Virus Evolution, 2021, 7(2), 1–4

DOI: https://doi.org/10.1093/ve/veab053
Advance access publication date: 5 June 2021
Reflections

Coordinating SARS-CoV-2 genomic surveillance in the United States

Martha I. Nelson[1,*,†] and Peter Thielen[2,‡]

[1]Laboratory of Parasitic Diseases, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 50 South Drive, Building 50 Room 1505, Bethesda, MD 20814, USA and [2]Biological Sciences Group, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Building 201, Laurel, MD 20723, USA

[†]http://orcid.org/0000-0003-4814-0179

[‡]https://orcid.org/0000-0003-1807-2785

*Corresponding author: E-mail: nelsonma@mail.nih.gov

Key words: SARS-CoV-2; pandemic; genomic; surveillance; public health; variant; evolution

The United States has rapidly responded to the emergence of new severe acute respiratory syndrome coronavirus 2 variants of concern by scaling up genomic surveillance. Tens of thousands of viral genomes are now sequenced in American labs each week to track the spread of variants originating in the United States (Annavajhala et al. 2021; Deng et al. 2021) or imported from other countries (Washington et al. 2021) to keep diagnostics, therapeutics, and vaccines up to date (Walensky, Walke, and Fauci 2021). An influx of Federal funding provides an unprecedented opportunity to build a new US genomic surveillance system from the ground up, informed by in-country expertise (National Academies of Sciences 2020; Black et al. 2020) as well as existing models of successful genomic surveillance systems established in other countries (COVID-19 Genomics UK (COG-UK) 2020; Seemann et al. 2020; Msomi, Mlisana, and Tulio 2020). Fully leveraging genetic data require a centrally coordinated national sampling strategy and consortiums for sharing valuable metadata, which are needed to study how new variants evade host immunity, cause severe disease, or transmit differently in human populations. However, US public and private labs have a history of autonomy and strong protections for patient privacy, presenting ongoing barriers to central coordination and data sharing.

1. Centrally coordinated sampling strategy

Routine, population-based sampling that provides an unbiased, representative survey of the genetic composition of viruses circulating over time and space is the gold standard for tracking how new variants relate to disease severity, population immunity, and epidemic trajectory. Instead, genomic surveillance is frequently performed opportunistically for practical reasons, introducing biases that limit the downstream utility of the data. America's vast network of public and private labs have independently generated large numbers of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes that provide highly resolved pictures of the genetic diversity and transmission chains underlying local epidemics (Bedford et al. 2020; Chu et al. 2020; Gonzalez-Reiche et al. 2020; Lemieux et al. 2021). But these pursuits often target local or hospital-based populations with detailed patient metadata and do not always align with broader population-based surveillance of national interest. Coordinating the efforts of America's diverse networks of state, commercially run, and academic labs within a nationwide surveillance consortium that standardizes population-based sampling is no small feat, but the success of the genomics program hinges on it. Carrots work better than sticks and one reason the UK consortium has been successful is because participants access user-friendly, customizable tools for visualizing local and national data trends over time and space (Argimón et al. 2016; Nicholls et al. 2020). Both cloud-hosted and locally implemented bioinformatics tools enable quick conversion from unprocessed sequence data to deposition in global data platforms (Connor et al. 2016; Grubaugh et al. 2019; Singer et al. 2020; Rambaut et al. 2020). Local officials also see benefits when real-time genomic data explain the necessity of unpopular policy reversals, such as the school closures that followed the spike of highly transmissible B.1.1.7 variants in the UK in December 2020 (Volz et al. 2021).

2. Building a research network for genomic epidemiology

High throughput bioinformatic pipelines allow state and local public health labs to flag new variants of concern (VOCs) emerging in communities (Hadfield et al. 2018). But fully leveraging genetic data to understand variants' epidemiological impact requires expertise in advanced phylodynamic methods that are still far from being automated (Lemey et al. 2020; du Plessis et al. 2021). Years ago, the Federal government had the foresight to establish two highly successful research networks, Research and Policy for Infectious Disease Dynamics and Models of Infectious Disease Agent Study, with experts in infectious disease epidemiology and modeling across academia and government (Nelson et al. 2019). The return on investment was high during outbreaks of Ebola, Zika, and pandemic influenza, when epidemiological expertise was on hand to guide vaccination strategies and other countermeasures (Merler et al. 2016). Moreover, the networks had a major downstream impact on developing a new workforce of talented young epidemiologists capable of dealing with highly complex epidemiological data for public health. The highly collaborative alumni network has continued to drive the science behind coronavirus disease 2019 (COVID-19) vaccine strategies, mask wearing, school closures, and social distancing (Borchering et al. 2021). Establishing a similar network for genomic epidemiology would ensure that material investments in generating genomic raw data translate into evidence-based guidance.

