EMERGING INFECTIOUS DISEASES®

U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION

Advances in Pathogen Genomics for Infectious Disease Surveillance, Control, and Prevention

May 2025

Vincent van Gogh (1853–1890), *The Starry Night* (1889). Oil on canvas, 29″ × 36 1/4″/ 73.7cm × 92.1 cm. Museum of Modern Art, New York, New York, USA. Public domain image from Google Art Project.

EMERGING INFECTIOUS DISEASES®

Emerging Infectious Diseases is published monthly by the Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H16-2, Atlanta, GA 30329-4018, USA. Telephone 404-639-1960; email eideditor@cdc.gov

EMERGING INFECTIOUS DISEASES®

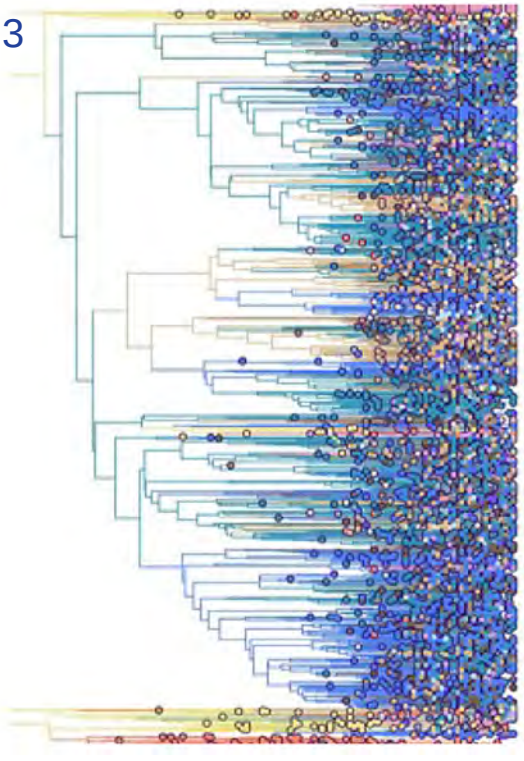
Advances in Pathogen Genomics for Infectious Disease Surveillance, Control, and Prevention

Supplement to May 2025

On the Cover

On the cover: Vincent van Gogh (1853–1890), *The Starry Night* (1889). Oil on canvas, 29 in × 36¼ in/73.7 cm × 92.1 cm. Museum of Modern Art, New York, New York, USA. Public domain image from Google Art Project.

S53



EMERGING INFECTIOUS DISEASES®
Supplement to May 2025

S63

About the Cover

A Decade of Partnerships and Progress in Pathogen Genomics in Public Health Practice

Duncan MacCannell, Bronwyn MacInnis, Scott Santibanez, Margaret A. Honein, Wendi Kuhnert, Christopher Braden

The field of public health is evolving continuously, driven in part by advances in science and biotechnology. Among the most transformative of those changes are the rapid acceleration of genomics, bioinformatics, and molecular epidemiology throughout the public health system and the growing list of applications of such methods for infectious diseases of both public health and economic concern (*1*). Those technologies have revolutionized the understanding, tracking, and prevention of many infectious diseases and are notable for their adaptability to different pathogens and changing diagnostic, surveillance, and response requirements, as well as to different levels of technical capacity in laboratory environments where they are in use.

To understand the effect of pathogen genomics on public health, we must consider the types of materials that can be sequenced, which include pathogens (viral, bacterial, fungal, parasitic) from host, vector, and environmental samples; the unique and complex features of pathogen genomes; and the thousands of potential datapoints their genomes can encode. Because molecular techniques generate massive amounts of raw sequence data, investments in bioinformatics, including data analysis tools, computing infrastructure, and skilled workforce, are critical to translate pathogen sequence data into actionable public health information. By combining new types of data with traditional epidemiologic approaches, sequencing provides an incredibly powerful new set of tools for investigating the spread and informing the prevention and control of infectious diseases.

Author affiliations: Centers for Disease Control and Prevention, Atlanta, Georgia, USA (D. MacCannell, S. Santibanez, M.A. Honein, W. Kuhnert, C. Braden); The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA (B. MacInnis).

DOI: https://doi.org/10.3201/eid3113.241670

The SARS-CoV-2 pandemic led to pathogen genomic sequencing being used at a previously unfathomable scale. By the end of 2024, >17 million virus genomes were catalogued globally, roughly one third from laboratories in the United States. Those sequences were assembled, annotated, and shared through several major international sequence repositories, including GenBank (*2*) and the GISAID (https://www.gisaid.org) EpiCoV database (*3*). This rapid sharing of pathogen genomic data enabled the global public health community to monitor and respond to the evolution and transmission of the virus and to begin critical work on technical and ethical requirements for better and more equitable data sharing.

Within the United States, SARS-CoV-2 sequence data provided a critical national baseline for variant surveillance; enabled reference characterization and risk assessment of emerging variants; aided in ongoing assessment of approved diagnostics, vaccines, and therapeutics; and guided overall pandemic response strategy and resource prioritization (*4*). The data also provided necessary insights at the state and local level, enabling many jurisdictions to implement comprehensive surveillance for unusual variants, phenotypes, or clinical outcomes; to guide clinical and therapeutic strategy; and to enhance outbreak response and infection control efforts on the basis of genomic and phylodynamic information (*5*).

For more than a decade, the Centers for Disease Control and Prevention (CDC) Advanced Molecular Detection (AMD) program has been at the forefront of integrating genomics and other novel laboratory technologies into routine public health practice. AMD is a cross-cutting program that works across CDC's infectious disease centers, state and local public health departments, and a growing network of academic, commercial, and global partners to help drive innovative, high complexity laboratory technologies and

the bioinformatic and workforce changes needed to implement the technologies effectively, sustainably, and at scale.

Initial AMD investments focused on building sequencing and bioinformatics capacity for CDC core programs, while expanding capacity in state and local public health departments. More recent activities have sought to strengthen and build the public health workforce and create opportunities for academic and private sector collaboration. Future efforts will focus on promoting public engagement around AMD technologies, developing a professionally diverse and resilient technical workforce, ensuring representativeness of AMD studies and leadership, and expanding access to the benefits of these investments. Global investments and partnerships in pathogen genomics have already shown great promise in building contextually relevant genomic sequencing and analysis capacity at a local and regional level in many low-to-middle income countries, including establishing locally managed networks of infrastructure, resources, and expertise.

This supplemental issue of *Emerging Infectious Diseases* brings together a collection of articles that showcase the progress and effect of pathogen genomics, bioinformatics, and genomic epidemiology on public health, illustrating both the current effects and the future promise of such technologies. The technologic contributions in this supplement issue highlight a range of different applications and collaborations, including case studies of outbreak responses and surveillance programs, necessary advances in bioinformatics tools and databases, practical considerations for capacity building and benefit to populations experiencing greater illness and death, and perspectives on the future direction of the field.

As public health continues to evolve and to adapt to new threats and challenges, the integration of new technologies and the data they provide will continue to shape the landscape of infectious disease surveillance and response. The insights gained from those new data will be crucial in addressing emerging and reemerging infectious diseases, ensuring we are better prepared to face public health challenges today and tomorrow.

About the Author

Dr. MacCannell is the director of the Office of Advanced Molecular Detection, Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases, at the Centers for Disease Control and Prevention. His interests include sustainable innovation in advanced laboratory technologies, bioinformatic capacity, and genomic epidemiology across the public health system.

References

1. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen genomics in public health. N Engl J Med. 2019;381:2569–80. https://doi.org/10.1056/NEJMsr1813907
2. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41(D1):D36–42. https://doi.org/10.1093/nar/gks1195
3. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017;1:33–46. https://doi.org/10.1002/gch2.1018
4. Karthikeyan S, Levy JI, De Hoff P, Humphrey G, Birmingham A, Jepsen K, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature. 2022;609:101–8. https://doi.org/10.1038/s41586-022-05049-6
5. Lambrou AS, Shirk P, Steele MK, Paul P, Paden CR, Cadwell B, et al.; Strain Surveillance and Emerging Variants Bioinformatic Working Group; Strain Surveillance and Emerging Variants NS3 Working Group; Strain Surveillance and Emerging Variants NS3 Working Group. Genomic surveillance for SARS-CoV-2 variants: predominance of the Delta (B.1.617.2) and Omicron (B.1.1.529) variants, United States, June 2021–January 2022. MMWR Morb Mortal Wkly Rep. 2022;71:206–11. https://doi.org/10.15585/mmwr.mm7106a4

Address for correspondence: Duncan MacCannell, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H24-11, Atlanta, GA 30329-4018, USA; email: fms2@cdc.gov

An Advanced Molecular Detection Roadmap for Nonlaboratorians

Jessica N. Ricaldi, J. Todd Parker, Nathelia Barnes, Hannah Turner, Scott Santibañez

This article, aimed at nonlaboratorians such as healthcare providers, public health professionals, and policymakers, provides basic concepts and terminology to enable better understanding of other manuscripts in this advanced molecular detection journal supplement. This article focuses on 3 aspects of advanced molecular detection: pathogen genomics, bioinformatics, and public health application, while providing additional resources for understanding.

Advanced molecular detection (AMD) combines next-generation sequencing (NGS), bioinformatics, and traditional epidemiology to provide detailed information on disease-causing microorganisms, or pathogens (*1*). AMD has become central to the US public health system's efforts to identify, track, and stop infectious diseases. The Centers for Disease Control and Prevention's (CDC) Office of Advanced Molecular Detection, part of the Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases, works to modernize the public health system's disease investigation capabilities by using the latest technologies and building AMD capacity in public health partner institutions (*1*).

Although AMD has empowered public health agencies across the United States to rapidly identify and solve outbreaks that were previously undetectable, the technical terminology can be challenging for many healthcare and public health professionals. This article aims to provide nonlaboratorians such as physicians, advanced practice providers, public health professionals, and policy makers with an overview of how advanced molecular approaches are used to detect and control infectious disease threats. This primer will assist nonlaboratory personnel to better understand the concepts and terminology used in this AMD journal supplement, as well as in the daily practice of clinical medicine and public health.

Author affiliation: Centers for Disease Control and Prevention, Atlanta, Georgia, USA

DOI: https://doi.org/10.3201/eid3113.241506

We will focus on 3 aspects of AMD: pathogen genomics, how laboratory scientists use technologies to study the genetic composition, or sequences, of infectious microorganisms; bioinformatics, how high-performance computing is used to analyze genetic sequence data; and public health application, how epidemiologists, clinicians, and other public health professionals combine information from field investigations with genetic sequence data to identify and stop outbreaks. Although much has been written about the use of NGS for mapping the human genome, the focus of this journal supplement is pathogen genomics, the sequencing of microorganism genomes that can cause infectious diseases.

Three Aspects of AMD

Pathogen Genomics

As recently as the late 20th Century, healthcare providers and clinical laboratories relied on established, culture-dependent techniques for the laboratory identification of bacteria and viruses and the reporting of such findings for disease surveillance. Sanger sequencing is a method for DNA sequencing of specific genes developed in the 1970s. Sanger sequencing is highly accurate but expensive and time-consuming, especially when sequencing an organism's entire genetic code or genome (*2*). The development of NGS in the early 2000s greatly advanced the field of genome sequencing and analysis, or genomics. NGS enabled the rapid, automated sequencing of many genetic fragments in parallel, providing a large amount of genetic information rapidly and at a lower cost compared with older methods. A wide range of approaches to sequencing have since been developed that can be targeted to look for a specific pathogen or be pathogen agnostic and sequence any microbial genetic material in a sample. Some examples of commercially available laboratory sequencing methods include detection of fluorescently labeled nucleotides (Illumina, https://www.illumina.com), detection of

hydrogen ions during polymerization (Ion Torrent, https://www.thermofisher.com), analysis of electrical signals from biologic molecules that have passed through nanometer-sized pores (Oxford Nanopore, https://www.nanoporetech.com), and direct observation of the sequencing process (PacBio, https://www.pacb.com). To date, available sequencing systems, or platforms, can be broadly grouped into short-read or long-read platforms on the basis of the length of sequence reads they produce, measured in base pairs. A base pair is a unit of double-stranded nucleic acids consisting of 2 complementary DNA nucleotide bases bound to each other by hydrogen bonds. Short-read platforms (<500 bp) fragment the genome to be sequenced into short fragments and are more reliable for detecting low frequency genetic variations and short insertions, deletions, and mutations (*3*). Long-read methods (3,500–11,000 bp) can read longer stretches of DNA, or complete regions of a gene, and are used when studying complex genomes such as in metagenomic sequencing, which involves analysis of genetic material of all organisms that may be within environmental or clinical samples. Short-read platforms are better for identifying the precise genome sequences, nucleotide by nucleotide, whereas long-read platforms are better for identifying large DNA insertions or deletions (*4*).

NGS involves both traditional laboratory components, such as sample collection, DNA extraction, and sequencing machines, and bioinformatics components, such as using computational models, also known as pipelines, to analyze the large volumes of data created by NGS. Analysis of these data can reveal epidemiologic patterns of disease transmission, genetic variations, antimicrobial resistance genes, and other information necessary to clinical care and public health. As NGS technologies became more widely available and affordable, sequencing whole bacterial and viral genomes to understand disease transmission became common in clinical and public health laboratories (*2*). Along with increased use, validation of NGS tests is both critical and difficult because of workflow variations across laboratories, such as differing sample types; operating procedures for extraction, amplification, and sequencing; and bioinformatic processes. Because of those differences, specific quality parameters are vital for both laboratory sequencing and bioinformatic technologies. CDC has invested in the development of quality management systems and quality system tools that are both technology and manufacturer specific.

Whole-genome sequencing (WGS), a type of NGS, enables scientists to determine a mostly complete sequence of an organism's genome and provides more data than methods that only sequence a portion of the genome. For example, in addition to providing information about the evolutionary history and relationships among streptococcal organisms, potential streptococcal drug resistance patterns and typing (e.g., M protein typing) can be genetically inferred by using the same WGS pipeline (*5*). WGS has also improved surveillance for foodborne pathogen outbreaks and enhanced the detection of trends in foodborne infections and antimicrobial resistance at the state public health laboratory level.

An additional application that has moved from research to clinical practice is 16S sequencing. The 16S ribosomal RNA gene is conserved and found in all bacteria and is the most widely used for phylogenetic identity of a bacteria and the most frequently ordered of advanced molecular tests (*6*). Scientists amplify, sequence, and compare it with other known 16S sequences, using 16S variable and conserved regions for clinical laboratory diagnosis. Whereas WGS is also used for viral sequencing, ribosomal RNA sequencing enables many bacteria to be identified at the genus or species level, including bacteria that are hard to cultivate or following the administration of antimicrobial therapy (*7*).

Bioinformatics

NGS provides a large amount of genetic information. Microbial bioinformatics is a data-driven approach that combines the use of sequencing data, machine learning, and artificial intelligence for rapid public health response. This field uses computational tools for disease surveillance, monitoring antimicrobial resistance, and outbreak investigations. By using computer science and statistical methods, such as high-performance supercomputing to organize and interpret the data, bioinformatic tools can track, identify, and monitor pathogens while tracing transmission pathways and phylogenetic origins. Phylogenetic methods play a crucial role in studying the evolutionary history and relationships among organisms. Bioinformatic pipelines are used to assemble genomes, detect genetic variants, and build phylogenies, which are visual representations of the evolutionary relationships among organisms. Those pipelines start with a defined set of files, such as FASTA sequences (a text-based format that represents nucleotide sequences). Connected software routines are then used to generate results, such as sequence alignment or tree figures. Different tools and workflows can also be used to assemble a genome or to perform variant calling

(i.e., detecting variants by comparing against a reference genome) (*8*).

Alignments can identify genetic variations such as single-nucleotide polymorphisms, which are variations in a single nucleotide at a specific place on the genome. Alignments of specific gene sequences or whole genomes aligned to a reference sequence are used as the input for the software or pipelines that generate phylogenies, which trace patterns of shared ancestry among organisms. By analyzing phylogenies, researchers can infer relatedness between pathogen sequences and describe them by using graphics and diagrams such as phylogenetic trees, which illustrate the genetic relationships among organisms. Phylogenetic trees are built by using different probability methods for analysis with various software developed to ensure computational efficiency (*8*). Phylogenetic trees that show relatedness among pathogens from different sources can provide additional information to complement traditional epidemiology data, determine associations, and help link human cases or establish a common source of infection.

Commercially available bioinformatics pipelines are often used for clinical diagnostic testing. Such pipelines must comply with patient safety, laboratory quality assurance, comparability across laboratories, and local and federal regulatory compliance requirements (*9*). Many open-source tools, such as Nextstrain (https://www.nextstrain.org), UShER (https://www.genome.ucsc.edu/cgi-bin/hgPhyloPlace), and MicrobeTrace (https://www.microbetrace.cdc.gov), and laboratory-developed code are also available and frequently used. Software containerization methods are used to package bioinformatics tools and pipelines into portable units (or containers), improving efficiency, reproducibility, and security (*10*), and to combine pathogen genomic information with other sources of information to estimate cases and predict the pathogen's origins, movements, and potential affect (*11*).

Scientists have a critical need to share information quickly and efficiently. As a part of the National Institutes of Health, the National Center for Biotechnology Information (NCBI) serves a key role by providing access to biomedical and genomic information, as well as developing software tools for bioinformatics and sequencing analysis. NCBI has served as a hub for sharing genomic information through the GenBank DNA sequence database. The availability of sequences through GenBank enables scientists to compare sequences from other laboratories. AMD is now used to track a wide array of disease agents, including antimicrobial-resistant foodborne bacterial and fungal pathogens, including many in the NCBI Pathogen Detection isolate browser (https://www.ncbi.nlm.nih.gov/pathogens). Other examples of resources that enable sharing of information include the Virus Pathogen Database and Analysis Resource (https://www.bv-brc.org), a platform which provides information about virus mutation, and GISAID (https://www.gisaid.org), an online platform that enables the sharing of information about viral genomic sequences. Those resources play a crucial role in monitoring viral pathogens, including novel strains and respiratory viruses, contributing to the understanding of their evolution and transmission patterns.

Public Health Application

Because sequencing costs decreased and platforms were created to manage and analyze larger sets of data, the use of those methods went from proof-of-concept and validating results against traditional epidemiology methods to becoming standard methodologies (*12*). AMD has become a central part of public health efforts to identify and control infectious diseases and is now incorporated into public health outbreak and emergency response, disease surveillance, drug resistance detection, clinical microbiology, and other public health applications (*13*).

In the United States, funding provided through CDC has been instrumental in building national capacity for AMD in state, local, and territorial public health laboratories, as well as in hospital and clinical laboratories. The use of AMD has evolved to include a wide array of infectious diseases, including respiratory diseases and antimicrobial drug resistant diseases (*14*,*15*). AMD has enhanced public health professionals' ability to rapidly identify pathogens across the country, track the spread and identify sources of outbreaks, detect drug resistance in US hospitals, inform vaccine development, conduct disease surveillance, and promote international collaborations. Other reports have provided examples of linkages of AMD to surveillance and epidemiology in traveler-based genomic surveillance, enhanced AMR surveillance, and outbreak investigation (*16*–*18*). The global relevance of AMD and presence of similar programs in other countries is also of note. For example, a report on the national genomic surveillance system for *Listeria monocytogenes* and the effect of implementing decentralized sequencing in Australia is included in this issue (*19*).

The following are a few examples of how AMD is used in the field of public health. First, PulseNet

is a national laboratory network that uses AMD diagnostics to detect and prevent foodborne outbreaks (*20*). PulseNet International involves implementation of whole WGS for global food-borne disease surveillance (*21*). Second, the MinION portable DNA sequencer was used during the 2014 Ebola virus outbreak in West Africa (*22*). Third, the Secure HIV TRAnsmission Cluster Engine has been used to identify clusters of highly similar HIV sequences, indicating rapid transmission (*23*). Fourth, AMD-supported diagnostics were used in the early detection of SARS-CoV-2 variants (*24*). Finally, rapid AMD-supported diagnostic testing was used during the mpox outbreak response (*25*).

A necessary part of applied public health is being aware of the caveats and limitations of methods and their applications, as well as understanding questions that public health practitioners should be asking when presented with data derived from NGS methods. Public health practitioners should consider several points when they are given data derived from NGS methods. First, practitioners should ask about the methods used when data are presented and be aware of the limitations of each. Each sequencing methodology has its own limitations. For example, if a lab provides sequencing data using nanopore, there are challenges in using those data for the detection of mutations. Second, practitioners should know a gene being detected doesn't indicate the gene is being expressed or is functional. Third, practitioners must realize different methods will yield different results; for example, genome assemblies of the same organism by 2 different methods may provide different single-nucleotide polymorphism counts when run on the same panel and could affect epidemiologic investigations, and different bacterial or viral classifiers will yield different results from the same metagenomic raw data. Finally, practitioners should be aware that increasing use of metagenomics presents many unique challenges with results interpretation. When you receive results of interest consider the following: what the method or pipeline was built to do versus what is it doing (i.e., using a SARS-CoV-2 method to look for bacterial DNA); what the result is based on, a single gene or part of a gene; and what databases were used for the analysis. Many pathogens of interest are overrepresented in databases and may turn up disproportionally in results.

Table. Education resources to further understanding of advanced molecular detection for nonlaboratorians

Resource title	Description	Link
Advanced Molecular Detection COVID-19 Genomic Epidemiology Toolkit	Toolkit to address topics related to the application of genomics to epidemiologic investigations and public health response to SARS-CoV-2.	COVID-19 Genomic Epidemiology Toolkit, https://www.cdc.gov/advanced-molecular-detection/php/training
American Society for Microbiology Comprehensive Training in Infectious Disease Applications of Next Generation Sequencing	Free training to educate the clinical microbiology workforce on next generation sequencing technologies to increase pathogen genomic sequencing capacity and increase preparedness for the next pandemic through enhanced molecular surveillance.	Training in NGS for Infectious Disease Applications, https://asm.org/Webinars/training-ngs-infectious-diseases
Association of Public Health Laboratories Advanced Molecular Detection	Resources for building the advanced molecular detection workforce and other related activities.	Advanced Molecular Detection, https://www.aphl.org/programs/infectious_disease/Pages/Advanced-Molecular-Detection.aspx
Association of Public Health Laboratories Learning Center	Multiple online courses on molecular testing and sequencing. Some courses are only available for members.	Learning Center, https://learn.aphl.org/learn/signin
Council of State and Territorial Epidemiologists Advanced Molecular Detection Workgroup	Information about public health professionals interested in using the latest next-generation genomic sequencing technologies at local, state, tribal, and territorial health departments for disease detection, surveillance, outbreak investigation, and prevention. The workgroup includes a listserv.	Advanced Molecular Detection Subcommittee, https://www.cste.org/page/AdvancedMolecularDetection
Council of State and Territorial Epidemiologists Webinar Library	Multiple webinars developed by council of state and territorial epidemiologists.	Webinar Library, https://www.cste.org/page/WebinarLibrary
Northwest Pathogen Genomics Center of Excellence	Website includes publications, situation reports, tools, workflows, and other educational resources.	Northwest Pathogen Genomics Center of Excellence, https://nwpage.org
Minnesota Pathogen Genomics Center of Excellence	Website describes projects completed by the Minnesota Pathogen Genomics Center of Excellence.	Infectious Disease Projects, https://www.health.state.mn.us/diseases/idlab/pathogen
Virginia Pathogen Genomics Center of Excellence	Videos and other teaching resources.	Virginia Pathogen Genomics Center of Excellence, https://va-pgcoe.org
State Public Health Bioinformatics Group	Resources and materials from previous trainings offered.	StaPH-B, https://staphb.org

Workforce Development and Capacity Building

Laboratory methods have evolved from relatively simple, culture-dependent techniques for identifying bacteria and viruses to expensive and time-consuming DNA sequencing to more rapid and less costly sequencing, resulting in vast amounts of genetic information. Information about those advances has not always been communicated clearly to healthcare and public health professionals or the public. For readers who are interested in learning more, CDC has enabled state, local, and territorial public health laboratories to provide regional training, including sequencing and molecular epidemiology tools trainings, Bioinformatics Regional Resources, and other opportunities designed to help clinicians, scientists, and public health practitioners to understand transmission chains, characterize emerging pathogens, and solve outbreaks (*1*) (Table).

Beyond regional trainings, examples of other resources include CDC's AMD Academy, hosted in partnership with the Association of Public Health Laboratories and the Council of State and Territorial Epidemiologists. The AMD Academy is a multiday training in molecular epidemiology and bioinformatics for epidemiologists and microbiologists from state, territorial, and local health departments. The COVID Genomic Epidemiology toolkit is another valuable resource that addresses topics related to the application of genomics to epidemiologic investigations and public health response to SARS-CoV-2 at state, territorial, and local levels. This toolkit provides introductory information such as how to interpret phylogenetic trees in the context of transmission, along with practical case studies demonstrating real-world applications (*26*).

In conclusion, this journal supplement aims to improve knowledge and awareness of advances in AMD. The progress of AMD techniques over the past few decades has had a considerable effect on public health. We hope that this article can begin to demystify the field of AMD by providing a brief introduction to other papers in this journal supplement and AMD in general.

The authors have not received additional financial support for the development of this manuscript.

About the Author

Dr. Ricaldi is a health scientist at the Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention. Her interests include emerging infections, molecular epidemiology, and emergency response.

References

1. Centers for Disease Control and Prevention. About CDC's advanced molecular detection program. Apr 3, 2024 [cited 2024 Nov 20]. https://www.cdc.gov/advanced-molecular-detection/php/about
2. Deharvengt SJ, Petersen LM, Jung HS, Tsongalis GJ. Nucleic acid analysis in the clinical laboratory. In: Clarke W, Marzinke MA, editors. Contemporary practice in clinical chemistry. 4th ed. New York: Academic Press; 2020. p. 215–34.
3. Carbo EC, Mourik K, Boers SA, Munnink BO, Nieuwenhuijse D, Jonges M, et al. A comparison of five Illumina, Ion Torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2. Eur J Clin Microbiol Infect Dis. 2023;42:701–13.
4. Cantu M, Morrison MA, Gagan J. Standardized comparison of different DNA sequencing platforms. Clin Chem. 2022;68:872–6.
5. Li Y, Rivers J, Mathis S, Li Z, Velusamy S, Nanduri SA, et al. Genomic surveillance of *Streptococcus pyogenes* strains causing invasive disease, United States, 2016–2017. Front Microbiol. 2020;11:1547.
6. Rao PS, Downie DL, David-Ferdon C, Beekmann SE, Santibanez S, Polgreen PM, et al. Pathogen-agnostic advanced molecular diagnostic testing for difficult-to-diagnose clinical syndromes-results of an emerging infections network survey of frontline us infectious disease clinicians, May 2023. Open Forum Infect Dis. 2024;11:ofae395. https://doi.org/10.1093/ofid/ofae395
7. Fida M, Khalil S, Abu Saleh O, Challener DW, Sohail MR, Yang JN, et al. Diagnostic value of 16S ribosomal RNA gene polymerase chain reaction/sanger sequencing in clinical practice. Clin Infect Dis. 2021;73:961–8. https://doi.org/10.1093/cid/ciab167
8. Janies D. Phylogenetic concepts and tools applied to epidemiologic investigations of infectious diseases. Microbiol Spectr. 2019;7:7.4.14. https://doi.org/10.1128/microbiolspec.AME-0006-2018
9. Association for Diagnostic and Laboratory Medicine. Next-generation sequencing bioinformatics pipelines. A guide to practical implementation for clinical laboratories. Mar 1, 2020 [cited 2024 Nov 20]. https://www.myadlm.org/CLN/Articles/2020/March/Next-Generation-Sequencing-Bioinformatics-Pipelines
10. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics. 2017;33:2580–2. https://doi.org/10.1093/bioinformatics/btx192
11. Engelthaler DM. Genomic surveillance and pathogen intelligence. Front Sci. 2024;2:1397048. https://doi.org/10.3389/fsci.2024.1397048
12. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011;364:730–9. https://doi.org/10.1056/NEJMoa1003176
13. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19:9–20. https://doi.org/10.1038/nrg.2017.88
14. Centers for Disease Control and Prevention, Office of Infectious Diseases, Board of Scientific Counselors. Teleconference of the board of scientific counselors, office of infectious diseases: December 6, 2018 [cited 2024 Nov 20]. https://stacks.cdc.gov/view/cdc/103777
15. Henao OL, Jones TF, Vugia DJ, Griffin PM; Foodborne Diseases Active Surveillance Network Workgroup.

