Automated Search of 16S Database
ABR Genes
Antibiotic Resistance Gene Annotation
WG Annotation
Whole Genome Annotation RAST/Prokka
WG Phylogeny
Align Reads to Reference Genome
Extract SNPs/Indels and Filter
Bayesian Inference of Phylogeny
Outbreak ID
Complete Genome/Plasmid Comparison
Identify SNPs, InDels, and rearrangements
MLST
Multilocus Sequence Typing Database
Plasmid Typing
Plasmid Replicon Typing Database
VIR Genes
Virulence Gene Annotation
Genome Sequencing Database

NOTES

The ARMoR-D database has table structures designed to store isolate-related data (demographic and results); and isolate location data (for the repository).

A number of dictionary tables are used to support user customization of values, and enhance data quality.

Personally identifying data (PERSONS above) is separated into its own table and encrypted.

Isolates from the same person are collected under a unique, de-identified SOURCES record.

Specialized RESULT SETS are linked to a generic RESULTs record, allowing simplified extension to additional specialized RESULT SETS.

FIGURE 2-1 Customized database for the Antimicrobial Resistance Monitoring and Research (ARMoR) Program.
SOURCE: Lesho et al., 2016.

### Outbreaks of Foodborne Pathogenic Bacteria

The value of genomic epidemiology is most evident for outbreaks of foodborne pathogenic bacteria. Genomic epidemiology is helpful for food safety regulators and is integrated into many surveillance systems such as the U.S. Food and Drug Administration's (FDA's) GenomeTrakr Network[4] (Ladner et al., 2019). This is demonstrated by one of many examples: a year-long investigation of 56 patients from 24 states who were identified by a national surveillance network as infected with shiga-toxin producing *E. coli* (Crowe et al., 2017). A lengthy diet questionnaire revealed an enrichment of patients who baked during the week before illness onset, eating or tasting homemade batter or dough. Across multiple states, patients provided the dry ingredients used in the recipe. It was found that the flour came from the same large domestic producer, *E. coli* was cultivated from this flour, and genome sequencing confirmed the match between patient and ingredient (Crowe et al., 2017). The ease of sharing and the richness and precision of genomic sequence data has catalyzed greater coordination across local, state, and national levels.

## BEST PRACTICES AND KEYS TO FUTURE SUCCESS

Efforts to date to apply genomic epidemiology to infectious disease outbreaks demonstrate the value of a multipronged, overlapping approach in which data from multiple sources—genomic, clinical, and epidemiological—are overlaid upon each other (Grubaugh et al., 2019; Houldcroft et al., 2017; Ladner et al., 2019). Collective use of this type of approach can enhance communication between different entities (e.g., public health laboratories, regional medical centers, academic hospitals, national organizations, and federal agencies) and provide a richer and more nuanced picture of the epidemiological landscape of an infectious disease outbreak, enabling the investigation and potential confirmation of hypotheses that would be difficult to pursue without the integration of genomic information into the approach. An instructive example is the surveillance framework developed to evaluate the circumstances of mosquito-borne transmission of yellow fever virus (Faria et al., 2018). By integrating rapid viral genomic surveillance with epidemiological and spatial data—from both humans and animals—it was shown that the outbreak originated from multiple sylvatic (forest-dwelling mosquito) transmission events from animals to humans, rather than urban human-to-human transmission by mosquitoes (Faria et al., 2018). Therefore, this type of framework can help to estimate the risk

[4] See https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network (accessed June 25, 2020).

of human viral exposure across space and time while also monitoring the likelihood of different routes of transmission.

Collaboration among stakeholders across multiple sectors is one of the keys for future success of these types of multipronged approaches. One established effort in this vein is FDA's GenomeTrakr, a distributed network of public, private, and academic laboratories that utilizes precision epidemiology, including whole genome sequencing for pathogen identification (Ladner et al., 2019). The network supports several real-time collaborative projects currently under way, including for analyzing samples of *Listeria monocytogenes* bacteria and *Salmonella enteritidis* bacteria (FDA, 2020). In total, GenomeTrakr has sequenced more than 462,000 isolates across a network spanning 15 federal laboratories and 25 state and university laboratories (FDA, 2020). As noted in the *E. coli* example described above, foodborne disease outbreaks can benefit from genomic analysis to allow the linking of cases to a point source of infection. The genomic data associated with the outbreaks, however, must be combined with relevant epidemiological data to identify risk that allow for specific interventions to be meaningful (Hill et al., 2017). As one 2016 review noted, despite the progress made around the use of whole genome sequencing in foodborne disease outbreaks and public health responses, there remains significant work to be done to ensure methodological consistency and standardization of sequencing approaches used by public health laboratories (Deng et al., 2016). In addition, even as collaboration across stakeholders in foodborne disease detection has advanced, the continued improvement of real-time, data-gathering efforts and incorporation of machine learning to inform predictive efforts represent the next frontier (Hill et al., 2017). By bringing together key stakeholders across the government and private sectors, the network is able to expedite real-time investigation of foodborne illness outbreaks and mitigate their effects on the public.

Expanding the global scope of genomic epidemiology as a practical method for timely and effective outbreak response will require building the technical capacities for genome sequencing and analysis in public agencies and private facilities. In particular, this will require collaborating with and supporting lower-resource settings that are disproportionately impacted by outbreaks of infectious diseases. Initiatives currently under way to support such efforts include the Association of Public Health Laboratories–CDC bioinformatics fellowship program,[5] the H3Africa initiative,[6] and public–private partnerships such as EcoHealth Alliance[7] and Metabiota[8] (Ladner et al., 2019).

[5] See https://www.aphl.org/fellowships/Pages/Bioinformatics.aspx (accessed June 25, 2020).

[6] See https://h3africa.org (accessed June 25, 2020).

[7] See https://www.ecohealthalliance.org/program/wab-net (accessed June 25, 2020).

[8] See https://metabiota.com (accessed June 25, 2020).

## REFERENCES

Andersen, K. G., B. J. Shapiro, C. B. Matranga, R. Sealfon, A. E. Lin, L. M. Moses, O. A. Folarin, A. Goba, I. Odia, P. E. Ehiane, M. Momoh, E. M. England, S. Winnicki, L. M. Branco, S. K. Gire, E. Phelan, R. Tariyal, R. Tewhey, O. Omoniwa, M. Fullah, R. Fonnie, M. Fonnie, L. Kanneh, S. Jalloh, M. Gbakie, S. Saffa, K. Karbo, A. D. Gladden, J. Qu, M. Stremlau, M. Nekoui, H. K. Finucane, S. Tabrizi, J. J. Vitti, B. Birren, M. Fitzgerald, C. McCowan, A. Ireland, A. M. Berlin, J. Bochicchio, B. Tazon-Vega, N. J. Lennon, E. M. Ryan, Z. Bjornson, D. A. Milner, Jr., A. K. Lukens, N. Broodie, M. Rowland, M. Heinrich, M. Akdag, J. S. Schieffelin, D. Levy, H. Akpan, D. G. Bausch, K. Rubins, J. B. McCormick, E. S. Lander, S. Günther, L. Hensley, S. Okogbenin, C. Viral Hemorrhagic Fever, S. F. Schaffner, P. O. Okokhere, S. H. Khan, D. S. Grant, G. O. Akpede, D. A. Asogun, A. Gnirke, J. Z. Levin, C. T. Happi, R. F. Garry, and P. C. Sabeti. 2015. Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell* 162(4):738–750.

CDC (U.S. Centers for Disease Control and Prevention). 2019. *Mumps*. https://www.cdc.gov/vaccines/pubs/pinkbook/mumps.html (accessed July 7, 2020).

CDC. 2020. *U.S. influenza surveillance system: Purpose and methods*. https://www.cdc.gov/flu/weekly/overview.htm (accessed June 24, 2020).

Chinese SARS Molecular Epidemiology Consortium. 2004. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* 303(5664):1666–1669.

Christie, A., G. J. Davies-Wayne, T. Cordier-Lassalle, D. J. Blackley, A. S. Laney, D. E. Williams, S. A. Shinde, M. Badio, T. Lo, S. E. Mate, J. T. Ladner, M. R. Wiley, J. R. Kugelman, G. Palacios, M. R. Holbrook, K. B. Janosko, E. de Wit, N. van Doremalen, V. J. Munster, J. Pettitt, R. J. Schoepp, L. Verhenne, I. Evlampidou, K. K. Kollie, S. B. Sieh, A. Gasasira, F. Bolay, F. N. Kateh, T. G. Nyenswah, and K. M. De Cock. 2015. Possible sexual transmission of Ebola virus—Liberia, 2015. *Morbidity and Mortality Weekly Report* 64(17):479–481.

Crowe, S. J., L. Bottichio, L. N. Shade, B. M. Whitney, N. Corral, B. Melius, K. D. Arends, D. Donovan, J. Stone, K. Allen, J. Rosner, J. Beal, L. Whitlock, A. Blackstock, J. Wetherington, L. A. Newberry, M. N. Schroeder, D. Wagner, E. Trees, S. Viazis, M. E. Wise, and K. P. Neil. 2017. Shiga toxin–producing *E. coli* infections associated with flour. *New England Journal of Medicine* 377(21):2036–2043.

Deng, X., H. C. d. Bakker, and R. S. Hendriksen. 2016. Genomic epidemiology: Whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annual Review of Food Science and Technology* 7(1):353–374.

Diallo, B., D. Sissoko, N. J. Loman, H. A. Bah, H. Bah, M. C. Worrell, L. S. Conde, R. Sacko, S. Mesfin, A. Loua, J. K. Kalonda, N. A. Erondu, B. A. Dahl, S. Handrick, I. Goodfellow, L. W. Meredith, M. Cotten, U. Jah, R. E. Guetiya Wadoum, P. Rollin, N. Magassouba, D. Malvy, X. Anglaret, M. W. Carroll, R. B. Aylward, M. H. Djingarey, A. Diarra, P. Formenty, S. Keïta, S. Günther, A. Rambaut, and S. Duraffour. 2016. Resurgence of Ebola virus disease in Guinea linked to a survivor with virus persistence in seminal fluid for more than 500 days. *Clinical Infectious Diseases* 63(10):1353–1356.

Diehl, W. E., A. E. Lin, N. D. Grubaugh, L. M. Carvalho, K. Kim, P. P. Kyawe, S. M. McCauley, E. Donnard, A. Kucukural, P. McDonel, S. F. Schaffner, M. Garber, A. Rambaut, K. G. Andersen, P. C. Sabeti, and J. Luban. 2016. Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell* 167(4):1088–1098.

Dietzel, E., G. Schudt, V. Krähling, M. Matrosovich, and S. Becker. 2017. Functional characterization of adaptive mutations during the West African Ebola virus outbreak. *Journal of Virology* 91(2).

Dudas, G., L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey, and A. Rambaut. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544(7650):309–315.

Faria, N. R., R. Azevedo, M. U. G. Kraemer, R. Souza, M. S. Cunha, S. C. Hill, J. Thézé, M. B. Bonsall, T. A. Bowden, I. Rissanen, I. M. Rocco, J. S. Nogueira, A. Y. Maeda, F. Vasami, F. L. L. Macedo, A. Suzuki, S. G. Rodrigues, A. C. R. Cruz, B. T. Nunes, D. B. A. Medeiros, D. S. G. Rodrigues, A. L. N. Queiroz, E. V. P. da Silva, D. F. Henriques, E. S. T. da Rosa, C. S. de Oliveira, L. C. Martins, H. B. Vasconcelos, L. M. N. Casseb, D. B. Simith, J. P. Messina, L. Abade, J. Lourenço, L. C. J. Alcantara, M. M. de Lima, M. Giovanetti, S. I. Hay, R. S. de Oliveira, P. D. S. Lemos, L. F. de Oliveira, C. P. S. de Lima, S. P. da Silva, J. M. de Vasconcelos, L. Franco, J. F. Cardoso, J. Vianez-Júnior, D. Mir, G. Bello, E. Delatorre, K. Khan, M. Creatore, G. E. Coelho, W. K. de Oliveira, R. Tesh, O. G. Pybus, M. R. T. Nunes, and P. F. C. Vasconcelos. 2016. Zika virus in the Americas: Early epidemiological and genetic findings. *Science* 352(6283):345–349.

Faria, N. R., J. Quick, I. M. Claro, J. Thézé, J. G. de Jesus, M. Giovanetti, M. U. G. Kraemer, S. C. Hill, A. Black, A. C. da Costa, L. C. Franco, S. P. Silva, C. H. Wu, J. Raghwani, S. Cauchemez, L. du Plessis, M. P. Verotti, W. K. de Oliveira, E. H. Carmo, G. E. Coelho, A. C. F. S. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L. S. Nascimento, H. A. O. Monteiro, I. C. Siqueira, M. G. de Queiroz, T. R. de Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, Jr., T. Jaenisch, G. L. Wallau, M. C. de Lima, V. Nascimento, E. M. de Cerqueira, M. M. de Lima, D. L. Mascarenhas, J. P. M. Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R. T. Nunes, E. C. Sabino, L. C. J. Alcantara, N. J. Loman, and O. G. Pybus. 2017. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546(7658):406–410.

Faria, N. R., M. U. G. Kraemer, S. C. Hill, J. Goes de Jesus, R. S. Aguiar, F. C. M. Iani, J. Xavier, J. Quick, L. du Plessis, S. Dellicour, J. Thézé, R. D. O. Carvalho, G. Baele, C. H. Wu, P. P. Silveira, M. B. Arruda, M. A. Pereira, G. C. Pereira, J. Lourenço, U. Obolski, L. Abade, T. I. Vasylyeva, M. Giovanetti, D. Yi, D. J. Weiss, G. R. W. Wint, F. M. Shearer, S. Funk, B. Nikolay, V. Fonseca, T. E. R. Adelino, M. A. A. Oliveira, M. V. F. Silva, L. Sacchetto, P. O. Figueiredo, I. M. Rezende, E. M. Mello, R. F. C. Said, D. A. Santos, M. L. Ferraz, M. G. Brito, L. F. Santana, M. T. Menezes, R. M. Brindeiro, A. Tanuri, F. C. P. Dos Santos, M. S. Cunha, J. S. Nogueira, I. M. Rocco, A. C. da Costa, S. C. V. Komninakis, V. Azevedo, A. O. Chieppe, E. S. M. Araujo, M. C. L. Mendonça, C. C. Dos Santos, C. D. Dos Santos, A. M. Mares-Guia, R. M. R. Nogueira, P. C. Sequeira,

R. G. Abreu, M. H. O. Garcia, A. L. Abreu, O. Okumoto, E. G. Kroon, C. F. C. de Albuquerque, K. Lewandowski, S. T. Pullan, M. Carroll, T. de Oliveira, E. C. Sabino, R. P. Souza, M. A. Suchard, P. Lemey, G. S. Trindade, B. P. Drumond, A. M. B. Filippis, N. J. Loman, S. Cauchemez, L. C. J. Alcantara, and O. G. Pybus. 2018. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* 361(6405):894–899.

FDA (U.S. Food and Drug Administration). 2020. *GenomeTrakr network*. https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network (accessed June 24, 2020).

Gardy, J. L., and N. J. Loman. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* 19(1):9–20.

GISAID (Global Initiative on Sharing All Influenza Data). 2020. *GISAID mission*. https://www.gisaid.org/about-us/mission (accessed June 24, 2020).

Grubaugh, N. D., J. T. Ladner, M. U. G. Kraemer, G. Dudas, A. L. Tan, K. Gangavarapu, M. R. Wiley, S. White, J. Thézé, D. M. Magnani, K. Prieto, D. Reyes, A. M. Bingham, L. M. Paul, R. Robles-Sikisaka, G. Oliveira, D. Pronty, C. M. Barcellona, H. C. Metsky, M. L. Baniecki, K. G. Barnes, B. Chak, C. A. Freije, A. Gladden-Young, A. Gnirke, C. Luo, B. MacInnis, C. B. Matranga, D. J. Park, J. Qu, S. F. Schaffner, C. Tomkins-Tinch, K. L. West, S. M. Winnicki, S. Wohl, N. L. Yozwiak, J. Quick, J. R. Fauver, K. Khan, S. E. Brent, R. C. Reiner, P. N. Lichtenberger, M. J. Ricciardi, V. K. Bailey, D. I. Watkins, M. R. Cone, E. W. Kopp, K. N. Hogan, A. C. Cannons, R. Jean, A. J. Monaghan, R. F. Garry, N. J. Loman, N. R. Faria, M. C. Porcelli, C. Vasquez, E. R. Nagle, D. A. T. Cummings, D. Stanek, A. Rambaut, M. Sanchez-Lockhart, P. C. Sabeti, L. D. Gillis, S. F. Michael, T. Bedford, O. G. Pybus, S. Isern, G. Palacios, and K. G. Andersen. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 546(7658):401–405.

Grubaugh, N. D., N. R. Faria, K. G. Andersen, and O. G. Pybus. 2018. Genomic insights into Zika virus emergence and spread. *Cell* 172(6):1160–1162.

Grubaugh, N. D., J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, and K. G. Andersen. 2019. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology* 4(1):10–19.

Guan, Y., B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, P. H. Li, L. J. Zhang, Y. J. Guan, K. M. Butt, K. L. Wong, K. W. Chan, W. Lim, K. F. Shortridge, K. Y. Yuen, J. S. Peiris, and L. L. Poon. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 302(5643):276–278.

Hill, A. A., M. Crotta, B. Wall, L. Good, S. J. O'Brien, and J. Guitian. 2017. Towards an integrated food safety surveillance system: A simulation study to explore the potential of combining genomic and epidemiological metadata. *Royal Society Open Science* 4(3):160721.

Holmes, E. C., G. Dudas, A. Rambaut, and K. G. Andersen. 2016. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nature* 538(7624):193–200.

Houldcroft, C. J., M. A. Beale, and J. Breuer. 2017. Clinical and biological insights from viral genome sequencing. *Nature Reviews Microbiology* 15(3):183–192.

Kan, B., M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, W. Liang, H. Zheng, K. Wan, Q. Liu, B. Cui, Y. Xu, E. Zhang, H. Wang, J. Ye, G. Li, M. Li, Z. Cui, X. Qi, K. Chen, L. Du, K. Gao, Y.-T. Zhao, X.-Z. Zou, Y.-J. Feng, Y.-F. Gao, R. Hai, D. Yu, Y. Guan, and J. Xu. 2005. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *Journal of Virology* 79(18):11892–11900.

Ladner, J. T., N. D. Grubaugh, O. G. Pybus, and K. G. Andersen. 2019. Precision epidemiology for infectious disease control. *Nature Medicine* 25(2):206–211.

Lesho, E. P., P. E. Waterman, U. Chukwuma, K. McAuliffe, C. Neumann, M. D. Julius, H. Crouch, R. Chandrasekera, J. F. English, R. J. Clifford, and K. E. Kester. 2014. The Antimicrobial Resistance Monitoring and Research (ARMoR) Program: The U.S. Department of Defense response to escalating antimicrobial resistance. *Clinical Infectious Diseases* 59(3):390–397.

Lesho, E., R. Clifford, F. Onmus-Leone, L. Appalla, E. Snesrud, Y. Kwak, A. Ong, R. Maybank, P. Waterman, P. Rohrbeck, M. Julius, A. Roth, J. Martinez, L. Nielsen, E. Steele, P. McGann, and M. Hinkle. 2016. The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. Department of Defense. *PLOS ONE* 11(5):e0155770.

Li, F., W. Li, M. Farzan, and S. C. Harrison. 2005. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* 309(5742):1864–1868.

Lu, H., Y. Zhao, J. Zhang, Y. Wang, W. Li, X. Zhu, S. Sun, J. Xu, L. Ling, L. Cai, D. Bu, and R. Chen. 2004. Date of origin of the SARS coronavirus strains. *BMC Infectious Diseases* 4:3.