3. Patient privacy

A pervasive barrier for US researchers tracking SARS-CoV-2 variants has been the difficulty linking a genomic sequence to contextual information about the patient (Black et al. 2020). Most SARS-CoV-2 sequences submitted to public repositories include accurate information about the US state and date (month, day, and year) of collection, as well as patient age and sex. However, clinicians, public health labs, and academic researchers routinely collect more detailed information, including the patient's county or zip code of residence, travel history, disease severity, as well as contact tracing data. Such information is vital to answer fundamental questions about how new variants transmit and cause disease across different ages (Davies et al. 2021)? However, sharing patient data publicly or with a collaborating research group in the United States is often blocked, even when de-identified, due to strong protections for patient personal health information and informed consent requirements for samples used in research (Beach et al. 2020). Removing a patient's name does not necessarily achieve de-identification when there are few COVID-19 cases in a community, patient names are reported in the news or known within organizations, and detailed spatial-temporal information is included with a sequence.

Genomic epidemiology has only come of age recently, and there has been little time to establish a regulatory framework. As a result, there is no national consensus for what granularity of patient data can be released with a genetic sequence, leaving state and local health departments and Institutional Review Boards (IRBs) to make decisions ad hoc. There are incentives to err on the side of caution when statistical uncertainty about identification combines with high penalties for violating the Health Insurance Portability and Accountability Act of 1996 (United States 1996). Even within the National Institutes of Health information on virus genetics is intentionally separated from epidemiological information on patients. Protecting individuals is a primary concern, but barriers to integrating genomic and epidemiological data across labs limits statistical power and impedes US-wide analyses. One proposed solution is to establish a new US database restricted to a consortium of researchers and public health workers, similar to the UK (Nicholls et al. 2020) so that sensitive data can be shared but not released publicly (Maxmen 2021a). Sharing data in a central database also requires clarified guidance from the Department of Health and Human Services (DHHS) on what level of patient data granularity can be shared in different contexts without violating Federal law.

4. Standardized metadata

From a practical standpoint, overstretched public health workers do not necessarily have the time to submit dozens of metadata fields that are potentially of interest to researchers. Many fields require time-consuming manual extraction from patient charts. National priorities need to be predefined and streamlined to scale up data integration nationally (Gardner et al. 2020). For example, a study based on tens of thousands of sequences for which basic clinical data are available (e.g. asymptomatic, mild, hospitalized, and death) along with simplified patient characteristics (age, sex, presence or absence of comorbidities) (Volz et al. 2020) is likely to have more power than a study with more detailed metadata (e.g. temperature, cough, ventilator use, and diabetes) but only hundreds of sequences (Thielen et al. 2021).

5. Global data sharing

The speed with which SARS-CoV-2 viruses spread across the world, particularly as international travel normalizes, means that all countries, even those as large as the United States, depend on quality genomic data from other countries to quickly identify and characterize new variants emerging globally and trace how they infiltrate the United States. Reaching the milestone of one million SARS-CoV-2 genomes submitted to GISAID reflects recent advances in genomic sequencing and data sharing on a global scale. However, enormous volumes of data introduce bioinformatic challenges that test scalability. The need to balance tradeoffs between the free flow of information and protecting data contributors has become dire as the sheer volume of genomic data becomes unwieldy for individual labs and standard bioinformatics software (Richard 2021). Providing more open access to the GISAID API would facilitate the mass flow of data into new high throughput bioinformatic tools and pipelines. Realistically, there needs to be a conceptual distinction between unethically using another group's data as the focus of a study versus including genomes from other countries as background data. The value of inclusion in a 30-page supplementary acknowledgment table listing thousands of submitters is less clear in the million-genome era. However, developing countries that shared valuable SARS-CoV-2 sequence data with the global community have received few COVID-19 vaccine doses to date, fueling concerns about exploitation and disincentivizing any relaxation of existing protections on virological data for the foreseeable future (Maxmen 2021b).