Foodborne diseases active surveillance network—2 decades of achievements, 1996–2015. Emerg Infect Dis. 2015;21:1529–36. https://doi.org/10.3201/eid2109.150581
16. Friedman CR, Morfino RC, Ernst ET. Leveraging a strategic public–private partnership to launch an airport-based pathogen monitoring program to detect emerging health threats. Emerg Infect Dis. 2025;13:S35–38. https://doi.org/10.3201/eid3113.241407
17. Torres LM, Johnson J, Valentine A, Brezak A, Schneider EC, D'Angeli M, et al. Integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA. Emerg Infect Dis. 2025;13::S25–34. https://doi.org/10.3201/eid3113.241227
18. Lloyd T, Khan SM, Heaton D, Shemsu M, Varghese V, Graham J, et al. Genomic modeling of an outbreak of multidrug-resistant Shigella sonnei, California, USA, 2023–2024. Emerg Infect Dis. 2025;13::S98–102. https://doi.org/10.3201/eid3113.241307
19. Andersson P, Dougall S, Mercoulia K, Horan KA, Seemann T, Lacey JA, et al. Effects of decentralized sequencing on national Listeria monocytogenes genomic surveillance, Australia, 2016–2023. Emerg Infect Dis. 2025;13:S89–97. https://doi.org/10.3201/eid3113.241357
20. Gerner-Smidt P, Hise K, Kincaid J, Hunter S, Rolando S, Hyytiä-Trees E, et al.; Pulsenet Taskforce. PulseNet USA: a five-year update. Foodborne Pathog Dis. 2006;3:9–19. https://doi.org/10.1089/fpd.2006.3.9
21. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al.; FWD-NEXT Expert Panel. PulseNet international: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017;22:30544. https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544
22. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore sequencing as a rapidly deployable ebola outbreak tool. Emerg Infect Dis. 2016;22:331–4. https://doi.org/10.3201/eid2202.151796
23. Oster AM, Lyss SB, McClung RP, Watson M, Panneer N, Hernandez AL, et al. HIV cluster and outbreak detection and response: the science and experience. Am J Prev Med. 2021;61(Suppl 1):S130–42. https://doi.org/10.1016/j.amepre.2021.05.029
24. Wegrzyn RD, Appiah GD, Morfino R, Milford SR, Walker AT, Ernst ET, et al. Early detection of severe acute respiratory syndrome coronavirus 2 variants using traveler-based genomic surveillance at 4 US airports, September 2021–January 2022. Clin Infect Dis. 2023;76:e540–3. https://doi.org/10.1093/cid/ciac461
25. Aden TA, Blevins P, York SW, Rager S, Balachandran D, Hutson CL, et al. Rapid diagnostic testing for response to the monkeypox outbreak—laboratory response network, United States, May 17–June 30, 2022. MMWR Morb Mortal Wkly Rep. 2022;71:904–7. https://doi.org/10.15585/mmwr.mm7128e1
26. Centers for Disease Control and Prevention. COVID-19 genomic epidemiology toolkit. Mar 28, 2024 [cited 2024 Nov 20]. https://www.cdc.gov/advanced-molecular-detection/php/training/index.html

Address for correspondence: Jessica N. Ricaldi, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H24-11, Atlanta, GA 30329-4018, USA; email: jricaldi@cdc.gov

Strategies and Opportunities to Improve Community Health through Advanced Molecular Detection and Genomic Surveillance of Infectious Diseases

Jazmyn Moore, Ruth Sanon, Yury Khudyakov, Nathelia Barnes

Advanced molecular detection (AMD) refers to the integration of next-generation sequencing, epidemiologic, and bioinformatics data to drive public health actions. As new AMD technologies emerge, it is critical to ensure those methods are used in communities that are most affected by disease-induced illness and death. We describe strategies and opportunities for using AMD approaches to improve health in those communities, which include improving access to pathogen sequencing, increasing data linkages, and using pathogen sequencing for those diseases where sequencing technologies can provide the best health outcome. Such strategies can help address and prevent differences in health outcomes in various populations, such as rural and tribal communities, persons with underlying health issues, and other populations that experience higher risks for infectious disease.

Advanced molecular detection (AMD) is the integration of next-generation pathogen sequencing, epidemiologic, and bioinformatics data to enable disease identification, responses, and public health actions (*1*). Specimens collected through public health activities, such as routine surveillance and outbreak investigations, are used by laboratory scientists to perform genomic sequencing of pathogens. Bioinformaticians and data scientists analyze and link the large amounts of genetic information obtained from pathogen DNA sequencing with epidemiologic and other data to better elucidate the origin and characteristics of a pathogen, including its potential response to countermeasures such as antiviral medications. This information is then shared with epidemiologists who use the data to determine transmission patterns and to identify and stop outbreaks (*2*). Those methods have been used to detect and monitor pathogens that cause infectious diseases, including those pathogens disproportionately affecting certain populations, such as hepatitis C virus (HCV) (*3*), HIV (*4*), *Mycobacterium tuberculosis* (*5*), and more recently, SARS-CoV-2 (L. Lyu et al., unpub. data, https://doi.org/10.1101/2023.12.28.23300535).

Epidemiology, which drives public health efforts, is the scientific study involved in identifying which populations are affected by different health conditions and why (*6*). The COVID-19 pandemic highlighted differences in health outcomes across various populations and the need to address factors that caused those differences. Here, we discuss strategies and opportunities for improving community health through AMD and genomic surveillance of pathogens.

Strategy 1—Ensuring Access to AMD Technologies

The US public health system has consistently faced funding challenges. For example, among local health departments, those located in rural areas are often understaffed and underfunded compared with their urban and suburban counterparts, hindering their ability to perform the 10 Essential Public Health Services originally enumerated in 1994 and updated in 2020 (https://www.cdc.gov/public-health-gateway/php/about/index.html) (*7*). Because of a series of factors, including challenges with accessing care, socioeconomic status, and living and working conditions, persons living in rural areas can have more barriers to care and worse health outcomes.

The Centers for Disease Control and Prevention (CDC) provides funding to health departments in all US

Author affiliation: Centers for Disease Control and Prevention, Atlanta, Georgia, USA

DOI: https://doi.org/10.3201/eid3113.241408

states, territories, and freely associated states and in 7 large cities to increase each jurisdiction's capacity to build and integrate laboratory, bioinformatics, and genomic epidemiology technologies as part of their infectious disease prevention capabilities (*8*). Technical assistance for AMD technologies can be provided to areas with fewer resources to modernize disease investigation techniques and can promote rapid public health responses; AMD methods provide faster, cost-effective results compared with older sequencing methods (*2*).

Although pathogen genome sequencing and other AMD technologies can expedite critical public health advances, it is critical to ensure that medically underresourced communities, including rural areas, are not left behind when new health technologies are adopted. When areas and communities with more resources are able to access health interventions that are generally inaccessible to groups with less resources, then existing differences in health outcomes can increase.

Example—Providing Support to Understand SARS-CoV-2 Transmission in Rural Texas

One method to distribute the benefits of AMD technologies is through larger health departments providing sequencing support to smaller health departments and communities that have no on-site sequencing or bioinformatics capacity; this support would aid outbreak investigation and surveillance efforts. For example, the Houston Health Department in Houston, Texas, USA, and academic partners used CDC funds and technologies to perform phylogeographic inference (a method that enables scientists to map how pathogens spread through space and time) to determine SARS-CoV-2 transmission in a rural community. The investigators were able to determine that SARS-CoV-2 outbreaks in rural areas were driven by repeated introductions of the virus from urban centers (L. Lyu et al., unpub. data). Those findings can help guide public health actions, such as resource allocation, contact tracing, and prevention efforts.

Opportunity—Improving Data Linkages

Factors influencing infectious disease outcomes can be better determined by ensuring relevant accompanying data, such as social, environmental, and demographic variables, are collected and linked to samples. Those data can be analyzed alongside test results and can highlight patterns of infectious disease transmission and identify populations at increased risk for illness and death. Integrating those data elements into systems and analyses can be done by strengthening linkages between epidemiologists and laboratory scientists. CDC has previously emphasized the importance of data integration and has supported efforts to increase staffing of personnel who perform those functions, such as laboratory scientists, genomic epidemiologists, data scientists, and bioinformaticians.

Strategy 2—Considering Public Health Effects When Prioritizing Pathogens for AMD Technologies

When public health resources and capacity are limited, outbreaks and public health problems are often assigned different levels of priority for funding and human resources. A strategy to reduce differences in health outcomes is to use pathogen genomic sequencing for populations disproportionately affected by illness or death when prioritizing those pathogens. Examples of pathogens that disproportionately affect certain populations are HIV, HCV, and *M. tuberculosis*.

Example—HIV in Men Who Have Sex with Men

In the United States, HIV has historically and disproportionately affected men who have sex with men (MSM), a population that also experiences social stigma and discrimination. AMD has been used to elucidate HIV transmission patterns to inform prevention, response, and public health program efforts (*4*). For example, the Georgia Department of Public Health, Atlanta, Georgia, USA, a part of CDC's Pathogen Genomics Center of Excellence, has used molecular cluster detection in routine surveillance to better determine HIV transmission patterns. The investigation found that Hispanic/Latino MSM were disproportionately represented in new HIV infections (*4*). That information guided subsequent qualitative investigations that revealed gaps in HIV prevention service coverage and informed culturally responsive HIV prevention activities focused on Hispanic/Latino MSM.

Some networks of persons living with HIV have raised concerns about using genomic surveillance for HIV prevention (*9*). Concerns have focused on privacy, confidentiality, consent, stigma and institutional bias, and criminalization (*9*). When conducting genomic surveillance, particularly for infectious diseases such as HIV and mpox that have been stigmatized (*10*,*11*), CDC prioritizes ethical and credible practices and advises all funded health departments to also uphold those practices. When considering implementation, the benefits of genomic surveillance activities should outweigh the risks; risks and benefits should be clearly communicated to potentially affected populations through community and clinical partners at the point of care and during other opportunities to engage affected persons. In addition to community engagement, other ethical considerations, including

data use agreements, plans for data storage, and data destruction protocols, should be evaluated and incorporated into surveillance plans (*12*,*13*). CDC has supported a series of information technology features that address data security and limit the misuse of any related AMD data (*1*). Finally, it is critical to communicate the actions taken to safeguard data and avoid unintentionally discouraging persons from seeking care (*14*). When possible, a collaborative approach should be taken that involves community members and community-based organizations to ensure that surveillance and outbreak investigation activities are conducted in a way that does not cause harm.

Example—Understanding HCV Transmission in Rural Indiana

Hepatitis C is another infectious disease that disproportionately affects certain populations and for which AMD has been used to guide public health action. Since 2004, cases of acute HCV infection associated with injected drug use have increased; persons who inject drugs are at highest risk for infection (*3*). An investigation of HIV and HCV infections among persons in rural Indiana who injected drugs revealed that widespread circulation of numerous HCV strains occurred long before an HIV outbreak began (*15*). A cloud-based AMD toolkit, the Global Hepatitis Outbreak and Surveillance Technology, helped detect a large HCV transmission network, which enabled the HIV outbreak (*15*). The HCV transmission network data were used to inform the development of tailored public health interventions to efficiently interrupt HCV transmission among persons who inject drugs (*16*). Results from the investigation suggested HCV transmission could be used as a warning sign for eventual HIV transmission, and HIV surveillance among persons with HCV infection and those who inject drugs could help to rapidly identify and halt both HIV and HCV transmission. The interruption of HIV and HCV transmission has critical health and economic implications, particularly in rural settings where access to healthcare might be limited. The results also underscored the need to increase knowledge of and access to HCV testing and treatment services among persons who inject drugs to prevent further disease transmission.

Example—Tuberculosis in Non–US-Born Populations

Tuberculosis (TB) disproportionately affects persons with HIV, those who report substance use, non–US-born persons within the United States, and persons with certain medical conditions, such as diabetes (*17*). Whole-genome sequencing is crucial to determine TB transmission networks and to detect drug-resistant cases, which require special treatment protocols (*18*). In addition, CDC developed MicrobeTrace, a web-based AMD tool, that helps users visualize genomic relationships and epidemiologic links to help track outbreaks of tuberculosis and other diseases (*19*). In 2022, most (96%) isolates from culture-positive TB cases were sequenced; this information was routinely analyzed, and the output was made available to public health partners for programmatic follow-up, including allocation of resources for investigation and intervention for specific cases (*17*). AMD technologies play a critical role in the identification and treatment of TB in the United States and globally.

In addition to considering traditional pathogen attributes, such as transmissibility and disease severity, prioritizing pathogens that disproportionately affect certain populations can help reduce health disparities. Priorities should be determined by jurisdictions and monitored and regularly reevaluated to ensure that they adequately address the most pressing public health challenges.

Across various populations and pathogens, it is critical to address privacy concerns, appropriately inform affected communities of applicable findings, and to use findings to optimize prevention efforts. To address those considerations, the AMD program relies on strong relationships between laboratory scientists and epidemiologists in health departments. Those relationships enable strategic interventions, engagement with communities, and strong disease outbreak investigations that promote more efficient and effective public health action.

Strategy 3—Ensuring a Robust Public Health Workforce

Public health workforces having varied training, expertise, and backgrounds can better address the needs of the communities they serve (*20*). Workforce heterogeneity across various dimensions permits the inclusion of differing opinions and perspectives and can improve organizational success (*21*).

Example—CDC Fellowships

Through various training opportunities, CDC supports the development of highly qualified and talented persons across the public health workforce, including those supporting AMD. The CDC's Office of Advanced Molecular Detection, National Center for Emerging and Zoonotic Infectious Diseases, has promoted the use of fellowships and collaborations with nontraditional partners, such as universities and other academia, to increase awareness of AMD practices and expand the skill and expertise of the AMD

workforce in both laboratory and health department settings. The Office of Advanced Molecular Detection has also demonstrated a long-standing commitment to training a professionally diverse cohort of next generation public health laboratory scientists through bioinformatics and genomic epidemiology fellowships in partnership with the Association of Public Health Laboratories. Open to students with varied academic and sociodemographic backgrounds, those programs are critical for providing highly qualified talent to the public health workforce.

Opportunity—Training

Training on how to collaborate with colleagues of various backgrounds, including topics such as working across generations and backgrounds and effective communication, can help generate cohesiveness in the workforce. For epidemiologists, training topics of interest might include collecting and analyzing demographic data variables, precisely measuring health outcomes in emerging genomic applications across different communities, and working with communities. For laboratory scientists, training topics of interest might include sampling strategies, data quality and timeliness, and using electronic health records as tools to identify and reduce differences in health outcomes. For bioinformaticians, training topics of interest might include recognizing and minimizing bias in algorithm design, privacy, and ethics.

It is critical that members of the workforce understand their roles in improving public health and the importance of integrating AMD technologies and approaches to reduce disease occurrence. As we better train our workforce, opportunities exist to create more effective public health interventions, multipathogen assessments, and guidelines to monitor and manage complex pathogens that affect communities.

Moving Forward

Without special consideration when introducing and expanding innovative public health technologies, existing disparities in infectious disease risk can be unintentionally exacerbated, and new disparities can arise. Public health efforts should consider populations that are disproportionately affected by illness or death caused by infectious diseases and mitigate disparate health outcomes in those groups. This mitigation can be achieved by prioritizing AMD resources, such as for rural or medically underserved areas or for pathogens that disproportionately affect different communities. Ensuring that the public health workforce reflects the population that it serves can improve overall effectiveness of control and prevention efforts. An ever-evolving field, AMD is transforming public health activities, deepening our knowledge of disease transmission and pathogen resistance mechanisms, and helping to resolve outbreaks. Ensuring that AMD technologies are used to close gaps in disparate health outcomes and prevent new gaps from arising will require strategic planning and implementation, community engagement, continued monitoring, and sustained resources.

Acknowledgments

We thank Shantrice Jones, Anna Llewellyn, Renee Calanan, Cathy Lansdowne, and Amanda Raziano for their thoughtful comments.

About the Author

Ms. Moore is a health scientist in the Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. Her primary research interests focus on parasites and other infectious diseases and their social and environmental determinants.

References

1. Gwinn M, MacCannell DR, Khabbaz RF. Integrating advanced molecular technologies into public health. J Clin Microbiol. 2017;55:703–14. https://doi.org/10.1128/JCM.01967-16
2. Centers for Disease Control and Prevention. What is advanced molecular detection (AMD)? [cited 2024 Jun 10]. https://www.cdc.gov/advanced-molecular-detection/about/index.html
3. Zibbell JE, Asher AK, Patel RC, Kupronis B, Iqbal K, Ward JW, et al. Increases in acute hepatitis C virus infection related to a growing opioid epidemic and associated injection drug use, United States, 2004 to 2014. Am J Public Health. 2018;108:175–81. https://doi.org/10.2105/AJPH.2017.304132
4. Saldana C, Philpott DC, Mauck DE, Hershow RB, Garlow E, Gettings J, et al. Public health response to clusters of rapid HIV transmission among Hispanic or Latino gay, bisexual, and other men who have sex with men—Metropolitan Atlanta, Georgia, 2021–2022. MMWR Morb Mortal Wkly Rep. 2023;72:261–4. https://doi.org/10.15585/mmwr.mm7210a3
5. Li Y, Regan M, Swartwood NA, Barham T, Beeler Asay GR, Cohen T, et al. Disparities in tuberculosis incidence by race and ethnicity among the U.S.–born population in the United States, 2011 to 2021: an analysis of national disease registry data. Ann Intern Med. 2024;177:418–27. https://doi.org/10.7326/M23-2975
6. Coggon D, Rose G, Barker DJP. What is epidemiology? 1997 [cited 2024 Dec 4]. https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/1-what-epidemiology
7. Harris JK, Beatty K, Leider JP, Knudson A, Anderson BL, Meit M. The double disparity facing rural local health departments. Annu Rev Public Health. 2016;37:167–84. https://doi.org/10.1146/annurev-publhealth-031914-122755
8. Centers for Disease Control and Prevention. Advanced molecular detection (AMD). National investment maps [cited 2024 Jun 5]. https://www.cdc.gov/advanced-molecular-detection/php/investments/index.html

9. Molldrem S, Smith AKJ, McClelland A. Advancing dialogue about consent and molecular HIV surveillance in the United States: four proposals following a federal advisory panel's call for major reforms. Milbank Memorial Fund. 2023 [cited 2024 Jun 28]. https://www.milbank.org/quarterly/articles/advancing-dialogue-about-consent-and-molecular-hiv-surveillance-in-the-united-states-four-proposals-following-a-federal-advisory-panels-call-for-major-reforms
10. Williams DR. The health of men: structured inequalities and opportunities. Am J Public Health. 2003;93:724–31. https://doi.org/10.2105/AJPH.93.5.724
11. Birch L, Bindert A, Macias S, Luo E, Nwanah P, Green N, et al. When stigma, disclosure, and access to care collide: an ethical reflection of mpox vaccination outreach. Public Health Rep. 2024;139:379–84. https://doi.org/10.1177/00333549231201617
12. Johnson S, Parker M. Ethical challenges in pathogen sequencing: a systematic scoping review. Wellcome Open Res. 2020;5:119. https://doi.org/10.12688/wellcomeopenres.15806.1
13. Dawson L, Benbow N, Fletcher FE, Kassaye S, Killelea A, Latham SR, et al. Addressing ethical challenges in US-based HIV phylogenetic research. J Infect Dis. 2020;222:1997–2006. https://doi.org/10.1093/infdis/jiaa107
14. Romero-Severson E, Nasir A, Leitner T. What should health departments do with HIV sequence data? Viruses. 2020;12:1018. https://doi.org/10.3390/v12091018
15. Ramachandran S, Thai H, Forbi JC, Galang RR, Dimitrova Z, Xia GL, et al.; Hepatitis C Investigation Team. A large HCV transmission network enabled a fast-growing HIV outbreak in rural Indiana, 2015. EBioMedicine. 2018;37:374–81. https://doi.org/10.1016/j.ebiom.2018.10.007
16. Campo DS, Khudyakov Y. Intelligent Network DisRuption Analysis (INDRA): a targeted strategy for efficient interruption of hepatitis C transmissions. Infect Genet Evol. 2018;63:204–15. https://doi.org/10.1016/j.meegid.2018.05.028
17. Centers for Disease Control and Prevention. Reported tuberculosis in the United States, 2022 [cited 2024 Jul 1]. https://www.cdc.gov/tb/statistics/reports/2022/default.htm
18. Dookie N, Khan A, Padayatchi N, Naidoo K. Application of next generation sequencing for diagnosis and clinical management of drug-resistant tuberculosis: updates on recent developments in the field. Front Microbiol. 2022;13:775030. https://doi.org/10.3389/fmicb.2022.775030
19. Campbell EM, Boyles A, Shankar A, Kim J, Knyazev S, Cintron R, et al. MicrobeTrace: retooling molecular epidemiology for rapid public health response. PLOS Comput Biol. 2021;17:e1009300. https://doi.org/10.1371/journal.pcbi.1009300
20. Coronado F, Beck AJ, Shah G, Young JL, Sellers K, Leider JP. Understanding the dynamics of diversity in the public health workforce. J Public Health Manag Pract. 2020;26:389–92. https://doi.org/10.1097/PHH.0000000000001075
21. Bresman H, Edmondson AC. Exploring the relationship between team diversity, psychological safety and team performance: evidence from pharmaceutical drug development. Harvard Business School working paper no. 22-055, February 2022 [cited 2024 Jun 28]. https://www.hbs.edu/faculty/Pages/item.aspx?num=61993

Address for correspondence: Jazmyn Moore, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H24-3, Atlanta, GA 30329-4018, USA; email: vin2@cdc.gov

The Next-Generation Sequencing Quality Initiative and Challenges in Clinical and Public Health Laboratories

Blake Cherney, Ariel Diaz, Camby Chavis, Christopher Ghattas, Diana Evans, Diego Arambula, Heather Stang, on behalf of the Next-Generation Sequencing Quality Initiative

The Next-Generation Sequencing (NGS) Quality Initiative addresses laboratory challenges faced when performing NGS by developing tools and resources to build a robust quality management system. Here, we illustrate how those products support laboratories in navigating complex regulatory environments and quality-related challenges while implementing NGS effectively in an evolving landscape.

Next-generation sequencing (NGS) technology has improved with the introduction of new platforms, updated chemistries, advancements in bioinformatic analyses, and computational innovations. As targeted and agnostic (e.g., metagenomic) sequencing approaches have been introduced, validation of NGS assays has increased in complexity, mostly because of sample type variability, stringent quality control criteria, intricate library preparation, and evolving bioinformatics tools (*1*–*3*). Complexity increases when validations are governed by the Clinical Laboratory Improvement Amendments of 1988 (CLIA) (*4*).

Performing NGS requires an experienced workforce to generate high-quality results. Retaining proficient personnel can be a substantial obstacle because of the unique and specialized knowledge required of them, which in turn increases costs for adequate staff compensation. Akkari et al. (*5*) found that some testing personnel held their positions for <4 years on average. In 2021, the Association of Public Health Laboratories (APHL) reported that 30% of surveyed public health laboratory staff indicated an intent to leave the workforce within the next 5 years (*6*). Additional barriers may arise when hiring and qualifying personnel under regulations such as CLIA and state hiring statutes (*4*).

The Study

In an effort to help clinical and public health laboratories, the Centers for Disease Control and Prevention and APHL collaborated to form the Next-Generation Sequencing Quality Initiative (NGS QI; https://www.cdc.gov/lab-quality/php/ngs-quality-initiative/index.html) to address challenges associated with implementing NGS in clinical and public health settings. NGS QI staff performed an initial assessment of needs and identified common challenges associated with personnel management, equipment management, and process management across NGS laboratories (Figure 1; Appendix, https://wwwnc.cdc.gov/EID/article/31/13/24-1175-App1.pdf). Among those challenges was a lack of high-quality guidance documents and standard operating procedures (SOPs) (*7*). The NGS QI found that laboratories were developing in-house resources that, although similar in content, contained varying levels of detail (*7*,*8*). The Initiative provides publicly available tools that can be used regardless of platform, agent, or application and that satisfy the needs of laboratories whether they are implementing NGS initially or refining existing workflows (Figure 2; Appendix).

A quality management system (QMS) enables continual improvement and proper document

Author affiliations: Centers for Disease Control and Prevention, Atlanta, Georgia, USA (B. Cherney, A. Diaz, C. Chavis, C. Gattas, D. Evans, D. Arambula, H. Stang); Booz Allen Hamilton Inc., Atlanta (A. Diaz, C. Chavis, C. Ghattas, D. Evans)

DOI: https://doi.org/10.3201/eid3113.241175

management in laboratories. All existing NGS QI products undergo a review period every 3 years to ensure they remain up to date relative to current technology, standard practice of care, and applicable changes in regulations (Figure 3). Previous internal surveys and workgroups identified validation tools as a high-priority task to assist laboratories in ensuring compliance with quality and regulatory standards (*9*). In response, the NGS QI created the Pathway to Quality-Focused Testing; although it is not a standalone document for developing an NGS-specific QMS, it complements other published tools and resources that address relevant topics in depth. Other recently released documents are tailored to validation and bioinformatic development, ranging from straightforward guidance to fillable templates. Since its establishment, the NGS QI has seen an increasing interest in NGS method validation because most clinical and public health laboratories are already using or are beginning to implement NGS within their workflows. For that reason, many of the NGS QI's resources assist with NGS assays for validation. The most widely used documents offered by the NGS QI are QMS Assessment Tool, Identifying and Monitoring NGS Key Performance Indicators SOP, NGS Method Validation Plan, and the NGS Method Validation SOP (Table). For example, the use of the Validation Plan document guided Orange County Public Health Laboratory (Santa Ana, California, USA) in generating a standard template containing NGS-related metrics, thereby reducing the burden on laboratories seeking to perform a validation (*10*).

The NGS QI develops and crosswalks its documents with regulatory, accreditation, and professional bodies (e.g., the US Food and Drug Administration [FDA], Centers for Medicare and Medicaid Services, and College of American Pathologists) to ensure they provide current and compliant guidance on Quality System Essentials (QSE) (Figure

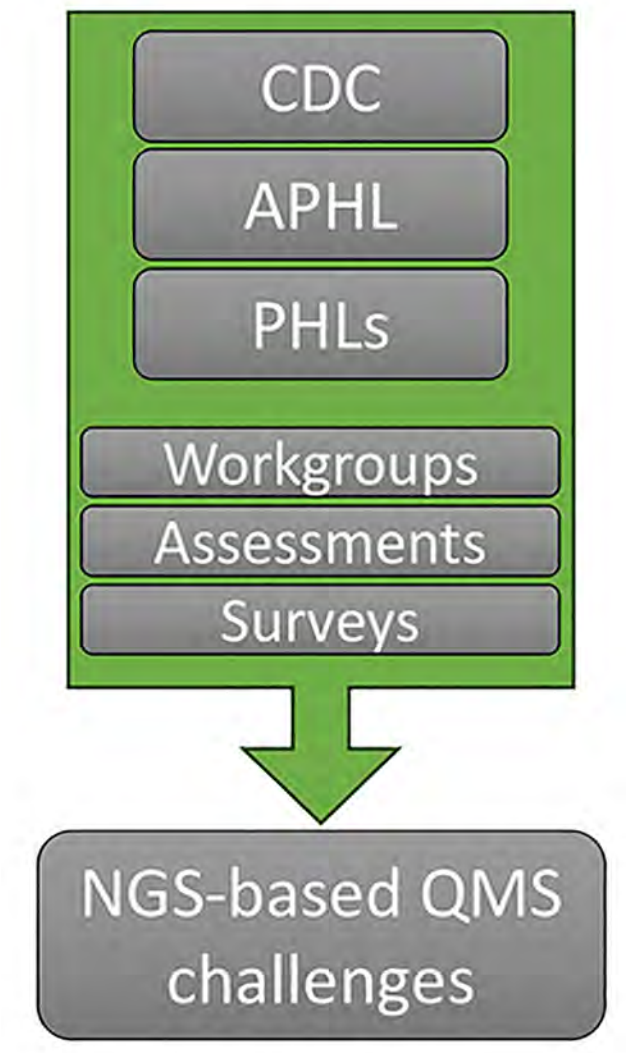


Figure 1. Partners and aim of the Next-Generation Sequencing Quality Initiative. Partners collaborate to identify and address NGS-specific challenges through development of a QMS. APHL, Association of Public Health Laboratories; CDC, Centers for Disease Control and Prevention; NGS, next-generation sequencing; PHL, public health laboratories; QMS, quality management system.

Figure 2. Depiction of Clinical and Laboratory Standards Institute's 12 QSEs as building blocks for tools and documents available on the website (https://www.cdc.gov/lab-quality/php/ngs-quality-initiative/qms-tools-resources.html). QSE, Quality System Essentials.

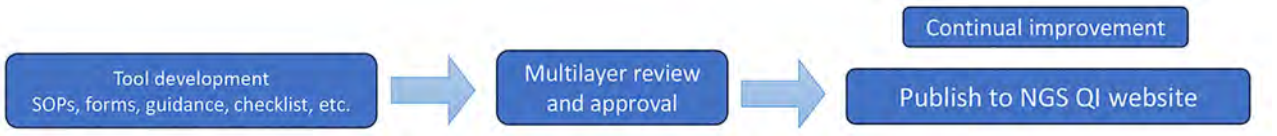


Figure 3. Depiction of the review and approval process for tools and documents published on the QI website (https://www.cdc.gov/lab-quality/php/ngs-quality-initiative/qms-tools-resources.html). NGS, Next-Generation Sequencing; QI, quality initiative; SOP, standard operating procedures.

3; Appendix) (*4*,*11*). To support challenges associated with staff training and competency assessment, the NGS QI has published 25 tools for the personnel management QSE (e.g., Bioinformatics Employee Training SOP) and 4 tools for the assessments QSE (e.g., Bioinformatician Competency Assessment SOP); the Initiative also works with partners to host or participate in online trainings (Appendix). A QMS must be able to adapt to an ever-changing environment, including improvements in software and chemistry, which can affect how validated NGS assays, pipelines, and results are developed, performed, and reported. Even as laboratories become more familiar with guidance documents and standard practices, there are other challenges: information technology cost, curated databases, developing standards, and newer platforms. For example, new kit chemistries from Oxford Nanopore Technologies (https://nanoporetech.com) that use CRISPR for targeted sequencing and improved basecaller algorithms using artificial intelligence, machine learning, and duplex data lead to increased accuracy (*2*). Other emerging platforms, such as Element Biosciences (https://www.elementbiosciences.com), also show increasing accuracies at Q40 with lower costs, which might encourage transition from older platforms to new platforms and chemistries (*12*). Although modernizing is beneficial, transitioning to new platforms requires additional resources and time to revalidate NGS workflows. Changes in policies and regulations can also create confusion and barriers for laboratories (*13*).

Table. Most frequently downloaded documents of the 113 posted on the Next-Generation Sequencing Quality Initiative website during January–June 2024*

Document	No. views
QMS Assessment Tool	548
Identifying and Monitoring NGS Key Performance Indicators SOP	410
NGS Method Validation Plan	410
NGS Method Validation SOP	199

*Total of 11,790 visits to website (https://www.cdc.gov/lab-quality/php/ngs-quality-initiative/qms-tools-resources.html) during July 2023–July 2024. NGS, next-generation sequencing; QMS, quality management system; SOP, standard operating procedure.