Mate, S. E., J. R. Kugelman, T. G. Nyenswah, J. T. Ladner, M. R. Wiley, T. Cordier-Lassalle, A. Christie, G. P. Schroth, S. M. Gross, G. J. Davies-Wayne, S. A. Shinde, R. Murugan, S. B. Sieh, M. Badio, L. Fakoli, F. Taweh, E. de Wit, N. van Doremalen, V. J. Munster, J. Pettitt, K. Prieto, B. W. Humrighouse, U. Ströher, J. W. DiClaro, L. E. Hensley, R. J. Schoepp, D. Safronetz, J. Fair, J. H. Kuhn, D. J. Blackley, A. S. Laney, D. E. Williams, T. Lo, A. Gasasira, S. T. Nichol, P. Formenty, F. N. Kateh, K. M. De Cock, F. Bolay, M. Sanchez-Lockhart, and G. Palacios. 2015. Molecular evidence of sexual transmission of Ebola virus. *New England Journal of Medicine* 373(25):2448–2454.

Metsky, H. C., C. B. Matranga, S. Wohl, S. F. Schaffner, C. A. Freije, S. M. Winnicki, K. West, J. Qu, M. L. Baniecki, A. Gladden-Young, A. E. Lin, C. H. Tomkins-Tinch, S. H. Ye, D. J. Park, C. Y. Luo, K. G. Barnes, R. R. Shah, B. Chak, G. Barbosa-Lima, E. Delatorre, Y. R. Vieira, L. M. Paul, A. L. Tan, C. M. Barcellona, M. C. Porcelli, C. Vasquez, A. C. Cannons, M. R. Cone, K. N. Hogan, E. W. Kopp, J. J. Anzinger, K. F. Garcia, L. A. Parham, R. M. G. Ramírez, M. C. M. Montoya, D. P. Rojas, C. M. Brown, S. Hennigan, B. Sabina, S. Scotland, K. Gangavarapu, N. D. Grubaugh, G. Oliveira, R. Robles-Sikisaka, A. Rambaut, L. Gehrke, S. Smole, M. E. Halloran, L. Villar, S. Mattar, I. Lorenzana, J. Cerbino-Neto, C. Valim, W. Degrave, P. T. Bozza, A. Gnirke, K. G. Andersen, S. Isern, S. F. Michael, F. A. Bozza, T. M. L. Souza, I. Bosch, N. L. Yozwiak, B. L. MacInnis, and P. C. Sabeti. 2017. Zika virus evolution and spread in the Americas. *Nature* 546(7658):411–415.

Pacciarini, F., S. Ghezzi, F. Canducci, A. Sims, M. Sampaolo, E. Ferioli, M. Clementi, G. Poli, P. G. Conaldi, R. Baric, and E. Vicenzi. 2008. Persistent replication of severe acute respiratory syndrome coronavirus in human tubular kidney cells selects for adaptive mutations in the membrane protein. *Journal of Virology* 82(11):5137–5144.

Ruan, Y. J., C. L. Wei, A. L. Ee, V. B. Vega, H. Thoreau, S. T. Y. Su, J.-M. Chia, P. Ng, K. P. Chiu, L. Lim, T. Zhang, C. K. Peng, E. O. L. Lin, N. M. Lee, S. L. Yee, L. F. P. Ng, R. E. Chee, L. W. Stanton, P. M. Long, and E. T. Liu. 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet (London, England)* 361(9371):1779–1785.

Siddle, K. J., P. Eromon, K. G. Barnes, S. Mehta, J. U. Oguzie, I. Odia, S. F. Schaffner, S. M. Winnicki, R. R. Shah, J. Qu, S. Wohl, P. Brehio, C. Iruolagbe, J. Aiyepada, E. Uyigue, P. Akhilomen, G. Okonofua, S. Ye, T. Kayode, F. Ajogbasile, J. Uwanibe, A. Gaye, M. Momoh, B. Chak, D. Kotliar, A. Carter, A. Gladden-Young, C. A. Freije, O. Omoregie, B. Osiemi, E. B. Muoebonam, M. Airende, R. Enigbe, B. Ebo, I. Nosamiefan, P. Oluniyi, M. Nekoui, E. Ogbaini-Emovon, R. F. Garry, K. G. Andersen, D. J. Park, N. L. Yozwiak, G. Akpede, C. Ihekweazu, O. Tomori, S. Okogbenin, O. A. Folarin, P. O. Okokhere,

B. L. MacInnis, P. C. Sabeti, and C. T. Happi. 2018. Genomic analysis of Lassa virus during an increase in cases in Nigeria in 2018. *New England Journal of Medicine* 379(18):1745–1753.

Snitkin, E. S., A. M. Zelazny, P. J. Thomas, F. Stock, D. K. Henderson, T. N. Palmore, and J. A. Segre. 2012. Tracking a hospital outbreak of carbapenem-resistant *klebsiella pneumoniae* with whole-genome sequencing. *Science Translational Medicine* 4(148):148ra116.

Snitkin, E. S., S. Won, A. Pirani, Z. Lapp, R. A. Weinstein, K. Lolans, and M. K. Hayden. 2017. Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant *klebsiella pneumoniae* in a regional outbreak. *Science Translational Medicine* 9(417).

Urbanowicz, R. A., C. P. McClure, A. Sakuntabhai, A. A. Sall, G. Kobinger, M. A. Müller, E. C. Holmes, F. A. Rey, E. Simon-Loriere, and J. K. Ball. 2016. Human adaptation of Ebola virus during the West African outbreak. *Cell* 167(4):1079–1087.

Whitmer, S. L. M., J. T. Ladner, M. R. Wiley, K. Patel, G. Dudas, A. Rambaut, F. Sahr, K. Prieto, S. S. Shepard, E. Carmody, B. Knust, D. Naidoo, G. Deen, P. Formenty, S. T. Nichol, G. Palacios, and U. Ströher. 2018. Active Ebola virus replication and heterogeneous evolutionary rates in EVD survivors. *Cell Reports* 22(5):1159–1168.

Wohl, S., H. C. Metsky, S. F. Schaffner, A. Piantadosi, M. Burns, J. A. Lewnard, B. Chak, L. A. Krasilnikova, K. J. Siddle, C. B. Matranga, B. Bankamp, S. Hennigan, B. Sabina, E. H. Byrne, R. J. McNall, R. R. Shah, J. Qu, D. J. Park, S. Gharib, S. Fitzgerald, P. Barreira, S. Fleming, S. Lett, P. A. Rota, L. C. Madoff, N. L. Yozwiak, B. L. MacInnis, S. Smole, Y. H. Grad, and P. C. Sabeti. 2020. Combining genomics and epidemiology to track mumps virus transmission in the United States. *PLOS Biology* 18(2):e3000611.

# 3

# Current Genomic Epidemiology Efforts Related to SARS-CoV-2

A collection of loosely affiliated and competing networks are acquiring and sharing genomic, clinical, and epidemiological data related to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Data are being generated at an escalating pace around the world, while networks for rapidly sharing those data are being established at the regional, national, and global levels. This chapter provides an overview of select SARS-CoV-2 data sources, identifies limitations of those sources, and highlights breakthrough efforts to combine and analyze genomic sequence data with clinical and epidemiological data for SARS-CoV-2. The proceeding chapter sets out the key considerations for such a framework to bring these data sources together.

## CURRENT SARS-CoV-2 DATA SOURCES

### U.S. Centers for Disease Control and Prevention SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance

The U.S. Centers for Disease Control and Prevention's (CDC's) SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance[1] (SPHERES) consortium was initiated in May 2020 to improve public health by coordinating a large-scale nationwide genomic sequencing effort across the United States. SPHERES aims to accelerate the generation and sharing of high-quality viral sequencing data from clinical and

[1] See https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html (accessed June 25, 2020).

public health laboratories as well as to set standards to streamline the collection of consistent metadata for integrated analyses. All 50 states now have next-generation sequencing capacity, but the scale up of genomic research efforts has been hampered by lack of workforce capacity and limited coordination of stakeholders across academia, nonprofit, private, and public entities. Using a crowdsourced model, SPHERES hopes to gather sequencing data in a high-quality, representative, and consistent manner that can be used to establish a national baseline needed to monitor trends and inform evidence-based public health responses to the public health emergency; current efforts are built with the architecture designed for influenza surveillance. SPHERES aims specifically to (1) maintain high-level monitoring, (2) use sequencing data to set national and regional baselines, (3) conduct sustainable, longitudinal data collection, (4) use sequencing data to support contact tracing efforts, (5) monitor viral genetic diversity over time, and (6) foster new collaborations and innovation through the public–private partnership model.

SPHERES currently has a patchwork funding environment, with sources including federal funding from CDC and the National Institutes of Health (NIH), academic laboratories, philanthropy, and other private partners, raising concerns of sustainability. In the context of uneven testing and sequencing practices across the country, SPHERES consortium sampling has been patchy and largely passive, with 21 states having submitted no sequencing data as of early June 2020. Consequently, the sampling is nonrepresentative in terms of geography and likely also viral genomic diversity. A further limitation is that SPHERES is not focused on collection and linkage of clinical data, in part because many laboratories do not provide detailed patient-level clinical metadata beyond a minimum set of identifiers.

### National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI)[2] at NIH serves as a primary repository for all genomic sequencing. In accordance with NIH Data Sharing policies, genomic scientists rapidly deposit and release assembled SARS-CoV-2 genomes and raw metagenomic reads in NCBI's GenBank and SRA databases, linked under BioProjects. NCBI provides a resource that includes links to SARS-CoV-2 reference sequences and several methods to explore the data. While different experimental methods have been established to obtain viral genomes, standards for completeness are in place. Standardization could be bolstered with SARS-CoV-2 samples distributed by the National Institute of Standards and Technology for sequencing and submission, thus benchmarking across centers. NCBI also runs a Pathogen Detection program that integrates bacterial pathogen

[2] See https://www.ncbi.nlm.nih.gov/sars-cov-2 (accessed June 25, 2020).

genomic sequences from patients, food, and environmental sources to track foodborne disease outbreaks, assess virulence, and detect antimicrobial resistance. This program might help to inform a similar integrated model for detecting and tracking viruses using genomic sequencing.

### Global Initiative on Sharing All Influenza Data

As described in Chapter 2, the Global Initiative on Sharing All Influenza Data[3] (GISAID) is a major global platform for rapidly and openly sharing data on all influenza viruses, and has been adapted to include SARS-CoV-2. GISAID collects geographical and species-specific genomic, clinical, and epidemiological data to contribute to understanding how viruses evolve and spread during outbreaks. The initial SARS-CoV-2 genome sequences from China were posted to GISAID; nearly every major SARS-CoV-2 study utilizes GISAID.

### Nextstrain

Nextstrain[4] is an open-source platform designed to harness the potential of genomic data from a variety of infectious disease pathogens to support epidemiological research and outbreak response. It hosts a GISAID-enabled interface of publicly available sequence data from every continent[5] and provides powerful analytic and visualization tools that can be used to explore data at various scales (e.g., global, continent, country, region). Nextstrain's work on SARS-CoV-2 has largely focused on identifying major clades—defined as clades that have reached 20 percent global frequency—as well as emerging clades. The Nextstrain tool can address questions about which regions of the genome are most variable, estimate the rate of infection, and apply some use of metadata (e.g., coloring sequences on the phylogenetic tree based on age of the person infected).

### National COVID Cohort Collaborative

The National COVID Cohort Collaborative (N3C)[6] is a centralized and secure portal for patient-level COVID-19 clinical data and a platform

[3] See https://www.gisaid.org (accessed June 25, 2020).

[4] See https://nextstrain.org (accessed June 25, 2020).

[5] See https://nextstrain.org/ncov/global (accessed June 25, 2020).

[6] The National COVID Cohort Collaborative is a collaboration among National Center for Advancing Translational Sciences–supported Clinical and Translational Science Awards Program hubs, the National Center for Data to Health, and distributed clinical data networks, with overall stewardship by NIH's National Center for Advancing Translational Sciences. More information is available at https://covid.cd2h.org/N3C (accessed June 25, 2020).

for deploying and evaluating methods and tools for clinicians, researchers, and health care professionals. N3C is a collaboration under the auspices of NIH's National Center for Advancing Translational Sciences[7] in response to COVID-19, joining resources of the National Center for Data to Health[8] and the Clinical and Translational Science Awards.[9] It also serves as a resource for clinicians and researchers with granular and complex clinical questions. N3C was developed to improve efficiency and accessibility of analyses with COVID-19 clinical data, to expand the ability to analyze coronavirus diseases, and to demonstrate novel approaches for collaborative data sharing. It is governed by a single, central Institutional Review Board (IRB) and has five workstreams: data partnership and governance; phenotype and data acquisition; data ingestion and harmonization; collaborative analytics; and synthetic data. The collaborative is currently building a dataset of patients with COVID-19 that can be securely linked to external datasets using coded identifiers in an enclave model that enables linking without the sharing of overtly identifiable health information.

## Examples of Regional Initiatives

### *Broad Institute*

The Broad Institute[10] in Boston, Massachusetts, has been a lead innovator in genomic research for 15 years through its Genomics Platform,[11] which aims to create foundational genomics resources and to facilitate large-scale, pioneering projects to understand the genomic basis of diseases. Currently, the viral genomics group at the Broad Institute is collaborating with Massachusetts General Hospital and the Massachusetts Department of Public Health to investigate the introduction and spread of COVID-19 in the Boston area. Their work suggests that more than 30 introduction events from both domestic and international sources occurred in the area (see Figure 3-1), with subsequent super-spreader events that involved rapid and initially asymptomatic transmission in a congregate living facility as well as widespread transmission at a pharmaceutical conference that underpinned the outbreak in the state (Virological, 2020).

Since 2016, the Broad Institute has partnered with the Massachusetts State Public Health Laboratory and CDC to build distributed capacity for genomic sequencing through a train-the-trainer program for regional- and

[7] See https://ncats.nih.gov (accessed June 25, 2020).

[8] See https://cd2h.org (accessed June 25, 2020).

[9] See https://ncats.nih.gov/ctsa (accessed June 25, 2020).

[10] See https://www.broadinstitute.org/coronavirus/epidemiology-surveillance#top (accessed June 25, 2020).

[11] See https://www.broadinstitute.org/genomics (accessed June 25, 2020).

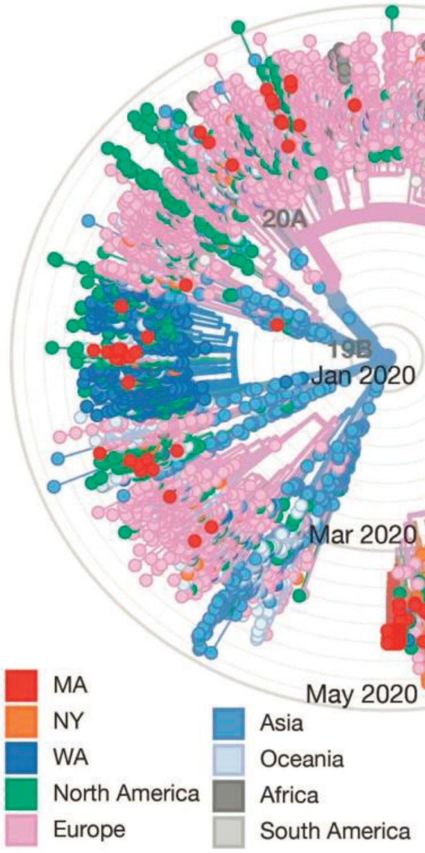


**FIGURE 3-1** A radial phylogenetic tree of a global set of 10,000 SARS-CoV-2 genomes available on GISAID, with distance from the center and color reflecting when and where the sample was collected, respectively.
NOTES: 331 complete SARS-CoV-2 genomes from Massachusetts (red) were sequenced at the Broad Institute and are spread throughout the tree. At least 30 putative introductions to the Boston area are posited based on the phylogenetic distribution of the earliest (most central) samples. The earliest documented case in Massachusetts was a subject who returned from Wuhan on January 29, 2020, with symptoms (single red dot located in Western inner circle of tree) with no detected transmissions (no descendent viruses observed). By contrast, clusters of SARS-CoV-2 genomic sequences seen in Southern and Eastern quadrants in February/March, together with epidemiologic data, support multiple local transmissions derived from single sources (super-spreading event).
SOURCES: Virological, 2020 (graphic), through adaptation from Nextstrain and data from GISAID.

state-level laboratory personnel. These programs could serve as a model for developing national coordination among state public health laboratories.

### *Chan Zuckerberg Biohub*

The Chan Zuckerberg Biohub[12] has three ongoing COVID-19 efforts being carried out in partnerships with the University of California, San Francisco; the California Department of Public Health; and local- and county-level departments of public health (DPHs). CLIAHUB is a Clinical Laboratory Improvement Amendments–certified laboratory that provides COVID-19 testing to counties and clinics across California. Samples from positive tests are routed directly for full genomic sequencing. COVID-Tracker is a minimum viable product built on Nextstrain by the Biohub for the visualization of data and importation of epidemiological metadata for use by county DPH officials. Last, COVIDNet is a collaboration with the state and county DPHs, academic partners, and commercial laboratories through which partners send positive samples to the Biohub, commercial laboratories, and other academic laboratories in California for the viral genomes to be sequenced. As of July 7, 2020, these efforts have yielded more than 650 genomes from 10 California counties deposited in GISAID, with an estimated 40,000 genomes expected to be completed over the next year.

### *Wadsworth Center*

The Wadsworth Center,[13] the public health laboratory for the New York State Department of Health, was the first laboratory in New York to receive Emergency Use Authorization for SARS-CoV-2 testing. While rapidly expanding its own capacity, Wadsworth supported other commercial and hospital laboratories in the state to develop and ramp up testing capacities for SARS-CoV-2. Building on its investment in statewide surveillance of antibiotic-resistant bacteria, food- and water-borne pathogens, and influenza virus, Wadsworth deployed a genomic epidemiological model to track the SARS-CoV-2 caseload in New York State. Thus far, more than 300 genomes from different time periods during the pandemic and covering many regions of the state—including New York City (NYC), the initial epicenter—have been identified. The Wadsworth Center is working closely with the NYC public health laboratory and regional academic and hospital laboratories to collect and integrate clinical metadata. The Governor of New York Andrew Cuomo utilized this genomic epidemiological data, along with daily analysis of the regional

[12] See https://www.czbiohub.org (accessed June 25, 2020).
[13] See https://www.wadsworth.org (accessed June 25, 2020).

test positivity rate, as part of the approach to manage New York State's response to SARS-CoV-2.

## Other Initiatives

### *COVID-19 Genomics UK Consortium*

The COVID-19 Genomics UK (COG-UK) Consortium[14] was established with £20 million ($25 million) of funding to increase the capacity to collect, sequence, and analyze whole genomes of SARS-CoV-2 virus samples collected in the United Kingdom with an explicit commitment to open science and data sharing. COG-UK was created to deliver clinically actionable data to local medical centers and the UK government to guide health interventions and policies. Collaborating partners include public health agencies, university laboratories, regional university hubs and health organizations, and sequencing centers, including the Wellcome Sanger Institute. More than 20,000 sequences have been published thus far on the COG-UK website, which includes links to all samples and metadata as well as protocols developed by consortium members for preparing, conducting, and analyzing sequences. COG-UK also regularly submits data to GISAID and can be viewed in Nextstrain. Among several workstreams, the project includes a working group on metadata, patient linkage, epidemiology, and health informatics. The UK Academy of Medical Sciences also launched an open-access database[15] to map and share UK preclinical research on COVID-19 therapies, as well as informing strategic decision making by policy makers and funders.