6. Raw sequence data

Consensus genomes are the primary unit of viral genome surveillance. High throughput sequencing platforms generate hundreds- or thousands-fold excess data to extract this consensus using a 'majority wins' strategy. The raw sequence data underlying consensus genomes can provide valuable insights into ambiguous base calls as well as minority variants. However, raw sequence reads submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) represent only a tiny fraction of genomes available on GenBank or GISAID. The low number of submissions is due to the logistical complexity of transferring large amounts of data, minimal observed benefit by submitters focused on generation of consensus genomes, and concerns about submitting sequence reads that may contain host DNA. Messaging needs to be improved that SRA has reduced barriers to upload and includes a process to remove host DNA. Additionally, the pathogen genomics community would benefit greatly from further development and adoption of standardized informatics workflows that easily transition from large core facilities to small independent laboratories, which often have minimal informatics support. Reducing the time between sequencing and data dissemination has been a primary bottleneck for many US groups that are establishing sequencing at scale for the first time, and often analytical capabilities are not in place to enable groups to both generate consensus sequences and upload the raw data that were used to generate them.

7. Conclusions

As countries ramp up genomics capacity to track a rapidly evolving virus (Geoghegan et al. 2020; Rockett et al. 2020; Munnink et al. 2020; Hammer et al. 2021; Tegally et al. 2021, Bugembe et al., 2021; Faria et al., 2021; Ranjan et al., 2021), the infrastructures built will hopefully outlast the current pandemic and improve outbreak response for decades to come. Domestically, the United States faces unusual structural challenges that go deeper than funding, but all stakeholders are incentivized by new systems that save time in managing large data streams, approving IRBs and providing useful information to guide decision-making. If you build it, they will come.

Acknowledgements

This work was supported in part by the Intramural Research Program of the NIH.

Funding

This work was supported in part by the Intramural Research Program of the NIH. This work was supported by the Centers of Excellence for Influenza Research and Surveillance, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services (DHHS), under contract numbers HHSN272201400007C (PT) and HHSN272201400008C (MIN).

Conflict of interest: None declared.

Disclaimer

The content does not necessarily reflect the views or policies of the U.S. Government or imply endorsement of any products.

References

Annavajhala, M. K. et al. (2021) 'A Novel SARS-CoV-2 Variant of Concern, B.1.526, Identified in New York', *medRxiv*.

Argimón, S. et al. (2016) 'Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography', *Microbial Genomics*, 2: e000093.

Beach, M. C. et al. (2020) 'Desperate Times: Protecting the Public from Research without Consent or Oversight during Public Health Emergencies', *Annals of Internal Medicine*, 173: 926–928.

Bedford, T. et al. (2020) 'Cryptic Transmission of SARS-CoV-2 in Washington State', *Science*, 370: 571–575.

Black, A. et al. (2020) 'Ten Recommendations for Supporting Open Pathogen Genomic Analysis in Public Health', *Nature Medicine*, 26: 832–41.

Borchering, R. K. et al. (2021) 'Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Non-pharmaceutical Intervention Scenarios - United States, April-September 2021', *MMWR Morbidity and Mortality Weekly Report*, 70: 719–24.

Bugembe, D. L. et al. (2021) 'A SARS-CoV-2 Lineage A Variant (A.23.1) with Altered Spike Has Emerged and Is Dominating the Current Uganda Epidemic', *medRxiv*.

Chu, H. Y. et al. (2020) 'Early Detection of Covid-19 through a Citywide Pandemic Surveillance Platform', *The New England Journal of Medicine*, 383: 185–7.

Connor, T. R. et al. (2016) 'CLIMB (The Cloud Infrastructure for Microbial Bioinformatics): An Online Resource for the Medical Microbiology Community', *Microbial Genomics*, 2: e000086.

COVID-19 Genomics UK (COG-UK). (2020) 'An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network', *The Lancet Microbe*, 1: e99–100.

Davies, N. G. et al. (2021) 'Increased Mortality in Community-Tested Cases of SARS-CoV-2 Lineage B.1.1.7', *Nature*.

Deng, X. et al. (2021) 'Transmission, Infectivity, and Antibody Neutralization of an Emerging SARS-CoV-2 Variant in California Carrying a L452R Spike Protein Mutation', *medRxiv*, 593: 270–274.

Faria, N. R. et al. (2021) 'Genomics and Epidemiology of the P.1 SARS-CoV-2 Lineage in Manaus, Brazil', *Science*, 372: 815–821.

Gardner, L. et al. (2020) 'A Need for Open Public Data Standards and Sharing in Light of COVID-19', *The Lancet Infectious Diseases*, e80.

Geoghegan, J. L. et al. (2020) 'Genomic Epidemiology Reveals Transmission Patterns and Dynamics of SARS-CoV-2 in Aotearoa New Zealand', *Nature Communications*, 11: 6351.