Conclusion

NGS is complex, and workflows often differ among specialties and sequencing approaches. Despite advancements in guidance, practice, and technology, NGS validation remains challenging. The NGS QI generates resources that are written broadly enough to benefit an array of laboratories and methods. Limitations in regulatory authority often prevent the development of prescriptive guidance. Although NGS QI's tools are applicable to most platforms and applications, the laboratories using each product may have additional quality assurance considerations. To keep up with evolving practices, the Initiative conducts cyclic review and performs regular or ad hoc (if significant changes warrant) updates. However, the rapid pace of changes in policy and technology means that regular updates do not always resolve challenges. Although completing a method validation or revalidation is resource intensive, it is important that, once validated, the entire workflow is locked down (*13*). Evaluating technological advancements is necessary; the shifts in testing needs for patient populations, the evolving public-health applications, and the ability to modify sequencing workflows depend heavily on institutional practices and regulatory bodies (i.e., local, state, federal, and accrediting organizations). Those factors indicate the need for bespoke practices among entities. On the path of creating high-quality, reproducible, and reliable results, obstacles will continuously arise. It is imperative to use a balanced review process to implement changes to sequencing workflows and stay current relative to the latest advancements, best practices, and regulatory requirements, which may not always align for practical implementation. As the pool of NGS QI's users continues to grow (Figure 4), the Initiative will continue adapting to needs by creating supporting documents and trainings focused on the application of the NGS QI documents, tools for emerging challenges (e.g., validation of machine learning algorithms, agnostic pathogen detection), curated databases, clinical decision tools, and frontline diagnostics for clinical and public health laboratories.

Figure 4. Trends in visits to the Next-Generation Sequencing Quality Initiative website (https://www.cdc.gov/lab-quality/php/ngs-quality-initiative/qms-tools-resources.html), by quarter, 2021–2024.

CDC/APHL NGS QUI members and partners have provided feedback or insights into quality practices and contributed to development of tools and resources. Members and partners include participants of the CDC NGS Quality Workgroup; Technical Coordinating Committee for the Advanced Molecular Detection Platform and NGS QI; CDC's Office of Advanced Molecular Detection; subject matter experts from CDC centers, institutes, and offices; and many state and local public health laboratories.

This work was supported by CDC's Office of Advanced Molecular Detection (Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases).

CDC internal chatbot (GPT 4) was used to check for grammar, synonyms, and punctuation.

About the Author

Mr. Cherney is part of the CDC's Quality and Safety Systems Branch, Division of Laboratory Systems, Office of Laboratory Systems and Response, and is the project lead for the NGS QI. His research interests include the evaluation of next-generation sequencing platforms, rapid detection and genetic engineering, structural variation, and antimicrobial resistance.

References

1. MacCannell D. Next generation sequencing in clinical and public health microbiology. Clin Microbiol Newsl. 2016;38:169–76. https://doi.org/10.1016/j.clinmicnews.2016.10.001
2. Bogaerts B, Van den Bossche A, Verhaegen B, Delbrassinne L, Mattheus W, Nouws S, et al. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. J Clin Microbiol. 2024;62:e0157623. https://doi.org/10.1128/jcm.01576-23
3. Arslan S, Garcia FJ, Guo M, Kellinger MW, Kruglyak S, LeVieux JA, et al. Sequencing by avidity enables high accuracy with low reagent consumption. Nat Biotechnol. 2024;42:132–8. https://doi.org/10.1038/s41587-023-01750-7
4. Laboratory requirements. 42 C.F.R. part 493. 1990. [cited 2024 Jun 19]. https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-G/part-493
5. Akkari Y, Dobin S, Best RG, Leung ML. Exploring current challenges in the technologist workforce of clinical genomics laboratories. Genet Med Open. 2023;1:100806. https://doi.org/10.1016/j.gimo.2023.100806
6. Association of Public Health Laboratories. Support public health laboratories. 2020 [cited 2024 Jun 17]. https://www.aphl.org/aboutAPHL/publications/Documents/WORK-2020-PHL-Advocacy.pdf
7. Hutchins RJ, Phan KL, Saboor A, Miller JD, Muehlenbachs A. Practical guidance to implementing quality management systems in public health laboratories performing next-generation sequencing: personnel, equipment, and process management (phase 1). J Clin Microbiol. 2019;57. 10.1128/jcm.00261-19
8. Clinical and Laboratory Standards Institute. A quality management system model for laboratory services. 5th QMS01. Wayne (PA): Clinical and Laboratory Standards Institute; 2019.
9. Mahoney S. A little help from the NGS Quality Initiative: validation of carbapenem-resistant *Acinetobacter baumannii* (CRAB) whole genome sequencing. Lab Matters. 2024;27–27.
10. Centers for Medicare & Medicaid Services. Clinical Laboratory Improvement Amendments (CLIA). 2024 [cited 2024 Jun 24]. https://www.cms.gov/medicare/quality/clinical-laboratory-improvement-amendments
11. Quality system regulation 21 C.F.R. part 820. 2024 [cited 2024 Jun 24]. https://www.ecfr.gov/current/title-21/chapter-I/subchapter-H/part-820
12. Food and Drug Administration. Laboratory developed tests. 2024 [cited 2024 Jun 24]. https://www.fda.gov/medical-devices/in-vitro-diagnostics/laboratory-developed-tests
13. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. J Clin Microbiol. 2016;54:2857–65. https://doi.org/10.1128/JCM.00949-16

Address for correspondence: Heather Stang or Diego Arambula, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop V24-3, Atlanta, GA 30329-4018, USA; email: btg0@cdc.gov or ouo4@cdc.gov

Advantages of Software Containerization in Public Health Infectious Disease Genomic Surveillance

Kelsey R. Florek, Erin L. Young, Kutluhan Incekara, Kevin G. Libuit, Curtis J. Kapsak

Bioinformatic software containerization, the process of packaging software that encapsulates an application together with all necessary dependencies to simplify installation and use, has improved the deployment and management of next-generation sequencing workflows in both clinical and public health laboratories. Containers have increased next-generation sequencing workflow reproducibility and broadened their usage across different laboratories. We highlight the value of the State Public Health Bioinformatics community's containerized software repository during the COVID-19 pandemic.

Since 2013, an increasing number of clinical and public health laboratories have adopted next-generation sequencing (NGS)–based assays (*1*). The genomic data generated from NGS assays often require a complex analysis workflow built from a variety of bioinformatic software. Because of the wide range of software used in workflows, challenges can arise when installing software and dependencies, increasing the time and cost of deploying NGS-based tests. Some software distribution tools, such as Conda (https://anaconda.org/anaconda/conda), provide a means to manage the software environment but do not provide an isolated and identical environment and often require additional steps of installing databases and dependencies. The emergence of software container applications has greatly improved NGS workflows by encapsulating software and dependencies into publicly available containers, providing a robust and controlled bioinformatic software solution (*2*,*3*). Ultimately, software containerization simplifies the process of creating and adopting NGS workflows and reduces maintenance issues and downtime, saving time and laboratory resources (*4*).

Author affiliations: Wisconsin State Laboratory of Hygiene, Madison, Wisconsin, USA (K.R. Florek); State of Utah Department of Health and Human Services, Utah Public Health Laboratory, Taylorsville, Utah, USA (E.L. Young); Connecticut Department of Public Health, Katherine A. Kelley State Public Health Laboratory, Rocky Hill, Connecticut, USA (K. Incekara); Theiagen Genomics, Highlands Ranch, Colorado, USA (K.G. Libuit, C.J. Kapsak)

DOI: https://doi.org/10.3201/eid3113.241363

Software Containerization

A container is a packaged unit of software that encapsulates an application with all necessary dependencies (*5*). The containers themselves are ephemeral and isolated from both the host environment and other containers. Those qualities ensure that changes occurring within a container are not shared across other containers and that they are not affected by any changes that might occur outside of the container. In that way, using containers increases reliability and reproducibility, even when multiple containers of the same software are running concurrently on the same system.

Software containerization approaches rely on a container engine, such as Docker (https://www.docker.com) or Apptainer (formerly Singularity) (https://apptainer.org), which oversees the tasks associated with creating and using these discrete software environments (*6*). Container images are created using a build file that provides a base image (an initial state for the environment) and stepwise instructions for bundling software code and dependencies. Once the image has been built, it can be used locally or hosted on a public resource where the container can be released and made publicly accessible. Those hosted container images can be downloaded and run on a variety of computing infrastructures, from laptops to high-performance computing (HPC) clusters, greatly simplifying the process of development to deployment (*7*).

Recognizing the value of containers, the State Public Health Bioinformatics (StaPH-B) community developed and actively maintains a repository of containerized software (https://github.com/StaPH-B/docker-builds). The StaPH-B docker-builds repository is a collection of dockerfiles used to build containers of bioinformatics software that are commonly used in public health genomic workflows. Those containers are publicly available and hosted at both Docker Hub (https://hub.docker.com/u/staphb) and Quay.io (https://quay.io/organization/staphb); total monthly downloads from Docker Hub range from 100,000 to >700,000 (Figure 1). The emphasis on quality assurance and quality control in the StaPH-B container repository separates the project from other container repositories. Anyone can submit a new or updated container to the StaPH-B repository following the contributing instructions on the GitHub repository (https://staphb.org/docker-builds/contribute). Each submission includes a test that confirms the functionality of the container image, and all pull requests are reviewed carefully by the repository maintainers to ensure that the software meets the needs and expectations of public health laboratories.

Containerization in Genomic Workflows

The reliable and reproducible nature of containers is advantageous for public health and clinical laboratories. Deploying bioinformatic analytical workflows and software in a clinical and public health setting is challenging and requires managing computer systems with capacity for scientific computing workloads under regulatory oversight (*8*). Software containerization approaches provide key advantages to genomic workflows.

Reproducibility

Containers provide isolated environments that enable strict control of both software versions and software dependencies (*9*). Containers follow a naming convention that is structured as <public-registry-name>/<organization>/<software>:<tag>, which enables quick identification of the software and a specific tag that often indicates version and owner information. In addition, some container engines enable containers to be referenced by a digest, which is a unique and immutable identifier of the container. Referencing containers by the digest ensures the container environment is unchanged and helps support regulatory compliance.

Isolation

Containers avoid version conflicts and enable multiple versions of software or software dependencies to be used on the same system. In addition, containers provide a separation of data in the container and data on the host system (*6*).

Replicable and Portable

Containerized software is easily distributed, reproducible, and readily scalable in Cloud or HPC (*10*).

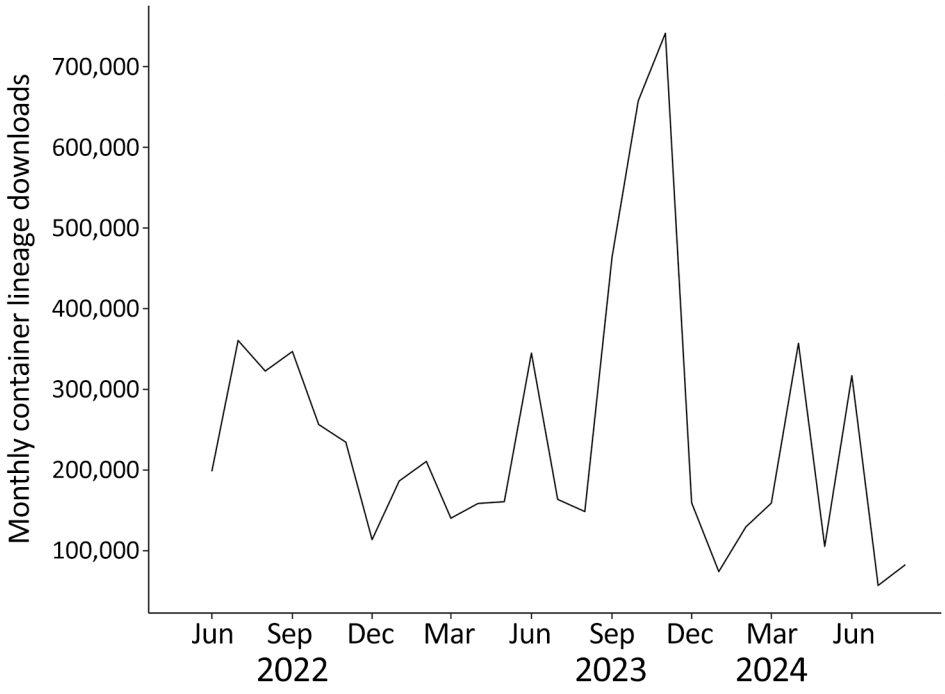


Figure 1. Monthly container image downloads across all State Public Health Bioinformatics containers hosted in Docker Hub (https://hub.docker.com/u/staphb) in study of advantages of software containerization in public health infectious disease genomic surveillance.

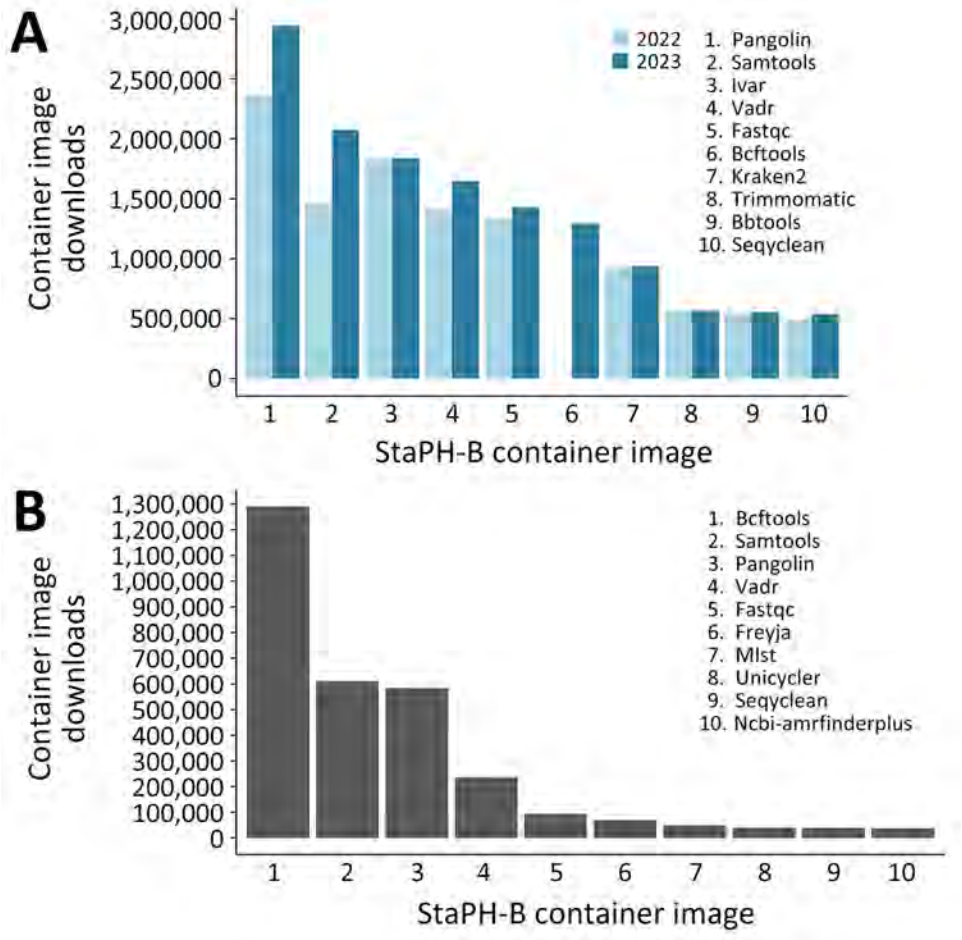


Figure 2. Top container image downloads across State Public Health Bioinformatics containers hosted in Docker Hub (https://hub.docker.com/u/staphb) in study of advantages of software containerization in public health infectious disease genomic surveillance. A) Top 10 overall container image downloads and their downloads in 2022 and 2023. B) Top 10 container images with the largest increase in downloads from 2022 to 2023.

Small databases or reference data can also be incorporated into containers, ensuring the database is maintained and controlled alongside the software.

Practical Application of Containers during the COVID-19 Pandemic

During the beginning of the COVID-19 pandemic, initiatives were created to enhance genomic surveillance (*11*), which led to many clinical and public health laboratories rapidly developing NGS-based surveillance testing to support variant detection. During that time, the StaPH-B docker project played a critical role in supporting the advancement of bioinformatic workflows by providing a reliable resource for SARS-CoV-2 sequence analysis containers. The StaPH-B container of Pangolin, a critical SARS-CoV-2 lineage tool, saw an addition of >500,000 downloads, from 2,360,607 downloads in 2022 to 2,944,235 downloads in 2023 (Figure 2) (*12*). Similarly, the StaPH-B container of iVar, a tool for amplicon-based viral sequencing, has been downloaded >1,836,046 times since it was added in 2020 (Figure 2) (*13*). The scale of downloads from this repository highlights the use of these programs in bioinformatic workflows and the effect of the project during the pandemic.

One of the largest effects on laboratories was a time savings in software installation and management, which grows substantially when scaling workflows to run in an HPC or Cloud environment. Installing Pangolin and its dependencies using Conda takes approximately 3 minutes on a new system, whereas downloading and running the StaPH-B container of Pangolin takes only 1 minute. Similarly, installing iVar takes 4.5 minutes, whereas downloading and running the StaPH-B container of iVar takes only 4 seconds. When using a distributed computing environment such as an HPC or Cloud environment, those time savings become a critical efficiency and cost savings. Those times also assume no installation issues or conflicts with other software and dependencies, a common occurrence with bioinformatics software that can stretch software deployment into days or weeks.

In summary, software containerization has rapidly changed the landscape of bioinformatics over the past 10 years and will continue to be a critical

component of public health genomic workflows for the future. The StaPH-B docker project represents a community-based effort providing a valuable resource of standardized bioinformatic tools, supporting reproducibility and regulatory compliance.

About the Author

Dr. Florek is a senior genomics and data scientist and leads the bioinformatics team in the Communicable Disease Division at the Wisconsin State Laboratory of Hygiene. Dr. Florek's work aims to democratize access to genomic data analytics and enhance the application of genomic data in public health. As a subject matter expert, Dr. Florek works closely with the Association of Public Health Laboratories and the CDC Advanced Molecular Detection program to enhance the public health workforce and improve access to actionable genomic data.

References

1. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen genomics in public health. N Engl J Med. 2019;381:2569–80. https://doi.org/10.1056/NEJMsr1813907
2. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics. 2017;33:2580–2. https://doi.org/10.1093/bioinformatics/btx192
3. Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A, Barton MD. Bioboxes: standardised containers for interchangeable bioinformatics software. Gigascience. 2015;4:47. https://doi.org/10.1186/s13742-015-0087-0
4. Tam JZ, Chua A, Gallagher A, Omene D, Okun D, DiFranzo D, et al. A containerization framework for bioinformatics software to advance scalability, portability, and maintainability. In: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. New York, NY, USA: Association for Computing Machinery; 2023. p. 1–5.
5. Kadri S, Sboner A, Sigaras A, Roy S. Containers in bioinformatics: applications, practical considerations, and best practices in molecular pathology. J Mol Diagn. 2022;24:442–54. https://doi.org/10.1016/j.jmoldx.2022.01.006
6. Alser M, Lawlor B, Abdill RJ, Waymost S, Ayyala R, Rajkumar N, et al. Packaging and containerization of computational methods. Nat Protoc. 2024;19:2529–39. https://doi.org/10.1038/s41596-024-00986-0
7. Boettiger C. An introduction to Docker for reproducible research. Oper Syst Rev. 2015;49:71–9. https://doi.org/10.1145/2723872.2723882
8. Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J, et al. Next-generation sequencing informatics: challenges and strategies for implementation in a clinical environment. Arch Pathol Lab Med. 2016;140:958–75. https://doi.org/10.5858/arpa.2015-0507-RA
9. Gruening B, Sallou O, Moreno P, da Veiga Leprevost F, Ménager H, Søndergaard D, et al.; BioContainers Community. Recommendations for the packaging and containerizing of bioinformatics software. F1000 Res. 2018;7:742. https://doi.org/10.12688/f1000research.15140.1
10. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. 2020;26:832–41. https://doi.org/10.1038/s41591-020-0935-z
11. National Academies of Sciences, Engineering, and Medicine; Division on Earth and Life Studies; Board on Life Sciences; Health and Medicine Division; Board on Health Sciences Policy; Committee on Data Needs to Monitor the Evolution of SARS-CoV-2. Genomic epidemiology data infrastructure needs for SARS-CoV-2: modernizing pandemic response strategies. Washington: National Academies Press; 2020.
12. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403–7. https://doi.org/10.1038/s41564-020-0770-5
13. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019;20:8. https://doi.org/10.1186/s13059-018-1618-7

Address for correspondence: Kelsey R. Florek, Wisconsin State Laboratory of Hygiene, 2601 Agriculture Dr, Madison, WI 53718, USA; email: kelsey.florek@slh.wisc.edu

Genomic Epidemiology for Estimating Pathogen Burden in a Population

W. Tanner Porter, David M. Engelthaler, Crystal M. Hepp

The role of genomics in outbreak response and pathogen surveillance has expanded and ushered in the age of pathogen intelligence. Genomic surveillance enables detection and monitoring of novel pathogens; case clusters; and markers of virulence, antimicrobial resistance, and immune escape. We can leverage pathogen genomic diversity to estimate total pathogen burden in populations and environments, which was previously challenging because of unreliable data. Pathogen genomics might allow pathogen burdens to be estimated by sequencing even a small percentage of cases. Deeper genomic epidemiology analyses require multidisciplinary collaboration to ensure accurate and actionable real-time pathogen intelligence.

The SARS-CoV-2 pandemic highlighted the importance and possibility of genomic surveillance for outbreak response and pathogen surveillance. The massive success of global SARS-CoV-2 sequencing projects, producing >17 million genomes (*1*), reflects the collective effort and dedication of the scientific and public health communities. That unparalleled dataset enabled identification of viral variants and case clusters, tracking of viral movements, and enhanced understanding of evolutionary principles. Driven by increased access to sequencing and analytic technologies, the age of pathogen intelligence has begun (*2*). That concept involves translating pathogen genomics into actionable knowledge, such as detecting outbreak clusters for transmission intervention (*3*,*4*), antimicrobial resistance markers to guide treatment (*5*), novel variants to prepare for new pandemic waves (*6*), and characterization of the evolutionary pathway of pathogens to identify mitigation opportunities (*7*). Although those applications are invaluable, modern genomics and computing power enable further expansion of genomic surveillance and the creation of large-scale pathogen intelligence.

Infectious disease trend estimation could benefit from large-scale pathogen intelligence. Case counts are often confounded by care-seeking behaviors, especially when persons experience mild illness or are asymptomatic or when diagnosis is challenging (e.g., environmental fungal diseases, such as coccidioidomycosis), leading to substantial underreporting. Statistical models can estimate undetected cases by using outside data to account for underreporting or nonreportable etiologies. However, accounting for underreporting is not a simple problem, especially when considering the role that social inequity has on reporting across space and time.

Pathogen tracking in wastewater was invaluable for proactively estimating case trends and tracking variants in near real-time across the SARS-CoV-2 pandemic. Although initially applied to sewersheds in London for tracking *Salmonella enterica* in the 1940s (*8*), the methodology continues to be extended to various pathogens. For example, wastewater surveillance for enterovirus D68, a nonreportable infection in the absence of acute flaccid paralysis, was successfully done in urban and rural communities and congregate living settings in the latter half of 2022 (D.E. Erickson et al., unpub. data, https://www.medrxiv.org/content/10.1101/2023.11.20.23297677v2). Knowledge of community-based trends for enterovirus D68 and other respiratory viruses could assist in mitigating potential albuterol shortages driven by viral-induced asthma exacerbations in children. However, wastewater surveillance is not a universal solution because accurate tracking has been less successful for organisms that are minimally shed through the gastrointestinal and urinary tracts or are highly susceptible to degradation, which results in a suboptimal genomic signal.

Author affiliations: Translational Genomics Research Institute, Flagstaff, Arizona, USA (W.T. Porter, D.M. Engelthaler, C.M. Hepp); Arizona State University, Phoenix, Arizona, USA (D.M. Engelthaler); Northern Arizona University, Flagstaff (C.M. Hepp)

DOI: https://doi.org/10.3201/eid3113.241203

With increased access to sequencing data, we can expand the possibilities of pathogen intelligence and usher in a second wave of genomic epidemiology. One promising method is phylodynamics, which involves leveraging pathogen genomic diversity and estimating coalescent rates to estimate disease trends (*9*). For example, our team worked with a remote Apache community in Arizona to track a largely isolated SARS-CoV-2 outbreak in 2020 that had a public health response driven by near-complete community sampling (*4*). Linear regression showed that genomically derived effective population size estimates from 36% of cases with sequenced genomes explained 86% of the variation in total case counts over time. However, we are investigating the role that sampling bias might have had on that correlation. Nonetheless, using phylodynamic methods to estimate disease burden could be invaluable for disease surveillance, enabling targeted and cost-effective programs that use remnant or prospective samples to estimate real-time disease dynamics, on the order of days or weeks, for pathogens that measurably evolve on those timescales (*10*). The genomic, public health, and bioinformatic communities must unite to clarify how we can routinely translate pathogen genomic signals into informative transmission trends and actionable insight.

At their core, phylodynamic estimations assume that, over time, pathogens accrue mutations at a consistent rate, which enables estimation of the evolutionary trajectory and rate of coalescence. That principle defines a theoretical minimum evolutionary rate combined with genome size or sequenced region relative to a pathogen's generation time. Previously, phylodynamic estimations were primarily confined to viral systems (*11*), where higher mutation rates, short replication periods, and large populations drive faster evolution. However, modern sequencing technologies provide larger sequenced regions, so those techniques have been used in bacterial systems (*12*) and will likely continue to expand to nonviral organisms.

In addition to evolutionary rates, the pathogen system is a critical consideration for phylodynamic inferences. In the simplest case, direct and successive human-to-human transmission enables phylodynamic estimates to be directly relatable to human disease trends (*10*,*13*); however, that model is complicated by pathogen introductions into populations and long-term infections. For pathogens with sylvatic cycles, phylodynamic estimates from nonhuman sources (e.g., vectors) reflect environmental population trends and can inform public health risks.

Sampling schemes must be considered because variations across space and time are unavoidable in most surveillance programs. Elucidating how that variation affects phylodynamic inferences and identifying optimal sampling strategies are critical for the larger community. Finally, numerous phylogenetic-based statistical models exist to conduct those analyses (*10*,*13*,*14*); however, our knowledge of how those programs perform on potentially biased or nonrepresentative datasets is limited. In addition to accuracy, computational efficiency and sustainability should be considered as genomic datasets continue to grow and require accurate and fast inferences to provide actionable insights. Large-scale multipathogen investigations are needed to compare the computational complexity, sensitivity, and specificity of phylodynamic estimates across sampling schemes, including genomic sequence subsampling and the creation of periods with increased or decreased sequencing efforts. Those analyses should benchmark findings across several phylogenetic-based statistical models and compare results to existing measures, including statistically modeled cases, because those analyses will enable the scientific and public health communities to precisely identify when phylodynamic inferences provide actionable intelligence.

In summary, genomic epidemiology will continue to transform the public health and outbreak response landscape and highlight the advantages of pathogen intelligence gathering. We have the ability and responsibility to further apply genomic principles to the public health world. That expansion of principles should involve well-characterized methods, which requires applied multidisciplinary investigations across pathogen systems and integration of real-world biases into their assessments.

Acknowledgments

We recognize the remarkable public health efforts of the White Mountain Apache Tribe during the COVID-19 pandemic, particularly the dedication of the community health representatives and the staff at Whiteriver Indian Hospital, Whiteriver, Arizona, USA, who worked tirelessly to ensure the tribal community's safety and well-being.

D.M.E. was funded by the City of Phoenix (award no. SLFRP1962), and C.M.H. was funded by the US Centers for Disease Control and Prevention (award nos. U01CK000649 and 75D30121C11191).

About the Author

Mr. Porter is a research associate at the Translational Genomics Research Institute's Pathogen & Microbiome Division. His research focuses on utilizing genomics to elucidate how pathogens move across space and time.