### *Global Alliance for Genomics and Health*

The Global Alliance for Genomics and Health[16] is a policy-framing and technical standards-setting organization seeking to enable responsible genomic data sharing within a human rights framework. Its actions to facilitate rapid sharing of high-quality data by the human genetics community can support timely and effective responses during global disease outbreaks. Composed of more than 500 organization members, including COG-UK, other members of the alliance include

- Canadian COVID Genomics Network,[17] a national collaboration to coordinate data sharing and analyses across the country and to

[14] See https://www.cogconsortium.uk (accessed June 25, 2020).

[15] See https://covidpipeline.acmedsci.ac.uk (accessed June 25, 2020).

[16] See https://www.ga4gh.org/covid-19 (accessed June 25, 2020).

[17] See https://www.genomecanada.ca/en/news/genome-canada-leads-40-million-genomics-initiative-address-covid-19-pandemic (accessed June 25, 2020).

accelerate genome sequencing to inform clinical and public health approaches;

- COVID-19 Beacon,[18] a tool to integrate sequencing data, identify specific genetic mutations, and create visualizations of the geographic and evolutionary origins of pathogens;
- COVID-19 Portal,[19] a European portal dedicated to the rapid collection and sharing of genomic data from the global research community;
- Galaxy COVID-19,[20] a resource that compiles best practices, infrastructure, and workflows to support genomic analyses of SARS-CoV-2 data; and
- Public Health Alliance for Genomic Epidemiology,[21] a global coalition working to develop consensus standards, share best practices, and advocate for open science and data sharing in public health microbial bioinformatics.

### Partnerships Among the Private Sector, Academia, and Public Health Agencies

New partnerships among the private sector, academia, and public health agencies are also collaborating to support data collection and genomic analysis of SARS-CoV-2. In Washington State, Microsoft's data science team is working with the Washington State Department of Health to build more efficient systems for collecting and managing large volumes of data about disease incidence and hospitalization (Edmond, 2020). This partnership was catalyzed when the Washington State Electronic Laboratory Reporting System was overwhelmed by the influx of negative SARS-CoV-2 test results reported early in the pandemic. In California, Amazon Web Services is partnering with the University of California, San Francisco, to perform genomic sequencing on samples from people with COVID-19 in the Bay area (Kent, 2020).

[18] See https://covid-19.dnastack.com/app/workspace/eyJrIjoiY292aWQtcHVibGljIiwiciI6ImFwcCIsImMiOiIwIn0/resource/sequences--eyJrIjoiY292aWQtcHVibGljIiwiciI6ImFwcCIsImMiOiIwIiwibiI6IjEifQ?filter=eyJmaWx0ZXJzIjp7fSwib3JkZXIiOlt7ImZpZWxkIjoiYWNjZXNzaW9uIiwiZGlyZWN0aW9uIjoiQVNDIn1dfQ%3D%3D (accessed June 25, 2020).

[19] See https://www.covid19dataportal.org (accessed June 25, 2020).

[20] See https://covid19.galaxyproject.org (accessed June 25, 2020).

[21] See https://pha4ge.github.io (accessed June 25, 2020).

## CURRENT EFFORTS TO INTEGRATE SARS-CoV-2 GENOME SEQUENCE DATA WITH CLINICAL AND EPIDEMIOLOGICAL DATA

The committee identified several breakthrough efforts to combine and analyze SARS-CoV-2 genome sequence data with clinical and epidemiological data that were conducted in Wuhan (China), NYC (United States), and Iceland. These efforts demonstrate the value of integrating data to support transmission tracking in real time as an outbreak unfolds.

In late 2019, a critically ill 41-year-old male presented at the Central Hospital of Wuhan, Hubei province, central China, with lung infiltrates and tested negative for all common respiratory pathogens. To identify the etiologic agent, direct metatranscriptomics were performed on bronchoalveolar lavage fluid with de novo assembly, which yielded a 30-kilobase genome with 89 percent sequence identity to a group of SARS-like coronaviruses previously identified in bats (Wu et al., 2020). Concurrently, critically ill patients with unidentified severe pneumonia disease were admitted to the intensive care unit of another hospital in Wuhan; it was noted that many worked at the same local seafood market. Because the outbreak was occurring with similar epidemiological features to SARS infections (e.g., occurring in winter, patients having links to a food market), researchers first tested and identified a positive signal for a coronavirus. Next, they used direct metagenomic sequencing of bronchoalveolar lavage fluid to identify non-human microbial DNA. With some additional targeted sequencing and analysis, they independently identified the complete SARS-CoV-2 genome (Zhou et al., 2020). Foreshadowing zoonotic investigations, both scientific research teams noted the genomic similarity of SARS-CoV-2 to the 2003 SARS-CoV and to published bat coronaviruses and epidemiological links of early cases to the indoor market selling seafood and other live wild animals. Importantly, both groups submitted SARS-CoV-2 genomes to NCBI/National Library of Medicine GenBank and GISAID, which immediately became the reference for an international effort to screen for new cases (Wu et al., 2020; Zhou et al., 2020).

In early 2020, increasing numbers of patients with clinical conditions consistent with a diagnosis of SARS-CoV-2 presented globally, including in the United States; as screening capacity increased, so too did the number of reported cases. NYC became an epicenter of SARS-CoV-2 infections—with 172,000 cases and 13,000 deaths reported in March and April 2020—prompting the Icahn School of Medicine at Mount Sinai in NYC to activate its extant Pathogen Surveillance Program,[22] which had been established the previous year to generate real-time genetic information on pathogens

[22] See https://icahn.mssm.edu/research/genomics/research/pathogen-surveillance (accessed June 25, 2020).

found to cause disease in its patients (Gonzalez-Reiche et al., 2020). With IRB approval, patient consent, and biospecimen handling already in place, investigators acquired clinical samples from 84 patients who tested positive for SARS-CoV-2 in the first weeks of March 2020. Its genome scientists accessed the reference genome and rapidly acquired full genome sequences of these clinical cases. Phylogenetic analyses were performed using the 2,363 other SARS-CoV-2 genomic sequences deposited in GISAID in March 2020. The NYC isolates were distributed throughout the phylogenetic tree, suggesting multiple independent introductions (Gonzalez-Reiche et al., 2020). By standardizing to Nextstrain's clade nomenclature, 87 percent of these NYC isolates were assigned to a clade that was the dominant clade in Europe at the time and suggested that travel from Europe accounted for the majority of those cases. Machine learning and Bayesian phylodynamic analyses generated an estimated period of untracked global transmission from late January to mid-February 2020. A few of the genomes closely matched strains from Washington State, which supported independent domestic introductions. These 84 patients were NYC residents from 21 neighborhoods across 4 boroughs in NYC and 2 towns in Westchester County. Based on zip code information, two monophyletic clusters—of 17 and 4 cases, respectively—were distributed across the NYC region, which suggested extensive, local, undetected transmission (Gonzalez-Reiche et al., 2020). As described previously in this chapter, many other academic centers, including the Broad Institute in Cambridge, Massachusetts, and the Chan Zuckerburg Biohub in San Francisco, California, established similar genomic tracking programs to identify the initial seeding and the subsequent spread of SARS-CoV-2 in their communities.

Many countries mounted a national response to control the spread of SARS-CoV-2, but Iceland's response was notable in that it immediately and actively leveraged genomic–epidemiological technology to develop innovative solutions. Although Iceland is geographically isolated and has a relatively small population of 360,000, the country welcomes 2 million tourists per year (Stofa, 2019). Importantly, deCODE Genetics-Amgen[23] of Iceland has been performing human population-level genomic sequencing to discover genetic risk factors for disease for 25 years. Iceland was able to conduct targeted testing on 9,199 persons at high risk for infection based on symptoms or recent travel history as well as population-based screening of 13,000 residents (Gudbjartsson et al., 2020). In total, 6 percent of the population was screened, with 13.3 percent of the targeted patients testing positive and 0.6–0.8 percent of the random-population screening testing positive. GenBank and GISAID received 581 genomes that were sequenced from these clinical samples. By leveraging the sequences in

[23] See https://www.decode.com (accessed June 25, 2020).

the GISAID repository, these genomes were assigned to 42 distinct clades, which provides a lower bound on the number of introductions to Iceland (Gudbjartsson et al., 2020). Genomic sequences revealed early virus importations followed by community and family spread of distinct viruses indicating other sources. Genomic sequencing also revealed some unanticipated links, such as a cluster of 14 people who were subsequently found to be linked through missing intermediates, which helped to explain community spread. Although the United States is 1,000 times more populous than Iceland and has 10 cities with populations greater than 1 million, Iceland provides an example of integration across sectors and entities, leading to improved public health surveillance and disease control.

## CONCLUDING REMARKS

The committee identified multiple limitations to current sources of genomic, clinical, and epidemiological data on SARS-CoV-2. Key limitations include insufficient funding, poor coordination, limited capacity for data integration, unrepresentative data, and lack of an adequately trained workforce with the multifaceted expertise needed to conduct this work. Funding to support these platforms and databases is inadequate both during and between outbreaks, and the funding that is available is not distributed uniformly across efforts. A passive and non-strategic system like SPHERES is inadequate. Fundamental governance and collaboration issues extending from the top down have led to the fragmentation of approaches and varying capacities at local and national levels. Many state public health laboratories are siloed, with disparate methods and widely varying levels of expertise in genomic sequencing for disease surveillance. Local jurisdictions are disproportionately resourced and have wide variability in their capacities to use data during an outbreak. Data architecture is similarly limited and fragmented at the local, state, and national levels, which is a barrier to rapid and open data sharing. Public health personnel often have no formal degree training in public health and thereby lack sufficient training in data science; more trainers in genomic epidemiology are needed to develop actionable interventions.

Robustly integrating genomic data with clinical and epidemiological data would require a health care and public health system with sufficient infrastructure, coordination, and capacity to integrate and analyze the data. However, integrating genomics with medical records, diagnostic outputs, and public health data is difficult in the United States due to fragmented record keeping as well as institutional and regulatory hurdles. Improving the representativeness of sampling will require greater local-level capacity and improving baseline surveillance, as the current reliance on cluster investigations is inadequate to identify viral genetic diversity. Data requests often pose further challenges to already overburdened clinical and reference laboratories, which conduct a majority of sample testing. Building national capacity to integrate genomic, clinical, and epidemiological data would have immediate impact for the nation's response to SARS-CoV-2 and position the United States to respond to any future microbial threat.

*Conclusion: Current sources of SARS-CoV-2 genome sequence data, and current efforts to integrate these data with relevant epidemiological and clinical data, are patchy, typically passive, reactive, uncoordinated, and underfunded in the United States. As a result, currently available data are unrepresentative of many important population features, biased, and inadequate to answer many of the pressing questions about the evolution and transmission of the virus, and the relationships of genome sequence variants with virulence, pathogenesis, clinical outcomes, and the effectiveness of countermeasures. Thus, the viral sequence data and associated data needed are not being collected.*

**RECOMMENDATION 1. The U.S. Department of Health and Human Services should ensure the generation of representative, high-quality full genome sequences of SARS-CoV-2 across the United States, and in the future, from emerging epidemic or pandemic pathogens, in order that these data can be used to meet key needs for genomic surveillance.**

- **Pathogen samples must be obtained from individuals who represent a broad diversity of factors such as race and ethnicity, gender, age, geography, and other demographic features such as housing type, clinical manifestations and outcomes, and transmissibility.**
- **Capacity for genomic sequencing should be developed and supported at many geographically distributed sites performing testing, including public health laboratories and academic and medical centers.**
- **Representative SARS-CoV-2 clinical samples from across the United States should be collected and sequenced on an ongoing basis to provide baseline data and facilitate near-real-time transmission tracking.**
- **Genome sequences should be shared openly on publicly accessible databases, such as the National Center for Biotechnology Information linked to the Global Initiative on Sharing All Influenza Data.**

## REFERENCES

Edmond, C. 2020. *How Microsoft is responding to the COVID-19 pandemic in Washington State*. https://news.microsoft.com/on-the-issues/2020/04/17/microsoft-covid-19-washington-state (accessed July 6, 2020).

Gonzalez-Reiche, A. S., M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, and H. van Bakel. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 369(6501):279–301.

Gudbjartsson, D. F., A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem, A. B. Agustsdottir, B. Eiriksdottir, R. Fridriksdottir, E. E. Gardarsdottir, G. Georgsson, O. S. Gretarsdottir, K. R. Gudmundsson, T. R. Gunnarsdottir, A. Gylfason, H. Holm, B. O. Jensson, A. Jonasdottir, F. Jonsson, K. S. Josefsdottir, T. Kristjansson, D. N. Magnusdottir, L. le Roux, G. Sigmundsdottir, G. Sveinbjornsson, K. E. Sveinsdottir, M. Sveinsdottir, E. A. Thorarensen, B. Thorbjornsson, A. Löve, G. Masson, I. Jonsdottir, A. D. Möller, T. Gudnason, K. G. Kristinsson, U. Thorsteinsdottir, and K. Stefansson. 2020. Spread of SARS-CoV-2 in the Icelandic population. *New England Journal of Medicine* 382(24):2302–2315.

Kent, J. 2020. *Amazon, UCSF partner for COVID-19 genome sequencing projects*. https://healthitanalytics.com/news/amazon-ucsf-partner-for-covid-19-genome-sequencing-projects (accessed July 7, 2020).

Stofa, F. M. 2019. *Tourism in Iceland in figures: 2018*. https://www.ferdamalastofa.is/static/files/ferdamalastofa/talnaefni/tourism-in-iceland-2018_2.pdf (accessed July 6, 2020).

Virological. 2020. *Introduction and spread of SARS-CoV-2 in the greater Boston area*. https://virological.org/t/introduction-and-spread-of-sars-cov-2-in-the-greater-boston-area/503 (accessed June 24, 2020).

Wu, F., S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579(7798):265–269.

Zhou, P., X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao, and Z.-L. Shi. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579(7798):270–273.

# 4

# Framework to Track and Correlate Viral Genome Sequences with Clinical and Epidemiological Data

To inform public health analysis of an infectious disease outbreak, the genomic sequence of the pathogen obtained from an infected person must be accurate and be linked with sufficient metadata for context. In this chapter, the committee lays out a framework to describe the types of clinical and epidemiological data that need to be linked to viral genome sequence data to answer specific questions related to transmission, evolution, treatment, and prevention of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and in the future, new emerging epidemic or pandemic pathogens. It concludes with a discussion about the data integration and infrastructure considerations for a system to track and correlate genomic, clinical, and epidemiological data. Demographic factors, such as age or occupation, are also important components to understand disease transmission and data collection needs for specific populations.

## CONSIDERATIONS FOR TRANSMISSION, EVOLUTION, AND CLINICAL DISEASE

### Overarching Data Collection Considerations

Acquisition of genomic data is one piece (see Recommendation 1 in Chapter 3) but will also be reliant on clinical and epidemiological data to understand the evolution of SARS-CoV-2 and the implications for transmission and clinical manifestations. The collection of clinical data is exceedingly important but also one of the biggest hurdles.

Temporal and geographic information (date and location of specimen collection) are essential for assessing spread of the pathogen in time and space throughout the epidemic, establishing transmission chains, developing predictions, and identifying clusters of similar sequences as an indication of a super-spreading event, for example. Similarly, any recent travel to places, gatherings, or events that might currently or subsequently be recognized as areas of high disease activity is fundamentally important for mitigation. Residence in a long-term care facility, recent (especially inpatient) clinical encounters, and close contact with a person known to have coronavirus disease 2019 (COVID-19) would be key variables for downstream use. Comorbid disease, immunosuppression, and disease severity may reveal associations with viral evolution that would otherwise be undecipherable. Preceding receipt of antiviral treatment, episode(s) of COVID-19, and any prior SARS-CoV-2 vaccination, are likely to become increasingly relevant in the future to contextualize SARS-CoV-2 evolution in response to selective pressure such as escape from antiviral responses.

A critical overarching consideration will be in ensuring representation through participatory parties (Gould et al., 2017). To ensure that epidemiological sampling is representative of populations at risk, basic demographics should be linkable with the genomic sequence. A mixture of public health, health care, tribal leaders, bioethics, community health leaders, and those working in genomic epidemiology would be beneficial to help determine how best to represent all critical parties. In fact, it may be helpful to establish a proactive "push" team that helps resource-challenged areas—such as tribal territories and critical access hospitals—ensure they are afforded representation. Adequate representation should go beyond geographical considerations, and should also include gender, race, ethnicity, living situation, and occupation.

Table 4-1 briefly outlines how viral genome sequence data, when combined with other types of data, can be used to inform questions related to transmission, evolution, and clinical disease.

### Transmission

Data on viral genomic sequences can answer questions related to the source(s) of the virus causing an outbreak. The "simplest use of genomic data" is used to show how viral spread happens when combined with phylogeographic approaches where it can be used to detect transmission hot spots and help direct interventions (Holmes et al., 2016). For the current situation with SARS-CoV-2, such data can determine how the virus is spreading between individuals and within a community. Once a vaccine is available, these data can determine whether new cases are due to virus importation or to local spread. For instance, genomic epidemiology with

**TABLE 4-1** Summary Table of Considerations for Transmission, Evolution, and Clinical Disease

| Goal | Question | Viral Genomic Sequence Data Needs | Clinical and/or Epidemiological Data Needs[a] |
|---|---|---|---|
| Transmission patterns | Is outbreak due to multiple introductions? Where is the virus coming from? | Pathogen samples from individuals who represent broad diversity from outbreaks and many regions/countries | Time and place of virus isolation and travel history of cases |
| | Is outbreak due to local spread? How and/or where is the virus being transmitted? | Sequences from local groups/areas with increased incidence rates | Local population-based information on sites of exposure, gatherings, isolated communities, and congregate living (long-term care facilities, hospitals, prisons) |
| | Is there evidence of super-spreading events and how important are they? | Sequences of virus from groups of people infected in the same setting | Information on sites of exposure, gatherings |
| Evolution/ influence of selective pressures | Is the virus changing in transmissibility? | Changes in viral genome sequence associated with increased spread | Calculations of $R_0$ (contact tracing data–number of people infected) |
| | Is resistance to antiviral drugs or other treatments changing? | Changes in viral genome associated with failure to respond to treatment | Hospital or health care center data on patients who do not respond to therapy or show failure of treatment |
| | Is there altered escape from the host immune response/within host evolution? | Changes in viral genome associated with persistence | Hospital data on patients who show prolonged shedding |
| | Is there changed protection from vaccine-induced immunity? | Changes in virus that affect epitopes important for protective immunity and sequences of viruses associated with vaccine failure | Vaccine trial databases and post-marketing vaccine failures |

*continued*

**TABLE 4-1** Continued

| Goal | Question | Viral Genomic Sequence Data Needs | Clinical and/or Epidemiological Data Needs[a] |
|---|---|---|---|
| Clinical disease | Are there strains/mutations associated with changes in disease severity? | Sequences of viruses from patients with different disease severity | Severity of symptoms, ICU, ventilation, mortality, length of hospitalization, and co-infections |
| | Are there strains/mutations that affect virus loads or clearance? | Sequences of viruses from patients with viral load data | RT-PCR data to measure viral load of respiratory secretions, blood, and feces over time |
| | Are there strains/mutations that affect response to different treatments? | Sequences of viruses from before and after treatment | Treatment type, duration, and outcome |
| | Are there strains/mutations that are associated with response to different treatments? | Sequences of viruses from different body sites and patients with and without specific complications | Clinical data on complications related to different organ systems (e.g., kidney, liver, nervous system) |
| | Are there strains/mutations that predispose to MIS-C? | Sequences of viruses from children in the same community/family with and without MIS-C | Clinical data over time on immune response, viral load, treatment, and response |

NOTE: ICU = intensive care unit; MIS-C = multisystem inflammatory syndrome in children; $R_0$ = basic reproductive number; RT-PCR = reverse transcription polymerase chain reaction.