Gonzalez-Reiche, A. S. et al. (2020) 'Introductions and Early Spread of SARS-CoV-2 in the New York City Area', *Science*, 369: 297–301.

Grubaugh, N. D. et al. (2019) 'An Amplicon-Based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and iVar', *Genome Biology*, 20: 8.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Hammer, A. S. et al. (2021) 'SARS-CoV-2 Transmission between Mink (Neovison vison) and Humans, Denmark', *Emerging Infectious Diseases*, 27: 547–51.

Lemey, P. et al. (2020) 'Accommodating Individual Travel History and Unsampled Diversity in Bayesian Phylogeographic Inference of SARS-CoV-2', *Nature Communications*, 11: 5110.

Lemieux, J. E. et al. (2021) 'Phylogenetic Analysis of SARS-CoV-2 in Boston Highlights the Impact of Superspreading Events'. *Science*, 371: eabe3261.

Plessis, L.D. et al. (2021) 'Establishment and Lineage Dynamics of the SARS-CoV-2 Epidemic in the UK', *Science*, 371: 708–12.

Maxmen, A. (2021a) 'Massive Google-Funded COVID Database Will Track Variants and Immunity', *Nature*, Epub ahead of print.

——— (2021b) 'Why Some Researchers Oppose Unrestricted Sharing of Coronavirus Genome Data', *Nature*, 593: 176–7.

Merler, S. et al. (2016) 'Containing Ebola at the Source with Ring Vaccination', *PLoS Neglected Tropical Diseases*, 10: e0005093.

Msomi, N., Mlisana, K., and Tulio, D. O. Network for Genomic Surveillance in South Africa writing group. (2020) 'A Genomics Network Established to Respond Rapidly to Public Health Threats in South Africa', *The Lancet Microbe*, 1: e229–30.

Munnink, O. et al. (2020) 'Rapid SARS-CoV-2 Whole-Genome Sequencing and Analysis for Informed Public Health Decision-Making in the Netherlands', *Nature Medicine*, 26: 1405–10.

Nelson, M. I. et al. (2019) 'Fogarty International Center Collaborative Networks in Infectious Disease Modeling: Lessons Learnt in Research and Capacity Building', *Epidemics*, 26: 116–27.

Nicholls, S. M. et al. (2020) 'MAJORA: Continuous Integration Supporting Decentralised Sequencing for SARS-CoV-2 Genomic Surveillance', *bioRxiv*.

Rambaut, A. et al. (2020) 'A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 5: 1403–7.

Ranjan, R., Sharma, A., and Verma, M. K. (2021) 'Characterization of the Second Wave of COVID-19 in India', *bioRxiv medRxiv*.

Richard, V. N. (2021) 'Scientists Call for Fully Open Sharing of Coronavirus Genome Data', *Nature*, 590: 195–6.

Rockett, R. J. et al. (2020) 'Revealing COVID-19 Transmission in Australia by SARS-CoV-2 Genome Sequencing and Agent-Based Modeling', *Nature Medicine*, 26: 1398–404.

Seemann, T. et al. (2020) 'Tracking the COVID-19 Pandemic in Australia Using Genomics', *Nature Communications*, 11: 4376.

Singer, J. et al. (2020) 'CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation', *Preprints*.

Tegally, H. et al. (2021) 'Detection of a SARS-CoV-2 Variant of Concern in South Africa' *Nature*, 592: 438–43.

Thielen, P. M. et al. (2021) 'Genomic Diversity of SARS-CoV-2 during Early Introduction into the Baltimore-Washington Metropolitan Area', *JCI Insight*, 6: e144350.

United States. (1996) 'Health Insurance Portability and Accountability Act of 1996. Public Law 104–191', *United States Statutes at Large*, 110: 1936–2103.

Volz, E. et al. (2020) 'Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity', *Cell*, 184: 64–75. e11.

——— (2021) 'Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from Linking Epidemiological and Genetic Data', *bioRxiv medRxiv*.

Walensky, R. P., Walke, H. T., and Fauci, A. S. (2021) 'SARS-CoV-2 Variants of Concern in the United States-Challenges and Opportunities', *JAMA: The Journal of the American Medical Association*, 325: 1037–8.

Washington, N. L. et al. (2021) 'Genomic Epidemiology Identifies Emergence and Rapid Transmission of SARS-CoV-2 B.1.1.7 in the United States', *medRxiv*.