References

1. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017;1:33–46. https://doi.org/10.1002/gch2.1018
2. Engelthaler DM. Genomic surveillance and pathogen intelligence. Front Sci. 2024;2:1397048. https://doi.org/10.3389/fsci.2024.1397048
3. Sundermann AJ, Chen J, Kumar P, Ayres AM, Cho ST, Ezeonwuka C, et al. Whole-genome sequencing surveillance and machine learning of the electronic health record for enhanced healthcare outbreak detection. Clin Infect Dis. 2022;75:476–82. https://doi.org/10.1093/cid/ciab946
4. Bowers JR, Yaglom HD, Hepp CM, Pfeiffer A, Jasso-Selles D, Bratsch N, et al. Unique genomic epidemiology of COVID-19 in the White Mountain Apache Tribe, April to August 2020, Arizona. MSphere. 2023;8:e0065922. https://doi.org/10.1128/msphere.00659-22
5. Bowers JR, Lemmer D, Sahl JW, Pearson T, Driebe EM, Wojack B, et al. KlebSeq, a diagnostic tool for surveillance, detection, and monitoring of *Klebsiella pneumoniae*. J Clin Microbiol. 2016;54:2582–96. https://doi.org/10.1128/JCM.00927-16
6. Callaway E. Heavily mutated Omicron variant puts scientists on alert. Nature. 2021;600:21. https://doi.org/10.1038/d41586-021-03552-w
7. Hepp CM, Cocking JH, Valentine M, Young SJ, Damian D, Samuels-Crow KE, et al. Phylogenetic analysis of West Nile virus in Maricopa County, Arizona: evidence for dynamic behavior of strains in two major lineages in the American Southwest. PLoS One. 2018;13:e0205801. https://doi.org/10.1371/journal.pone.0205801
8. Sikorski MJ, Levine MM. Reviving the "Moore swab": a classic environmental surveillance tool involving filtration of flowing surface water and sewage water to recover typhoidal *Salmonella* bacteria. Appl Environ Microbiol. 2020;86:e00060-20. https://doi.org/10.1128/AEM.00060-20
9. Frost SDW, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. Philos Trans R Soc Lond B Biol Sci. 2010;365:1879–90. https://doi.org/10.1098/rstb.2010.0060
10. Hill V, Baele G. Bayesian estimation of past population dynamics in BEAST 1.10 using the skygrid coalescent model. Mol Biol Evol. 2019;36:2620–8. https://doi.org/10.1093/molbev/msz172
11. Bedford T, Cobey S, Pascual M. Strength and tempo of selection revealed in viral gene genealogies. BMC Evol Biol. 2011;11:220. https://doi.org/10.1186/1471-2148-11-220
12. Steinig E, Duchêne S, Aglua I, Greenhill A, Ford R, Yoannes M, et al. Phylodynamic inference of bacterial outbreak parameters using nanopore sequencing. Mol Biol Evol. 2022;39:msac040. https://doi.org/10.1093/molbev/msac040
13. Smith MR, Trofimova M, Weber A, Duport Y, Kühnert D, von Kleist M. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. Nat Commun. 2021;12:6009. https://doi.org/10.1038/s41467-021-26267-y
14. Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T. Estimating epidemic incidence and prevalence from genomic data. Mol Biol Evol. 2019;36:1804–16. https://doi.org/10.1093/molbev/msz106

Address for correspondence: David M. Engelthaler, Translational Genomics Research Institute, 3051 Shamrell Blvd, Flagstaff, AZ 86005, USA; email: dengelthaler@tgen.org

Integrating Genomic Data into Public Health Surveillance for Multidrug-Resistant Organisms, Washington, USA

Laura Marcela Torres,[1] Jared Johnson,[1] Audrey Valentine,[1] Audrey Brezak, Emily C. Schneider, Marisa D'Angeli, Jennifer Morgan, Claire Brostrom-Smith, Chi N. Hua, Michael Tran, Darren Lucas, Joenice Gonzalez De Leon, Drew MacKellar, Philip Dykema, Kelly J. Kauber, Allison Black

Mitigating antimicrobial resistance (AMR) is a public health priority to preserve antimicrobial treatment options. The Washington State Department of Health in Washington, USA, piloted a process to leverage longitudinal genomic surveillance on the basis of whole-genome sequencing (WGS) and a genomics-first cluster definition to enhance AMR surveillance. Here, we outline the approach to collaborative surveillance and describe the pilot using 6 carbapenemase-producing organism outbreaks of 3 species: *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, and *Klebsiella pneumoniae.* We also highlight how we applied the approach to an emerging outbreak. We found that genomic and epidemiologic data define highly congruent outbreaks. By layering genomic and epidemiologic data, we refined linkage hypotheses and addressed gaps in traditional epidemiologic surveillance. With the accessibility of WGS, public health agencies must leverage new approaches to modernize surveillance for communicable diseases.

Multidrug resistance threatens modern medicine and public health by limiting our ability to effectively treat serious infections (*1*). Accordingly, reducing and preventing antimicrobial resistance (AMR) is a high priority. Of particular concern are carbapenemase-producing organisms (CPOs), a subset of multidrug-resistant organisms (MDROs) that are resistant to carbapenems—an important class of antibiotics typically reserved as a last resort—and associated with high mortality rates (*2*). CPOs can transfer their resistance genes via mobile genetic elements, like plasmids, across multiple species, contributing to the proliferation of AMR (*3,4*). CPOs and plasmids carrying carbapenemase genes have the potential to make all current antimicrobial drugs ineffective; as such, public health prioritizes surveillance and containment of AMR. Comprehensive strategies critical to mitigate AMR, including antimicrobial stewardship; prompt, accurate diagnosis, and treatment; and infection prevention and control to limit transmission, depend on AMR surveillance data (*5–7*). Recognizing those needs, global and national public health agencies advocate for robust AMR surveillance systems providing timely, high-quality data to inform global, regional, and local containment strategies (*5,8*). By incorporating complementary data sources, robust AMR surveillance systems enable early warning of pathogen emergence, enhance monitoring of epidemiologic trends, improve detection of outbreaks, and deepen understanding of transmission events.

AMR surveillance and cluster investigations rely on epidemiology of person, place, and time, coupled with the genetic and phenotypic characteristics of suspected pathogens. Whole-genome sequencing (WGS) has become a standard method for determining genetic characteristics of pathogens because it enables more comprehensive AMR gene detection compared with traditional PCR-based methods. WGS also enables full-genome

Author affiliations: Washington State Department of Health, Shoreline, Washington, USA (L.M. Torres, J. Johnson, A. Valentine, A. Brezak, E.C. Schneider, M. D'Angeli, C.N. Hua, M. Tran, D. Lucas, J. Gonzalez De Leon, D. Mackellar, P. Dykema, K.J. Kauber, A. Black); Public Health Seattle and King County, Seattle, Washington, USA (J. Morgan, C. Brostrom-Smith)

DOI: https://doi.org/10.3201/eid3113.241227

[1]These first authors contributed equally to this article.

Table 1. Overview of 6 outbreaks of multidrug-resistant organism outbreaks, Washington, USA*

Outbreak ID	Pathogen	No. linked cases	No. health facilities
1	*Pseudomonas aeruginosa*	3	1
2	*Acinetobacter baumannii*	5	1
3	*A. baumannii*	6	1
4	*A. baumannii*	6	1
5	*A. baumannii*	5	5
6	*Klebsiella pneumoniae*	5	1

*We determined linkages between cases by epidemiologic evidence. ID, identification.

comparisons between isolates through core-genome single-nucleotide polymorphism (SNP) analysis. This wider view offers superior resolution over traditional methods that only consider a fraction of the genome, such as multilocus sequence typing (MLST). That resolution reduces misclassification and other biases when making inferences about transmission events (*7*,*9*) and improves our ability to differentiate related and unrelated cases (*10*). Taken together, WGS data enable us to detect MDRO clusters earlier and deploy infection control interventions more quickly (*11*,*12*), detect genes associated with AMR, determine whether resistance is due to chromosomal mutations or to mobile resistance genes (*7*,*13*,*14*), and build genomic datasets that provide context for prospective analyses (*11*).

Given the potential of WGS to advance routine AMR surveillance, we developed and integrated a genomics-first approach into our AMR surveillance system at the Washington State Department of Health. Select MDROs, including CPOs, *Candida auris*, and vancomycin-resistant *Staphylococcus aureus*, are the focus of MDRO surveillance in Washington. Within this system, MDRO sequencing data, generated by Washington Public Health Laboratory (WAPHL) via the Centers for Disease Control and Prevention (CDC)–funded Antibiotic Resistance Laboratory Network (ARLN) testing activities, are ingested into recombination-aware bioinformatics pipelines to identify genomic relationships. Then, data are passed through a workflow that sources and combines surveillance and genomic data. Central to that approach, we established communication and reporting protocols to foster collaborative discussion between laboratory and epidemiology programs about inferences derived from the different data sources. We piloted the approach on 6 historical MDRO outbreaks to explore congruence between genomically and epidemiologically defined clusters and to assess the additive effect of integrating genomic information. Here, we present the results of the pilot and show how to use the integrated surveillance system to support MDRO outbreak investigations prospectively.

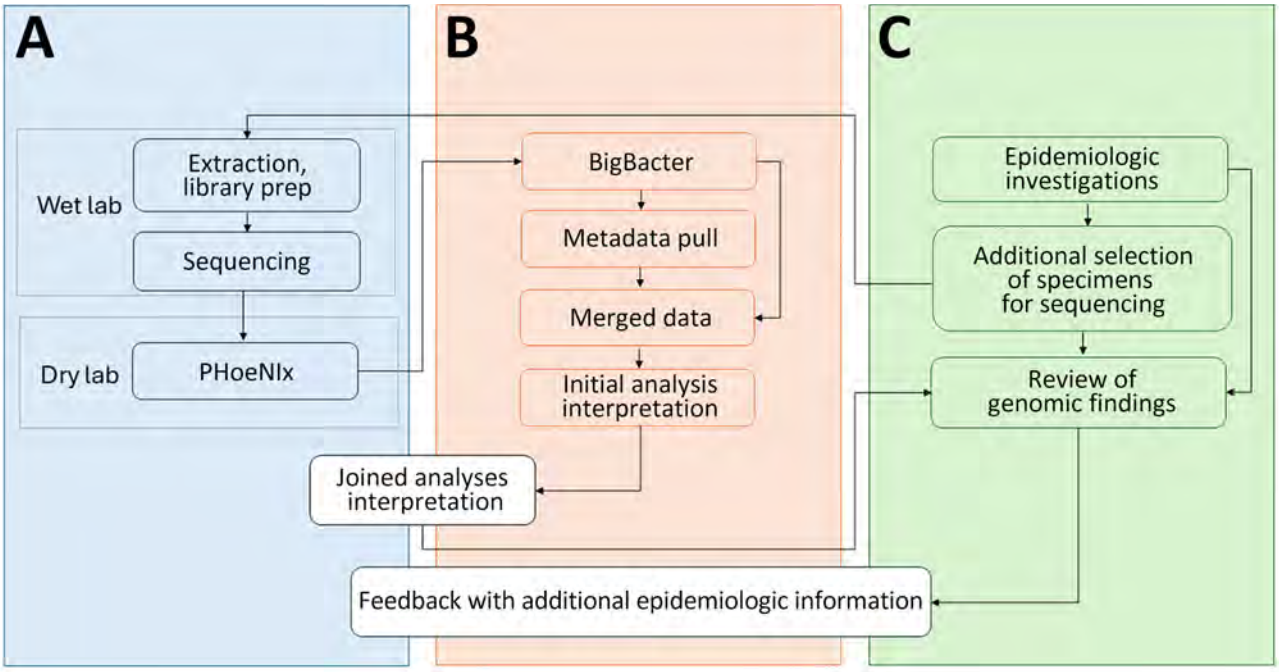


Figure 1. Data flow and cross-team communication channels for our system for integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA. This diagram shows how a sample, then data, move through our integrated surveillance program. Tasks that are handled jointly across programs are highlighted in white. A) Tasks conducted by Washington Public Health Laboratory. B) Tasks conducted by Molecular Epidemiology Program. C) Tasks conducted by Multidrug-Resistant Organism Program and local health jurisdictions.

Table 2. Results of pilot study of genomic and epidemiologic surveillance of outbreaks of multidrug-resistant organism infections, Washington, USA*

Outbreak ID	Pathogen	No. health facilities	No. cases, n = 36	No. isolates sequenced, n = 43	Epidemiologically linked only, n = 5	Epidemiologically and genomically linked, n = 32	Genomically linked only, n = 6
1	*Pseudomonas aeruginosa*	1	5	8†	0	6	2
2	*Acinetobacter baumannii*	1	5	6‡	0	6	0
3	*A. baumannii*	1	6	6	0	6	0
4	*A. baumannii*	1	7	10†	3	6	1
5	*A.baumannii*	5	8	8	0	5	3
6	*Klebsiella pneumoniae*	1	5	5	2§	3	0

*ID, identification.
†One case had 3 isolates sequenced and 1 had 2 isolates sequenced.
‡One case had 2 isolates sequenced.
§Sample 5 was placed into a separate genomic cluster due to relatively large pairwise genetic differences between this isolate and the remaining outbreak 6 isolates, as determined by PopPUNK (*16*).

Methods

WAPHL, the Multidrug-Resistant Organism Program (MDROP), and the Molecular Epidemiology Program (MEP) teams at the Washington State Department of Health analyzed 6 known MDRO outbreaks across 3 species, *A. baumannii*, *P. aeruginosa*, and *K. pneumoniae*, and multiple health facilities (Table 1). The outbreaks were identified through laboratory detection of targeted CPOs by clinical laboratories or WAPHL through the ARLN; methods are summarized on CDC's ARLN Testing Web site (*15*). Cases were identified by detection of a carbapenemase in clinical isolates and through colonization screening performed for MDRO containment response or admission screening. Using epidemiologic investigation methods, MDROP and local health jurisdictions identified linked cases.

Epidemiologic Data

CPOs are reportable in Washington; public health staff investigate all CPOs in partnership with affected healthcare facilities and manage patient screening among epidemiologically linked healthcare contacts. MDROP partners with local health jurisdictions to perform longitudinal surveillance using an Antimicrobial Resistance Information Exchange (ARIE), investigate potential clusters, perform containment responses, and document reported outbreaks. For this pilot, MDROP provided MEP a master list for each of the 6 outbreaks, including epidemiologic information about index cases, facility admissions, known epidemiologic linkages, and isolate identifiers to link case and sequencing data.

Sequencing and Genomic Analysis

We performed WGS using DNA extracted with the MagNA Pure 96 Small Volume Kit on an MP96 system (both Roche, https://ww.roche.com) from bacterial cultures grown on blood agar (Thermo Fisher Scientific, https://www.thermofisher.com) for 24 hours at 35–37°C. We prepared paired-end DNA libraries using the Illumina DNA Prep kit with Nextera DNA CD indexes sequenced on a MiSeq System (all Illumina, https://www.illumina.com) using the 2 × 250 bp (500-cycle) v2 kit. We used the CDC PHoeNIx pipeline (https://zenodo.org/record/8147510) to perform general bacterial analysis, including quality control, de novo assembly, taxonomic classification, and AMR gene detection. We repeated sequencing for samples with ≤40× average read depth, ≤1 Mb genome size, ≥500 assembly scaffolds, or ≥2.58 assembly ratio SD (Appendix 1 Table 1, https://wwwnc.cdc.gov/EID/article/31/13/24-1277-App1.pdf). PHoeNIx outputs feed into the WAPHL BigBacter pipeline (https://github.com/DOH-JDJ0303/bigbacter-nf), which enables bacterial genomic surveillance by performing phylogenetic analysis and differentiating clusters of closely related bacteria that are maintained in a personalized database. We clustered samples genomically using PopPUNK version 2.6.0 as described (*16*) and calculated accessory distances and core SNPs within each genomic cluster using the PopPUNK sketchlib functions and Snippy version 4.6.0 (https://github.com/tseemann/snippy). We identified and masked recombinant regions in the Snippy output using Gubbins version 3.3.1 as described (*17*). We generated phylogenetic trees and distance matrices using IQTREE2 version 2.2.2.6 as described (*18*) and custom scripts in R (The R Project for Statistical Computing, https://www.r-project.org) and Bash (Free Software Foundation, Inc., https://www.gnu.org/software/bash).

We linked the BigBacter genomic outputs to metadata attributes queried from our laboratory information and surveillance systems, enabling joint analysis and visualization in R and Nextstrain Auspice (*19*). We used the phylogenetic trees, SNP matrices, and BigBacter's cluster designation to identify genomic clusters. To explore congruence between genomic clusters and epidemiologically defined clusters in our pilot, we identified the subset of

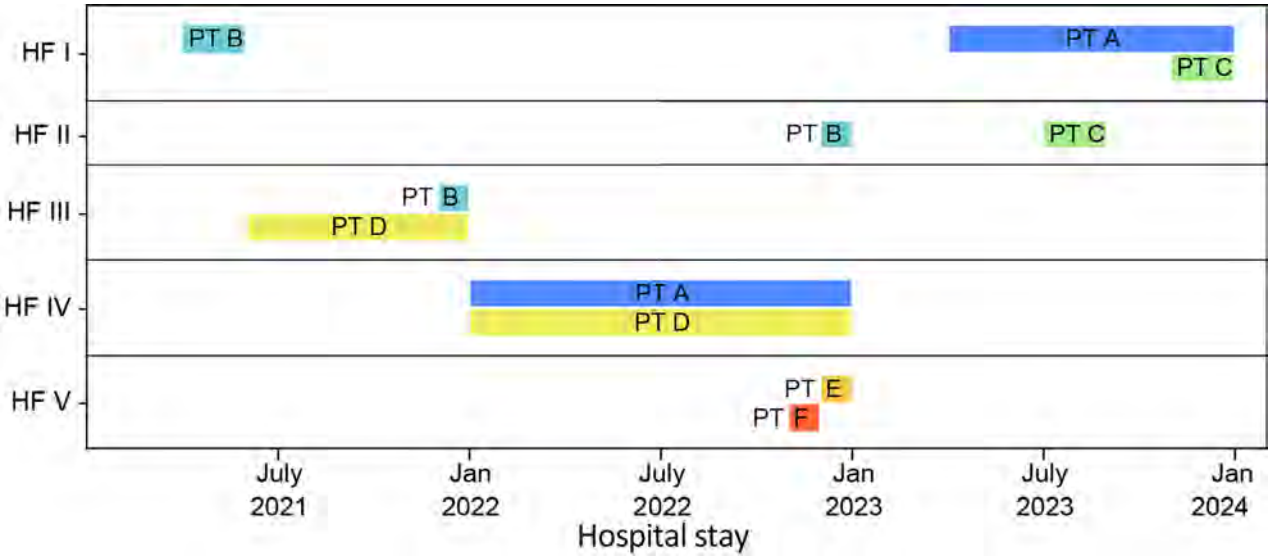


Figure 2. Timeline showing overlap of *Klebsiella pneumoniae* carbapenemase–producing *K. pneumoniae* infection patients in healthcare facilities in Washington, USA, as part of study of integrating genomic data into public health surveillance for multidrug-resistant organisms. PTs A, B, and C stayed in HF I. PT A might have overlapped with PT C in HF I in 2023; PT B stayed in HF I in 2021. PTs B and C both stayed in HF II but at different times, in 2022 and in 2023. PT D stayed at HF III in 2021, where an overlap with PT B might have occurred, and in 2022, PT D might have overlapped with PT A in HF IV. PTs E and F who had stayed in HF V could also be related to this outbreak. HF, health facility; PT, patient.

genomic clusters that grouped cases associated with 6 outbreaks defined by MDROP. Then, we looked at the union of all sequenced samples in relevant genomic clusters (n = 43) and all cases identified as part of the 6 epidemiologically defined outbreaks (n = 36). We defined samples as follows: genomically linked only, meaning that the sequenced sample grouped in a relevant genomic cluster and either the core genome sequences were closely related (≤10 SNPs) or a larger SNP distance could be explained by differences in sample collection dates; epidemiologically linked only, meaning that MDROP had linked a case to an outbreak, but that the sequence did not meet the genomically linked definition; or epidemiologically and genomically linked, meaning that both MDROP epidemiologists' assessment and sequencing data grouped the case as part of the relevant outbreak. MEP, WAPHL, and MDROP met to discuss the findings. Communication between our programs helped address perceived utility of routine genomic analyses and enabled us to develop processes for ongoing data production, analytics, interpretation, and cross-program communication.

Results

Cluster Detection Using a Genomics-First Approach

To pilot integrated surveillance, we evaluated whether genomic data and epidemiologic investigations grouped the same cases for 6 known, epidemiologically defined outbreaks. We analyzed 221 sequences of *P. aeruginosa*, *A. baumannii*, and *K. pneumoniae*, collected during December 2017–May 2024; those sequences grouped into 48 genomic clusters. Six of the genomic clusters were largely concordant with the 6 epidemiologically defined outbreaks (n = 36 cases). The 6 genomic clusters grouped 42 sequences, of which 32 were classified as epidemiologically and genomically linked (Table 2; Appendix 1 Figures 1–6). One epidemiologically linked case grouped into a seventh genomic cluster with no other linked cases, indicating that genomic data did not support the linkage. Although BigBacter groups related sequences, pairwise genetic divergence within a cluster can still exceed the SNP distance threshold we use to define genomic linkage. Indeed, 4 epidemiologically linked cases grouped into outbreak-related genomic clusters but were not considered genomically linked because they diverged from other sequenced cases by 14–56 SNPs; that distance could not be explained by differences in sample collection dates (Table 2; Appendix 1 Figures 4, 6). Six sequences grouped into relevant genomic clusters with minimally divergent core genome sequences, but those cases had not been linked to the outbreaks through epidemiologic information; the cases were genomically linked only (Table 2; Appendix 1 Figures 1, 4, 5). Our findings show general concordance between epidemiologic and genomic clusters and demonstrate instances where genomic data may refine cluster definitions.

Development of Standard Integrated Genomic Epidemiology Reports

We sought to develop mechanisms to jointly analyze genomic and epidemiologic data and communicate across teams about the inferences. MEP, MDROP, and WAPHL discussed the pilot study findings, including the utility and limitations of genomic analyses, and collectively designed a new data and communication workflow. The workflow required us to bridge siloed data sources (Figure 1); to do so, we programmatically ingest laboratory identifiers and query the surveillance database. Working with MDROP, we determined which epidemiologic information are most important for contextualizing genomic information (e.g., submitter facility name, submitter county, collection date, etc.). We source, format, and export this information as a metadata file that can be overlaid onto phylogenetic trees.

MEP and WAPHL iteratively refined the information included in the reports to meet MDROP's needs. The current version of the report includes 3 components. The first component is an automated R markdown-based report that parses the BigBacter output and metadata to summarize key information, such as the total number of sequences per genomic cluster, number of new sequences added to previously identified clusters, submitting health facilities and counties, and sequences with close or intermediate genomic linkage (Appendix 2, https://wwwnc.cdc.gov/EID/article/31/13/24-1227-App2.pdf). The

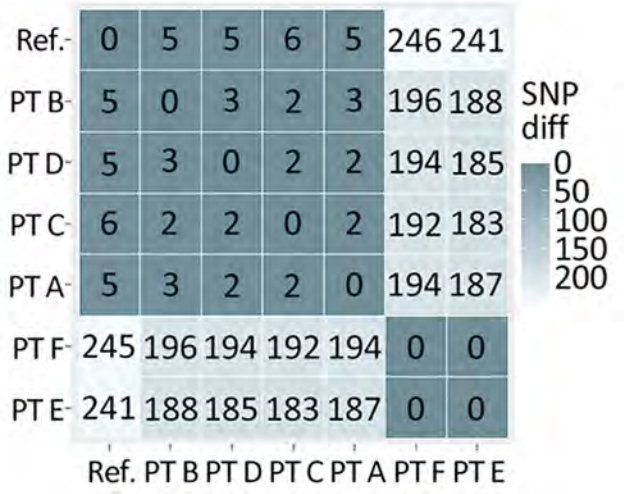


Figure 3. SNP matrix showing number of polymorphic sites observed when making pairwise comparisons between the core genome of the sequences in a cluster of *Klebsiella pneumoniae* carbapenemase–producing *K. pneumoniae* infection isolates as part of study of integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA. Dark gray represents lower SNP distances and light gray larger SNP distances. Diff, difference; PT, patient; ref., reference; SNP, single-nucleotide polymorphism.

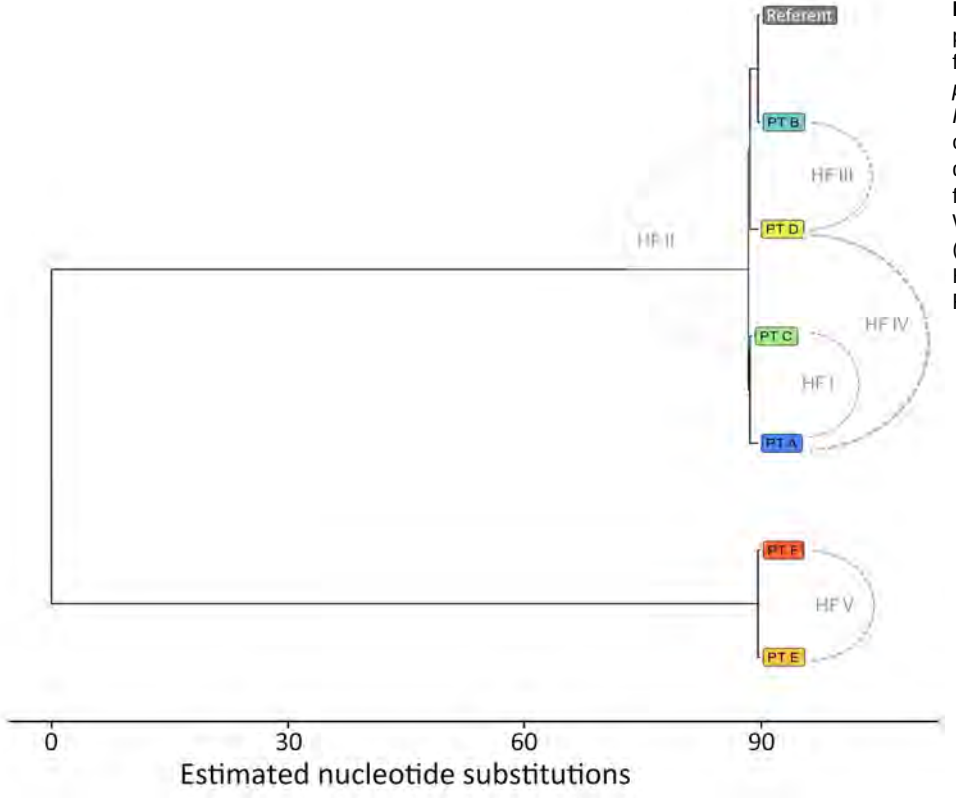


Figure 4. Maximum-likelihood phylogenetic tree of sequences from patients with *Klebsiella pneumoniae*–producing *K. pneumoniae* infection as part of study of integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA. Five patients (A–E) are shown, and relevant HFs are noted. HF, health facility; PT, patient.

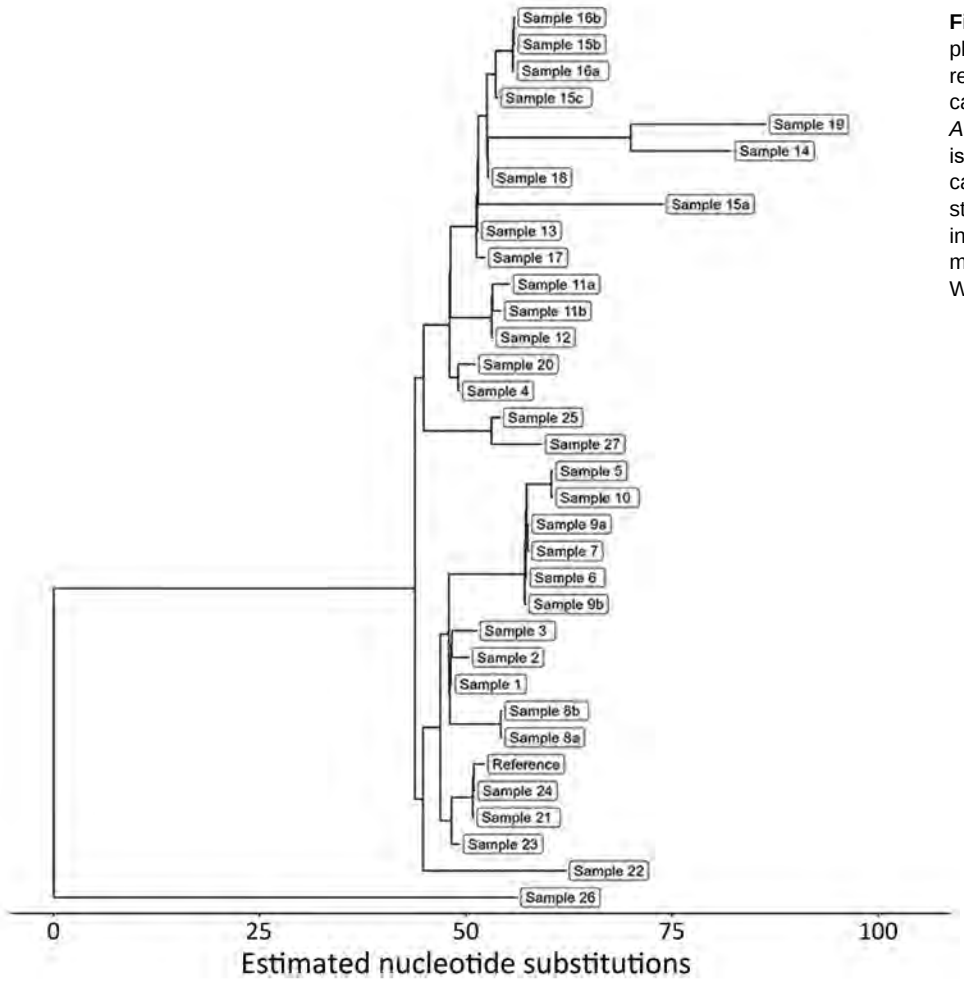


Figure 5. Maximum-likelihood phylogenetic tree showing relationships among 33 carbapenemase-producing *Acinetobacter baumannii* isolates with the OXA-235–like carbapenemase gene as part of study of integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA.

second component is a narrative interpretation of the genomic data written by MEP epidemiologists; that component alerts MDROP epidemiologists to transmission dynamics consistent with the genomic data, such as detection of new introductions or ongoing transmission of an outbreak. The final component of the report is a Microreact (*20*) dashboard, where we share interactive multipanel figures including SNP distance matrices and phylogenetic trees; this type of reporting is a standard feature of Washington's AMR surveillance. Among other outcomes, the approach has improved our understanding of *K. pneumoniae* transmission within a multifacility outbreak and helped us ascertain linkages between carbapenemase-producing *A. baumannii* (CRAB) cases that were previously unknown.