[a] The committee recognizes that clinical and epidemiological data often come from very different data collection sources and efforts, but for the purposes of this table these data needs have been incorporated into one column.

knowledge of viral sequences from different regions is regularly used to determine whether cases of measles virus infection are due to introduction from countries with continued endemic measles or to chains of transmission within the community due to inadequate population immunity (Harvala et al., 2015; Penedos et al., 2015).

### *Where Is the Virus Coming From?*

As described in Chapter 3, sequencing 87 SARS-CoV-2 genomes from infected patients early in the spread of COVID-19 in New York City demonstrated multiple independent introductions of dominant strains circulating in Europe followed by undetected local transmissions (Gonzalez-Reiche et al., 2020). In a hypothetical scenario, the reader should imagine the first group of college students arriving to a college campus in August 2020. If cases of COVID-19 begin to be detected in the days and weeks that follow, administrators and health care providers will need to respond in near real time. Important to their mitigation strategy will be distinguishing multiple independent introductions from local transmission. To understand what proportion of students came to campus carrying SARS-CoV-2 strains from their home regions will require national and international baseline data. Moreover, to know which events to discourage, students will need to provide accurate data of their activities and contacts—many of whom they will not know.

Genomic data linked to time, place, and exposure history will help to cluster cases, delineate local transmissions, and illuminate which epidemiological links need not be investigated further.

### *Where Is the Virus Being Transmitted?*

Of particular epidemiological importance for SARS-CoV-2 is identification of route of transmission, asymptomatic spread, and super-spreading events. Virus sequence data can help identify transmission via different pathways, both expected and unexpected (Holmes et al., 2017). For instance, SARS-CoV-2 RNA is frequently found in stool samples as well as respiratory secretions with more persistent shedding from the gastrointestinal tract (Xu et al., 2020). Viral RNA in stool and as aerosols in the toilet areas of communal living facilities (Liu et al., 2020) may or may not represent infectious virus (Wölfel et al., 2020). Identification of fecal–oral transmission will require epidemiological information on exposures linked to virus sequence information and could have a substantial effect on public health interventions. Likewise, knowledge of transmission from sites of virus persistence (particularly semen as now recognized for Zika and Ebola viruses) provides opportunities for late transmission to reignite outbreaks after apparent control, which affect public health interventions.

Super-spreading events and identification of the settings where they occur are of particular epidemiological importance. These events can only be identified with viral sequence data from multiple individuals involved in an outbreak linked to information on participant activities, such as religious services, sporting events, or concerts (Holmes et al., 2016).

## Evolution and Influence of Selective Pressures

To better understand the evolution of SARS-CoV-2 in the United States or elsewhere, it would be ideal to integrate patient clinical data and genomic sequence data, with representation of both abundant and rare viral genotypes, representative of geographic, gender, racial, ethnic, and other demographics. Of course, the difficulty of this goal is the challenge in obtaining such data, given the current lack of an efficient and reliable network to connect data drawn from local regions across the United States. Thus, it remains challenging to elucidate how the virus is currently evolving, which suggests poor ability to predict its future potential for evolution in the face of ongoing and novel selection pressures, such as vaccine development.

A brief comparison of the evolution patterns of SARS-CoV, Middle East respiratory syndrome (MERS)-CoV, and SARS-CoV-2 reveals interesting similarities and differences. Although SARS-CoV-2 studies suggest an emergence event involving single lineage, it is clear that multiple introductions of SARS-CoV and MERS-CoV occurred early in the expanding epidemic (Liya et al., 2020). During the SARS-CoV epidemic, distinct mutations in the receptor-binding domain were critically associated with the emergence of middle- and late-phase isolates that spread geographically, but transiently throughout the world (Hu et al., 2017). Other interesting differences include the high transmissibility of SARS-CoV-2, prior to disease symptom onset, while both SARS-CoV and MERS-CoV are primarily transmitted after clinical disease onset. Mortality rates of the three emerging coronaviruses are estimated at 1, 10, and 35 percent for SARS-CoV-2, SARS-CoV, and MERS-CoV, respectively. While asymptomatic infections were and are rare in the 2003 SARS-CoV epidemic and the ongoing MERS-CoV outbreak, asymptomatic infections are common in SARS-CoV-2 infections, recently estimated to represent 40–50 percent of all cases (Feaster and Goh, 2020). What are the genetic differences between SARS-CoV and SARS-CoV-2 that regulate these fundamental differences in transmissibility, virulence, and pathogenesis? Could highly virulent, highly transmissible coronavirus strains emerge from zoonotic sources or during an expanding epidemic or pandemic? How does virulence evolve after a zoonotic emergence event? What is the relationship between the evolution of virulence and coronavirus transmissibility? Using model organisms, the evolutionary relationships between virulence and transmissibility are thought to be complex traits and include examples of synergistic and antagonistic relationships (Geoghegan and Holmes, 2018). Given the large diversity of novel coronaviruses harbored in bats and other animals, it is therefore conceivable that many worse highly transmissible and highly virulent zoonotic coronaviruses may exist in nature that threaten human populations in the future. Consequently, fundamental insights into the evolutionary trade-offs and genetic relationships between SARS-CoV-2 evolution, virulence, and transmissibility may better inform global preparedness efforts, designed to minimize the impact of consequential coronavirus disease outbreaks of the future (Messenger et al., 1999).

For example, if the mutation rate (replication fidelity) changes such that more allele substitutions occur per round of genome replication, it would indicate that greater variation and adaptive potential is available to the virus as raw fuel for evolution by natural selection (Duffy et al., 2008; Elena and Sanjuán, 2007). In turn, this scenario could lead to adaptive change whereby the major virus variants become either more or less dangerous (virulent), such that increased or decreased mortality risk becomes associated with COVID-19. Thus, evolution of a higher mutation rate in the virus may not be necessarily problematic from a clinical or public health perspective, because viral adaptation may coincide with greater or lesser host morbidity and mortality. The adaptive potential of any biological system relies on a positive correlation between increased mutation rate and a larger number of useful (beneficial) mutations occurring in the population (Orr, 2000). That is, mutation rate can create more changes per unit of time, but there is no guarantee that this will also create a larger fraction of beneficial mutations, because the latter is determined by how well spontaneous mutations provide an adaptive match to the selective challenges faced by the population. Nevertheless, even if the mutation rate of SARS-CoV-2 remains unchanged, the short generation times of the virus—coupled with the very large number of infected human hosts—create ample opportunity for rare spontaneous mutations to arise and spread over short periods of time, indicating enormous virus evolutionary potential.

### *Is Virus Transmissibility Changing?*

To date, about a dozen mutations in the gene encoding the spike protein are accumulating and being evaluated for positive selection. A prominent D614G mutation identified both in China and Europe in January 2020 is expanding in geographic range and frequency across the world (Korber et al., 2020). The mutation is located on the interface between spike protomers where it may alter stability that enhances infectivity. Identification of this mutation is leading to more detailed studies aimed at unraveling the importance of this mutation in the biology of SARS-CoV-2 and its relationship to other mutations in the genome that may contribute to the selective sweep of this genotype across the globe. In addition to the in vitro data on competitive cell entry and growth rates that have been released recently (Grubaugh et al., 2020; Hu et al., 2020; Korber et al., 2020; Zhang et al., 2020), it will be important to examine the time course and natural history of mutant and isogenic parental strain experimental infections in relevant whole animal models, as well as in naturally infected humans, and households.

In addition, several other spike mutations have been observed in smaller clusters of cases. However, none have risen to global prominence. For example, signal peptide mutations (L5F and L8V) could potentially affect posttranslational modifications, folding, abundance, and glycosylation, while residue changes V367F, G476S, and V483A are found within the RBD domain. However, only G476S is located at the RBD binding interface. The functional significance of these mutations in mammalian angiotensin-converting enzyme 2 interaction networks (primate and animal) remain unknown and warrant additional study. Finally, several other mutations occur in regions of unknown function (H49Y, Y145H/del, Q239K) and appear to be diminishing, or are remaining stable in the population and located in and about the fusion machinery (A831V and D839Y/N/E) or c-terminal end (P1263L). Other mutations have been recorded in ORF1ab and ORF8 regions, although their functional significance remains unknown (Chang et al., 2020).

### *Is the Virus Evolving in Response to Selective Pressures?*

Many selective pressures could lead to evolved changes in viral traits that improve the success and spread of SARS-CoV-2 infection. For instance, increased virus particle stability in aerosols or on surfaces could promote transmission opportunities (van Doremalen et al., 2020) and methods exist to interrogate how spontaneous mutations can improve virus stability against environmental degradation (Ogbunugafor et al., 2013). Relatedly, increased time between transmission events should select for infectious viruses with increased particle stability, although evolution of increased particle stability tends to trade-off with rate of genome replication, such that more durable viruses suffer slower reproduction (Goldhill and Turner, 2014). However, it is unclear whether the stability–reproduction trade-off would reduce viral load in an infected human as examined in influenza virus (Handel et al., 2014), a highly relevant clinical concern for COVID-19. Therefore, prolonged social distancing could select for SARS-CoV-2 variants with increased particle stability that may or may not affect viral load during infection. If genomic epidemiology studies point to virus evolution at loci that affect SARS-CoV-2 particle stability, it would provide motivation to closely study whether clinical symptoms in infected patients are changing as well.

Additional selective pressures (e.g., antiviral and immune modulating treatments and eventually vaccination) are being introduced with unknown consequences for virus evolution. Antibody responses and drug activities are dependent on specific regions of the viral proteins. For instance, the drug remdesivir requires certain regions of the viral RNA-dependent polymerase and exoribonuclease (Agostini et al., 2018), and the receptor-binding domain of the spike protein is the main target for neutralizing antibody and most vaccines in development (Premkumar et al., 2020). Mutation in these proteins could affect treatment and vaccine efficacy.

More recently, host susceptibility loci on human chromosomes 3 and 9 may be unevenly distributed globally, potentially selecting for new variants of SARS-CoV-2 that interacts specifically well with this human genotype (Ellinghaus et al., 2020). Identification of these viral changes and their importance requires linkage of patient metadata to genome sequencing, to determine response to treatments and identify instances of vaccine failure. It will be critical to the control of this pandemic to recognize both types of evolution in real time to be able to institute corrective action.

Together, these data reveal a critical need for reverse genetic strategies and well-developed models to investigate the role of these mutations in SARS-CoV-2 biology and disease. Moreover, these data reveal the critical need for linkage of patient metadata to genome sequencing, without which definitive causal epidemiological associations between genotype and phenotype cannot be determined.

### Clinical Disease

There has been relatively little sequence variability among human SARS-CoV-2 isolates so far, so compelling associations between sequence variant/mutations and specific clinical outcomes or features have not yet been identified. Nonetheless, identification of virus strains with different clinical features would provide insights into disease pathogenesis and potentially identify patients requiring specific interventions. Linking virus sequence data with data on patient demographics, hospitalization, duration of hospitalization, clinical complications, intensive care unit (ICU) stay, co-infections, ventilation/duration, use of extracorporeal membrane oxygenation (MacLaren et al., 2020), duration of positive reverse transcription polymerase chain reaction tests for SARS-CoV-2 RNA, and exposure/response to experimental treatments (e.g., remdesivir, convalescent plasma, dexamethasone, immune modulators) would facilitate identification of:

- Strains/mutations associated with changes in disease severity, viral loads, and viral shedding periods
- Strains/mutations associated with more co-infections or response to certain medical interventions, such as convalescent plasma
- Strains/mutations associated with specific complications, for example, neurologic manifestations (Wood, 2020), vascular complications (Ackermann et al., 2020; Lang et al., 2020), hypercoagulable states, gastrointestinal manifestations (Ong et al., 2020), or fecal shedding
    - A question that underlies all of these organ system complications: Are there SARS-CoV-2 mutations that affect organ tropism in humans?

- Strains or mutations associated with the pediatric multisystem inflammatory syndrome (Feldstein et al., 2020)

## OPPORTUNITIES TO SUPPORT DATA INTEGRATION

An essential component of all of this involves the timely integration of data. Information flows comprising viral genomic data, as well as associated clinical and epidemiological aspects, will be useless unless there are proactive efforts to distill the data into the most useful components and then to organize the data into an integrated format that can be used by researchers, health care providers, public health practitioners, and policy makers. Given the potential large scale of the universe of data that might be available, it will be important to determine what types of information are thought to be the most relevant. Given the uncertainty as to how the current or future pandemics might progress, however, it will also be important to build flexibility and expansion capability in the resultant data management system in order to accommodate additional sources and types of data.

A national system for integrating genomic, clinical, and epidemiological data collected during an infectious disease outbreak would receive large volumes of data coming in from multiple sources, including federal, state, and local public health agencies; health care networks; and public health and clinical laboratories. Currently, no central repository exists for these different types of data and the entities contributing the information do not have dedicated staff to curate the data. Efforts to build an infrastructure to facilitate integration will likely face multiple challenges related to coordination, interoperability, flexibility, and privacy. For instance, interoperability may be a challenge because incoming data will likely be shared in a range of different formats. Constraints across existing databases, such as differences in the content fields for inputting information, can preclude the ability to record and share the full range of relevant data. Laws and regulations that govern data sharing and privacy—as well as their local-level interpretations—can also potentially impact the data in variable ways, depending on the data source, relevant regulatory or legal restrictions, and concerns related to protected health information. Regulatory and governance considerations are discussed further in Chapter 5.

In terms of encouraging participation, tying participation into existing Medicare and Medicaid financial incentives—similar to the efforts of the BioSense Platform (Gould et al., 2017)—can ensure a wide group of participants, including those in ambulatory settings. This effort would link into hospital data and help answer several important clinical questions. Hospital grants could also be made available for data agreements. With competing data reporting requirements, and the fact that a new system may create new expectations for laboratories that would normally dispose of samples, data collection must find a middle ground with a return on the investment. Providing real-time data pushes to participating parties, such as infection prevention programs or hospitals, is also an important consideration.

A prime opportunity to address these barriers is to develop agreed-upon standard data packages for submission to the system. Importantly, these packages should allow for some degree of variation for different sources and types of data. For example, a hospital laboratory would submit a comprehensive package of clinical and diagnostic data, while a commercial clinical laboratory would submit a larger population-based data package lacking clinical details. State and local public health agencies would likely have a variety of data types. To support this work, scoping of those data packages should be factored into any analytical plans. An actual data repository, along with the requisite support and analytical staff, also needs to be established. A key component of building this repository is to establish countrywide reporting relationships across all levels to ensure that comprehensive data are being submitted. This data repository should be flexible—for example, it should ensure ease of adding new data types and fields—as well as be accessible for advanced analytical methods, such as machine learning and artificial intelligence analyses to inform disease and epidemiology models.

## INFRASTRUCTURE NEEDS

Most previous efforts to integrate genomic, clinical, and epidemiological data in response to viral or microbial outbreaks have been conducted on a small scale. To optimize the application of integrated data to inform the response to SARS-CoV-2 and future outbreaks, these efforts will need to be scaled up to nationwide infrastructure through which data can be shared and reported. A primary role of the U.S. Centers for Disease Control and Prevention (CDC) is epidemiological surveillance. The agency has links to each state-level health department as well as a global network of other national agencies, in which CDC serves as the country's representative in international cooperation to fight emerging infectious diseases. The fields of clinical microbiology and epidemiology have now largely embraced genomic sequencing. Although there have been successful efforts in applying genomic epidemiology to influenza and outbreaks of foodborne bacteria (see the case studies in Chapter 2), CDC has lagged behind in incorporating genomics to its full potential. CDC is responsible for funding public health laboratories nationwide to facilitate the integration of data; however, most of those laboratories remain substantially under-resourced.

To enable larger-scale collaboration and coordination of data in a national system, insights can be gleaned from the innovative elements and constraints of CDC's ongoing efforts and from other existing regional networks of data integration. CDC's PulseNet,[1] established in 1996, allows members of the network to compare whole genome sequencing of bacterial DNA to help detect and mitigate foodborne outbreaks. The National Action Plan for Combating Antibiotic-Resistant Bacteria[2] (CARB), a national strategy (PCAST, 2014) to track antibiotic-resistant bacteria, led to the establishment of CDC's Antibiotic Resistance Laboratory Network.[3] The network strengthens national laboratory capacity to rapidly perform genomic epidemiological studies, as well as providing a mechanism for coordination and reporting. This served as the impetus for the coordination of all reporting across New York State that was leveraged for SARS-CoV-2.

### Enclave Model in the National COVID Cohort Collaborative to Enable Linkage of Detailed Clinical Metadata

The National COVID Cohort Collaborative (N3C) embodies a massive, scalable collection of medical record data from people infected with SARS-CoV-2 in a centralized, secure enclave (see Chapter 3).[4] N3C uses a project-specific hashed identifier constructed using data security standards to support linking data from disparate sources without revealing the personal identifiers used to generate the hashed ID (N3C, 2020). To support linking SARS-CoV-2 genome sequences to clinical metadata in N3C, viral genome sequences, or links (e.g., accession numbers) to their records in GenBank or the Global Initiative on Sharing All Influenza Data, would need to be deposited into N3C.

N3C is expected to contain data from 2–3 million people with confirmed SARS-CoV-2 infection by the end of 2020, and is designed with the potential to accommodate data from the U.S. population. Data in N3C are converted to the Observational Medical Outcomes Partnership (OMOP) standard (version 5.3.1, currently) after ingestion, mapping, and harmonization from multiple supported data standards. Because data accessible in the N3C are a limited dataset under terms preventing re-identification, important epidemiological activities such as contact tracing are not supported; nonetheless, inclusion of SARS-CoV-2 genomic data into N3C would represent a clinically phenotyped collection of viral genomic sequences that could scale to the U.S. population.

[1] See https://www.cdc.gov/pulsenet/index.html (accessed June 25, 2020).

[2] See https://aspe.hhs.gov/pdf-report/national-action-plan-combating-antibiotic-resistant-bacteria-progress-report-year-3 (accessed June 25, 2020).

[3] See https://www.cdc.gov/drugresistance/solutions-initiative/ar-lab-network.html (accessed June 25, 2020).

[4] See https://ncats.nih.gov/news/releases/2020/NIH-launches-analytics-platform-to-harness-nationwide-COVID-19-patient-data-to-speed-treatments (accessed June 25, 2020).

### Using Influenza Infrastructure to Integrate SARS-CoV-2 Data

Linking genomic data for SARS-CoV-2 with clinical and epidemiological data might be possible by utilizing pre-existing systems for tracking changes in the genomic structure of the influenza virus. CDC collaborates with many partners in state, local, and territorial health departments and laboratories, offices of vital statistics, health care providers, clinics, and emergency departments to monitor influenza on an annual basis (CDC, 2020). The U.S. influenza surveillance system is designed to find out when and where influenza activity is occurring; determine what influenza viruses are circulating; detect changes in influenza viruses; and measure the impact influenza is having on outpatient illness, hospitalizations, and deaths (CDC, 2020). These goals are in line with what the committee proposes for the use of genomic data on SARS-CoV-2.

Approximately 100 public health and 300 clinical laboratories in all 50 states, Puerto Rico, Guam, and the District of Columbia participate in surveillance for influenza viruses through either the U.S. World Health Organization Collaborating Laboratories System or through the National Respiratory and Enteric Virus Surveillance System.

Data from clinical laboratories provide useful information on the timing and intensity of influenza activity from respiratory specimens largely obtained for diagnostic purposes. Public health laboratories provide data useful to understand what influenza virus types, subtypes, and lineages are circulating and the age groups being affected as test specimens are collected primarily for the purposes of surveillance. For genetic characterization, all influenza-positive surveillance samples are submitted for genomic sequencing by CDC to determine the genetic characteristics of circulating influenza viruses and to monitor the course of evolution of viruses circulating in the population under surveillance. Phylogenetic analysis classifies virus gene segments into genetic clades or subclades. CDC also tests a sample of the influenza viruses collected by public health laboratories for susceptibility to antiviral, such as neuraminidase inhibitors using genomic sequencing analysis and/or a functional assay.