Differentiation of Outbreak and Nonoutbreak Samples Using Genomic Data

In a prospective analysis of a *K. pneumoniae* carbapenemase–producing *K. pneumoniae* outbreak involving multiple healthcare facilities, epidemiologic investigation data alone could not clarify how transmission had occurred; recent healthcare during the exposure period involved multiple cases, some with overlapping healthcare stays (Figure 2). Integrating genomic and case-level data helped us refine relationships between cases and formulate a hypothesis for how cases were connected across facilities. MEP and WAPHL reported that sequences from patients A, B, and C were closely related (2–3 SNPs) (Figure 3). MDROP confirmed epidemiologic linkages among some of those patients (Figures 2, 3), but a common link was missing. MDROP hypothesized that patients D, E, or F could be the missing link and requested a review of their sequencing results, pending sequencing for patient D. MDROP's reasoning was that patient D might have overlapped with patients A and B. Sequencing revealed that patients E and F had identical core-genome sequences but diverged greatly from the other sequenced cases (Figure 4). MDROP confirmed an epidemiologic link between patients E and F, noting they received care at

the same facility and shared staff. The genomic and epidemiologic information helped confirm these patients were connected to each other but not related to the outbreak in question. The sequence from patient D, however, was genomically linked (2–3 SNPs) to sequences from patients A, B, and C (Figure 3). The close genomic distances and the overlap in healthcare stays with patients A and B supported the hypothesis that patient D was one of the missing links. Patient C's relationship to the outbreak remains unclear; patient C tested positive upon admission but reported no healthcare encounters before August 2023. Despite that remaining question, genomic analyses helped confirm 1 missing link, excluded 2 patients from this outbreak, and revealed that the outbreak was larger than originally thought.

Genomic Data Linking Historical Carbapenemase-Producing *A. baumannii* Cases

We assessed the congruence between epidemiologic surveillance data and genomic clustering for a retrospective set of CRAB isolates with the OXA-235–like carbapenemase gene. Two outbreaks were known to MDROP at healthcare facilities I and IV. First, we reviewed all 33 sequenced CRAB OXA-235 isolates representing 27 cases collected during August 2019–December 2023. We compiled healthcare encounters for cases from MDROP's linelist and ARIE and matched 137 admissions across 29 facilities from July 2020–May 2024. We visualized genomic analyses and epidemiologic data using vistime and ggtree (https://shosaco.github.io/vistime) (*21*) in R.

We used PopPUNK (*16*) for genomic clustering; all 33 isolates were assigned to the same genomic cluster (Figure 5). The cluster had a maximum pairwise divergence of 119 SNPs. To identify closer genetic relationships indicative of clonal transmission, we used BigBacter to partition the cluster into groups of sequences separated by ≤10 SNPs (*22,23*), resulting in 12 partitions (Figure 6). Seven partitions contained multiple sequences. We defined sequences within a partition

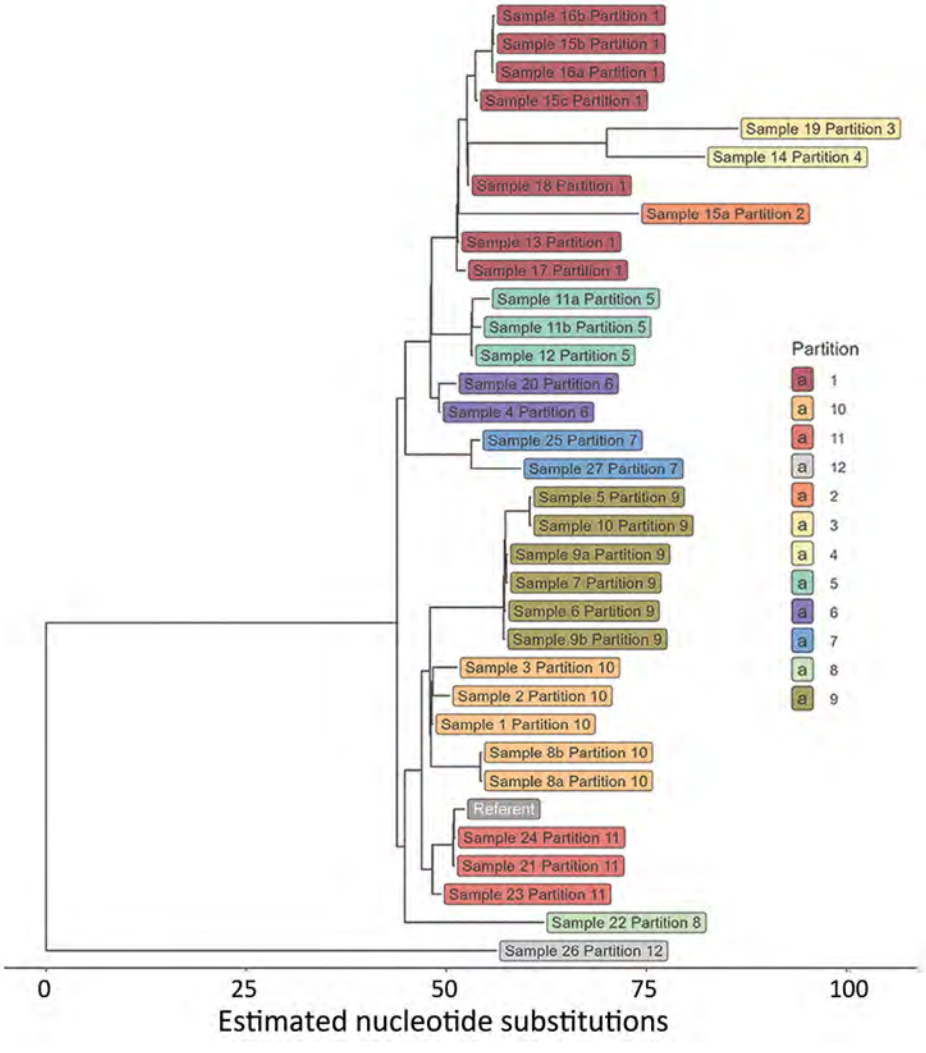


Figure 6. Maximum-likelihood phylogenetic tree showing partitions of 33 carbapenemase-producing *Acinetobacter baumannii* isolates with the OXA-235–like carbapenemase gene as part of study of integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA. Colors indicate 12 partitions demarcating seq uences separated by ≤10 SNPs. Seven of the partitions contain multiple sequences and 5 (2, 3, 4, 8, and 12) contain 1 sequence.

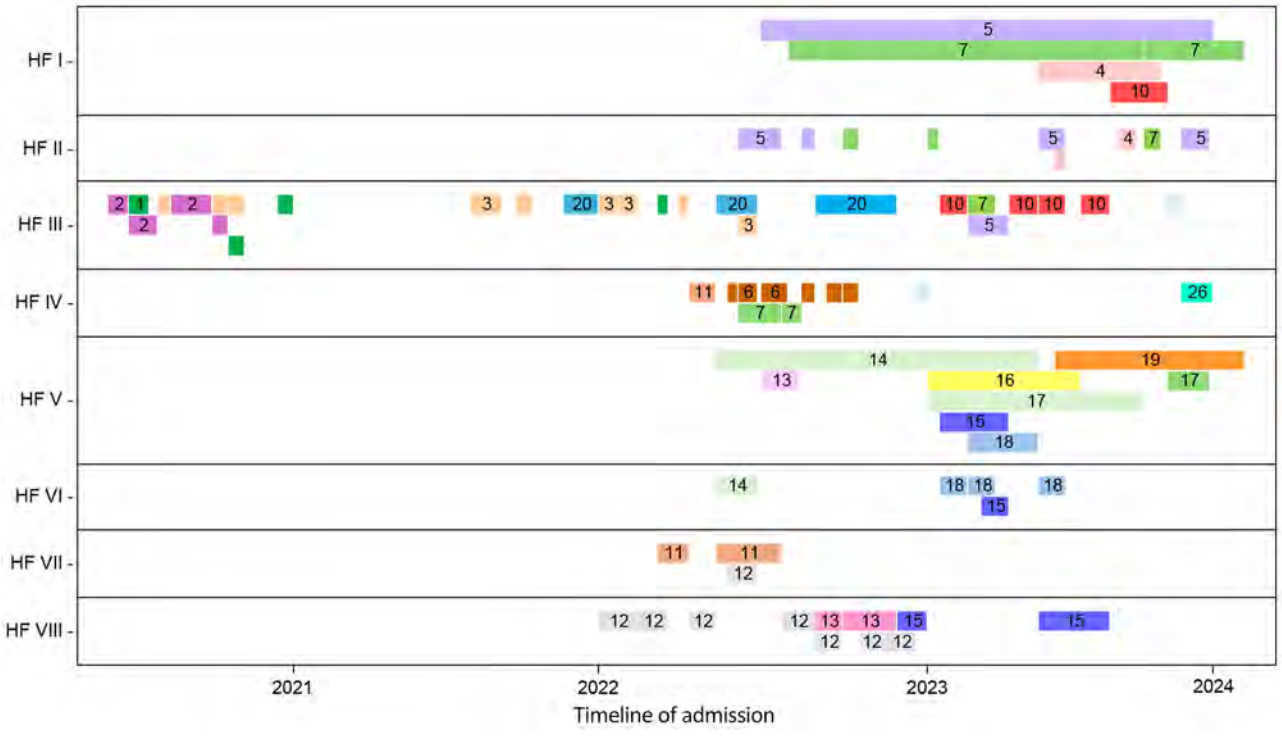


Figure 7. Healthcare encounters at facilities of interest among carbapenemase-producing *Acinetobacter baumannii* OXA-235 cases as part of study of integrating genomic data into public health surveillance for multidrug-resistant organisms, Washington, USA. Six cases (14, 15, 16, 17, 18, and 19) are linked to a screening event at HF V, and 3 cases (5, 7, and 10) are linked to a screening event at HF I. Cases cannot be in 2 or more health facilities simultaneously. However, we only have access to admission and discharge dates; therefore, the figure may show some cases in multiple locations at the same time if transfers occurred without an associated admission and discharge. HF, health facility.

as genomically linked to each other. In this analysis, MDROP defined epidemiologic linkage between cases as temporally overlapping visits at the same healthcare facility. We considered 8 facilities that had cases with overlapping visits to be facilities of interest (Figure 7). We categorized cases that were epidemiologically linked and belonged to the same genomic partition as epidemiologically and genomically linked. We evaluated concordance between genomic and epidemiologic data by categorizing sequences from the 7 partitions as epidemiologically and genomically linked, epidemiologically linked only, or genomically linked only. Four partitions (1, 5, 9, and 10) included 21 sequences; we considered 17 of those epidemiologically and genomically linked and 4 genomically linked only. We classified the sequences in the remaining 3 multisequence partitions (6, 7, and 11) as genomically linked only; partition 6 contained 2 sequences from cases that were not epidemiologically linked, and sequences in partitions 7 and 11 were from cases missing epidemiologic data (Appendix 1 Table 1). Five partitions (2, 3, 4, 8, and 12) contained only 1 sequence and thus had no evidence of genomic linkage. Of those 5 sequences, we considered 3 epidemiologically linked (Appendix 1 Table 2); 2 sequences lacked epidemiologic data.

Our results highlight the consistency that genomically and epidemiologically defined clusters can have, as well as how our definition for epidemiologic linkage may lack sensitivity and specificity. Indeed, detailed retrospective case review prompted by genomic linkages described by our analysis yielded 10 epidemiologic links unknown to MDROP.

Discussion

Here, we describe our approach to integrating genomics into our AMR surveillance system and transitioning from a pilot assessment to a repeatable workflow. Integrating genomic data into AMR surveillance has helped us identify additional outbreak cases, sensitively classify outbreak or nonoutbreak cases, and confirm hypothesized linkages. Furthermore, we reduced silos between programs, fostering collective discussion to guide data interpretation and next steps. Building on this success, we now perform automated genomic cluster detection for all MDRO bacterial pathogens sequenced at WAPHL, and we plan to expand this approach to other surveillance programs.

Our approach has some notable benefits. First, our system characterizes genomic relationships using distance-based analysis of sequence data. Although national surveillance systems in the United States such as PulseNet (*24*) and TB GIMS (*25*) have transitioned from MLST, only a predefined set of loci within the core genome are considered, and the set of relevant loci cannot expand on an outbreak-by-outbreak basis. Although sequence types delineate whether sequences are nearly identical or not, they do not allow epidemiologists to directly estimate genetic distances between sequences. Second, BigBacter by default stores genomic cluster information in a running database, providing historical context when analyzing new sequences. This is one of the beneficial features of systems such as PulseNet, as it enables detection of reemerging outbreaks or strains (*26*), but to our knowledge such approaches are rarely implemented and maintained by a single state agency. Finally, our system mitigates the bias that can arise when sequencing is prompted solely by epidemiologic hypotheses. By sequencing MDRO detections regardless of outbreak status and identifying clusters given genetic relatedness only, we draft genomics-informed hypotheses independent of hypotheses derived from epidemiologic investigation data. When findings from both data streams are consistent, it strengthens our belief that we understand transmission within the cluster, whereas discrepancies prompt us to reinvestigate or evaluate gaps specific to each data source. This approach stands in contrast to targeted sequencing efforts where sequencing occurs only upon request, such as when surveillance epidemiologists have defined an outbreak.

Despite those benefits, our integrated AMR surveillance system has some limitations. Ideally, our system would include environmental and nonhuman isolates to clarify risk for zoonotic and environmental transmission of CPOs to humans (*13*,*14*). However, we lack access to those sample types, and our system's slow turnaround time limits its utility. In our system, bacterial sequencing proceeds from cultured isolates, resulting in genomic analysis being shared ≈1 month after carbapenemase detection. By then, WGS only provides post hoc confirmation about links that have been already identified, rather than real-time information to inform infection control practices. Finally, WGS is expensive, which could make this program unsustainable in the absence of stable and appropriate funding.

Through our efforts to develop, test, and deploy an integrated AMR surveillance system, MDROP can leverage pathogen genomics for public health response. During active investigations, MDROP can intervene when genomic links are identified, guiding actions to improve infection control practices. Furthermore, by developing this system collectively, our system includes perspectives from surveillance epidemiology, molecular epidemiology, and bioinformatics and reduces silos between teams. Building on initial successes, we continue to refine this system to increase the timeliness of genomic inferences and identify best practices to engage local health jurisdictions.

This work is supported by funding from the CDC Pathogen Genomics Centers of Excellence cooperative agreement (no. NU50CK000630) and the CDC Enhancing Epidemiology and Laboratory Capacity AMD Sequencing and Analytics cooperative agreement (no. NU50CK000515). This study and report were supported in part by an appointment to the Applied Epidemiology Fellowship Program administered by the Council of State and Territorial Epidemiologists and funded by CDC cooperative agreement no. 1NU38OT000297-03-00.

About the Author

Ms. Torres is an epidemiologist with the Molecular Epidemiology Program at the Washington State Department of Health. She is interested in the application of pathogen genomics as a core domain of infectious disease detection and response.

References

1. Centers for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2019. 2019 Nov [cited 2024 Apr 8]. https://stacks.cdc.gov/view/cdc/82532
2. Bonomo RA, Burd EM, Conly J, Limbago BM, Poirel L, Segre JA, et al. Carbapenemase-producing organisms: a global scourge. Clin Infect Dis. 2018;66:1290–7. https://doi.org/10.1093/cid/cix893
3. Hardiman CA, Weingarten RA, Conlan S, Khil P, Dekker JP, Mathers AJ, et al. Horizontal transfer of carbapenemase-encoding plasmids and comparison with hospital epidemiology data. Antimicrob Agents Chemother. 2016;60:4910–9. https://doi.org/10.1128/AAC.00014-16
4. Mathers AJ, Cox HL, Kitchel B, Bonatti H, Brassinga AKC, Carroll J, et al. Molecular dissection of an outbreak of carbapenem-resistant Enterobacteriaceae reveals intergenus KPC carbapenemase transmission through a promiscuous plasmid. MBio. 2011;2:e00204–11.
5. Siegel JD, Rhinehart E, Jackson M, Chiarello L. Management of multidrug-resistant organisms in healthcare settings, 2006. 2006 [cited 2024 Jul 17]. https://www.cdc.gov/infection-control/media/pdfs/Guideline-MDRO-H.pdf
6. Tacconelli E, Sifakis F, Harbarth S, Schrijver R, van Mourik M, Voss A, et al.; EPI-Net COMBACTE-MAGNET Group. Surveillance for control of

antimicrobial resistance. Lancet Infect Dis. 2018;18:e99–106. https://doi.org/10.1016/S1473-3099(17)30485-1
7. Waddington C, Carey ME, Boinett CJ, Higginson E, Veeraraghavan B, Baker S. Exploiting genomics to mitigate the public health impact of antimicrobial resistance. Genome Med. 2022;14:15. https://doi.org/10.1186/s13073-022-01020-2
8. World Health Organization. Global antimicrobial resistance and use surveillance system (GLASS) report 2022. 2022 [cited 2024 Jul 17]. https://iris.who.int/bitstream/handle/10665/364996/9789240062702-eng.pdf
9. Branch-Elliman W, Sundermann AJ, Wiens J, Shenoy ES. The future of automated infection detection: innovation to transform practice (Part III/III). Antimicrob Steward Healthc Epidemiol. 2023;3:e26. https://doi.org/10.1017/ash.2022.333
10. Black A, Dudas G. The applied genomic epidemiology handbook: a practical guide to leveraging pathogen genomic data in public health. 1st edition. Boca Raton (FL): Chapman & Hall/CRC Press; 2024.
11. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. Microbiology (Reading). 2018;164:1213–9. https://doi.org/10.1099/mic.0.000700
12. Sherry NL, Gorrie CL, Kwong JC, Higgs C, Stuart RL, Marshall C, et al.; Controlling Superbugs Study Group. Multi-site implementation of whole genome sequencing for hospital infection control: a prospective genomic epidemiological analysis. Lancet Reg Health West Pac. 2022;23:100446. https://doi.org/10.1016/j.lanwpc.2022.100446
13. Baker KS, Jauneikaite E, Nunn JG, Midega JT, Atun R, Holt KE, et al.; SEDRIC Genomics Surveillance Working Group. Evidence review and recommendations for the implementation of genomics for antimicrobial resistance surveillance: reports from an international expert group. Lancet Microbe. 2023;4:e1035–9. https://doi.org/10.1016/S2666-5247(23)00281-1
14. Wareth G, Brandt C, Sprague LD, Neubauer H, Pletz MW. WGS based analysis of acquired antimicrobial resistance in human and non-human *Acinetobacter baumannii* isolates from a German perspective. BMC Microbiol. 2021;21:210. https://doi.org/10.1186/s12866-021-02270-7
15. Centers for Disease Control and Prevention. Antimicrobial Resistance Laboratory Network testing. 2024 [cited 2024 Jul 17]. https://www.cdc.gov/antimicrobial-resistance-laboratory-networks/php/about/testing-services.html
16. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019;29:304–16. https://doi.org/10.1101/gr.241455.118
17. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15–15. https://doi.org/10.1093/nar/gku1196
18. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74. https://doi.org/10.1093/molbev/msu300
19. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3.
20. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom. 2016;2. 10.1099/mgen.0.000093
21. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
22. Mangioni D, Fox V, Chatenoud L, Bolis M, Bottino N, Cariani L, et al. Genomic Characterization of carbapenem-resistant *Acinetobacter baumannii* (CRAB) in mechanically ventilated COVID-19 patients and impact of infection control measures on reducing CRAB circulation during the second wave of the SARS-CoV-2 pandemic in Milan, Italy. Microbiol Spectr. 2023;11:e00209–23.
23. Fitzpatrick MA, Ozer EA, Hauser AR. Utility of whole-genome sequencing in characterizing *Acinetobacter* epidemiology and analyzing hospital outbreaks. J Clin Microbiol. 2016;54:593–612.
24. Centers for Disease Control and Prevention. Implementing whole genome sequencing for foodborne disease surveillance. PulseNet. 2024 [cited 2024 Dec 4]. https://www.cdc.gov/pulsenet/php/wgs/wgs-vision.html
25. Centers for Disease Control and Prevention. Tuberculosis whole-genome sequencing. 2024 [cited 2024 Dec 4]. https://www.cdc.gov/tb/php/genotyping/whole-genome-sequencing.html
26. Tolar B, Joseph LA, Schroeder MN, Stroika S, Ribot EM, Hise KB, et al. An overview of PulseNet USA databases. Foodborne Pathog Dis. 2019;16:457–62. https://doi.org/10.1089/fpd.2019.2637

Address for correspondence: Laura Marcela Torres, Washington State Department of Health, 1610 NE 150th St, Shoreline, WA 98155, USA; email: marcela.torres@doh.wa.gov

Leveraging a Strategic Public–Private Partnership to Launch an Airport-Based Pathogen Monitoring Program to Detect Emerging Health Threats

Cindy R. Friedman, Robert C. Morfino, Ezra T. Ernst

Airport-based pathogen monitoring is a critical tool that can contribute to early detection and characterization of existing and new pathogen threats. A novel public–private partnership between an airport spa group, a biotech company, and the Centers for Disease Control and Prevention was instrumental in establishing a multimodal pathogen genomic surveillance program at US international airports. That public–private partnership addressed critical challenges that neither party could overcome independently, resulting in the development and deployment of a scalable, flexible early warning system for pathogen detection and public health monitoring.

The COVID-19 pandemic demonstrated the need for large-scale public health response mechanisms including surveillance, laboratory testing, and genomic sequencing (*1*). Public–private partnerships can be used to deploy those measures rapidly, especially during public health emergencies. During the pandemic, the Centers for Disease Control and Prevention (CDC) partnered with an airport spa company (XpresCheck, https://xprescheck.com) and a biotech firm (Ginkgo Biosecurity, https://biosecurity.ginkgo.bio) to develop an innovative traveler-based genomic surveillance system for early detection of new SARS-CoV-2 variants. The program expanded from an initial proof-of-concept pilot launched during September 2021 to a dynamic multiairport monitoring system during a few days in November 2021, just as the Omicron variant was identified (*2*).

International air travelers move rapidly across the globe, bringing the potential to spread communicable diseases, thereby making traveler-based public health surveillance a critical tool to gain early knowledge about the emergence and spread of existing and new pathogens. During the 20 years before the COVID-19 pandemic, global traveler-based public health surveillance programs were driven by more traditional collaborations with academic institutions and travel medicine clinics to conduct sentinel surveillance among travelers (*3*). Those programs provided insights and guidance for traveling populations and travel medicine clinicians (*3*). However, those programs relied on symptomatic travelers seeking medical care and clinicians ordering appropriate laboratory tests, which, if positive, might undergo whole-genome sequencing and be reported to public health. That screening process takes time and, early in the COVID-19 pandemic, its effectiveness was limited because nearly half of patients with SARS-CoV-2 infections were asymptomatic (*4*). As the pandemic evolved, many symptomatic persons might not have sought healthcare or might have used antigen-based self-tests that would not yield a sample for genomic sequencing; therefore, new variants emerging in one part of the world could go undetected while spreading globally (*5*).

In early 2021, US public health authorities sought to quickly sequence samples from incoming international travelers at US airports—not for case identification or contact tracing, but to gain an early snapshot of new SARS-CoV-2 variants entering the country. Rapid detection of those variants, which had

Author affiliations: Centers for Disease Control and Prevention, Atlanta, Georgia, USA (C.R. Friedman); Gingko Bioworks, Boston, Massachusetts, USA (R.C. Morfino); Xwell, New York, New York, USA (E.T. Ernst)

DOI: https://doi.org/10.3201/eid3113.241407

previously been challenging, was crucial for timely analysis and to help adjust mitigation strategies.

The preferences for the proposed genomic surveillance program included infrastructure for recruiting and sampling travelers in airports, voluntary traveler participation, and a rapid turnaround time for

Table. Leveraging public–private partnerships to expand CDC's Traveler-based Genomic Surveillance airport-based pathogen monitoring program, September 2021–August 2024*

Milestones	2021		2022	2023	Jan–Aug 2024
	Sep 29–Nov 27	Nov 28–Dec 31			
Launch	Launched 6-week pilot, demonstrating operational feasibility and detection and genomic sequencing of SARS-CoV-2 in samples from travelers	Expanded pilot for Omicron surge; identified Omicron subvariants BA.2 and BA.3 six weeks before those variants were reported globally (*2*)	Launched airplane wastewater pilot at JFK (*5*); demonstrated retroactively that US predeparture test requirement during COVID-19 pandemic reduced postarrival positivity by 50% (*8*); enhanced surveillance for 2022 FIFA World Cup (*9*)	Expanded coverage of flights from China and surrounding hubs during China's removal of its "zero-COVID" policy and subsequent surge of cases; detected first BA.2.86 in a traveler from Japan (*10*); detected FLiRT† mutations in wastewater samples 3 weeks before reported globally	Launched transatlantic airplane wastewater pilot in collaboration with United Kingdom Health Security Agency; enhanced surveillance during Hajj and 2024 Summer Olympics
Airports involved	EWR, JFK T4, SFO	ATL, EWR, JFK T4, SFO	ATL, EWR, IAD, JFK T4, SFO	ATL, BOS, EWR, IAD, JFK T4, JFK T8, LAX, SEA, SFO	BOS, EWR, IAD, JFK T4, JFK T8, LAX, MIA, SEA, SFO
Modalities	Nasal sampling in airport; at-home saliva sampling with questionnaire	Nasal sampling in airport; at-home saliva sampling with questionnaire	Nasal sampling in airport and traveler questionnaire; discontinued at-home saliva sampling; airplane wastewater sampling	Nasal sampling in airport and traveler questionnaire; airplane wastewater sampling; airport triturator;‡ air monitoring	Nasal sampling in airport and traveler questionnaire; airplane wastewater sampling; airport triturator; air monitoring
Median (range) participants per week§	535 (19–1395)	1,434 (1,334–1,746)	1,217 (325–3,490)	6,320 (1,689–9,321)	7,249 (4,366–12,628)
Median (range) traveler countries of origin per week§	1	6	43 (6–87)	123 (56–138)	143 (116–161)
Wastewater samples collected	0	0	89	417	783
Air samples collected	0	0	0	95	438
Laboratory methods used	RT-PCR, amplicon-based sequencing	RT-PCR, amplicon-based sequencing	RT-PCR, amplicon-based sequencing, target enrichment sequencing	RT-PCR, dRT-PCR, amplicon-based sequencing, target enrichment sequencing	RT-PCR, dRT-PCR, amplicon-based sequencing, target enrichment sequencing
Pathogen targets	SARS-CoV-2	SARS-CoV-2	SARS-CoV-2, influenza A and B pilot	SARS-CoV-2, influenza A and B, RSV testing of nasal samples, air, and wastewater; *Mycoplasma pneumoniae* testing of nasal samples in response to global outbreak reports; mpox testing of airplane and triturator‡ wastewater	Expanded multipathogen enrichment sequencing panel for up to 66 viruses deployed for wastewater samples

*ATL, Atlanta Hartsfield-Jackson International Airport, Atlanta, Georgia, USA; BOS, Logan Airport, Boston, Massachusetts, USA; CDC, Centers for Disease Control and Prevention; dRT-PCR, digital reverse transcription PCR; EWR, Newark Liberty International Airport, Newark, New Jersey, USA; FIFA, Fédération Internationale de Football Association; JFK T4 and T8, John F. Kennedy International Airport Terminal 4 and Terminal 8, Queens, New York, USA; IAD, Washington Dulles International Airport, Dulles, Virginia, USA; LAX, Los Angeles International Airport, Los Angeles, California, USA; MIA, Miami International Airport, Miami, Florida, USA; RSV, respiratory syncytial virus; RT- PCR, reverse transcription PCR; SEA, Seattle-Tacoma International Airport, Seattle, Washington, USA; SFO, San Francisco International Airport, San Francisco, California, USA.
†SARS-CoV-2 variants characterized by specific spike mutations-F to L at position 456 and R to T at position 346-enhancing their transmissibility and immune evasion capabilities.
‡A consolidation point that captures wastewater samples from multiple flights and does not include airport terminal waste.
§Nasal swab sampling.

reporting results, including genomic sequences. The Centers for Disease Control and Prevention (CDC) learned about an airport-based company that pivoted from offering spa services, such as manicures and massages, to providing rapid COVID-19 testing for outbound US travelers needing to meet international entry requirements. A biotechnology company joined the initiative to contribute approaches for genomic sequencing. The resulting public–private partnership, the Traveler-based Genomic Surveillance Program, provided flexibility to adjust methods quickly—a key advantage of the collaboration. The partnership addressed a critical gap that none of the parties could fill independently: developing and deploying a scalable early warning system for public health genomic surveillance.

Public health agencies often lack the capacity to manage operational complexity at the scale that industry partners bring to the table. The collaboration we describe established a traveler-based genomic surveillance system that could adapt to the evolving pandemic, respond swiftly to emerging threats, and serve as a novel tool for outbreak detection and pandemic preparedness. For example, collecting samples from volunteer travelers required access to specific areas of the airport, security clearance for staff, developing a customized process for recruiting consenting travelers, and collecting and registering samples without interfering with airport operations, all while seamlessly integrating into the travelers' journeys.

When the program expanded to include environmental testing, collecting wastewater samples from aircraft necessitated creating a collection device and sampling process (*6*). That creation involved several design and testing cycles and negotiations with multiple stakeholders, including ground handlers, airport authorities, airlines, operations teams, and local public health agencies. The dynamic pandemic landscape required rapid operational scaling, including quick staff recruitment and increased testing capacity within hours in response to catalysts, such as rising case numbers in certain global regions, countries with limited sequencing capability, or newly identified variants (*7*). In addition, the program needed the ability to revert to baseline operations when the acute event concluded. In all scenarios, the private sector's ability to rapidly develop a prototype, pilot test it, and execute new solutions at scale was crucial in enabling CDC to achieve its vision and objectives in this arena.

Over the next 3 years, this public–private partnership enabled expansion of traveler-based genomic surveillance that had tested >600,000 travelers and >1,200 wastewater samples across 10 airports by August 2024 (Table). During the expansion, fast turnaround time from sample collection to reporting was critical, so the partners built a process that provided reporting of PCR results within 24–48 hours and whole-genome sequencing within 10 days of collection. The Traveler-based Genomic Surveillance Program evolved into a multimodal platform that included nasal, aircraft wastewater, and air sampling and a comprehensive approach for multipathogen detection (*11*). The private sector played a crucial role in the evolution of the program through scaled technology deployment, rapid iterations in response to changing conditions, and extending reach into areas typically beyond the traditional scope of public health.