The U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) collects information on outpatient visits to health care providers in all 50 states, Puerto Rico, the District of Columbia, and the U.S. Virgin Islands for influenza-like illness (ILI). More than 2,500 outpatient health care providers around the country report data to CDC every week recording the total number of patients seen, including specifically the number of those patients with ILI by age group (0–4 years, 5–24 years, 25–49 years, 50–64 years, and ≥65 years).

The Influenza Hospitalization Surveillance Network (FluSurv-NET) monitors laboratory confirmed influenza-associated hospitalizations in children younger than 18 years of age (since the 2003–2004 influenza season) and adults (since the 2005–2006 influenza season). High-risk medical conditions are extracted from patient medical charts at the time of hospitalization, including cardiovascular disease, chronic lung disease, immunocompromised condition, obesity, and pregnancy status that match similar underlying conditions of interest in patients with COVID-19.

## Health Information Exchanges

In the wake of the Health Information Technology for Economic and Clinical Health Act of 2009[5] and continuing financial incentives from the Centers for Medicare & Medicaid Services, there is widespread adoption of electronic medical records systems by hospitals and physician practices (CDC, 2019). In addition, health information exchanges, built largely to facilitate the exchange of digital health information for clinical treatment purposes, exist across the country; according to one survey, 7 out of 10 hospitals in the United States belong to at least one nationwide health data sharing network (Johnson et al., 2018).

In the 21st Century Cures Act, the U.S. Congress required the U.S. Department of Health and Human Services (HHS) to establish a voluntary network to facilitate nationwide digital sharing of electronic health information, and a public–private partnership has been launched, led by the Sequoia Project, to create a national health information sharing "trusted exchange framework" pursuant to a common agreement (HealthIT.gov, 2020b). In addition, on May 1, 2020, the HHS Office of the National Coordinator for Health Information Technology finalized rules to prohibit health care providers, certified electronic medical record vendors, and health information exchanges from "blocking" the sharing of information, including for public health purposes; these rules will go into effect on November 2, 2020 (HealthIT.gov, 2020a). Although this national network is still in formation, certified electronic medical record vendors and health information exchanges across the country could be leveraged today to facilitate the sharing of clinical metadata that will help public health departments and researchers answer critical questions related to SARS-CoV-2 and COVID-19. Implementation of interoperable health records systems must be cognizant of the potential for sensitive and private personal health information to be inadvertently shared, en masse. This not only risks violating individuals' rights to privacy and non-discrimination, but also undermines public trust and as a result, the potential accuracy of public health data gathered. Upgrading existing records systems will likely be necessary to allow for options to protect privacy such as segmentation

[5] See https://www.govinfo.gov/content/pkg/PLAW-111publ5/pdf/PLAW-111publ5.pdf (accessed June 25, 2020).

of data in a manner that achieves both the goals of sharing and irrelevant or sensitive individual personal health data (Rothstein and Tovino, 2019). While unrestricted access to shared data would incur privacy risks, the enclave model of N3C described above illustrates sharing of national-scale health data with strong privacy protections.

### Participatory Surveillance

The growing field of participatory surveillance allows individuals to report symptoms of illness through crowd-sourced, voluntary systems that allow for community-level health monitoring (Smolinski et al., 2017). A number of participatory surveillance systems already exist worldwide. Most of these systems collect epidemiological data that are provided to public health authorities and research institutions and used to analyze trends and broaden surveillance beyond the traditional, sentinel surveillance approach. For instance, participatory surveillance provides a mechanism to collect information on influenza in the community at large. Because the majority of persons with influenza each year do not seek medical care (Biggerstaff et al., 2014; Van Cauteren et al., 2012; van Noort et al., 2007), a large number of self-reporting systems collect information on ILI. Boston Children's Hospital's Flu Near You[6] is a self-reported ILI system that has combined laboratory testing for diagnosis of influenza or other respiratory pathogens with the epidemiological data collected by the open-source GoViral Study (Li, 2016). Individuals who report symptoms of illness compatible with influenza are provided with a home test kit and asked to collect a sputum sample for testing; the results of the test are then compared to the symptom data shared in the open-source system. Such participatory surveillance systems could potentially be expanded through the use of home test kits that would allow for genomic sequencing of pathogens in the community.

Numerous community-based surveillance systems have arisen during the COVID-19 pandemic. For example, the Flu Near You system was adapted with expanded symptoms related to SARS-CoV-2 infection into the COVID Near You[7] system. Other systems have arisen as longitudinal research projects that are collecting epidemiological data along with testing results for COVID-19. All of these systems offer opportunities to incorporate genomic sequencing, which could target data collection from specific subsets of the population or from people in specific geographic regions. Genomic sequencing could be added to existing systems for COVID-19 symptom reporting, tracking, and contact tracing in the United States.

[6] See https://flunearyou.org (accessed June 25, 2020).

[7] See https://www.covidnearyou.org/us/en-US (accessed June 25, 2020).

## PARTNERSHIPS, COORDINATION, AND CAPACITY CONSIDERATIONS

Fostering partnerships across laboratories in different sectors and at different levels—from state and local public health to clinical, academic, and commercial laboratories—will be critical for developing capacity and facilitating coordination necessary for national-level genomic data that can be integrated with clinical and epidemiological data. These efforts should seek to partner with a range of different laboratories to better represent the entire population. CDC's SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance will likely cover a large proportion of the population, but better coverage could be achieved by partnering with the third-party private laboratories that are often contracted with health care systems (e.g., LabCorp or Quest Diagnostics). Hospital and clinical laboratories are also valuable sources of information, especially in rural or critical access areas. Tests are going unused in many of these settings, where facilities often lack laboratory capacity and local health systems face multiple barriers to utilizing their samples (Maxmen, 2020). In such settings, partnerships can also provide access to important data on hard-to-reach patient populations. These same partnerships should exist with academic and commercial laboratories, albeit with an awareness of capacity considerations.

Data usability, capacity considerations, and key outputs—to include content and periodicity of reports—are important aspects of coordinating partnerships, particularly for smaller hospital and public health laboratories operating with limited resources. For example, close support, coordination, and capacity building are all valuable for mitigating the stress experienced by laboratories in the context of an infectious disease outbreak. Many of these laboratories may have valuable data on vulnerable patient populations that are not being reported into broader databases or larger systems due to a range of barriers, such as bureaucratic red tape, dependency on limited resources, and often outdated tools for communication and data sharing (e.g., faxing). Tying data collection and integration into hospitals' meaningful use standards could be a beneficial approach for biosurveillance. Ultimately, however, the utilization of a low-cost approach that mitigates such barriers to collecting, analyzing, and sharing data is critical. Systems should be in place that enable hospitals to seamlessly push data to their key stakeholders, such as public health agencies, infection prevention programs, and clinical partners. Furthermore, partnerships can help to ensure these data are presented in a way that is beneficial to the end user. For instance, genomic data are of great value for many purposes, but clinical and epidemiological data may be relevant and applicable to patient care and public health outcomes.

### Workforce Capacity Development for Genomic Epidemiology

An important consideration in the development of such a system is the building of genome sequencing and analysis capabilities within public health agencies and health care systems. Even with advancing technology, multidisciplinary and highly trained teams remain the most valuable asset for combining genomic, clinical, and epidemiological data into actionable knowledge (Lesho et al., 2016). Since 2016, the Broad Institute has partnered with the Massachusetts State Public Health Laboratory and CDC to build distributed capacity for genomic sequencing through a train-the-trainer program for regional- and state-level laboratory personnel. This program, for example, could serve as a model for developing national coordination among state public health laboratories.

## CONCLUDING REMARKS

There remains no central repository to house the large volume of individually identifiable data from various actors involved in the public health response to SARS-CoV-2 in the United States, just as none existed for prior infectious disease outbreaks or for hypothetical future outbreaks. While it is important to learn from several of the smaller scale examples described in this report, building out successful elements from these success stories remains a major challenge at the national scale. The committee recognizes that advancing beyond the current small-scale efforts to a national or even global repository is a challenging undertaking, but the current pandemic puts the lack of such a system in stark relief. Incremental efforts, such as establishing regional repositories, can be taken now and leveraged in the future for a large-scale effort. As noted above, leveraging and expanding existing infrastructure and planning—through programs such as N3C, PulseNet, CARB, ILINet, and health information exchanges—will be crucial to addressing the data infrastructure challenge in a way that is both innovative and iterative. The creation of a system of data infrastructure built on a standard data package could cultivate a more interoperable data environment, a challenge of paramount importance when principles such as flexibility and privacy remain priorities. Ultimately, a data management and infrastructure system with investment in the proper resources, staff, and storage will be critical for the coordination of data needs in response to SARS-CoV-2 and future outbreak responses.

**RECOMMENDATION 2. The U.S. Department of Health and Human Services should develop and invest in a national data infrastructure system that constructively builds on existing programmatic infrastructure with the ability to accurately, efficiently, and safely link genomic data, clinical data, epidemiological data, and other relevant data across**

**multiple sources critical to a public health response such as the current SARS-CoV-2 outbreak. Such a system should:**

- **Allow for the linkage of genomic data, clinical data, epidemiological data, and other relevant data in a way that is not overly burdensome to laboratories that collect data regularly.**
- **Create and foster safe data-sharing practices to ensure that individuals' personal identifying information remains unexposed when data are being used and shared across the system.**
- **Be grounded in the pursuit of standardization, interoperability, flexibility, and the practical linkage of data, including consideration of a potential national patient identifier.**
- **Consider not only the data required to create such a system, but also investment in mechanisms supporting the collection and analysis of such data, including promoting formal education in "data wrangling" at the intersection of data science and infectious disease epidemiology.**
- **Conduct regular annual reviews—including scenario-based simulations—to identify capacity gaps, promote process improvement (based on existing U.S. infrastructure to assess the annual risk of seasonal influenza, work could improve usability and coverage of health information exchanges, and other initiatives), and ensure inclusion of entities with supporting functions across scales—including private health care systems that provide data or state and local public health laboratories that collect data—in ongoing system development and evaluation.**

## REFERENCES

Ackermann, M., S. E. Verleden, M. Kuehnel, A. Haverich, T. Welte, F. Laenger, A. Vanstapel, C. Werlein, H. Stark, A. Tzankov, W. W. Li, V. W. Li, S. J. Mentzer, and D. Jonigk. 2020. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in COVID-19. *New England Journal of Medicine* 383:120–128.

Agostini, M. L., E. L. Andres, A. C. Sims, R. L. Graham, T. P. Sheahan, X. Lu, E. C. Smith, J. B. Case, J. Y. Feng, R. Jordan, A. S. Ray, T. Cihlar, D. Siegel, R. L. Mackman, M. O. Clarke, R. S. Baric, and M. R. Denison. 2018. Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. *mBio* 9(2):e00221-18.

Biggerstaff, M., M. A. Jhung, C. Reed, A. M. Fry, L. Balluz, and L. Finelli. 2014. Influenza-like illness, the time to seek healthcare, and influenza antiviral receipt during the 2010-2011 influenza season-united states. *The Journal of Infectious Diseases* 210(4):535–544.

CDC (U.S. Centers for Disease Control and Prevention). 2019. *Public health and promoting interoperability programs: Introduction*. https://www.cdc.gov/ehrmeaningfuluse/introduction.html (accessed July 7, 2020).

CDC. 2020. *U.S. influenza surveillance system: Purpose and methods*. https://www.cdc.gov/flu/weekly/overview.htm (accessed June 24, 2020).

Chang, T. J., D. M. Yang, M. L. Wang, K. H. Liang, P. H. Tsai, S. H. Chiou, T. H. Lin, and C. T. Wang. 2020. Genomic analysis and comparative multiple sequences of SARS-CoV-2. *Journal of the Chinese Medical Association* 83(6):537–543.

Duffy, S., L. A. Shackelton, and E. C. Holmes. 2008. Rates of evolutionary change in viruses: Patterns and determinants. *Nature Reviews Genetics* 9(4):267–276.

Elena, S. F., and R. Sanjuán. 2007. Virus evolution: Insights from an experimental approach. *Annual Review of Ecology, Evolution, and Systematics* 38(1):27–52.

Ellinghaus, D., F. Degenhardt, L. Bujanda, M. Buti, A. Albillos, P. Invernizzi, J. Fernández, D. Prati, G. Baselli, R. Asselta, M. M. Grimsrud, C. Milani, F. Aziz, J. Kässens, S. May, M. Wendorff, L. Wienbrandt, F. Uellendahl-Werth, T. Zheng, X. Yi, R. de Pablo, A. G. Chercoles, A. Palom, A.-E. Garcia-Fernandez, F. Rodriguez-Frias, A. Zanella, A. Bandera, A. Protti, A. Aghemo, A. Lleo, A. Biondi, A. Caballero-Garralda, A. Gori, A. Tanck, A. Carreras Nolla, A. Latiano, A. L. Fracanzani, A. Peschuck, A. Julià, A. Pesenti, A. Voza, D. Jiménez, B. Mateos, B. Nafria Jimenez, C. Quereda, C. Paccapelo, C. Gassner, C. Angelini, C. Cea, A. Solier, D. Pestaña, E. Muñiz-Diaz, E. Sandoval, E. M. Paraboschi, E. Navas, F. García Sánchez, F. Ceriotti, F. Martinelli-Boneschi, F. Peyvandi, F. Blasi, L. Téllez, A. Blanco-Grau, G. Hemmrich-Stanisak, G. Grasselli, G. Costantino, G. Cardamone, G. Foti, S. Aneli, H. Kurihara, H. ElAbd, I. My, I. Galván-Femenia, J. Martín, J. Erdmann, J. Ferrusquía-Acosta, K. Garcia-Etxebarria, L. Izquierdo-Sanchez, L. R. Bettini, L. Sumoy, L. Terranova, L. Moreira, L. Santoro, L. Scudeller, F. Mesonero, L. Roade, M. C. Rühlemann, M. Schaefer, M. Carrabba, M. Riveiro-Barciela, M. E. Figuera Basso, M. G. Valsecchi, M. Hernandez-Tejero, M. Acosta-Herrera, M. D'Angiò, M. Baldini, M. Cazzaniga, M. Schulzky, M. Cecconi, M. Wittig, M. Ciccarelli, M. Rodríguez-Gandía, M. Bocciolone, M. Miozzo, N. Montano, N. Braun, N. Sacchi, N. Martínez, O. Özer, O. Palmieri, P. Faverio, P. Preatoni, P. Bonfanti, P. Omodei, P. Tentorio, P. Castro, P. M. Rodrigues, A. Blandino Ortiz, R. de Cid, R. Ferrer, R. Gualtierotti, R. Nieto, S. Goerg, S. Badalamenti, S. Marsal, G. Matullo, S. Pelusi, S. Juzenas, S. Aliberti, V. Monzani, V. Moreno, T. Wesse, T. L. Lenz, T. Pumarola, V. Rimoldi, S. Bosari, W. Albrecht, W. Peter, M. Romero-Gómez, M. D'Amato, S. Duga, J. M. Banales, J. R. Hov, T. Folseraas, L. Valenti, A. Franke, and T. H. Karlsen. 2020. Genomewide association study of severe COVID-19 with respiratory failure. *New England Journal of Medicine*. doi: 10.1056/NEJMoa2020283.

Feaster, M., and Y.-Y. Goh. 2020. High proportion of asymptomatic SARS-CoV-2 infections in 9 long-term care facilities, Pasadena, California, USA, April 2020. *Emerging Infectious Diseases* 26(10).

Feldstein, L. R., E. B. Rose, S. M. Horwitz, J. P. Collins, M. M. Newhams, M. B. F. Son, J. W. Newburger, L. C. Kleinman, S. M. Heidemann, A. A. Martin, A. R. Singh, S. Li, K. M. Tarquinio, P. Jaggi, M. E. Oster, S. P. Zackai, J. Gillen, A. J. Ratner, R. F. Walsh, J. C. Fitzgerald, M. A. Keenaghan, H. Alharash, S. Doymaz, K. N. Clouser, J. S. Giuliano, Jr., A. Gupta, R. M. Parker, A. B. Maddux, V. Havalad, S. Ramsingh, H. Bukulmez, T. T. Bradford, L. S. Smith, M. W. Tenforde, C. L. Carroll, B. J. Riggs, S. J. Gertz, A. Daube, A. Lansell, A. Coronado Munoz, C. V. Hobbs, K. L. Marohn, N. B. Halasa, M. M. Patel, and A. G. Randolph. 2020. Multisystem inflammatory syndrome in U.S. children and adolescents. *New England Journal of Medicine* 383:334–346.

Geoghegan, J. L., and E. C. Holmes. 2018. The phylogenomics of evolving virus virulence. *Nature Reviews Genetics* 19(12):756–769.

Goldhill, D. H., and P. E. Turner. 2014. The evolution of life history trade-offs in viruses. *Current Opinion in Virology* 8:79–84.

Gonzalez-Reiche, A. S., M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, and H. van Bakel. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 369(6501):279–301.

Gould, D. W., D. Walker, and P. W. Yoon. 2017. The evolution of biosense: Lessons learned and future directions. *Public Health Reports (Washington, DC: 1974)* 132(1 Suppl):7S–11S.

Grubaugh, N. D., W. P. Hanage, and A. L. Rasmussen. 2020. Making sense of mutation: What D614G means for the COVID-19 pandemic remains unclear. *Cell* 182(4):794–795.

Handel, A., C. Lebarbenchon, D. Stallknecht, and P. Rohani. 2014. Trade-offs between and within scales: Environmental persistence and within-host fitness of avian influenza viruses. *Proceedings of the Royal Society B: Biological Sciences* 281(1787).

Harvala, H., Å. Wiman, A. Wallensten, K. Zakikhany, H. Englund, and M. Brytting. 2015. Role of sequencing the measles virus hemagglutinin gene and hypervariable region in the measles outbreak investigations in Sweden during 2013–2014. *The Journal of Infectious Diseases* 213(4):592–599.

HealthIT.gov. 2020a. *Information blocking*. https://www.healthit.gov/topic/information-blocking (accessed July 7, 2020).

HealthIT.gov. 2020b. *Trusted exchange framework and common agreement*. https://www.healthit.gov/topic/interoperability/trusted-exchange-framework-and-common-agreement (accessed July 7, 2020).

Holmes, E. C., G. Dudas, A. Rambaut, and K. G. Andersen. 2016. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nature* 538(7624):193–200.

Hu, B., L.-P. Zeng, X.-L. Yang, X.-Y. Ge, W. Zhang, B. Li, J.-Z. Xie, X.-R. Shen, Y.-Z. Zhang, N. Wang, D.-S. Luo, X.-S. Zheng, M.-N. Wang, P. Daszak, L.-F. Wang, J. Cui, and Z.-L. Shi. 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLOS Pathogens* 13(11):e1006698.

Hu, J., C.-L. He, Q.-Z. Gao, G.-J. Zhang, X.-X. Cao, Q.-X. Long, H.-J. Deng, L.-Y. Huang, J. Chen, K. Wang, N. Tang, and A.-L. Huang. 2020. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity and decreases neutralization sensitivity to individual convalescent sera. *bioRxiv.* https://doi.org/10.1101/2020.06.20.161323.

Johnson, C., Y. Pylypchuk, and V. Patel. 2018. *Methods used to enable interoperability among U.S. non-federal acute care hospitals in 2017*. The Office of the National Coordinator for Health Information Technology.

Korber, B., W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, and M. D. Wyles. 2020. Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182(4):812–827.