To address skepticism about motives of private firms engaging in public health partnerships and the safeguards needed to secure public trust (*12*,*13*), it is essential to acknowledge concerns openly, emphasize shared goals, implement ethical oversight, prioritize long-term commitments, and highlight successful partnerships like those implemented during the COVID-19 pandemic (*14*). The World Health Organization's Global Genomic Surveillance Strategy for Pathogens with Pandemic Potential 2022–2032 underscores the value of multisectoral partnerships for its successful implementation (*15*). Furthermore, a report by the National Academies of Science, Engineering, and Medicine on optimizing public–private partnerships for clinical cancer research highlights that successful partnerships should focus on the public good, address large-scale problems and unmet needs, leverage the strengths of each sector beyond what either could achieve on its own, and promote the generation of information, knowledge, or data for the public use (*16*). The Traveler-based Genomic Surveillance program's public–private partnership exemplifies those criteria, demonstrating that multisectoral partnerships can be vital to public health before, during, and after crises.

About the Author

Dr. Friedman is an infectious disease physician and senior advisor in the Division of Global Migration Health, National Center for Zoonotic and Emerging Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. Her research interests include strengthening pathogen biosurveillance to improve global biosecurity, genomic surveillance and epidemiology, and wastewater and environmental surveillance.

References

1. Berman P, Cameron MA, Gaurav S, Gotsadze G, Hasan MZ, Jenei K, et al. Improving the response to future pandemics requires an improved understanding of the role played by institutions, politics, organization, and governance. PLOS Glob Public Health. 2023;3:e0001501. https://doi.org/10.1371/journal.pgph.0001501
2. Wegrzyn RD, Appiah GD, Morfino R, Milford SR, Walker AT, Ernst ET, et al. Early detection of severe acute respiratory syndrome coronavirus 2 variants using traveler-based genomic surveillance at 4 US airports, September 2021–January 2022. Clin Infect Dis. 2023;76:e540–3. https://doi.org/10.1093/cid/ciac461
3. Hamer DH, Rizwan A, Freedman DO, Kozarsky P, Libman M. GeoSentinel: past, present and future. J Travel Med. 2020;27:taaa219. https://doi.org/10.1093/jtm/taaa219
4. Oran DP, Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. Ann Intern Med. 2020;173:362–7. https://doi.org/10.7326/M20-3012
5. Galloway SE, Paul P, MacCannell DR, Johansson MA, Brooks JT, MacNeil A, et al. Emergence of SARS-CoV-2 B.1.1.7 lineage—United States, December 29, 2020–January 12, 2021. MMWR Morb Mortal Wkly Rep. 2021;70:95–9. https://doi.org/10.15585/mmwr.mm7003e2
6. Morfino RC, Bart SM, Franklin A, Rome BH, Rothstein AP, Aichele TWS, et al. Aircraft wastewater surveillance for early detection of SARS-CoV-2 variants—John F. Kennedy International Airport, New York City, August–September 2022. MMWR Morb Mortal Wkly Rep. 2023;72:210–1. https://doi.org/10.15585/mmwr.mm7208a3
7. World Health Organization. WHO TAG-VE statement on the meeting of 3 January on the COVID-19 situation in China. 2023 Jan 4 [cited 2023 Aug 19]. https://www.who.int/news/item/04-01-2023-tag-ve-statement-on-the-3rd-january-meeting-on-the-covid-19-situation-in-china
8. Bart SM, Smith TC, Guagliardo SAJ, Walker AT, Rome BH, Li SL, et al. Effect of predeparture testing on postarrival SARS-CoV-2–positive test results among international travelers—CDC traveler-based genomic surveillance program, four U.S. airports, March–September 2022. MMWR Morb Mortal Wkly Rep. 2023;72:206–9. https://doi.org/10.15585/mmwr.mm7208a2
9. Byrd KM, Bart SM, Smith TC, Loh SM, Rothstein AP, Guagliardo SAJ, et al. Goal! Goal! Goal! Detection of SARS-CoV-2 variants in travelers during the FIFA World Cup, Qatar 2022—CDC Traveler-Based Genomic Surveillance Program, November 2022–January 2023. Presented at: 2024 Epidemic Intelligence Service (EIS) Conference; April 23–26, 2024; Atlanta, GA, USA.
10. Bart SM, Rothstein AP, Philipson CW, Smith TC, Simen BB, Tamin A, et al. Notes from the field: early identification of the SARS-CoV-2 Omicron BA.2.86 variant by the Traveler-based Genomic Surveillance Program—Dulles International Airport, August 2023. MMWR Morb Mortal Wkly Rep. 2023;72:1168–9. https://doi.org/10.15585/mmwr.mm7243a3
11. Centers for Disease Control and Prevention (CDC). Traveler-based genomic surveillance for early detection of new SARS-CoV-2 variants [cited 2024 Dec 10]. https://wwwnc.cdc.gov/travel/page/travel-genomic-surveillance
12. Reich MR. Public-private partnerships for public health. Nat Med. 2000;6:617–20. https://doi.org/10.1038/76176
13. Horn R, Merchant J, Horn R, Merchant J, Bale M, Banner N, et al.; UK-FR+GENE (Genetics and Ethics Network) Consortium. Ethical and social implications of public-private partnerships in the context of genomic/big health data collection. Eur J Hum Genet. 2024;32:736–41. https://doi.org/10.1038/s41431-024-01608-9
14. Schmitt K. Why public health needs the private sector. Hopkins Bloomberg Public Health. 2023;Spring/Summer:2023 [cited 2024 Aug 9]. https://magazine.publichealth.jhu.edu/2023/why-public-health-needs-private-sector
15. Carter LL, Yu MA, Sacks JA, Barnadas C, Pereyaslov D, Cognat S, et al. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022–2032. Bull World Health Organ. 2022;100:239–239A. https://doi.org/10.2471/BLT.22.288220
16. National Academies of Sciences, Engineering, and Medicine. Optimizing public-private partnerships for clinical cancer research: proceedings of a workshop. Washington (DC): The National Academies Press; 2024.

Address for correspondence: Cindy R. Friedman, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H16-4, Atlanta, GA 30329-4018, USA; email: ccf6@cdc.gov

Respiratory Virus Detection and Sequencing from SARS-CoV-2–Negative Rapid Antigen Tests

Emmanuela Jules,[1] Charlie Decker,[1] Brianna Jeanne Bixler,[1] Alaa Ahmed, Zijing (Carol) Zhou, Itika Arora, Henok Tafesse, Hannah Dakanay, Andrei Bombin, Ethan Wang, Jessica Ingersoll, Kathy Bifulco, Jennifer K. Frediani, Richard Parsons, Julie Sullivan, Morgan Greenleaf, Jesse J. Waggoner, Greg S. Martin, Wilbur A. Lam, Anne Piantadosi

Genomic epidemiology offers insight into the transmission and evolution of respiratory viruses. We used metagenomic sequencing from negative SARS-CoV-2 rapid antigen tests to identify a wide range of respiratory viruses and generate full genome sequences. This process offers a streamlined mechanism for broad respiratory virus genomic surveillance.

The COVID-19 pandemic highlighted the importance of genomic epidemiology in understanding virus transmission and evolution, informing essential countermeasures from nonpharmaceutical interventions to vaccines. Massive global efforts in SARS-CoV-2 genomic surveillance were made possible by widespread diagnostic testing and the growth of new infrastructure and methods for sequencing and analysis (*1*). Most genomic surveillance pipelines in the United States obtained residual SARS-CoV-2–positive samples from clinical, public health, and commercial laboratories. That strategy was effective during the pandemic but difficult to maintain with the rise of at-home rapid antigen tests (*2*,*3*). As traditional sample sources declined, our group and others demonstrated that residual samples from rapid antigen tests could be used to generate and analyze full SARS-CoV-2 sequences for genomic surveillance (*4*–*6*).

In this study, we build upon that work by identifying, sequencing, and analyzing other respiratory viruses using residual swab samples from negative BinaxNOW COVID-19 antigen tests (Abbott, https://www.abbott.com). This multivirus approach is key because SARS-CoV-2 has transitioned to an endemic virus whose symptoms resemble those of other respiratory viruses (*7*). Thus, there is both a need for broad testing and an opportunity to expand genomic surveillance for respiratory viruses using self-collected samples.

Methods

In brief, participants were enrolled in a parent study evaluating novel viral diagnostic tests through the RADx program at the Atlanta Center for Microsystems Engineered Point-of-Care Technologies (Atlanta, GA, USA). The study protocol was approved by the Emory University Institutional Review Board and the Grady Health Research Oversight Committee (both in Atlanta). We performed RNA metagenomic sequencing as described (*8*), obtaining a median of 5.8 million reads per sample (Appendix 1, https://wwwnc.cdc.gov/EID/article/31/5/24-1191-App1.pdf; Appendix 2, https://wwwnc.cdc.gov/EID/article/31/5/24-1191-App2.xlsx). We used a 3-step

Author affiliations: Emory University School of Medicine, Atlanta, Georgia, USA (E. Jules, C. Decker, B.J. Bixler, A. Ahmed, Z.(C.) Zhou, I. Arora, H. Tafesse, H. Dakanay, A. Bombin, E. Wang, J. Ingersoll, J. Sullivan, J.J. Waggoner, W.A. Lam, A. Piantadosi); Emory Integrated Genomics Core, Winship Cancer Institute of Emory University, Atlanta (A. Ahmed); Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta (K. Bifulco, J.K. Frediani, R. Parsons); The Atlanta Center for Microsystems Engineered Point-of-Care Technologies, Atlanta (J.K. Frediani, R. Parsons, J. Sullivan, M. Greenleaf); Georgia Clinical and Translational Science Alliance, Atlanta (M. Greenleaf); Emory University Division of Pulmonary, Allergy, Critical Care Medicine and Sleep Medicine, Atlanta (G.S. Martin); Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta (W.A. Lam); Aflac Cancer and Blood Disorders Center of Children's Healthcare of Atlanta, Atlanta (W.A. Lam)

DOI: https://doi.org/10.3201/eid3113.241191

[1]These first authors contributed equally to this article.

bioinformatic approach to detect viruses (Appendix 1 Figure 1) using KrakenUniq (https://gihub.com/fbreitwieser/krakenuniq), blastn (https://blast.ncbi.nlm.nih.gov, and reference mapping (https://github.com/briannajeanne/metagen/tree/main). Our final criterion required coverage of ≥10% of the viral genome, or reads mapping to ≥3 nonoverlapping regions of the viral genome with ≥80% identity, similar to clinical diagnostic criteria that have previously been used for metagenomic sequencing (*9*).

Results

We collected negative BinaxNOW test samples from 53 persons during April–August 2023 (Appendix 1 Table), a period during which 68% of the BinaxNOW tests in the parent study were negative. All persons were symptomatic at the time of testing (Table), and the median interval between symptom onset and testing was 2 (range 0–9) days. Reverse transcription PCR (RT-PCR) was positive for influenza B in 3 samples and negative for influenza A, respiratory syncytial virus, and SARS-CoV-2 in all samples (Appendix 2).

Metagenomic sequencing identified a low level of SARS-CoV-2 in 1 sample and a different pathogenic human respiratory virus in 17 (33%) of the other 52 samples (Appendix 2). We detected parainfluenza viruses (n = 7), rhinoviruses (n = 5), influenza B (n = 3), seasonal coronaviruses (n = 2), and adenovirus (n = 1) (Figure 1). In 1 sample, we detected both influenza B and parainfluenza 2. In another sample positive for influenza B by RT-PCR, metagenomic sequencing did not identify influenza but identified human mastadenovirus E. Thus, excluding SARS-CoV-2, we detected a total of 18 viruses across 17 samples.

The duration of time between sample collection and nucleic acid extraction was similar for samples with a virus detected (median 6 [range 4–12] days) and samples with no virus detected (median 7 [range 5–19] days). RT-PCR for RNase P was positive in all samples tested, and the percentage of human reads was similar between samples with and without viruses detected (p = 0.07 by Mann-Whitney U test) (Appendix 2). We saw no difference in the total number of reads obtained for samples with and without viruses detected (p = 0.29 by Mann-Whitney U test).

We compared potential differences in symptoms between persons in whom a virus was detected and those in whom no virus was detected and observed the following disparities: congestion (83% vs. 63%), sore throat (78% vs. 54%), chills (61% vs. 37%), and headache (72% vs. 49%) (Table). However, none of those differences were statistically significant. The time between symptom onset and testing was similar between persons with a virus detected (median 2 [range 0–9] days) and those without a virus detected (median 2 [range 0–6] days).

Of the 18 viruses detected, we generated full viral genome sequences from 11 (61%) with >90% coverage and 71- to 24,000-fold depth (Appendix 2). Those 11 sequences consisted of parainfluenza

Table. Participant symptoms in study of respiratory virus detection and sequencing from SARS-CoV-2–negative rapid antigen tests*

Symptom	Total participants, N = 53	Participants with a virus detected, n = 18†	Participants with no virus detected, n = 35
Upper respiratory	47 (88.7)	17 (94.4)	30 (85.7)
Congestion/runny nose	37 (69.8)	15 (83.3)	22 (62.9)
Sore throat	33 (62.3)	14 (77.8)	19 (54.3)
Loss of sense of taste or smell	7 (13.2)	2 (11.1)	5 (14.3)
Lower respiratory	43 (81.1)	15 (83.3)	28 (80.0)
Cough	39 (73.6)	15 (83.3)	24 (68.6)
Shortness of breath	23 (43.4)	6 (33.3)	17 (48.6)
Gastrointestinal	15 (28.3)	6 (33.3)	9 (25.7)
Vomiting	4 (7.6)	3 (16.7)	1 (2.9)
Nausea	11 (20.8)	2 (11.1)	9 (25.7)
Diarrhea	2 (3.8)	1 (5.6)	1 (2.9)
Abdominal pain	7 (13.2)	2 (11.1)	5 (14.3)
Systemic	35 (66.0)	13 (72.2)	22 (62.9)
Fever, temperature >100.4 °F	18 (34.0)	7 (38.9)	11 (31.4)
Chills	24 (45.3)	11 (61.1)	13 (37.1)
Fatigue	27 (50.9)	11 (61.1)	16 (45.71)
Other	41 (77.4)	16 (88.9)	25 (71.4)
Headache	30 (56.6)	13 (72.2)	17 (48.6)
Joint pain	14 (26.4)	3 (16.7)	11 (31.4)
Muscle pain	31 (58.5)	10 (55.6)	21 (60.0)

*Values are no. (%) participants reporting each symptom at the time of testing. For symptom categories, the number of participants with ≥1 symptom in that category is reported.
†Includes 1 person in whom SARS-CoV-2 was detected at a low level and 17 persons in whom an alternative human pathogenic respiratory virus was detected.

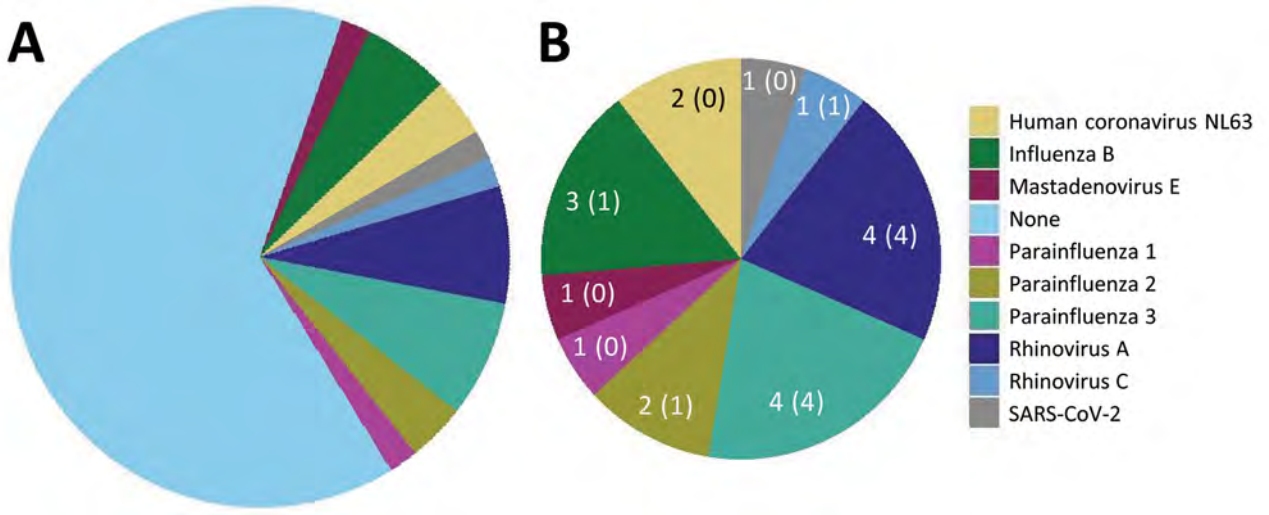


Figure 1. Frequency of human pathogenic respiratory viruses found in 53 residual samples from SARS-CoV-2–negative BinaxNOW tests (Abbott, https://www.abbott.com) in study of respiratory virus detection and sequencing from negative rapid antigen tests. Pie charts indicate the number of samples positive for each virus among all samples (A) and among the 18 positive samples (B). Numbers indicate the number of samples with a virus identified, followed in parentheses by the number of samples with a >90% complete genome sequence assembled.

3 (4/4 samples), parainfluenza 2 (1/2), rhinovirus (5/5), and influenza B (1/3).

We performed phylogenetic analysis of parainfluenza 3 as a proof-of-concept for genomic epidemiology studies and found substantial diversity. Using the lineage classification system described in Lee et al. (*10*), 2 of our sequences clustered with lineage A1 sequences from 2019–2023 (Figure 2, panel A), another clustered with lineage C sequences from Japan in 2023, and the fourth with lineage C sequences from the United States collected during 2015–2017 (Figure 2, panel B), all with high bootstrap support

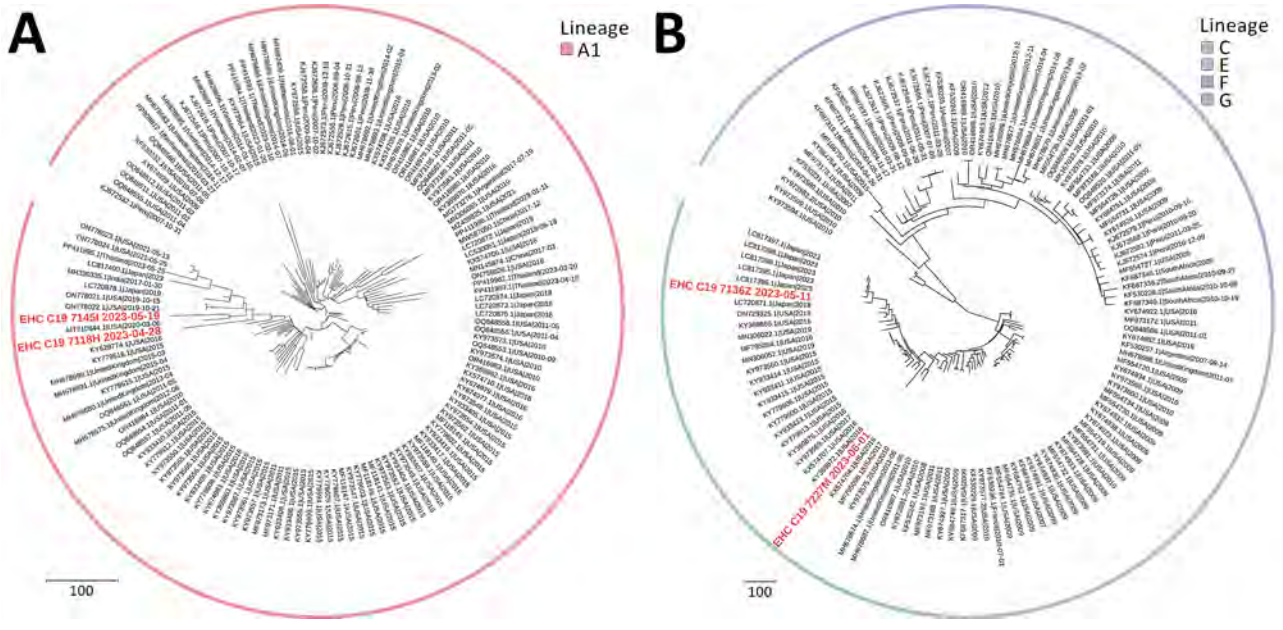


Figure 2. Phylogenetic analysis of parainfluenza 3 virus sequences in study of respiratory virus detection and sequencing from SARS-CoV-2–negative rapid antigen tests. The names of sequences obtained in this study are bold and in red, and reference sequences are in black. The outer ring indicates virus lineage. A) Representative sequences from lineage A1; B) representative sequences from lineages C, E, F, and G. Each tree is a maximized parsimony subtree using downsampled data from the full analysis in Appendix 1 Figure 2 (https://wwwnc.cdc.gov/EID/article/31/5/24-1191-App1.pdf), for ease of visualization. GenBank accession numbers are provided for reference sequences. Scale bars indicate number of substitutions per site.

(Appendix 1 Figure 2). Of note, only ≈450 complete parainfluenza 3 virus sequences are available; the data from our small study represent nearly 1% of this number, underscoring the opportunity to easily expand genomic surveillance using this approach.

In addition to human pathogenic respiratory viruses, we detected >100 viruses of no clinical significance, including bacteriophages and plant viruses, many of which were also detected in our negative controls (Figure 3; Appendix 1 Figures 3, 4). Similarly, we found mastadenovirus C in about one third of all samples and negative controls, all with low genome coverage (Appendix 3, https://wwwnc.cdc.gov/EID/article/31/5/24-1191-App3.xlsx). Those findings are all consistent with environmental or reagent contaminants. Herpesviruses were reported in many samples by KrakenUniq and blastn but generally were not confirmed by reference mapping. One adult participant had confirmed detection of human herpesvirus 6, which, given the participant's age, more likely reflects latent virus than acute infection. Overall, 1,367 viral taxa were identified by Kraken Uniq, only 254 (18.6%) were confirmed by blastn, and only 137 (53.9% [10% of total]) met our criteria for detection (Appendix 3), highlighting the importance of confirmatory steps in metagenomic analysis.

Discussion

Our study demonstrates that RNA metagenomic sequencing of residual swab samples from negative BinaxNOW COVID-19 tests can be used to detect a broad range of respiratory viruses, including rhinoviruses, parainfluenza viruses, influenza B, seasonal coronaviruses, and adenovirus. All of those viruses have overlapping symptoms, both with one another and with SARS-CoV-2, underscoring the need for

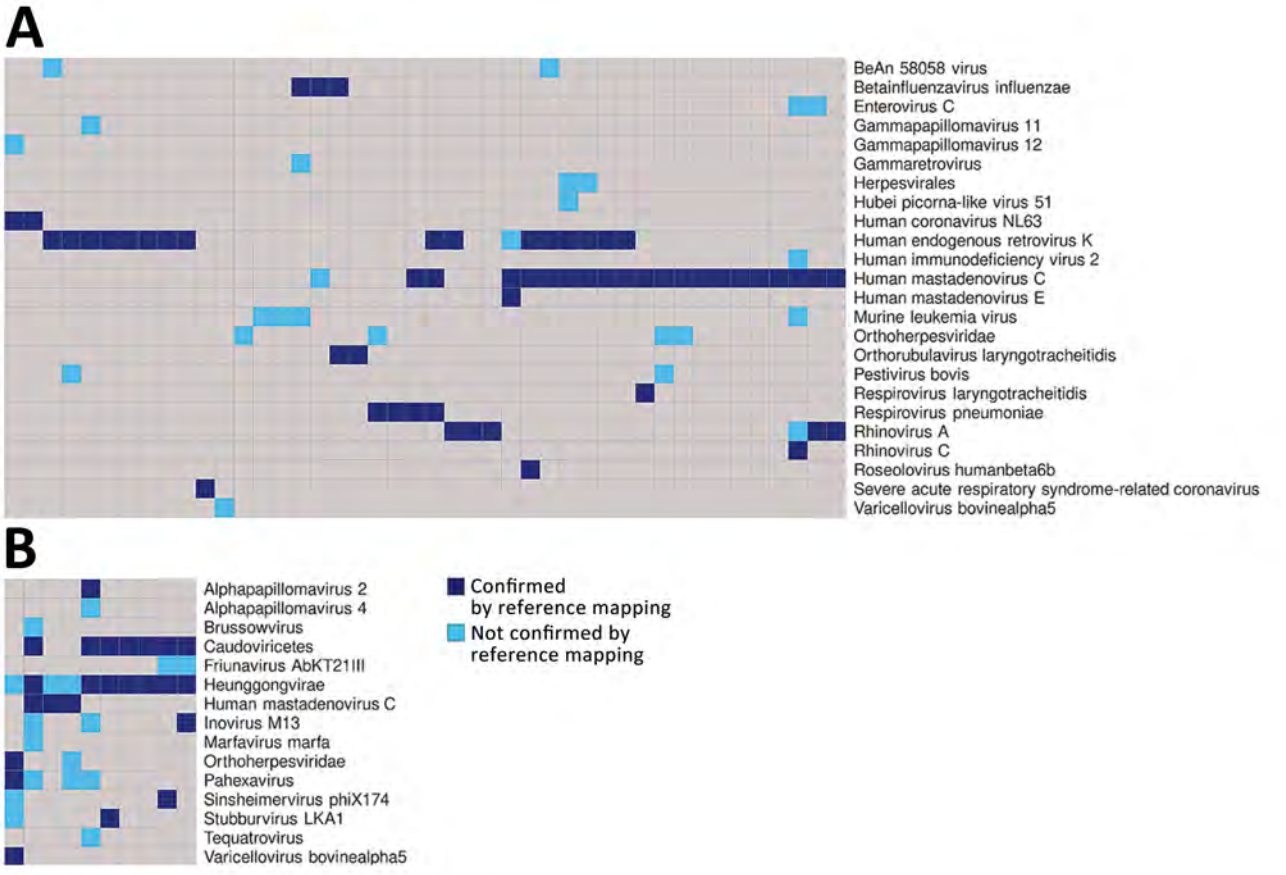


Figure 3. Plot of the viral taxa (rows) that were detected in each sample (columns) in study of respiratory virus detection and sequencing from SARS-CoV-2–negative rapid antigen tests. A) Results from samples used in this study; B) results from negative controls. Dark blue boxes indicate viruses that were detected by both KrakenUniq (https://github.com/fbreitwieser/krakenuniq) and blastn (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) and were confirmed by reference mapping (covering ≥3 distinct regions or 10% of the reference virus genome). Light blue boxes indicate viruses that were detected by both KrakenUniq and blastn but not confirmed by reference mapping. This figure only includes results for which ≥1 read mapped to a reference genome sequence. Further detail including sample identifiers is shown in Appendix 1 Figures 3,4 (https://wwwnc.cdc.gov/EID/article/31/5/24-1191-App1.pdf).

multivirus testing approaches. Although our study was not designed for clinical diagnosis, metagenomic sequencing is increasingly used clinically, and our results illustrate the need for rigorous analysis techniques and careful interpretation.

Of note, only 33% of samples had a human pathogenic respiratory virus. This finding is similar to that of our previous study, in which alternative respiratory viruses were detected in only 40% of SARS-CoV-2–negative persons using residual clinical samples early in the pandemic (*8*). Possible explanations include persons with a noninfectious syndrome, a bacterial or other nonviral infection, or a virus present at a low level. Some persons could also have been infected with a DNA virus not optimally captured by RNA sequencing. However, we detected adenovirus, the most prevalent respiratory DNA virus. Among common RNA viruses, we did not detect influenza A or respiratory syncytial virus, which we attribute to the winter-predominant seasonality of these viruses, whereas our samples were collected in spring and summer.

Of note, of the 18 viruses detected, we were able to generate full viral genome sequences from 11 (61%) using moderate sequencing depths. Thus, the single laboratory technique of metagenomic sequencing can not only identify diverse respiratory viruses but also contribute to their genomic surveillance. The surprisingly high depth of genome coverage achieved for many sequences indicates that throughput and cost can be improved by reducing total sequencing reads from each sample in future studies.

By combining metagenomic sequencing with the use of residual antigen test samples, we demonstrate a mechanism for convenient and broad respiratory virus surveillance. Our study used BinaxNOW tests, which conveniently preserve the used swab within the kit cassette; future work is needed to evaluate this approach using rapid antigen test strips themselves, as previously demonstrated for SARS-CoV-2 sequencing (*5*). In addition, future studies would benefit from a regulatory framework in which, after rigorous analysis and careful interpretation, clinically significant results can be returned to study participants, who are likely curious about the presence of other respiratory viruses when rapid antigen testing is negative for COVID-19. In conclusion, our study illustrates that residual samples from self-collected antigen tests can be a powerful sample source for investigating the genomic epidemiology of a broad range of respiratory viruses, building upon the strong foundations for viral surveillance established during the COVID-19 pandemic.

Acknowledgments

We thank the study participants.

All raw sequencing data (cleaned of human reads) is available in the National Center for Biotechnology Information Sequence Read Archive under BioProject PRJNA1144955, and assembled virus genome sequences are available in GenBank with accession numbers listed in Appendix 2.

This work was supported by National Institutes of Health (NIH) awards U54 EB027690 02S1, U54 EB027690 03S1, and U54EB027690 03S2 UL1 TR002378 and by the Centers for Disease Control and Prevention–funded Georgia Pathogen Genomics Center of Excellence contract 40500-050-23234506. B.B. was supported by NIH award F31ES031845. This study was supported in part by the Emory Integrated Genomics Core (EIGC) (RRID:SCR_023529), which is subsidized by the Emory University School of Medicine and is one of the Emory Integrated Core Facilities. Additional support was provided by the Georgia Clinical & Translational Science Alliance of the NIH under award UL1TR002378. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NIH.