Lang, M., A. Som, D. P. Mendoza, E. J. Flores, N. Reid, D. Carey, M. D. Li, A. Witkin, J. M. Rodriguez-Lopez, J. O. Shepard, and B. P. Little. 2020. Hypoxaemia related to COVID-19: Vascular and perfusion abnormalities on dual-energy ct. *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(20)30367-4.

Lesho, E., R. Clifford, F. Onmus-Leone, L. Appalla, E. Snesrud, Y. Kwak, A. Ong, R. Maybank, P. Waterman, P. Rohrbeck, M. Julius, A. Roth, J. Martinez, L. Nielsen, E. Steele, P. McGann, and M. Hinkle. 2016. The challenges of implementing next generation sequencing across a large healthcare system, and the molecular epidemiology and antibiotic susceptibilities of carbapenemase-producing bacteria in the healthcare system of the U.S. Department of Defense. *PLOS ONE* 11(5):e0155770.

Li, K. 2016. *Dr. Rumi Chunara and Sofia Ahsanuddin: The GoViral Study*. https://www.ghjournal.org/the-goviral-study (accessed July 6, 2020).

Liu, Y., Z. Ning, Y. Chen, M. Guo, Y. Liu, N. K. Gali, L. Sun, Y. Duan, J. Cai, D. Westerdahl, X. Liu, K. Xu, K.-f. Ho, H. Kan, Q. Fu, and K. Lan. 2020. Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature* 582(7813):557–560.

Liya, G., W. Yuguang, L. Jian, Y. Huaiping, H. Xue, H. Jianwei, M. Jiaju, L. Youran, M. Chen, and J. Yiqing. 2020. Studies on viral pneumonia related to novel coronavirus SARS-CoV-2, SARS-CoV, and MERS-CoV: A literature review. *Apmis* 128(6):423–432.

MacLaren, G., D. Fisher, and D. Brodie. 2020. Preparing for the most critically ill patients with COVID-19: The potential role of extracorporeal membrane oxygenation. *JAMA* 323(13):1245–1246.

Maxmen, A. 2020. Thousands of coronavirus tests are going unused in US labs. *Nature* 580(7803):312–313.

Messenger, S. L., I. J. Molineux, and J. J. Bull. 1999. Virulence evolution in a virus obeys a trade off. *Proceedings of the Royal Society B: Biological Sciences* 266(1417):397–404.

N3C (National COVID Cohort Collaborative). 2020. *National COVID Cohort Collaborative (N3C): A national resource for shared analytics*. https://ncats.nih.gov/n3c/about (accessed August 20, 2020).

Ogbunugafor, C., B. W. Alto, T. M. Overton, A. Bhushan, N. M. Morales, and P. E. Turner. 2013. Evolution of increased survival in RNA viruses specialized on cancer-derived cells. *The American Naturalist* 181(5):585–595.

Ong, J., B. E. Young, and S. Ong. 2020. COVID-19 in gastroenterology: A clinical perspective. *Gut* 69(6):1144–1145.

Orr, H. A. 2000. The rate of adaptation in asexuals. *Genetics* 155(2):961–968.

PCAST (President's Council of Advisors on Science and Technology). 2014. *Report to the President on combatting antibiotic resistance.* Washington, DC.

Penedos, A. R., R. Myers, B. Hadef, F. Aladin, and K. E. Brown. 2015. Assessment of the utility of whole genome sequencing of measles virus in the characterisation of outbreaks. *PLOS ONE* 10(11):e0143081.

Premkumar, L., B. Segovia-Chumbez, R. Jadi, D. R. Martinez, R. Raut, A. J. Markmann, C. Cornaby, L. Bartelt, S. Weiss, Y. Park, C. E. Edwards, E. Weimer, E. M. Scherer, N. Rouphael, S. Edupuganti, D. Weiskopf, L. V. Tse, Y. J. Hou, D. Margolis, A. Sette, M. H. Collins, J. Schmitz, R. S. Baric, and A. M. de Silva. 2020. The receptor-binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Science Immunology* 5(48):eabc8413.

Rothstein, M., and S. Tovino. 2019. Privacy risks of interoperable health records: Segmentation of sensitive information will help. *Journal of Law, Medicine & Ethics* 47:771–777.

Smolinski, M. S., A. W. Crawley, J. M. Olsen, T. Jayaraman, and M. Libel. 2017. Participatory disease surveillance: Engaging communities directly in reporting, monitoring, and responding to health threats. *JMIR Public Health and Surveillance* 3(4):e62.

Van Cauteren, D., S. Vaux, H. de Valk, Y. Le Strat, V. Vaillant, and D. Lévy-Bruhl. 2012. Burden of influenza, healthcare seeking behaviour and hygiene measures during the A(H1N1)2009 pandemic in France: A population based study. *BMC Public Health* 12(1):947.

van Doremalen, N., T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, S. I. Gerber, J. O. Lloyd-Smith, E. de Wit, and V. J. Munster. 2020. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England Journal of Medicine* 382(16):1564–1567.

van Noort, S. P., M. Muehlen, H. Rebelo de Andrade, C. Koppeschaar, J. M. Lima Lourenço, and M. G. Gomes. 2007. GripeNet: An Internet-based system to monitor influenza-like illness uniformly across Europe. *Eurosurveillance* 12(7):5–6.

Wölfel, R., V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, and C. Wendtner. 2020. Virological assessment of hospitalized patients with COVID-2019. *Nature* 581(7809):465–469.

Wood, H. 2020. New insights into the neurological effects of COVID-19. *Nature Reviews Neurology* 16(8):403.

Xu, Y., X. Li, B. Zhu, H. Liang, C. Fang, Y. Gong, Q. Guo, X. Sun, D. Zhao, J. Shen, H. Zhang, H. Liu, H. Xia, J. Tang, K. Zhang, and S. Gong. 2020. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nature Medicine* 26(4):502–505.

Zhang, L., C. B. Jackson, H. Mou, A. Ojha, E. S. Rangarajan, T. Izard, M. Farzan, and H. Choe. 2020. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv.* https://doi.org/10.1101/2020.06.12.148726.

# 5

# Governance and Regulatory Considerations

In the United States, federal or state laws do not protect or mandate the sharing of viral sequence data or virus samples—hence, any sharing of such data and samples is done voluntarily and without concerns about possible regulatory barriers. In contrast, there are federal and state laws protecting clinical and epidemiological data. This chapter addresses governance and regulatory considerations related to these data, both in the United States and globally.

U.S. federal law does not currently require the sharing of clinical and epidemiological data with federal agencies to fight the pandemic. This is a consequence of the United States as a federal system, with primary public health legal authority inherent to states' police powers. Depending on the jurisdiction, state law may require certain specified entities—such as health care professionals and laboratories—to report certain diseases to public health agencies and the form of the reports. While all states and territories regularly share aggregate data with the U.S. Centers for Disease Control and Prevention (CDC) for national public health surveillance purposes, sharing such data is voluntary and is typically limited to epidemiological data, not viral sequence data.

Consequently, absent a state law mandate, such sharing takes place voluntarily. Without funding to support the costs associated with gathering and preparing data to be disclosed to others (e.g., public health authorities and researchers), and without clear regulatory pathways and established infrastructure to support sharing, the sharing of data and biosamples is suboptimal. Additional potential barriers exist when obtaining or sharing information or biosamples from—and with—international sources.

Entities often cite privacy laws as a barrier to the sharing of data and biospecimens. However, as explained below, federal law expressly permits the sharing of data and biospecimens for public health and research purposes. Nevertheless, misconceptions and confusion about the law, hesitation to seek legal advice due to perceived or actual time or financial constraints, as well as general risk aversion, can create obstacles to sharing. Clear guidance on what federal law permits with respect to sharing, and examples of how sharing of data and biospecimens to fight coronavirus disease 2019 (COVID-19) in the United States is already occurring pursuant to these existing legal pathways can help eliminate those obstacles.

In addition to clarity, removing barriers to rapid and comprehensive clinical, epidemiological, and sequence data and biosamples will be dependent on national-level leadership and governance. This will likely require national-level leadership and planning to create supportive legal or strategic frameworks that instill principles of good governance, including accountability (clarifying authorities and responsibilities), transparency, equity, participation, and clear legal protections for individuals' rights, including privacy and non-discrimination, and certainty as to their scope. This challenge is not only one faced by the United States. As COVID-19 has demonstrated, the rapid and comprehensive sharing of clinical, epidemiological, and sequence data is vital to global response.

## FEDERALISM BARRIERS AND OPPORTUNITIES

The need for rapid and comprehensive viral sequence sharing may also serve as an opportunity for addressing barriers inherent in the federal system to data sharing. Given the national and interstate threat posed by a pandemic like COVID-19, there is an important rationale for shifting sharing from purely intrastate and voluntary to federal authorities, to interstate and federal data sharing, either through formalized voluntary agreements contingent on state consent (similar to arrangements in other federations), or through the U.S. Congress passing legislation that can be appropriately fixed under an enumerated federal head of power. Alternatively, this may be facilitated by terms of federal funding requirements for needed systems and infrastructure for data sharing or federal data sharing floor preemption laws, limited to the period during a public health emergency (explored in greater detail in the section on governance below).

## INTERNATIONAL SHARING BARRIERS

The International Health Regulations (IHR, 2005) is a legally binding treaty that sets out countries' obligations for preventing, detecting, and responding to international public health threats. Under the IHR, countries have the obligation to share with the World Health Organization (WHO) "timely, accurate, and sufficiently detailed public health information" including "case definitions, laboratory results, source and type of the risk, number of cases and deaths, [and] conditions affecting the spread of the disease" about potential public health emergencies of international concern (WHO, 2005). While not expressly required, in practice countries may interpret public health information to include viral genome sequences. At minimum, and as demonstrated in the first 2 weeks of January during the emergence of COVID-19, there is a normative expectation that countries share the genetic sequence of pathogens that may constitute a potential public health emergency of international concern (Rourke et al., 2020). The United States is a State Party to the IHR; however, it adopted a reservation that its obligations under the IHR are subject to the limitations of federalism, where obligations fall within state jurisdictions (Rourke et al., 2020). However, under international law this is unlikely to displace the federal government's obligations under the IHR.

While the IHR do not currently contain an express obligation to share virus genome sequences, likely reforms of the IHR following the COVID-19 pandemic may serve as an opportunity for crystalizing potential obligations to share virus genome sequence data among countries and WHO (Rourke et al., 2020). In addition, ongoing negotiations in other international forums may have future implications for viral genome sharing, namely discussions at the Meeting of Parties for the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization* (2010).[1] Researchers should be aware that more than 30 countries have implemented access and benefit-sharing legislation, which may include some countries' requirements to obtain prior informed consent and negotiate mutually agreed terms for benefit sharing in exchange for access to pathogen samples, and may increasingly require similar arrangements for sequence data sharing (WHO, 2020).

Under the IHR, a country is required to keep confidential and process anonymously any information it collects or receives from another country or WHO, which refers to an identified or identifiable person, as required by national laws. The IHR permit countries to disclose and process personal data when it is essential for "assessing and managing" a public health risk, but countries must ensure that any personal data are processed fairly, lawfully, and consistently with a public health purpose; adequate, relevant, and not excessive; accurate; and not retained longer than necessary for the public health purpose. The IHR, including obligations to conduct surveillance, share information, and implement control measures, must also be

[1] See https://www.cbd.int/abs/about (accessed June 25, 2020).

implemented in a manner consistent with human rights, which indirectly include the right to privacy and non-discrimination.

## PERCEIVED VERSUS ACTUAL DOMESTIC LEGAL BARRIERS

While there is no federal obligation to share virus genome sequences, there is a range of perceived and actual legal barriers under U.S. domestic law regarding clinical and epidemiological data sharing. Clarifying the application of these laws and communication with potential stakeholders (including public health departments and researchers) should be a priority. Below, the committee examines two federal laws commonly cited as barriers to data sharing—the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule—and clarifies the scope of their application as it relates to viral sequence sharing.

## THE HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT (HIPAA)

The regulations under HIPAA apply to many of the sources of data identified in this report. For example, HIPAA "covered entities" include most physician practices, hospitals, clinics, clinical laboratories, pharmacies, and health plans (CMS, 2020).[2] HIPAA also covers contractors to those covered entities (also known as "business associates") who also may be potential sources of data identified in this report (HHS, 2020d). Examples of potential business associate sources of clinical and epidemiological data are electronic medical record vendors and health information exchanges.

The HIPAA Privacy Rule establishes rules for the use and disclosure of identifiable health information (known as protected health information [PHI]). (Of note: HIPAA does not govern biospecimens; however, HIPAA would cover any PHI associated with such biospecimens.) The HIPAA Security Rule requires that any digital PHI be stored and transmitted securely, among other security safeguards; the Privacy Rule calls for the adoption of reasonable safeguards to protect PHI not subject to the Security Rule (e.g., paper PHI).

### Disclosures to Public Health Authorities

The rules expressly permit the disclosure of PHI—identifiable information—to public health authorities (or contractors working on their behalf), without the need to first obtain the consent or authorization of the

[2] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 160.103.

data subject. Specifically, covered entities may use or disclose PHI to "public health authorities authorized by law to collect or receive such information for preventing or controlling disease, injury, or disability, including but not limited to, ... the conduct of public health surveillance, public health investigations, and public health interventions."[3] When the public health authority is also a HIPAA-covered entity, the authority is expressly permitted to use PHI for the above purposes.[4] Ordinarily, business associates may similarly disclose PHI to public health authorities if their contracts with covered entities (known as business associate agreements) permit them to do so; however, the U.S. Department of Health and Human Services' (HHS's) Office for Civil Rights (OCR) recently exercised enforcement discretion to enable business associates to disclose information related to COVID-19 for public health purposes notwithstanding conflicting or unclear terms in their business associate agreements (HHS, 2020c). This announcement from OCR led to a private-sector effort to facilitate the collection and reporting of COVID-19-relevant data to public health authorities by a type of business associate—health information exchanges (HIEs)—across the country (HIEs across the nation largely facilitate the exchange of clinical information among health care providers for treatment purposes) (McClellan and Mostashari, 2020).

Uses and disclosures of PHI to public health authorities must meet the HIPAA Privacy Rule's "minimum necessary" standard. Covered entities are required to make "reasonable efforts" to use, disclose, or request "only the minimum amount of PHI needed to accomplish the intended purpose of the use, disclosure, or request."[5] Covered entities must develop policies to define what constitutes the minimum necessary for routine disclosures, or requests for information; however, entities are permitted to reasonably rely on requests from public officials as meeting the minimum necessary standard. Consequently, if public health authorities request PHI for COVID-19 purposes, covered entities (or business associates) may disclose PHI consistent with those requests.

Although HIPAA expressly permits the disclosure of identifiable PHI to public health authorities, many entities still feel more comfortable disclosing information stripped of identifiers out of an abundance of caution (or potentially due to a misunderstanding of what HIPAA allows). The HIPAA Privacy Rule also permits the use and disclosure of a limited dataset, which is still PHI but has been rendered less identifiable by the removal of 16 com-

[3] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. CFR § 164.512(b)(1).

[4] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.512(b)(2).

[5] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.502(b) and 164.514(d).

mon identifiers, for public health purposes.[6] The recipient must enter into a data use agreement that establishes the permitted uses and disclosures of the information and prohibits re-identification or contact of individuals in the dataset. The committee heard testimony of one example of sharing of data for COVID-19 efforts using a HIPAA limited dataset.

Data that have been "de-identified" in accordance with the HIPAA Privacy Rule are no longer covered by HIPAA and can be used and disclosed without limitation. As a result, entities often seek to meet the HIPAA de-identification standards prior to disclosing data, even to public health authorities. HIPAA establishes two methodologies for de-identifying PHI. The safe harbor methodology requires the removal of 18 categories of identifiers and no actual knowledge that the recipient of the data can re-identify it. The statistical or expert methodology requires someone with statistical expertise to determine that the dataset, in the hands of the intended recipient, is at very low risk of re-identification. De-identification per either methodology does not require execution of a data use agreement.

The committee heard testimony regarding the challenges of reliably linking data across multiple sources when the data are de-identified using the safe harbor methodology. Because the safe harbor requires the removal of fields that can be useful for linking data, this suggests that the disclosure of limited datasets, or identifiable PHI, may be more effective for sharing data to combat COVID-19. Also, linking of data, de-identified using statistician or expert methodology, may facilitate linking (Datavant, 2018).

## Disclosures Required by Law

HIPAA also expressly permits PHI to be disclosed where such disclosure is required by another state or federal law.[7] For example, if a state or locality were to enact an order mandating disclosure of information related to severe acute respiratory coronavirus 2 (SARS-CoV-2), a covered entity (or business associate) could make such a disclosure to the extent consistent with such law. The minimum necessary standard does not apply to such disclosures;[8] as a result, entities may disclose in accordance with legal mandates without needing to consider whether such disclosure meets the minimum necessary standard.

[6] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.514(e).

[7] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.512(a).

[8] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.514(d).

### Uses and Disclosures for Research

HIPAA permits PHI to be used or disclosed for "research" purposes.[9] "Research" is defined as a "systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge."[10] Disclosures of PHI for research typically require prior authorization of the individual; however, this requirement can be waived or altered by a Privacy Board or an Institutional Review Board (IRB) if the following conditions are met:

- The use or disclosure of PHI involves no more than minimal risk to individual privacy;
- The research could not practicably be conducted without the waiver or alteration; and
- The research could not practicably be conducted without access to and use of PHI.[11]

HIPAA also permits limited datasets (see above) to be used or disclosed for research purposes, as long as the researcher has executed the required data use agreement. Similarly, data de-identified per HIPAA's standards can be used or disclosed for any purpose, including research. The same issues identified above regarding linking de-identified data across data sources could apply here as well.

### Distinction Between Public Health and Research

Although HIPAA defines what constitutes a "research" use or disclosure, HIPAA does not define what constitutes a "public health" use or disclosure. This could cause obstacles due to confusion over which rules to follow in disclosing data related to COVID-19 and SARS-CoV-2. However, OCR, which has oversight over the HIPAA privacy and security regulations, has issued guidance making clear that any disclosures to public health authorities (which includes the National Institutes of Health)—whether for public health practice or for research purposes—would be permitted under the Privacy Rule provisions governing disclosures to public health authorities (HHS, 2020a). As a result, HIPAA's research provisions would govern uses or disclosures of information of COVID-19-related data for analytics purposes when those disclosures are for "generalizable knowledge" and

[9] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.512(i).

[10] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.501.

[11] Health Insurance Portability and Accountability Act of 1996, H.R. 3103, 110 Stat. 1936, 104 P.L. 191. 45 CFR § 164.512(i)(2)(ii).

to entities *other than public health authorities* or contractors working on their behalf.

## COMMON RULE

The federal rules governing research on human subjects—otherwise known as the Common Rule—apply to research using identifiable biospecimens and data.[12] The Common Rule applies to federally funded human subjects research but also governs health care entities that receive federal funding for some research and have agreed to follow the Common Rule for all research on human subjects, regardless of funding source (HHS, 2020b).

The Common Rule applies only to research—defined similarly as that term is defined in HIPAA; it does not apply to uses and disclosures of data to public health authorities (or their contractors) for "public health surveillance."[13] Unlike with respect to HIPAA, the definition of "public health surveillance" is not interpreted to mean any activity conducted by a public health authority, including research activities. However, the definition of public health surveillance is quite broad and arguably supports the collection and analysis of data and biospecimens by public health authorities. Public health surveillance includes

> The collection and testing of information or biospecimens, conducted, supported, requested, ordered, required, or authorized by a public health authority. Such activities are limited to those necessary to allow a public health authority to identify, monitor, assess, or investigate potential public health signals, onsets of disease outbreaks, or conditions of public health importance (including trends, signals, risk factors, patterns in diseases, or increases in injuries from using consumer products). Such activities include those associated with providing timely situational awareness and priority setting during the course of an event or crisis that threatens public health (including natural or man-made disasters).