About the Author

Ms. Jules is currently a research specialist in the Department of Pathology and Laboratory Medicine in the Emory University School of Medicine. She will be applying to medical school with the aspiration of becoming a family doctor and expanding healthcare to underserved communities.

References

1. Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B, Fouchier RAM, et al. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. Nat Med. 2021;27:1518–24.
2. Ritchey MD, Rosenblum HG, Del Guercio K, Humbard M, Santos S, Hall J, et al. COVID-19 self-test data: challenges and opportunities—United States, October 31, 2021–June 11, 2022. MMWR Morb Mortal Wkly Rep. 2022;71:1005–10.
3. Rader B, Gertz A, Iuliano AD, Gilmer M, Wronski L, Astley CM, et al. Use of at-home COVID-19 tests—United States, August 23, 2021–March 12, 2022. MMWR Morb Mortal Wkly Rep. 2022;71:489–94.
4. Nguyen PV, Carmola LR, Wang E, Bassit L, Rao A, Greenleaf M, et al. SARS-CoV-2 molecular testing and whole genome sequencing following RNA recovery from used BinaxNOW COVID-19 antigen self tests. J Clin Virol. 2023;162:105426.
5. Martin GE, Taiaroa G, Taouk ML, Savic I, O'Keefe J, Quach R, et al. Maintaining genomic surveillance using whole-genome sequencing of SARS-CoV-2 from rapid antigen test devices. Lancet Infect Dis. 2022;22:1417–8.

6. Hassouneh SA, Trujillo A, Ali S, Cella E, Johnston C, DeRuff KC, et al. Antigen test swabs are comparable to nasopharyngeal swabs for sequencing of SARS-CoV-2. Sci Rep. 2023;13:11255.
7. Geismar C, Nguyen V, Fragaszy E, Shrotri M, Navaratnam AMD, Beale S, et al. Symptom profiles of community cases infected by influenza, RSV, rhinovirus, seasonal coronavirus, and SARS-CoV-2 variants of concern. Sci Rep. 2023;13:12511.
8 . Babiker A, Bradley HL, Stittleburg VD, Ingersoll JM, Key A, Kraft CS, et al. Metagenomic sequencing to detect respiratory viruses in persons under investigation for COVID-19. J Clin Microbiol. 2020;59:e02142-20.
9. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 2019;29:831–42.
10. Lee K, Park K, Sung H, Kim MN. Phylogenetic lineage dynamics of global parainfluenza virus type 3 post-COVID-19 pandemic. MSphere. 2024;9:e0062423. https://doi.org/10.1128/msphere.00624-23

Address for correspondence: Anne Piantadosi, Woodruff Memorial Research Building, 101 Woodruff Cir, Atlanta, GA 30322, USA; email: anne.piantadosi@emory.edu

Large-Scale Genomic Analysis of SARS-CoV-2 Omicron BA.5 Emergence, United States

Kien Pham,[1] Chrispin Chaguza, Rafael Lopes, Ted Cohen, Emma Taylor-Salmon, Melanie Wilkinson, Volha Katebi, Nathan D. Grubaugh, Verity Hill

The COVID-19 pandemic has been marked by continuous emergence of novel SARS-CoV-2 variants. Questions remain about the mechanisms with which those variants establish themselves in new geographic areas. We performed a discrete phylogeographic analysis on 18,529 sequences of the SARS-CoV-2 Omicron BA.5 sublineage sampled during February–June 2022 to elucidate emergence of that sublineage in different regions of the United States. The earliest BA.5 sublineage introductions came from Africa, the putative variant origin, but most were from Europe, matching a high volume of air travelers. In addition, we discovered extensive domestic transmission between different US regions, driven by population size and cross-country transmission between key hotspots. We found most BA.5 virus transmission within the United States occurred between 3 regions in the southwestern, southeastern, and northeastern parts of the country. Our results form a framework for analyzing emergence of novel SARS-CoV-2 variants and other pathogens in the United States.

SARS-CoV-2, the causative virus of the COVID-19 pandemic, has demonstrated the ability to evolve into novel variants. The Omicron (B.1.1.529) variant, detected in late 2021 in southern Africa, was deemed a variant of concern by the World Health Organization and soon became dominant in the United States and worldwide (*1*). Omicron lineages are defined by ≈60 mutations, 32 of those in the spike protein, that have granted evolutionary advantage over co-circulating variants because they enhance intrinsic transmissibility and immune escape (*1*–*4*). New Omicron sublineages, as well as recombinants, have subsequently emerged (*5*,*6*). Furthermore, the complex mosaic of immunity in the human population, likely caused by different levels of vaccination or previous infection, indicates the landscape for SARS-CoV-2 variant emergence has changed since the start of the pandemic. With ongoing variant emergence, changing patterns of spread must be elucidated, because those patterns have considerable implications in prevention and mitigation plans.

Recent advances in virus sequencing and phylogenetics has enabled the timely use of large-scale phylogenetic analyses to determine SARS-CoV-2 dynamics (*7*). Studies have been conducted globally, including in Brazil (*8*), The Gambia (*9*), and New Zealand (*10*), to explore the origins, emergence, and dynamics of SARS-CoV-2 variants. In the United Kingdom, multiple analyses of national-level spread from major population centers have been conducted, showing early spread from the origin(s) of introduction and the seeding and subsequent local transmission to new locations (*11*–*14*). Furthermore, studies in the United States have shown the increased risk for virus importation among states compared with international origin (*15*), the importance of superspreading events promoting early transmission (*16*), and effects of international introductions of the Alpha variant (*17*).

We used a Bayesian discrete phylogeographic framework to determine the introduction and spread of a novel SARS-CoV-2 lineage into different regions of the United States. We focused on Omicron sublineage BA.5 during its global emergence period within the first 6 months of 2022 because of its rapid national spread, long-term persistence, and public

Author affiliations: Yale School of Public Health, New Haven, Connecticut, USA (K. Pham, C. Chaguza, R. Lopes, T. Cohen, E. Taylor-Salmon, N.D. Grubaugh, V. Hill); Yale School of Medicine, New Haven (E. Taylor-Salmon); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (M. Wilkinson, V. Katebi); Yale University, New Haven (N.D. Grubaugh)

DOI: https://doi.org/10.3201/eid3113.240981

[1]Current affiliation: Program for Appropriate Technology in Health (PATH) Southeast Asia, and Hanoi Medical University, Hanoi, Vietnam.

health importance (Figures 1, 2). Omicron sublineage BA.5 established itself during times of lower SARS-CoV-2 incidence and remained prominent until the end of 2022 (Figure 1) (*18*). However, BA.5 never achieved complete dominance in the United States, co-circulating instead with other major Omicron sublineages, such as BA.2.12.1, BA.4, and XBB.1 (*5*). Moreover, BA.5 dissemination occurred on the background of a highly immune population because of vaccination and previous infections with other Omicron sublineages (*5*,*19*). Newer variants are likely to be introduced onto a similar immune landscape; thus, the dynamics of BA.5 introductions and dissemination offer a useful case study for how new lineages might spread across the United States. Furthermore, because most social and travel restrictions have been lifted and data streams have become more limited, clarifying within-country spread will enable targeted surveillance activities in the future.

Methods

Dataset Generation

To define our study period, we balanced having a large enough time period to cover key events with avoiding an intractably large final dataset. Therefore, we compared the frequencies of Omicron BA.5 with other variants in each continent and selected the week for which every continent had a BA.5 frequency of ≥25% (week commencing June 13, 2022). That cutoff is somewhat arbitrary, but the speed of BA.5 spread on a continental level meant that changing the threshold only resulted in a few weeks' difference either way (e.g., changing it to 50% added 2 weeks to the dataset; changing to 10% resulted in 1 week less).

We assembled a dataset of BA.5 whole-genome sequences sampled in the United States and globally during the inferred emergence period, estimated to be during February–June 2022. First, we downloaded all sequences that had complete location and collection date metadata from GISAID (https://www.gisaid.org) and had the BA.5 pango lineage designation (*20*). We then used Nextclade (*21*) to filter for low-quality control score and genome coverage of <70%. To mitigate sampling bias, we categorized global BA.5 data by continent.

Within the United States, the genomic surveillance policy is largely decided by the individual state, causing potential bias in data from each region (*22*,*23*). To ameliorate that disparity, we divided the country into 10 regions according to the locations of the 10 regional offices of the US Department of Health and Human Services (DHHS) (Figure 3). To account for possible selection bias from that heterogeneity, we subsampled the full dataset in 1-week windows proportional to the population of each region. We chose to use population because case counts are also biased between and within countries (especially those as large as the United States) because of varying availability of resources and case definitions. We felt this choice was appropriate for SARS-CoV-2 because so much of each country's population was infected; thus, in this specific case, we decided that population was a less biased metric on which to base our subsampling scheme than case counts. Specifically, we used the population proportion of the region, either global continent or US region, and multiplied by the total number of BA.5 genomes to find 1 fixed number of genomes (selected every week for that region). The final dataset selected for analysis consisted of 18,529 sequences, 9,350 from the United States and 10,258 from non-US countries (Table). For the emergence period, the earliest sample was collected on February 25, 2022, whereas the latest sample was collected on June 19, 2022.

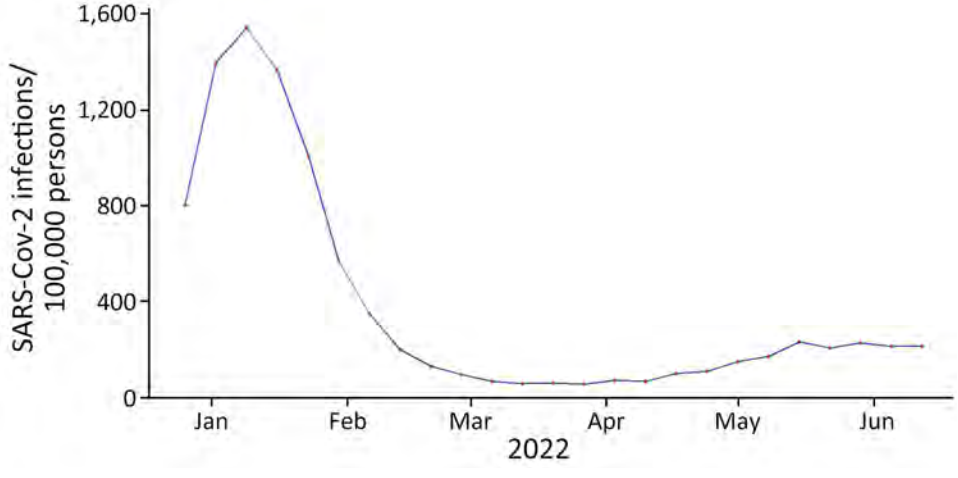


Figure 1. Number of estimated weekly SARS-CoV-2 infections in study of large-scale genomic analysis of Omicron BA.5 emergence, United States, January 2022–June 2022. Source: https://covidestim.org

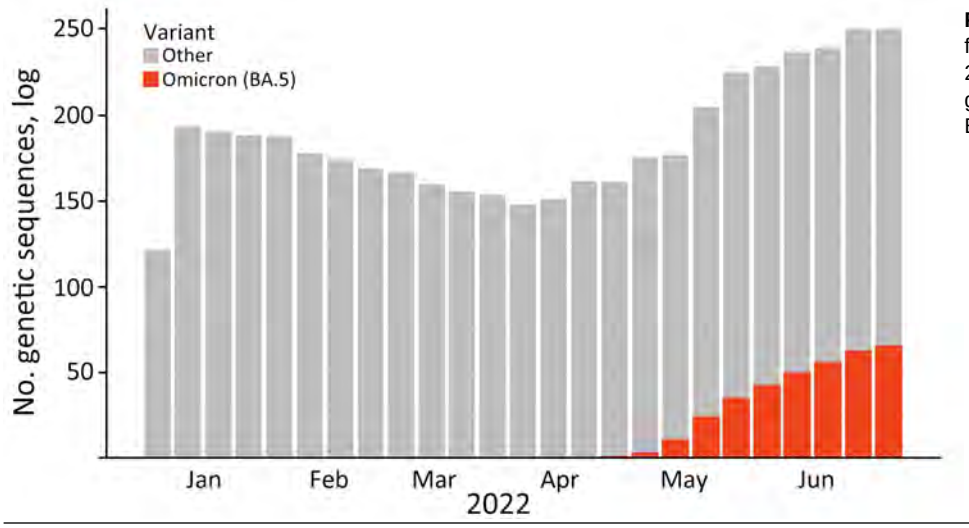


Figure 2. SARS-CoV-2 variant frequency during January–June 2022 in study of large-scale genomic analysis of Omicron BA.5 emergence, United States.

Phylogeographic Analysis

We performed multiple sequence alignments by using the Nextclade tool, Nextalign (*21*), and Wuhan-Hu-1/2019 as the reference genome (GenBank accession no. MN908947.3). We then constructed a maximum-likelihood phylogenetic tree by using IQ-TREE version 2.2.2 (*24*), the Hasegawa-Kishino-Yano nucleotide substitution model (*25*), and outgroup rooting on the MN908947 reference genome. We assessed the temporal signal by using TempEST version 1.5.3 (*26*) and found the timeframe for the dataset was too short to have a strong temporal signal (Appendix Figure 1, https://wwwnc.cdc.gov/EID/article/31/13/24-0981-App1.pdf). We were still able to prune molecular clock outliers (*27*) by using jclusterfunk version 0.0.25 (https://github.com/snake-flu/jclusterfunk).

Because of the large size of the genomic dataset, we used the alternative tree likelihood function in BEAST version 1.10.4, which was developed for efficient estimation of large phylogenies (*13*,*28*). We used maximum-likelihood trees described previously for the topologic estimation and time-calibrated those trees approximately by using TreeTime version 0.9.4 (*29*) to reduce the percentage of states that needed to be discarded for burn-in.

Because of the low temporal signal in the dataset, we fixed the clock rate at 8×10^{-4} substitutions/site/year, as previously described (*30*–*32*). We used the nonparametric Skygrid coalescent model (*33*) with 23 grid points defined according to approximately equal intervals within the global emergence period. We ran 2 Markov chain Monte Carlo chains for 1 billion iterations each to ensure convergence with the same part of the posterior distribution. We used Tracer version 1.7.1 d to assess convergence after run completion and discarded 10% of Markov model states for burn-in (*34*).

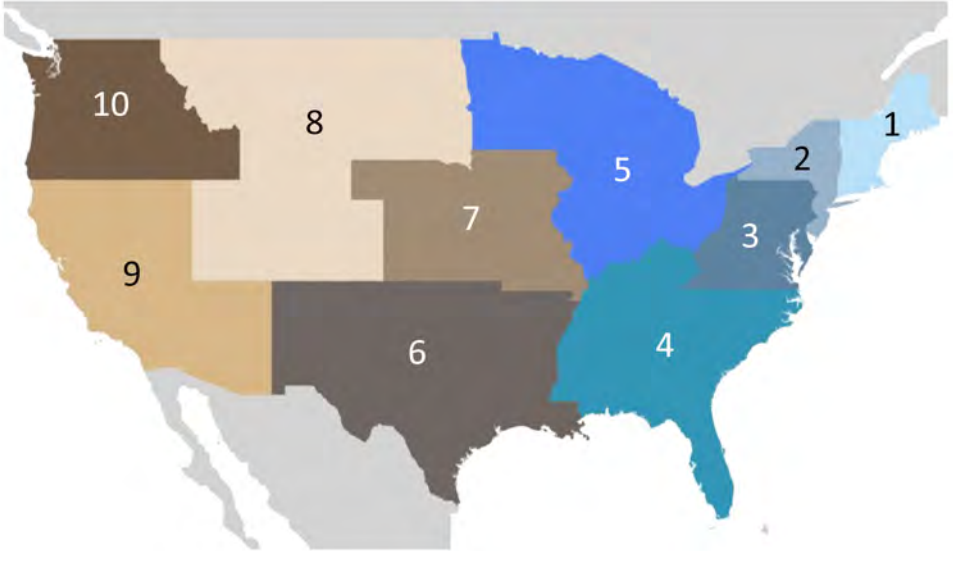


Figure 3. Ten regions of the United States evaluated in large-scale genomic analysis of SARS-CoV-2 Omicron BA.5 emergence. Regions have been designated by the US Department of Health and Human Services (https://www.hhs.gov/about/agencies/regional-offices/index.html).

Table. Final genomic sequence dataset for BA.5 discrete phylogeographical analysis in study of SARS-CoV-2 Omicron BA.5 emergence, United States

Locations	No. sequences
Global	
Africa	1,579
Asia	3,273
Europe	1,589
North America	1,281
Oceania	455
South America	1,388
US regions*	
1	389
2	1,050
3	637
4	1,814
5	1,339
6	1,029
7	314
8	376
9	1,637
10	379

*Regions are shown in Figure 3.

We chose 1 random tree from the post–burn-in posterior distribution from the previous analysis to use as the fixed tree in a discrete trait analysis. We analyzed 2 separate geographic scales: 1 analysis at the global level, which included 6 continents (Africa, Asia, Europe, North America without the United States, Oceania, and South America) and the United States as a country; and 1 analysis at the US national level, which included 10 DHHS regions using the continental dataset as background global context. In both analyses, we used an asymmetric continuous-time Markov chain to estimate transition rates between locations. Each chain ran for 2 million states; we discarded 10% of Markov model states for burn-in.

For the international analysis, we used custom Python scripts to estimate the average number of introductions across each tree in the posterior distribution and then selected a final tree closest to that average number. For the domestic analysis, we chose a random tree in the posterior distribution to maintain stability of clades within the United States across the analysis. We defined an introduction event as the point in which a node is in a different location than its parent, either originating from another continent into the United States (for international introduction) or between US regions (for domestic analysis). We did not account for reintroduction within the same clade; thus, once the location changed to the United States, that clade was counted as only 1 introduction event. If a node in 1 subtree coincided with a node in another subtree, we only counted the node that had the older root and eliminated the other. We determined the size of an introduction to be the number of sequences that immediately followed a change in location within a node. We estimated the time of introduction as halfway between the first US/domestic location node and its parent. We generated figures by using custom Python scripts and trees by using the Baltic Python package (https://github.com/evogytis/baltic).

To examine drivers of BA.5 domestic spread, we constructed a linear regression model that incorporated geographic proximity and population. Within the model, the proportion of directional domestic introductions between a pair of US regions was the outcome; the 2 independent variables were the binary neighboring relationship between that pair and the numeric total population of the 2 regions. We obtained population data from the US Census Bureau (https://www.census.gov).

Travel Data

To examine possible factors affecting BA.5 spread in different US regions, we collected data for monthly international and domestic air travel into US states during February–June 2022 (*34*). Those data were adjusted air passenger estimates, sampled according to ticket sales and reporting from airline carriers and assumed to represent 100% of the market. Adjusted travel volume represents the aggregate number of passenger journeys, not necessarily unique persons. We defined passenger journeys as airline transport between original embarkment and disembarkment in the United States. Both direct and indirect (i.e., connecting) flights were included.

Data Availability

The flight travel volume data were provided by OAG Aviation Worldwide Ltd. OAG Traffic Analyser, version 2.6.1 (http://analytics.oag.com/analyser-client/home; accessed 2023 Apr 24). The data were used under the US Centers for Disease Control and Prevention license for the current study and so are not publicly available. The authors are available to share the air passenger data upon reasonable request and with the permission of OAG Aviation Worldwide Ltd.

We obtained all genomic data from GISAID (acknowledgements table at https://doi.org/10.55876/gis8.240620dg). The XML files and outputs from the BEAST analyses are also available (https://github.com/grubaughlab/2025_paper_BA.5_United-States).

Results

International Introductions of Omicron BA.5 Sublineage into the United States

We examined the dynamics of BA.5 global introductions into the United States by using a discrete

phylogeographic analysis at the continent level and between US regions (Figure 3) and reconstructed introductions across the resulting phylogeny (Figure 4). An average of 1,168 (95% CI 1,137–1,198) introductions occurred from other continents into the United States across the posterior distribution of the entire period (January 1, 2022, through the week of June 13, 2022). The inferred time of the first introduction into the United States was the second week of February 2022, nearly 3 weeks before the collection date of the earliest US sequence on February 26, 2022 (Figure 4). During the earlier part of this emergence period (until mid-May 2022), most (68%) introductions were from international importation (Figure 5), despite air travel in the United States being predominantly between US regions during the study period (domestic volume was ≈80% of all air travel volume) (Figure 6). After BA.5 became established in the United States in mid-May 2022 (Figure 2), 72% of the between-region introductions came from domestic sources (Figure 7). During the entire study period, most international introductions came from Asia (27.8%), Europe (26.3%), and Africa (14.7%) (Figure 8, panel A).

We observed a chronological change in the relative dominance of continents as origins of BA.5 introductions into the United States (Figure 8). Introductions from Africa, despite only representing 14.7% of total BA.5 international introductions, comprised 41.9% of all international introductions before mid-May 2022. A high rate of introductions

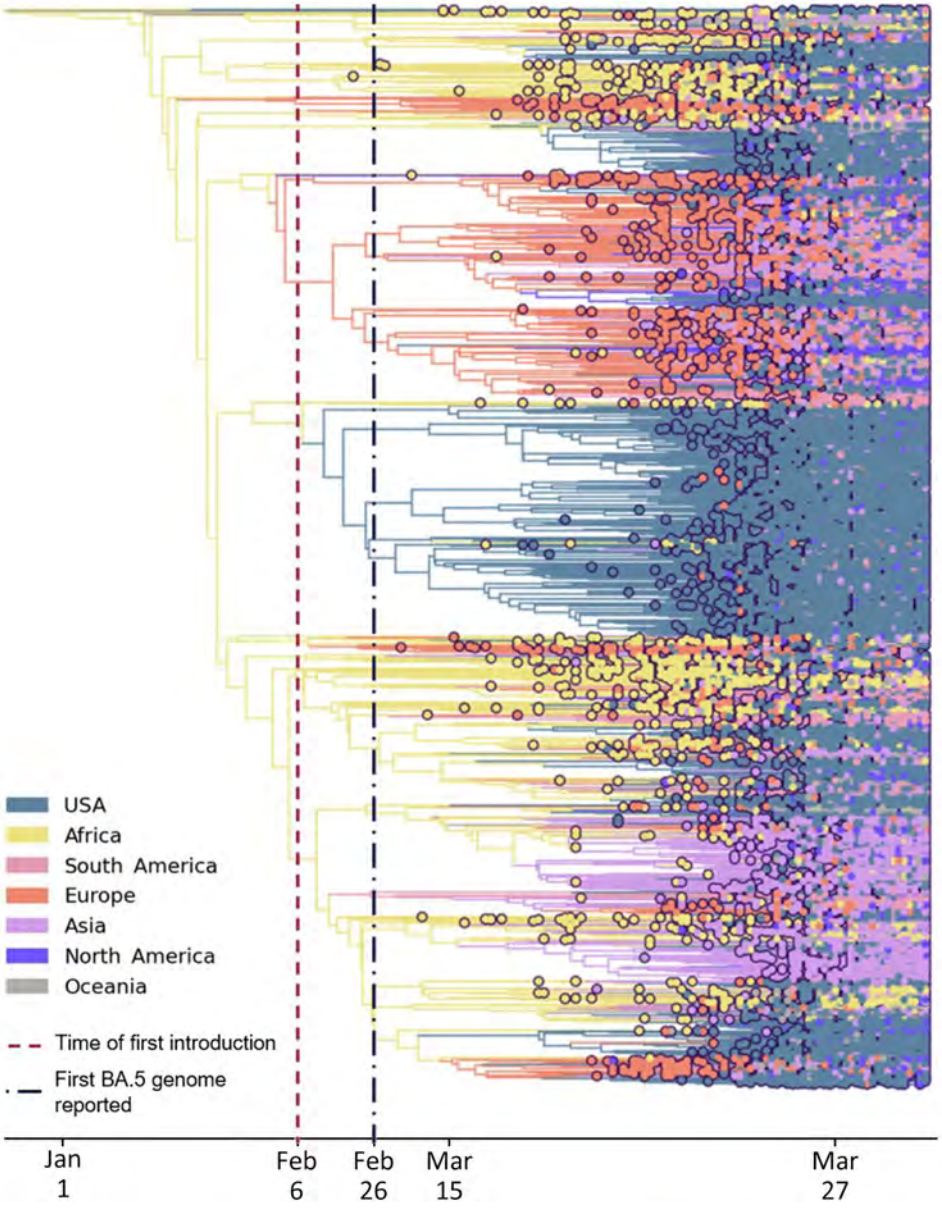


Figure 4. Time-scaled phylogeographic analysis of SARS-CoV-2 Omicron BA.5 sequences in the United States during January–June 2022. Analysis of BA.5 emergence was conducted by using 18,529 sequences collected globally and in the United States. Purple dotted line indicates the inferred date of the first introduction. The blue dotted line indicates the first sample of BA.5 sequenced in the United States. Colors indicate origin of the BA.5 variant.

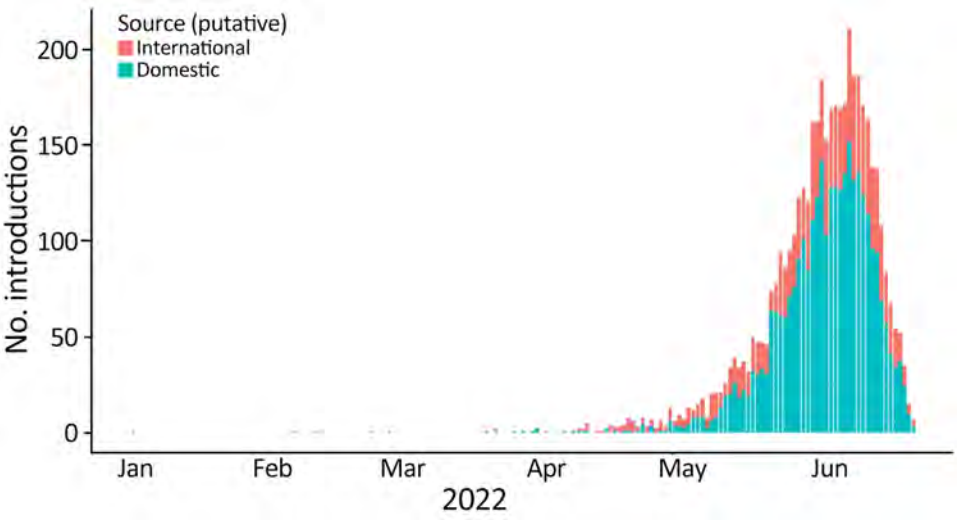


Figure 5. Numbers and timeline of domestic and international introductions of SARS-CoV-2 Omicron BA.5 in the United States during January–June 2022.

from Africa into all 10 US regions occurred, despite low travel volumes (Figure 9; Appendix Figures 2, 3). Indeed, Africa had the highest ratio of BA.5 introductions per travel volume, at ≈0.3 introductions per 1,000 passengers (Figure 9, panel A), likely because the BA.5 sublineage originated in Africa. As BA.5 prevalence increased globally, introductions from Europe, Asia, and North America became more critical (Figures 4, 5, 8), matching high travel volumes from those areas. Therefore, early emergence was determined by the variant's geographic origin, whereas later introductions were connected to travel volume.

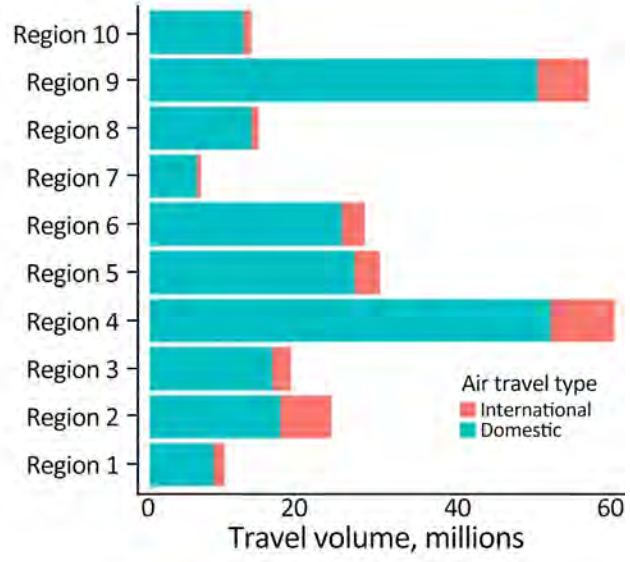


Figure 6. Air travel volume into different regions of the United States in study of large-scale genomic analysis of SARS-CoV-2 Omicron BA.5 emergence, January–June 2022. Domestic and international air travel volume are indicated. Regions designated by the US Department of Health and Human Services are shown in Figure 3.

We examined the effect of timing on the size of international introduction events. During the first BA.5 introduction into the United States in early February 2022 and its detection ≈3 weeks later, 5 total introductions occurred (Figures 4, 8). Although 4 of those were singletons, 1 introduction from Africa during late February contained 3,980 sequences, the largest during the entire study period (Figures 8, panel B; Figure 9, panel B). Cluster size was highest during early introductions and decreased over time (Figure 8, panel B). Introduction events from Africa, most occurring earlier during the study period, tended to have high outbreak clade sizes; 9 clusters had >100 sequences (Figure 9, panel B). Introductions from Europe had only 4 clusters with >100 sequences; no other global regions had clusters of that size (Figure 9, panel B).

We found 2 main phases of BA.5 emergence in the United States. Large introductions from Africa dominated the early emergence phase before May 2022. As prevalence increased globally, international introductions had greater ties to air travel volume; hence, more introductions came from Europe, Asia, and North America. Because of a decrease in the susceptible population and possible behavior changes after an uptick in Omicron BA.5 cases, introductions from Europe, Asia, and North America did not expand as much as the earlier events from Africa.