To the extent it is necessary for data sources or recipients to distinguish between public health surveillance or practice and research for purposes of compliance with law, resources exist to help inform this distinction (Hodge and Gostin, 2004).

The Common Rule requires that research using identifiable biospecimens and data be approved by an IRB. In general, the Common Rule also requires prior informed consent for research uses of identifiable biospecimens and data—but the requirement can be waived or altered by an IRB using criteria similar (but not the same as) those for HIPAA:

[12] The Common Rule. 45 CFR § 46.

[13] The Common Rule. 45 CFR § 46.102(1)(2).

- The research involves no more than minimal risk to the participants;
- The research could not practicably be carried out without the waiver or alteration;
- If the research involves identifiable information, the research could not practicably be carried out without the identifiable information;
- The waiver will not adversely affect the rights and welfare of the participants; and
- Whenever appropriate, the participants will be provided with additional pertinent information after participating.[14]

Under the Common Rule certain types of secondary research uses of identifiable biospecimens and data are exempt from the requirements of the Common Rule (although an IRB determines whether the conditions for these exemptions have been met). Secondary research involves research on biospecimens or data that were originally collected or generated for a non-research purpose—such as for clinical purposes and/or for reporting to public health (the information collected to study SARS-CoV-2 is likely to all be secondary data, not generated solely for purposes of SARS-CoV-2 research). As a result, these exemptions may be particularly useful for conducting SARS-CoV-2 and COVID-19 research. For example, secondary research uses of identifiable information or biospecimens are exempt if:

> The information (including information about biospecimens) is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify the subjects.[15]

Also exempt is research using identifiable information that is governed by HIPAA (e.g., research conducted by a covered entity).[16]

The Common Rule was recently revised to include a new pair of exemptions that allow for the creation of research-ready databases of identifiable biospecimens and information under a broad consent; then subsequent research uses would need to be submitted for IRB review but would not require re-consent of the participants.[17] The researcher must obtain a limited IRB determination that:

- The researcher has obtained broad consent meeting new Common Rule requirements;

[14] The Common Rule. 45 CFR § 46.117.

[15] The Common Rule. 45 CFR § 46.111(4)(ii).

[16] The Common Rule. 45 CFR § 46.111(4)(iii).

[17] The Common Rule. 45 CFR § 46.111(7)–(8).

- Such consent is documented (or the need for documentation is waived by the IRB); and
- If there is a change made in the way the identifiable information is stored and maintained, there are adequate measures taken to protect participants' privacy and the confidentiality of the data.[18]

The broad consent for storing and maintaining identifiable information or biospecimens for secondary research must meet the following requirements:

- Provides a general description of the types of research that may be conducted with data or biospecimens from the participant (which must be sufficient that a reasonable person would expect the types of research to be conducted with the data);
- Describes the information or biospecimens that might be used in research, whether the information or biospecimens might be shared for research purposes, and the types of institutions or researchers who might conduct the research;
- Describes the period of time the database is to be maintained;
- Includes a statement that the participant will not be informed of the details of any specific research studies using his or her data or biospecimens, including purposes that he or she might not have chosen to consent to if he or she had the option to do so;
- Includes a statement that no clinically relevant research results will be shared with the participant unless it is certain that such results will always be shared;
- Includes a statement that participation in the database is voluntary and that there is no penalty or loss of benefits for refusal to participate (and that the individual can discontinue participation prospectively at any time);
- Includes disclosure of any reasonably foreseeable risks or discomforts, as well as disclosure of any benefits to the subject or others;
- Includes the extent to which confidentiality of records will be maintained; and
- Includes information about whom to contact with further questions or if the participant thinks he or she may have suffered a research-related injury.[19]

If identifiable biospecimens are collected, the broad consent must include the following (if appropriate):

[18] The Common Rule. 45 CFR § 46.111(a)(8).

[19] The Common Rule. 45 CFR § 46.116(d)(1).

- A statement that the participant's biospecimens (even if identifiers are removed) may be used for commercial profit and whether the participant will or will not share in this profit; and
- Whether the research might include whole genome sequencing (i.e., sequencing of a human germline or somatic specimen with the intent to generate the genome or exome sequence of that specimen). Note that this requirement refers to the sequencing of human DNA, not the sequencing of viral genetic material.

Once these data are stored and maintained pursuant to the above exemption, subsequent research uses of that data are also exempt from the Common Rule, so long as an IRB, under limited review, determines that the research to be conducted is within the scope of the broad consent obtained for storage and maintenance (or approves a waiver or alteration of consent), and the investigator does not include return of research results in the study plan (although such individual results can be returned where required by law).[20]

Finally, data and biospecimens that are not identifiable—for example, where the identity of the participant cannot be readily ascertained—are not considered to be human subjects research under the Common Rule.[21] The Common Rule did not adopt the HIPAA de-identification standards; consequently, there is often some question about when data and biospecimens are not considered to be subject to the Common Rule. Researchers will typically seek a determination from an IRB regarding whether research is not covered by the Common Rule. (Of note: the Common Rule calls for federal departments and agencies covered by the Common Rule to reexamine the meaning of identifiable information and identifiable biospecimens,[22] but no additional guidance has been issued.)

### Atypical Data Sources and State Law

Other sources of information on SARS-CoV-2 and COVID-19 may be useful sources of data, albeit less likely for sources of genomic information on SARS-CoV-2. The ability of these other data sources (e.g., private laboratories not subject to HIPAA or applications used by consumers) to share information for either public health or research purposes is likely to depend on the privacy policies or terms of use for these sources. State laws may also govern the ability of these atypical data sources—as well as sources covered by HIPAA and the Common Rule—to share data. State health privacy laws typically cover more sensitive types of information such as mental health, human genome, substance abuse treatment, and repro-

[20] The Common Rule. 45 CFR § 46.111(8).

[21] The Common Rule. 45 CFR § 46.101–102.

[22] The Common Rule. 45 CFR § 46.102(7).

ductive and sexual health data—information that may be less important to studying SARS-CoV-2 and COVID-19 (of note: federal law provides heightened privacy protections for identifiable substance abuse treatment information when maintained by federally supported substance abuse treatment programs), requiring informed consent prior to disclosure for most purposes.[23] California's new privacy law—the California Consumer Privacy Act[24]—covers a broad scope of information (not just particularly sensitive types of data); however, its applicability is limited to for-profit companies that meet certain thresholds for data collection or monetization (Gold and Hennessey, 2019; Wolf et al., 2019).

## GOVERNANCE

Sharing of data and viral genome sequences is crucial during a national public health emergency. There is a critical role for the federal government to play in coordinating and leading the sharing of such data between states and to the federal government. Given the interstate threat posed by a national public health emergency, this could serve as a moment to shift sharing from purely intrastate and voluntary to federal authorities, to interstate and federal sharing—through state consent and agreement (similar to arrangements in other federations like Australia and Canada) or if such an arrangement would appropriately fit under the scope of constitutionally granted federal authorities. An example potential trigger for this shift is the declaration of a federal public health emergency by the Secretary of HHS under 42 USC § 319, which waives certain other laws as the U.S. Congress determines to enable a national response to a public health emergency. These data-sharing and reporting processes should be clearly established and resourced as an urgent matter, and prior to an emergency. Without a clear and urgent public health rationale, changing reporting processes during an emergency should be avoided, and emergencies should not justify not complying with principles of good governance, including data transparency. Current examples include the release of goods from the national stockpile, permitting emergency use authorizations for medications, or waivers of Medicaid requirements. This would require the U.S. Congress to adopt legislation enabling this and could be an opportunity to provide a legislative framework, cognizant of states' public health powers and the constitutional limits on the federal government's law-making powers, that streamlines data sharing.

The sharing of viral sequence data and associated information should be guided by national-level leadership to create supportive legal or strategic

[23] Confidentiality of Substance Use Disorder Patient Records. 42 CFR § 2.

[24] See https://oag.ca.gov/privacy/ccpa (accessed June 25, 2020).

frameworks that instill principles of good governance. This includes embedding clear accountability processes that clarify authorities and responsibilities; the principles of transparency, equity, and participation; and clear and certain legal protections for public health agencies, researchers, and individuals' rights.

**RECOMMENDATION 3. The U.S. Department of Health and Human Services should establish an effective and sustainable science-driven leadership and governance structure for the use of SARS-CoV-2 genome sequences in addressing critical national public health and basic science issues, develop a national strategy, and ensure the funding needed for successful execution of the strategy.**

- **Leaders of this effort must have sufficient authorities and responsibilities to ensure that key issues are identified and prioritized, representative data are generated, and barriers to data sharing are diminished.**
- **A national strategy for SARS-CoV-2 genome sequences linked to clinical and epidemiological data should be developed that articulates goals, priorities, and a path for achieving them.**
- **A board with diverse relevant expertise should be established with broad authority to oversee and advise the national strategy for SARS-CoV-2 genome sequences linked to clinical and epidemiological data, and the delivery of actionable data for related investigations.**

## CONCLUDING REMARKS

Although federal law permits the sources of data and biospecimens to disclose these materials to public health authorities, and to make them available for research—often without the need to obtain an individual's consent—confusion about the law, and conservative interpretations due to fear of running afoul of the law, translate into genuine obstacles to sharing. This includes how shared data, or public health research findings from shared data, are managed, reported, and described. This is particularly relevant where such information may contribute to stigma or discrimination against individuals (such as when a case is the first identified in an outbreak) or groups of individuals (such as individuals from one geographic location or population group).

The committee heard testimony about the reluctance to collect "identifiable" data and biospecimens due to concerns about being responsible for stewarding such highly sensitive data. At the same time, the committee also heard testimony about the challenges of linking data across data silos due to lack of a universal patient identifier and because methods to de-identify data (e.g., the HIPAA safe harbor method) can impact data utility and also create obstacles to linkage. In a pandemic, when time is of the essence, uncertainty provides an unacceptable drag on our national response.

As noted above, federal law already provides HHS with authority to waive certain aspects of HIPAA when the Secretary of HHS declares a public health emergency (ASPR, 2019a). However, the HIPAA provisions that are permitted to be waived largely impact the delivery of clinical care; waiver of such provisions will not create more efficient reporting of data or sharing of biospecimens. In response to the limitations on its waiver authority, HHS has instead selectively issued enforcement discretion to facilitate greater data sharing to fight the pandemic; for example, OCR issued guidance to permit HIPAA business associates to share PHI with public health authorities, notwithstanding any conflicting terms in their business associate agreements (HHS, 2020c).

Detailed guidance from HHS—both with respect to the application of HIPAA and the Common Rule—has the potential to significantly improve sharing of data and biospecimens. Official communications from regulators can significantly reduce uncertainty about how such sharing can occur. Ideally, such guidance should be detailed, describing clear pathways for data and biospecimen sharing and providing illustrative examples, and be updated promptly over time to respond to new questions and concerns. In the absence of the U.S. Congress granting broader waiver authority to HHS, the department could use its enforcement discretion more broadly to cut through red tape to achieve greater sharing of clinical data and biospecimens to power the national response to threats posed by SARS-CoV-2 and COVID-19 in the short term—and to establish pathways that enable the sharing of biospecimens and data to combat the next national infectious disease threat. Such enforcement discretion should have a clear trigger—and an end point.

For example, upon a triggering event such as the declaration of a national public health emergency by the Secretary of HHS, sharing of a limited dataset of PHI with public health authorities or their designees, by covered entities or business associates, could be expressly deemed to be in compliance with the HIPAA Privacy Rule, either without the need for a data use agreement or upon execution of a single standard data use agreement provided by HHS, and without regard to contrary provisions in a business associate agreement. Such disclosures need to be required to be done securely, including by leveraging existing secure pathways for public health and clinical data reporting (e.g., HIEs and certified electronic medical records). To assure the data that are shared are sufficient (and to meet minimum necessary determinations), CDC could publish from time to time the minimum datasets to be shared—and OCR should deem these datasets to constitute "minimum necessary" for purposes of the Privacy Rule.

Similarly, the HHS Office for Human Research Protections could issue written guidance to confirm such reporting and disclosures of biospecimens or viral sequence data to public health authorities (or their designees) pursuant to a declared public health emergency to be not human subjects research (and thus not covered by the Common Rule). These are but two examples; HHS could work with SARS-CoV-2 experts to determine needed sharing pathways and issue guidance that facilitates such sharing.

Public health authorities (or contractors acting on their behalf) could subsequently make available de-identified (per HIPAA standards) data for the further conduct of research by public and private entities into SARS-CoV-2 and COVID-19. To enable such data to be linkable across data silos, HHS could work with the private sector to quickly disseminate best practices on linking data in privacy protective ways, such as through the use of one-way hashing techniques.

To assure that such data are handled in ways that can be trusted by the public, public health authorities should commit to abiding by applicable privacy, confidentiality, security, non-discrimination, and civil rights laws, and be transparent with the public about their data collection initiatives (while still being allowed to take advantage of the waiver of the Paperwork Reduction Act requirements that is frequently granted during national emergencies) (ASPR, 2019b). There is also an important opportunity to assess whether other existing federal laws insufficiently enable viral sequence sharing with a federal authority, and appropriate voluntary agreements between states and the federal government, federal funding requirements, or federal data sharing floor preemption laws, during a public health emergency.

## REFERENCES

ASPR (Office of the Assistant Secretary for Preparedness and Response). 2019a. *1135 waivers*. https://www.phe.gov/Preparedness/legal/Pages/1135-waivers.aspx (accessed July 7, 2020).

ASPR. 2019b. *Public health emergency declaration*. https://www.phe.gov/Preparedness/legal/Pages/phedeclaration.aspx (accessed July 7, 2020).

CMS (Centers for Medicare & Medicaid Services). 2020. *Are you a covered entity?* https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/HIPAA-ACA/AreYouaCoveredEntity (accessed July 7, 2020).

Datavant. 2018. *Overview of Datavant's de-identification and linking technology for structured data*. Datavant. https://datavant.com/wp-content/uploads/2018/05/Datavant_De-Identifying-and-Linking-Structured-Data-Whitepaper.pdf (accessed August 20, 2020).

Gold, K., and J. Hennessey. 2019. *CCPA: What health care, biotech and life sciences companies should know now*. https://iapp.org/news/a/ccpa-round-up-what-health-care-biotech-and-life-sciences-companies-should-know-now (accessed July 7, 2020).

HHS (U.S. Department of Health and Human Services). 2020a. *Does the HIPAA Privacy Rule's public health provision permit covered entities to disclose protected health information to authorities such as the National Institutes of Health (NIH)?* https://www.hhs.gov/hipaa/for-professionals/faq/297/does-the-hipaa-public-health-provision-permit-covered-entities-to-disclose-information-to-authorities/index.html (accessed July 7, 2020).

HHS. 2020b. *Human subject regulations decision charts*. https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts/index.html (accessed July 7, 2020).

HHS. 2020c. *OCR announces notification of enforcement discretion to allow uses and disclosures of protected health information by business associates for public health and health oversight activities during the COVID-19 nationwide public health emergency*. https://www.hhs.gov/about/news/2020/04/02/ocr-announces-notification-of-enforcement-discretion.html (accessed July 7, 2020).

HHS. 2020d. *Summary of the HIPAA Privacy Rule*. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (accessed July 7, 2020).

Hodge, J., and L. Gostin. 2004. *Public health practice vs. research*. Atlanta, GA: Council of State and Territorial Epidemiologists.

McClellan, M., and F. Mostashari. 2020. *Data interoperability and exchange to support COVID-19 containment*. Washington, DC: Duke-Margolis Center for Health Policy.

Rourke, M., M. Eccleston-Turner, A. Phelan, and L. Gostin. 2020. Policy opportunities to enhance sharing for pandemic research. *Science* 368(6492):716–718.

WHO (World Health Organization). 2005. *International health regulations*. Geneva, Switzerland: WHO Press.

WHO. 2020. *Pandemic influenza preparedness (PIP) framework: Draft report on Decision WHA72 (12) 1(b)*. Geneva, Switzerland: World Health Organization.

Wolf, L., E. Brown, R. Kerr, G. Razick, G. Tanner, B. Duvall, S. Jones, J. Brackney, and T. Posada. 2019. *The web of legal protections for participants in genomic research*. Health Matrix: Journal of Law-Medicine. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3328892 (accessed July 27, 2020).

# Appendix A

# Committee Biosketches

**Diane Griffin, M.D., Ph.D.** (*Chair*), is a University Distinguished Service Professor and the Alfred and Jill Sommer Chair of the W. Harry Feinstone Department of Molecular Microbiology and Immunology at the Johns Hopkins Bloomberg School of Public Health. Dr. Griffin is a virologist recognized for her work on the pathogenesis of viral infections. She is known particularly for her studies on measles and alphavirus encephalomyelitis that have delineated the role of the immune response in virus clearance, vaccine-induced protection from infection, tissue damage, and immune suppression. Dr. Griffin was born in Iowa City, Iowa, and grew up in Oklahoma City, Oklahoma. She graduated from Augustana College, Rock Island, Illinois, with a degree in biology and from the Stanford University School of Medicine in 1968 with a Ph.D. in immunology and an M.D., followed by a residency in internal medicine. She was a postdoctoral fellow in virology and infectious diseases at the Johns Hopkins University School of Medicine and joined the faculty in 1974. She has been the president of the American Society for Virology and the American Society for Microbiology and is a member of both the National Academy of Sciences and the National Academy of Medicine.

**Ralph Baric, Ph.D.,** is the William R. Kenan, Jr. Distinguished Professor in the Department of Epidemiology and a professor in the Department of Microbiology and Immunology. He is a Harvey Weaver Scholar from the National Multiple Sclerosis Society and an Established Investigator Awardee from the American Heart Association. In addition, he is a World Technology Award Finalist and a fellow of the American Association for Microbiology. He has spent the past three decades as a world leader in the study of coronaviruses and is single-handedly responsible for the University of North Carolina at Chapel Hill's world leadership in coronavirus research. For these past three decades, Dr. Baric has warned that the emerging coronaviruses represent a significant and ongoing global health threat, particularly because they can jump, without warning, from animals into the human population, and they tend to spread rapidly. The Baric Lab uses coronaviruses as models to study the genetics of RNA virus transcription, replication, persistence, pathogenesis, genetics, and cross-species transmission. He has used alphavirus vaccine vectors to develop novel candidate vaccines. Dr. Baric has led the world in recognizing the importance of zoonotic viruses as a potentially rich source of new emerging pathogens in humans, with detailed studies of the molecular, genetic, and evolutionary mechanisms that regulate the establishment and dissemination of such a virus within a newly adopted host. Specifically, he works to decipher the complex interactions between the virion and cell surface molecules that function in the entry and cross-species transmission of positive-strand RNA viruses. In 2017, 2018, and 2019, Dr. Baric was named to Clarivate Analytics' Highly Cited Researchers list, which recognizes researchers from around the world who published the most widely cited papers in their field. Also in 2017, he was awarded a grant for more than $6 million from the National Institute of Allergy and Infectious Diseases to accelerate the development of a promising new drug in the fight against deadly coronaviruses, which is currently in clinical trials to reverse coronavirus disease 2019 in humans. In this collaboration, he continued his partnership between the Gillings School and Gilead Sciences Inc. to focus on an experimental antiviral treatment that he had previously shown to prevent the development of severe acute respiratory syndrome coronavirus in mice. The drug was also shown to inhibit Middle East respiratory syndrome coronavirus and multiple other coronaviruses, suggesting that it may actually inhibit all coronaviruses. He continues to work with this drug.

**Kent Kester, M.D.,** is currently the vice president and the head of translational science and biomarkers at Sanofi Pasteur. In this capacity, he leads a team of more than 200 research and clinical professionals in the United States and France focused on the translational development of new vaccines. During a 24-year career in the U.S. Army, he worked extensively in clinical vaccine development and led multiple research platforms at the Walter Reed Army Institute of Research, the U.S. Department of Defense's (DoD's) largest and most diverse biomedical research laboratory with a major emphasis on emerging infectious diseases, an institution he later led as its Commander/Director. His final military assignment was as the associate dean for clinical research in the School of Medicine at the Uniformed Services University of the Health Sciences (USUHS). During his military service, Dr. Kester was appointed as the lead policy advisor to the U.S. Army Surgeon General in both infectious diseases and in medical research and development. In these capacities, he worked extensively in the interagency environment and developed a variety of U.S. Army and DoD medical policies related to infectious diseases, both clinical and research aspects. Dr. Kester holds an undergraduate degree from Bucknell University and an M.D. from Jefferson Medical College, completing his internship and residency in internal medicine at the University of Maryland and a research fellowship in infectious diseases at the Walter Reed Army Medical Center. Currently a member of the U.S. Government Presidential Advisory Council on Combating Antibiotic-Resistant Bacteria and the U.S. Department of Veterans Affairs Health Services Research & Development Service Merit Review Board, he previously chaired the Steering Committee of the National Institute of Allergy and Infectious Diseases (NIAID)/USUHS Infectious Disease Clinical Research Program, and has served as a member of the U.S. Food and Drug Administration's Vaccines & Related Biologics Products Advisory Committee, the NIAID Advisory Council, and the U.S. Centers for Disease Control and Prevention's Office of Infectious Diseases Board of Scientific Counselors. He is the vice chair of the National Academies of Sciences, Engineering, and Medicine's Forum on Microbial Threats. Board-certified in both internal medicine and infectious diseases, Dr. Kester holds faculty appointments at USUHS and the University of Maryland; and is a fellow of the American College of Physicians, the Royal College of Physicians of Edinburgh, the Infectious Disease Society of America, and the American Society of Tropical Medicine and Hygiene. He is a member of the clinical faculty at the University of Maryland Shock Trauma Center in Baltimore.

**Deven McGraw, J.D., M.P.H.,** is the general counsel and the chief regulatory officer for Ciitizen, a consumer health technology start-up. Previously she directed U.S. health privacy and security as the deputy director of health information privacy at the U.S. Department of Health and Human Services' Office for Civil Rights and the chief privacy officer (acting) of The Office of the National Coordinator for Health Information Technology. Widely recognized for her expertise in health privacy, she directed the Health Privacy Project at the Center for Democracy & Technology for 6 years and led the privacy and security policy work for the Health Information Technology for Economic and Clinical Health Health IT Policy Committee. She also served as the chief operating officer of the National Partnership for Women and Families. She advised health industry clients on Health Information Portability and Accountability Act compliance and data governance while a partner at Manatt, Phelps & Phillips, LLP. She graduated magna cum laude from the Georgetown University Law Center and has an M.P.H. from Johns Hopkins University.

**Alexandra Phelan, S.J.D., LL.M., LL.B.,** is an assistant professor at the Center for Global Health Science and Security in the Department of Microbiology and Immunology at the Georgetown University School of Medicine. Dr. Phelan also holds an appointment as an adjunct professor of law at the Georgetown University Law Center. Dr. Phelan works on legal and policy issues related to infectious diseases, with a particular focus on emerging and reemerging infectious disease outbreaks and international law. She has worked as a consultant for the World Health Organization, the World Bank, and Gavi, the Vaccine Alliance, and has advised on matters including international law and pathogen sharing, human rights law and Zika, intellectual property law, and contract law. She previously worked for a number of years as a solicitor at a firm in Melbourne, Australia, and was admitted to practice to the Supreme Court of Victoria and High Court of Australia in 2010. Dr. Phelan's doctorate examined how overlap between fields of international law—in particular, global health law, international human rights law, and international environmental law—can serve as the catalyst to progressively develop international law to prevent and respond to infectious diseases. She also holds a Master of Laws, specializing in international law, from the Australian National University and a Bachelor of Biomedical Science/Bachelor of Laws (Honours) double degree from Monash University. She also holds a Diploma of Languages (Mandarin Chinese). Dr. Phelan is a General Sir John Monash Scholar and was recognized as an associate fellow of the Royal Commonwealth Society in 2015 for her human rights advocacy during the 2013–2016 Ebola outbreak.

**Saskia Popescu, Ph.D.,** is an infectious disease epidemiologist and a senior infection preventionist. In her current role at HonorHealth, she is responsible for high-consequence disease preparedness for a five-hospital system with dozens of clinics. She performs daily surveillance, risk assessments and review of microbiology reports, investigates outbreaks and health care–associated infections, and guides policy to ensure safety for patients and staff during medical care. More recently, she has been spearheading the coronavirus disease 2019 response for the hospital system and worked on the frontlines to ensure health care workers have the proper training and supplies needed to safely care for infectious patients. During these efforts, she was made a 2017 fellow of the Johns Hopkins Center for Health Security Emerging Leaders in Biosecurity Initiative and has served as an external expert for the European Centre for Disease Control. More recently, Dr. Popescu also serves as an adjunct professor for the University of Arizona's Mel and Enid Zuckerman College of Public Health Epidemiology and Biostats program. Prior to joining HonorHealth, Dr. Popescu was an infection preventionist at Phoenix Children's Hospital, where she led its Ebola readiness and communicable disease surveillance efforts. Dr.

Popescu's graduate studies began with an M.P.H. in epidemiology from the University of Arizona, where she was awarded the Frontier Interdisciplinary eXperience Homeland Security Career Development Grant in Food Protection and Defense for her research and thesis on food system vulnerability to terrorism. Her research also focused on infectious disease modeling and the airborne transmission of microorganisms, like smallpox, in pressurized air cabins. Following this, she completed an M.A. in international security studies from the University of Arizona, where her research addressed the roadblocks for non-state actors to build bioweapons. This research assessed tacit knowledge and manufacturing barriers, as well as dispersal and diagnostic limitations. Following her master's degree, Dr. Popescu received her Ph.D. in biodefense from George Mason University. Her honors include receiving the Presidential Scholarship, the Outstanding Doctoral Student award, the Frances Harbour Award, and serving as a George Mason Global Health Security Ambassador for the 5th Annual Global Health Security Agenda Ministerial meeting in Bali, Indonesia. Dr. Popescu's research and doctoral dissertation examined the political and economic obstacles for U.S. hospitals to invest in infection prevention and control efforts and the resulting impact on national biodefense. Her analysis included case studies including outbreaks of Middle East respiratory syndrome coronovirus, severe acute respiratory syndrome coronavirus, Ebola virus disease, and health care–associated infections. Dr. Popescu has served as a health care liaison for the Arizona Department of Health Services BioSense efforts to ensure the syndromic surveillance system is effectively adapted to health care. She has published several peer-reviewed articles and book chapters on the importance of infection prevention and biopreparedness. She is a certified infection preventionist through the Certification Board of Infection Control and Epidemiology.

**Stuart Ray, M.D., FACP, FIDSA,** serves as the vice chair of medicine for Data Integrity and Analytics and as a professor in the Division of Infectious Diseases within the Department of Medicine, with secondary appointments in viral oncology and health sciences informatics, at the Johns Hopkins University (JHU) School of Medicine. He directs the virology laboratory and is a clinical investigator in the Center for Viral Hepatitis Research in the Division of Infectious Diseases. He is a faculty member of the graduate immunology program, the graduate pharmacology program, and of the Janeway Firm of the Osler Medical Service. Dr. Ray received his M.D. from the Vanderbilt University School of Medicine in 1990. After an internship and residency at Johns Hopkins Hospital, he continued there as the assistant chief of service in medicine (1995–1996) and completed a fellowship in infectious diseases. In 1997, Dr. Ray joined the JHU faculty. His laboratory work has focused on viral sequence variation during acute and chronic infections, developing and applying computational and molecular biology tools to underlying mechanisms including stochastic variation, immune selection, and viral fitness. He continues to care for inpatients and outpatients with HIV, hepatitis C virus, and other infectious diseases. He is an elected member of the American Society for Clinical Investigation, and a fellow of the American College of Physicians and the Infectious Diseases Society of America. During the coronavirus disease 2019 (COVID-19) pandemic, Dr. Ray has served as an attending physician on a dedicated COVID-19 ward at Johns Hopkins Hospital, and as a leader of JHU's severe acute respiratory syndrome coronavirus 2 viral genomic sequencing efforts.

**David Relman, M.D.,** is the Thomas C. and Joan M. Merigan Professor in Medicine, a professor of microbiology and immunology, and a senior fellow at the Freeman Spogli Institute for International Studies at Stanford University. He is also the chief of infectious diseases at the Veterans Affairs Palo Alto Health Care System in Palo Alto, California. Dr. Relman was an early pioneer in the modern study of the human indigenous microbiota (microbiome). A landmark paper in 1999 and another in 2005 were among the first to describe the human oral and gut microbiota, respectively, with modern molecular methods. Most recently, his work has focused on human microbial community assembly, and community stability and resilience. Principles of disturbance and landscape ecology are tested in clinical studies of the human microbiome. Previous work included the development of methods for pathogen discovery, and the identification of several historically important and novel microbial disease agents. He has advised the U.S. government on emerging infectious diseases, human–microbe interactions, and future biological threats. He is a member of the Intelligence Community Studies Board at the National Academies of Sciences, Engineering, and Medicine, and served as the chair of the Boards of Scientific Counselors at the National Institute of Dental and Craniofacial Research and the National Center for Biotechnology Information, both at the National Institutes of Health, and as the president of the Infectious Diseases Society of America (2012–2013). He is a fellow of the American Academy of Microbiology and a member of the National Academy of Medicine.

**Julie Segre, Ph.D.,** is a senior investigator at the National Human Genome Research Institute within the National Institutes of Health (NIH) Intramural Research Program, and a newly elected member of the National Academy of Medicine. She is an expert in microbial genomic sequence and analysis, with a focus on emerging infectious disease and antimicrobial resistance. Dr. Segre integrated bacterial genomic sequencing methods into hospital epidemiological investigations to track a carbapenem-resistant *Klebsiella pneumoniae* outbreak at the NIH Clinical Center in 2011. She served on the

2014 President's Committee of Advisors in Science and Technology commissioned report to address combating antimicrobial resistance. She serves on the Board of Directors for the American Society of Microbiology and the National Insitute of Allergy and Infectious Diseases' executive committee overseeing the antimicrobial resistance program. She has spent the past 2 years collaborating with U.S. Centers for Disease Control and Prevention colleagues and hospital epidemiologists to track the spread of the human fungal pathogen *Candida auris* in U.S. nursing homes.

**Mark Smolinski, M.D., M.P.H.,** currently serves as the president of Ending Pandemics. Dr. Smolinksi brings 25 years of experience in applying innovative solutions to improve disease prevention, response, and control across the globe. Dr. Smolinski is leading a well-knit team, bringing together technologists; human, animal, and environmental health experts; and key community stakeholders to co-create tools for early detection, advanced warning, and prevention of pandemic threats. Since 2009, he has served as the chief medical officer and the director of global health at the Skoll Global Threats Fund (SGTF), where he developed the Ending Pandemics in Our Lifetime Initiative in 2012. His work at SGTF created a solid foundation for the work of Ending Pandemics, which branched out as an independent entity on January 1, 2018. Prior to SGTF, Dr. Smolinksi developed the Predict and Prevent Initiative at Google.org, as part of the starting team at Google's philanthropic arm. Working with a team of engineers, Google Flu Trends (a project that had tremendous impact on the use of big data for disease surveillance) was created in partnership with the U.S. Centers for Disease Control and Prevention (CDC). Dr. Smolinski has served as the vice president for biological programs at the Nuclear Threat Initiative (NTI), a public charity directed by CNN founder Ted Turner and former U.S. Senator Sam Nunn. Before NTI, he led an 18-member expert committee of the Institute of Medicine on the 2003 landmark report *Microbial Threats to Health: Emergence, Detection, and Response*. Dr. Smolinksi served as the sixth Luther Terry Fellow in Washington, DC, in the Office of the U.S. Surgeon General and as an epidemic intelligence officer with CDC. He received his B.S. in biology and M.D. from the University of Michigan in Ann Arbor. He is board-certified in preventive medicine and public health and holds an M.P.H. from the University of Arizona.

**Paul Turner, Ph.D.,** is the Rachel Carson Professor of Ecology and Evolutionary Biology at Yale University and a faculty member in microbiology at the Yale School of Medicine. He studies the evolutionary genetics of viruses, particularly bacteriophages that specifically infect bacterial pathogens, and RNA viruses that are vector-transmitted by mosquitoes. Dr. Turner received a biology degree (1988) from the University of Rochester and a Ph.D. (1995) in zoology from Michigan State University. He did postdoctoral training at the National Institutes of Health (NIH), the University of Valencia in Spain, and the University of Maryland, College Park, before joining Yale's Ecology and Evolutionary Biology Department in 2001. His service to the profession includes chair of the American Society for Microbiology (ASM) Division on Evolutionary and Genomic Microbiology, as well as membership on the National Science Foundation's (NSF's) Biological Sciences Advisory Committee, ASM Committee on Minority Education, and multiple National Research Council advisory committees. Dr. Turner was elected a member of the National Academy of Sciences, fellow of the American Academy of Arts & Sciences, fellow of the American Academy of Microbiology, councilor of the American Genetic Association, chair of the Gordon Research Conference on Microbial Population Biology, and chair of the Jacques Monod Conference on Viral Emergence. He chaired the Watkins Graduate Research Fellowship award committee for ASM, and received the E.E. Just Endowed Research Fellowship and William Townsend Porter Award from the Marine Biological Laboratory, and fellowships from the Woodrow Wilson Foundation, NSF, NIH, and the Howard Hughes Medical Institute. Dr. Turner has served as the director of graduate studies and as the chair of the Ecology and Evolutionary Biology Department at Yale, as well as Yale's dean of science and chair of the Biological Sciences Advisory and Tenure Promotion Committees.

**Deborah Zarin, M.D.,** is the program director, Advancing the Clinical Trials Enterprise, as part of the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard. Dr. Zarin was the director of ClinicalTrials.gov at the National Library of Medicine, National Institutes of Health (NIH), between 2005 and 2018. In that capacity, she oversaw the world's largest clinical trials registry, as well as the development and implementation of the first public database for summary clinical trial results. She played a major role in the development and implementation of key legal and policy mandates for clinical trial reporting, including regulations under the Food and Drug Administration Amendments Act (42 CFR Part 11) and the NIH trial reporting policy. Dr. Zarin's current focus is on the quality and reporting of clinical trials, as well as related efforts to improve the functioning of the clinical research enterprise overall. Previous positions held by Dr. Zarin include the director of the Technology Assessment Program at the Agency for Healthcare Research and Quality and the director of the Practice Guidelines program at the American Psychiatric Association. In these positions, Dr. Zarin conducted systematic reviews and related analyses in support of evidence-based clinical and policy recommendations. Dr. Zarin graduated from Stanford University and received her doctorate in medicine from Harvard Medical School. She completed a clinical decision-making fellowship and a pediatric internship, and is board-certified in general psychiatry as well as in child and adolescent psychiatry.

# Appendix B

# Public Committee Meeting Agendas

The committee held three virtual meetings in June 2020, and portions of two of the meetings were open to the public. The agendas for these open sessions are included in this appendix. To inform the committee's deliberations, various information-gathering mechanisms were used throughout the study:

1. The committee's first virtual meeting in June 2020 included an open session where a representative from the U.S. Centers for Disease Control and Prevention's severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance Program provided its perspectives on the charge to the committee and state and local monitoring of SARS-CoV-2 in order to provide additional background information and context for the study.
2. The committee's second virtual meeting in June 2020 included a public session where the committee heard from researchers operating in three major areas: (1) those studying genomic data, precision epidemiology, and the intersection of emerging infectious disease data sources; (2) those with in-depth knowledge regarding the considerations around the sharing, linkage, and public health application of genomic, epidemiologic, and clinical data; and (3) those who can speak to the lessons learned from previous approaches and existing initiatives. These helped the committee to better understand the nature of the data needs space, as well as the challenges experienced.

## PUBLIC AGENDAS

**Friday, June 12, 2020**
**Virtual Meeting**

### OPEN SESSION

**SESSION III Sponsor Briefing: Discussion of the Committee's Charge**
Objective: To hear from the sponsors of the study regarding their perspectives on the charge to the committee.

4:15 p.m. **Welcome and Introductions**
Diane Griffin, *Committee Chair*
Vice President
National Academy of Sciences
University Distinguished Service Professor
W. Harry Feinstone Department of Molecular Microbiology and Immunology
Johns Hopkins Bloomberg School of Public Health

4:20 p.m. **Sponsor Perspective on Charge to the Committee**
David (Chris) Hassell, *Sponsor*
Acting Principal Deputy Assistant Secretary
Office of the Assistant Secretary for Preparedness and Response
U.S. Department of Health and Human Services

4:30 p.m. **Discussion with Committee**

4:40 p.m. **SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance (SPHERES) Program**
Duncan MacCannell
Chief Science Officer
Office of Advanced Molecular Detection
National Center for Emerging and Zoonotic Infectious Diseases
U.S. Centers for Disease Control and Prevention

4:50 p.m. **Discussion with Committee**

5:00 p.m. **ADJOURN OPEN SESSION**

**Friday, June 19, 2020**
**Virtual Meeting**

**OPEN SESSION**

3:15 p.m. **Welcome and Introductions**

DIANE GRIFFIN, *Committee Chair*
Vice President
National Academy of Sciences
University Distinguished Service Professor
W. Harry Feinstone Department of Molecular Microbiology and Immunology
Johns Hopkins Bloomberg School of Public Health

3:20 p.m. **Speaker Session 1: Genomic Data, Precision Epidemiology, and Working at the Intersection of Emerging Infectious Disease Data Sources**

JOSEPH DERISI
Professor, Biochemistry and Biophysics
University of California, San Francisco
Co-President
Chan Zuckerberg Biohub

ALEX GRENINGER
Assistant Professor, Laboratory Medicine
University of Washington School of Medicine
Assistant Director, Clinical Virology Laboratory
University of Washington Medical Center

3:30 p.m. **Discussion with Committee**

3:50 p.m. **Speaker Session 2: Considerations Around the Sharing, Linkage, and Public Health Application of Genomic, Epidemiological, and Clinical Data**

THERESA COLECCHIA
Senior Associate General Counsel
Office of the Vice President and General Counsel
Johns Hopkins University

Christopher Chute
Bloomberg Distinguished Professor of Health Informatics
Professor of Medicine, Public Health, and Nursing
Chief Research Information Officer, Johns Hopkins Medicine
Deputy Director, Institute for Clinical and Translational Research
Division of General Internal Medicine
Johns Hopkins University

Jill Taylor
Director, Wadsworth Center
New York State Department of Health

4:05 p.m. **Discussion with Committee**

4:30 p.m. **Speaker Session 3: Learning from Previous Approaches and Existing Initiatives**

Pardis Sabeti
Professor, Immunology and Infectious Diseases
Harvard T.H. Chan School of Public Health
Professor, Harvard FAS Center for Systems Biology
Department of Organismic and Evolutionary Biology
Institute Member, Broad Institute
Investigator, Howard Hughes Medical Institute

Emil Lesho
Infectious Disease Specialist
Rochester Regional Health
Professor of Medicine
Uniformed Services University of the Health Sciences

4:40 p.m. **Discussion with Committee**

5:00 p.m. **ADJOURN OPEN SESSION**