Domestic Movement of Omicron BA.5 in the United States

To evaluate BA.5 transmission within the United States, we performed a discrete phylogeographic analysis using 10 DHHS-defined regions (Figure 3; Appendix Table 1). We inferred 3,137 within-country introductions across a single posterior tree, ≈70% of total introductions across the entire study period. Early

international introduction events were followed by substantial domestic transmission (Figures 4, 5), and all 10 DHHS regions received >50% of their introductions from domestic sources (Figure 6). Those domestic movements grew in proportion throughout the study period and overtook the number of international introductions (Figures 5), aligning with the high (80%) proportion of domestic air travel (Figure 6).

No noticeable geographic structure within the phylogeny was observed (i.e., sequences from different locations were intermixed, implying frequent interregion transmission during the emergence period) (Figure 4). Inspection of the 3 largest and earliest US clades, rooted in region 2 (several northeastern states, including New York), region 9 (southwestern states, including California), and region 4 (southeastern states, including Florida) (Figure 10), indicated that geographically close locations tended to have more interregion movement. Clades from region 2 and 4 were primarily transmitted to other East Coast regions, and clades from region 9 were transmitted to other West Coast and West/Central regions (Figures 10, 11). Nonetheless, interactions between regions 4 and 9, and to a lesser extent regions 2 and 5 (including Illinois), indicated coast-to-coast spread was a critical BA.5 emergence mechanism.

Several key hotspots for transmission existed. All DHHS regions had considerably higher introduction counts originating from regions 4 and 9 (Figure 12). The interaction rate between regions 4 and 9 and other regions represented 71.6% of total domestic BA.5 movements (Figure 12). Correspondingly, regions

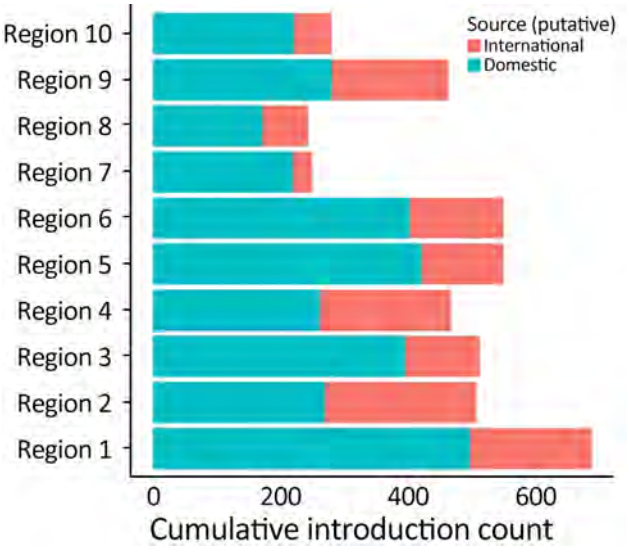


Figure 7. Total number of introductions of Omicron BA.5 into regions of the United States in study of large-scale genomic analysis of SARS-CoV-2 BA.5 emergence, January–June 2022. Cumulative numbers during the study period are indicated according to domestic or international origin. Regions designated by the US Department of Health and Human Services are shown in Figure 3.

4 and 9 also had the highest volumes of both international and domestic air travel (Figure 6). Region 1 (New England) had the highest (≈70%) number of incoming domestic introductions originating from regions 4 and 9 (Figure 12). Therefore, the strong transmission from regions 4 and 9 likely underpinned BA.5 emergence in the United States. We also theorize that region 1 was the top recipient of domestic introduction events because of the higher rate of interstate

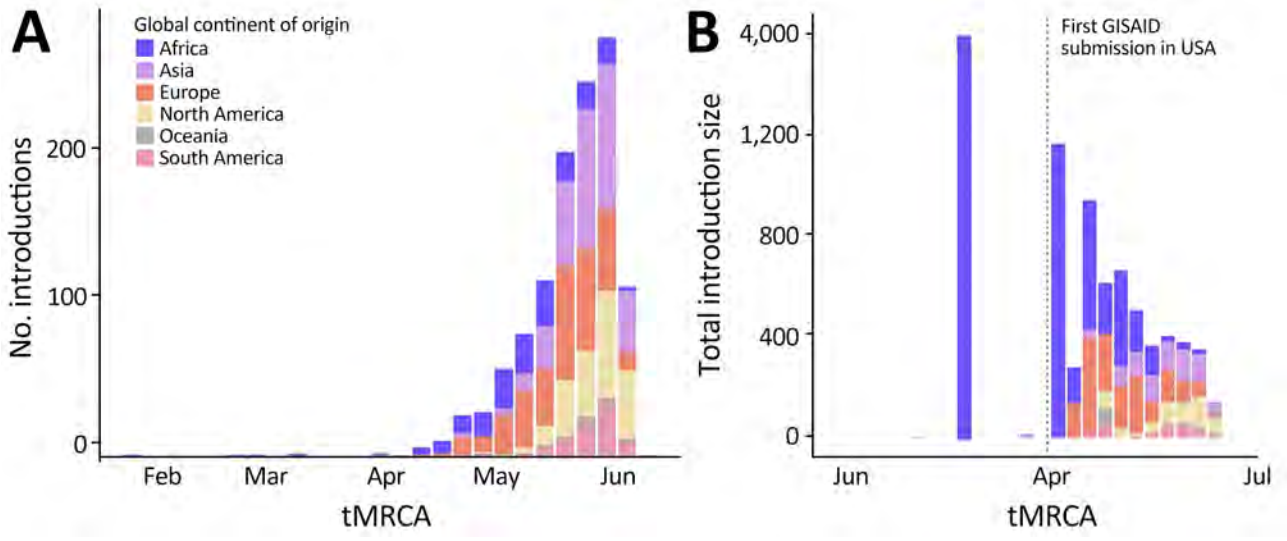


Figure 8. Spatiotemporal dynamics of international introductions of SARS-CoV-2 Omicron BA.5 lineage into the United States during February–June 2022. A) Numbers and timeline of BA.5 introduction events according to continent. B) Total introduction cluster size (number of sequences) of BA.5 international introduction events into the United States during the entire study period. Size was determined by the number of sequences per introduction. GISAID, https://www.gisaid.org; tMRCA, time to most recent common ancestor.

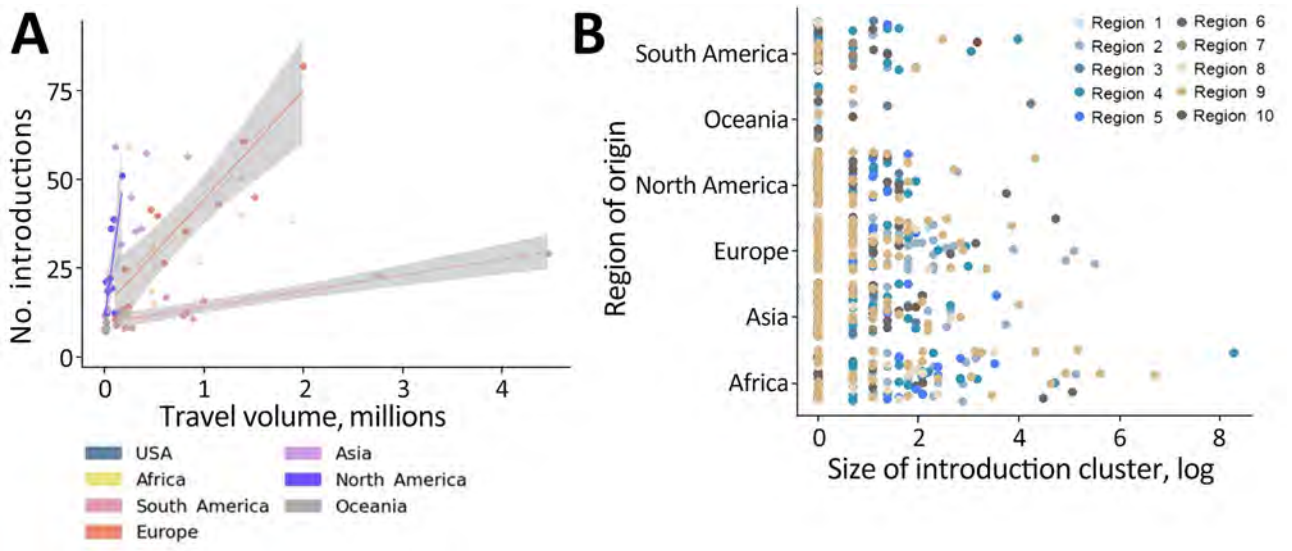


Figure 9. Associations between travel from different countries and number and cluster size of SARS-CoV-2 Omicron BA.5 introductions into the United States, February–June 2022. A) Linear regresssions indicating associations between the number of introductions into the United States from different continents and international travel volume according to that continent. (B) Cluster sizes of BA.5 introductions originating from different continents into the 10 Department of Health and Human Services regions of the United States. Regions designated by the US Department of Health and Human Services are shown in Figure 3.

travel between regions 1–3, as well as incoming air travel from other regions (Figures 6, 11).

To explore possible underlying drivers of virus movement across the United States, we performed linear regressions between pairs of locations, using population sizes and whether those locations shared a border as predictors. We found that the population size of the origin location was a significant predictor for the number of virus movements between a pair of locations ($p<0.0001$). In comparison, the destination population combined with whether the 2 locations shared a land border was not a significant predictor for virus movement ($p>0.1$) (Appendix Table 2).

Discussion

As SARS-CoV-2 continues to spread in the United States and globally, it will be essential to elucidate how new variants disseminate. We found that Omicron BA.5 was first introduced into the United States primarily from its geographic origin in Africa and then spread domestically from large populations and key hotspots, which are common between variants.

The earliest BA.5 introductions into the United States came from Africa despite low rates of air travel, indicating the importance of a variant's geographic origin. Early introduction events were also much larger than later introductions, which is a common thread among the waves of SARS-CoV-2 across the globe, despite different demographic and intervention contexts (*13*). As prevalence rose globally in the later half of the study period, a higher proportion of introductions from Europe and Asia occurred, potentially corresponding to higher travel volume (*35*). Similar dynamics have occurred with Delta variant introductions into the United Kingdom (*11*). The combination of the earliest introductions being the most important and later introductions coming from many locations makes international travel restrictions challenging to implement, even aside from ethical concerns (*36*); the speed required to prevent the most critical early introductions from a particular origin, if it is even known, is unachievable in most settings.

Domestic transmission played a substantial role in BA.5 dissemination in the United States. Whereas rates of interregion transmission exceeded those of global importation across the entire study period, most domestic virus movement occurred during the later phase. We show widespread secondary transmission occurred across the United States after the initial international introduction, which corroborates previous findings indicating SARS-CoV-2 transmission is driven by domestic dynamics (*15*,*17*). The domestic BA.5 spread was significantly associated with population size of the origin location, which fits with previous descriptions of SARS-CoV-2 transmission starting from large urban centers into other areas (*37*,*38*). Along with geographic proximity being somewhat essential, that finding fits a classical gravity model of disease transmission (*39*).

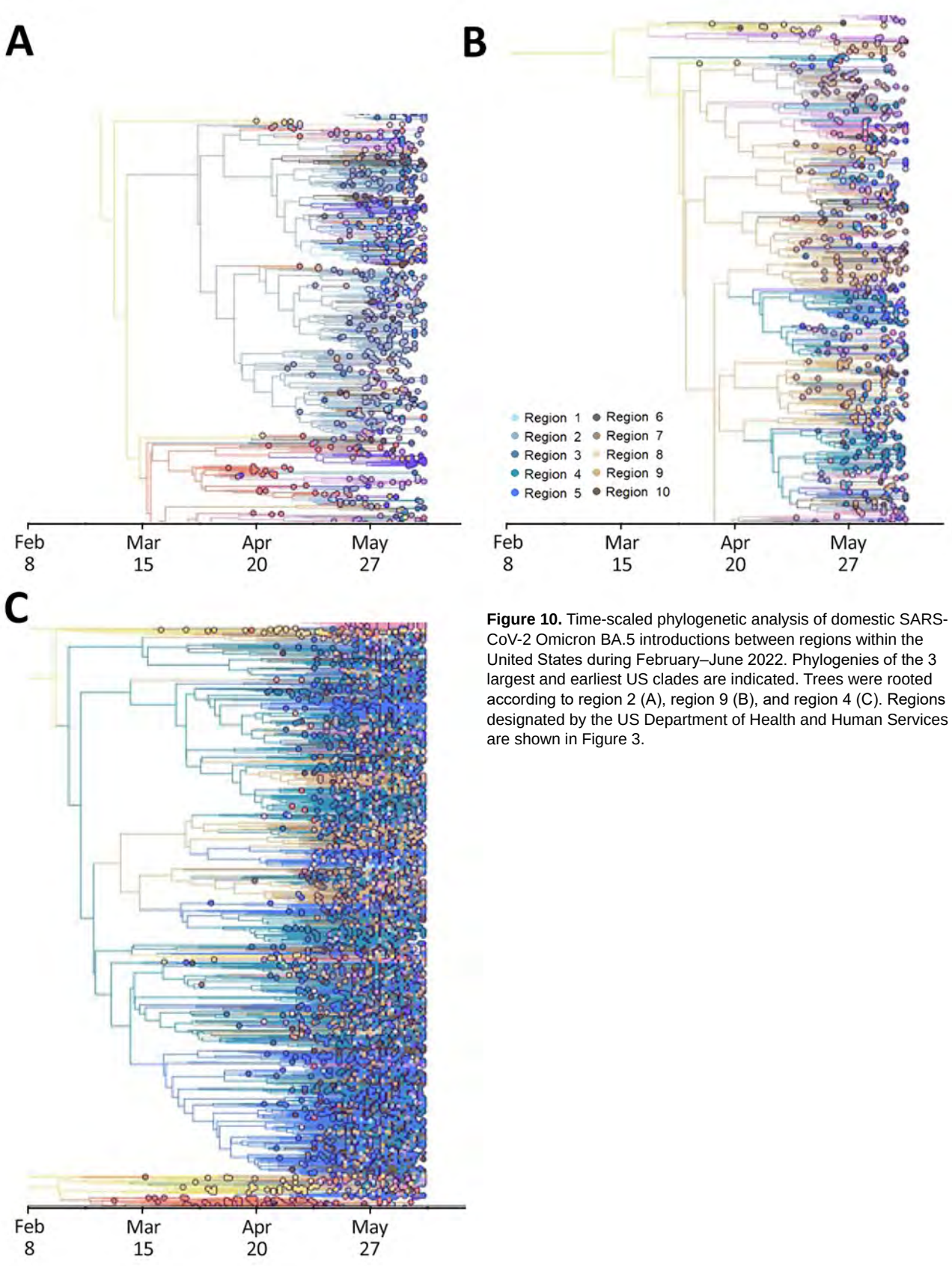


Figure 10. Time-scaled phylogenetic analysis of domestic SARS-CoV-2 Omicron BA.5 introductions between regions within the United States during February–June 2022. Phylogenies of the 3 largest and earliest US clades are indicated. Trees were rooted according to region 2 (A), region 9 (B), and region 4 (C). Regions designated by the US Department of Health and Human Services are shown in Figure 3.

Figure 11. SARS-CoV-2 Omicron BA.5 movements between regions within the United States from study of BA.5 emergence during January–June 2022. Thickness of the lines indicates the prevalence of the movement across the maximum clade credibility tree; arrows indicate direction of introduction.

Cross-country BA.5 spread between DHHS regions 2, 4, and 9 highlight the role of specific hotspots in promoting BA.5 emergence. Those 3 regions received the most introductions from Africa and had the 3 largest and earliest US clades, playing a critical role in receiving and disseminating early BA.5 introductions. That finding is similar to the dissemination of the SARS-CoV-2 Alpha variant (*17*); New York, New York, received the most introductions from the Alpha variant's origin, followed by California and Florida. Therefore, we might expect those regions to be critical during future variant introductions. Furthermore, we found that region 1 (New England) was the highest recipient of domestic introductions, likely from high interaction rates with 2 of the key hotspots (regions 2 and 4). We suggest that regions 2, 4, and 9 were primary hotspots because of their major urban centers (e.g., New York, Atlanta, and Los Angeles). Those findings fit the description of early virus lineage movements between larger cities, followed by spatial expansion into nearby areas (*14*).

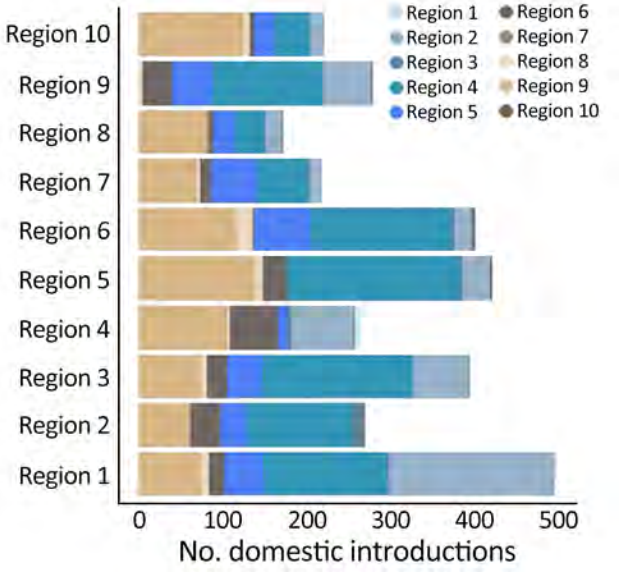


Figure 12. Number of domestic SARS-CoV-2 Omicron BA.5 introductions into each region of the United States in study of BA.5 emergence during January–June 2022. Regions designated by the US Department of Health and Human Services are indicated in Figure 3.

The first limitation of our study is that our subsampling method reflected the broader inequality in genomic surveillance worldwide (*22*,*23*). We attempted to minimize those biases through subsampling and categorization into broader continents and US DHHS regions. Rooting our tree in Africa, despite sequences from Europe overwhelming the global dataset, suggests that our attempts to mitigate this international bias were somewhat successful. Our categorization into larger regions (within and outside the United States) might have introduced residual confounding, preventing exploration of interstate introduction events. We also chose to use population size to subsample, rather than case-based metrics that might appear more relevant. However, obtaining unbiased incidence/hospitalization/death estimates during an outbreak is challenging, especially when comparing large geographic areas, such as the United States or entire continents. All data are imperfect sources of information in this context because large amounts of heterogeneity exist in how those data are recorded because of resource limitations, varying case definitions, and political concerns. We therefore used population size, which we concluded should be less biased. Second, geographic variation in sequencing efforts might have affected our cluster size results by artificially increasing the size of introductions from Africa compared with Europe (i.e., there might be missing sequences from Africa, which would split clusters into smaller introductions). Our downsampling scheme should have helped mitigate

this limitation, and the pattern of early and large introductions fits with other settings (*13*). Third, we defined the variant emergence phase according to a frequency growth curve to filter for early BA.5 sequences, which we deemed essential to our research; that definition might not properly reflect the true emergence time for a novel variant, although this only changes the length of our study period. Finally, we did not test other factors that might have driven the international introduction of Omicron BA.5 into the United States, such as distance through air networks or income levels.

In conclusion, our findings support the role of phylogenetics in SARS-CoV-2 surveillance and contribute a phylogeographic framework for studying the emergence of other infectious pathogens in the United States. Countries have lifted pandemic restrictions and the general population has a mosaic of immunity; thus, the epidemiologic landscape presents opportunities for positive selection of novel SARS-CoV-2 variants. Determining the different dynamics of introduction in US regions will be critical for timely and cost-effective policymaking, particularly for health authorities. Our methods can be used to extend beyond SARS-CoV-2 analyses and can form a framework for phylogeographic analysis of large datasets to discern the spatiotemporal spread of other novel pathogens.

Acknowledgments

We thank Anne Hahn and Nicholas Chen for their help with this study and everyone who has contributed genomic data to GISAID, which makes work like this possible. We gratefully acknowledge all data contributors; that is, the authors and their originating laboratories responsible for obtaining the specimens and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

This work was supported by the Centers for Disease Control and Prevention Broad Agency Announcement (contract nos. 75D30122C14697 and 75D30120C09570) (to N.D.G.).

N.D.G. is a paid consultant for Pfizer-BioNTech.

About the Author

Mr. Pham holds an MPH from Yale School of Public Health in Connecticut and works as a technical officer/data analyst for the Program for Appropriate Technology in Health (PATH), Southeast Asia regional hub and adjunct lecturer at Hanoi Medical University in Vietnam. His research interests focus on infectious disease modeling and phylogenetic methods.

References

1. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 2022;603:679–86. https://doi.org/10.1038/s41586-022-04411-y
2. Chaguza C, Coppi A, Earnest R, Ferguson D, Kerantzas N, Warner F, et al. Rapid emergence of SARS-CoV-2 Omicron variant is associated with an infection advantage over Delta in vaccinated persons. Med. 2022;3:325–334.e4. https://doi.org/10.1016/j.medj.2022.03.010
3. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature. 2022;602:657–63. https://doi.org/10.1038/s41586-021-04385-3
4. Tegally H, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, et al.; NGS-SA consortium. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. Nat Med. 2022;28:1785–90. https://doi.org/10.1038/s41591-022-01911-2
5. Ma KC, Shirk P, Lambrou AS, Hassell N, Zheng XY, Payne AB, et al. Genomic surveillance for SARS-CoV-2 variants: circulation of Omicron lineages—United States, January 2022–May 2023. MMWR Morb Mortal Wkly Rep. 2023;72:651–6. https://doi.org/10.15585/mmwr.mm7224a2
6. Chakraborty C, Bhattacharya M, Chopra H, Islam MA, Saikumar G, Dhama K. The SARS-CoV-2 Omicron recombinant subvariants XBB, XBB.1, and XBB.1.5 are expanding rapidly with unique mutations, antibody evasion, and immune escape properties—an alarming global threat of a surge in COVID-19 cases again? Int J Surg. 2023;109:1041–3. https://doi.org/10.1097/JS9.0000000000000246
7. Hill V, Githinji G, Vogels CBF, Bento AI, Chaguza C, Carrington CVF, et al. Toward a global virus genomic surveillance network. Cell Host Microbe. 2023;31:861–73. https://doi.org/10.1016/j.chom.2023.03.003
8. Giovanetti M, Slavov SN, Fonseca V, Wilkinson E, Tegally H, Patané JSL, et al. Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil. Nat Microbiol. 2022;7:1490–500. https://doi.org/10.1038/s41564-022-01191-z
9. Kanteh A, Jallow HS, Manneh J, Sanyang B, Kujabi MA, Ndure SL, et al. Genomic epidemiology of SARS-CoV-2 infections in The Gambia: an analysis of routinely collected surveillance data between March, 2020, and January, 2022. Lancet Glob Health. 2023;11:e414–24. https://doi.org/10.1016/S2214-109X(22)00553-8
10. Douglas J, Winter D, McNeill A, Carr S, Bunce M, French N, et al. Tracing the international arrivals of SARS-CoV-2 Omicron variants after Aotearoa New Zealand reopened its border. Nat Commun. 2022;13:6484. https://doi.org/10.1038/s41467-022-34186-9
11. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, et al.; COVID-19 Genomics UK (COG-UK) Consortium. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. Nature. 2022;610:154–60. https://doi.org/10.1038/s41586-022-05200-3
12. Kraemer MUG, Hill V, Ruis C, Dellicour S, Bajaj S, McCrone JT, et al.; COVID-19 Genomics UK (COG-UK) Consortium. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. Science. 2021;373:889–95. https://doi.org/10.1126/science.abj0113
13. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al.; COVID-19 Genomics UK (COG-UK) Consortium. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371:708–12. https://doi.org/10.1126/science.abf2946

14. Tsui JLH, McCrone JT, Lambert B, Bajaj S, Inward RPD, Bosetti P, et al.; COVID-19 Genomics UK (COG-UK) Consortium. Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. Science. 2023;381:336–43. https://doi.org/10.1126/science.adg6605
15. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, et al. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell. 2020;181:990–996.e5. https://doi.org/10.1016/j.cell.2020.04.021
16. Zeller M, Gangavarapu K, Anderson C, Smither AR, Vanchiere JA, Rose R, et al. Emergence of an early SARS-CoV-2 epidemic in the United States. Cell. 2021;184:4939–4952.e15. https://doi.org/10.1016/j.cell.2021.07.030
17. Alpert T, Brito AF, Lasek-Nesselquist E, Rothman J, Valesano AL, MacKay MJ, et al. Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. Cell. 2021;184:2595–2604.e13. https://doi.org/10.1016/j.cell.2021.03.061
18. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20:533–4. https://doi.org/10.1016/S1473-3099(20)30120-1
19. Klaassen F, Chitwood MH, Cohen T, Pitzer VE, Russi M, Swartwood NA, et al. Changes in population immunity against infection and severe disease from severe acute respiratory syndrome coronavirus 2 Omicron variants in the United States between December 2021 and November 2022. Clin Infect Dis. 2023;77:355–61. https://doi.org/10.1093/cid/ciad210
20. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403–7. https://doi.org/10.1038/s41564-020-0770-5
21. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6:3773. https://doi.org/10.21105/joss.03773
22. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun. 2022;13:7003. https://doi.org/10.1038/s41467-022-33713-y
23. Abbasi J. How the US failed to prioritize SARS-CoV-2 variant surveillance. JAMA. 2021;325:1380–2. https://doi.org/10.1001/jama.2021.3368
24. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015
25. Hasegawa M, Kishino H, Yano T. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22:160–74. https://doi.org/10.1007/BF02101694
26. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2:vew007. https://doi.org/10.1093/ve/vew007
27. Hill V, Baele G. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. Mol Biol Evol. 2019;36:2620–8. https://doi.org/10.1093/molbev/msz172
28. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4:vey016. https://doi.org/10.1093/ve/vey016
29. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4:vex042. https://doi.org/10.1093/ve/vex042
30. Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, et al.; SeqCOVID-SPAIN consortium. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature. 2021;595:707–12. https://doi.org/10.1038/s41586-021-03677-y
31. Aggarwal D, Warne B, Jahun AS, Hamilton WL, Fieldman T, du Plessis L, et al.; Cambridge Covid-19 testing Centre; University of Cambridge Asymptomatic COVID-19 Screening Programme Consortium; COVID-19 Genomics UK (COG-UK) Consortium. Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission. Nat Commun. 2022;13:751. https://doi.org/10.1038/s41467-021-27942-w
32. Ghafari M, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus OG, et al. Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. Mol Biol Evol. 2022;39:msac009. https://doi.org/10.1093/molbev/msac009
33. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2013;30:713–24. https://doi.org/10.1093/molbev/mss265
34. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67:901–4. https://doi.org/10.1093/sysbio/syy032
35. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10:e1003932. https://doi.org/10.1371/journal.ppat.1003932
36. Jecker NS, Atuire C. Who's in? Who's out? The ethics of COVID-19 travel rules. The Conversation, Nov 30, 2021 [cited 2024 May 22]. http://theconversation.com/whos-in-whos-out-the-ethics-of-covid-19-travel-rules-172053
37. McBroome J, Martin J, de Bernardi Schneider A, Turakhia Y, Corbett-Detig R. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. Virus Evol. 2022;8:veac048. https://doi.org/10.1093/ve/veac048
38. Barajas-Carrillo VW, Covantes-Rosales CE, Zambrano-Soria M, Castillo-Pacheco LA, Girón-Pérez DA, Mercado-Salgado U, et al. SARS-CoV-2 transmission risk model in an urban area of Mexico, based on GIS analysis and viral load. Int J Environ Res Public Health. 2022;19:3840. https://doi.org/10.3390/ijerph19073840
39. Truscott J, Ferguson NM. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. PLOS Comput Biol. 2012;8:e1002699. https://doi.org/10.1371/journal.pcbi.1002699

Address for correspondence: Verity Hill, Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510, USA; email: verity.hill@yale.edu

Detection and Tracking of SARS-CoV-2 Lineages through National Wastewater Surveillance System Pathogen Genomics

Dorian J. Feistel, Rory Welsh, Jeffrey Mercante, Miguella Mark-Carew, Jason Caravas, Arun Boddapati, Samantha Sevilla, Matthew H. Seabolt, Dhwani Batra, Suchitra Chavan, Shatavia Morrison, Jesse Yoder, Hannah Long, Satvik Mishra, Benjamin Lorentz, Andi Dhroso, Iryna V. Goraichuk, Seonghye Jeon, Daniel M. Cornforth

We conducted retrospective analysis of the emergence of the SARS-CoV-2 JN.1 variant in US wastewater during November 2023–July 2024 using Aquascope, a bioinformatics pipeline for the National Wastewater Surveillance System. This study highlights the value of open-source bioinformatics tools in tracking pathogen variants for public health monitoring.

The emergence and rapid global spread of SARS-CoV-2 emphasized the need for efficient methods of identifying and tracking viral changes as they circulate within communities. Wastewater pathogen genomic surveillance offers a timely, noninvasive, and cost-effective method for detecting pathogen genetic material in sewersheds, providing a comprehensive snapshot of community transmission dynamics to monitor infection trends (*1*). Wastewater surveillance complements clinical surveillance and can identify viruses shed by persons who are presymptomatic, asymptomatic, or not tested in healthcare facilities, making it a robust measure of overall prevalence of SARS-CoV-2 lineages in circulation, the early geographic spread of emerging variants already detected in humans, and novel variants of SARS-CoV-2 not yet detected in humans (*2*).

In 2020, the Centers for Disease Control and Prevention (CDC) established the National Wastewater Surveillance System (NWSS) to track the spread of SARS-CoV-2 in wastewater at the local level (*3*). Since then, laboratories in academia and public health have made considerable advancements in tools for characterizing pathogen genomic variation in wastewater (*3*–*6*). Through wastewater sequencing, NWSS monitors genetic variation in SARS-CoV-2, identifying variants and mutations that may affect disease severity or efficacy of PCR-based diagnostics, vaccines, or therapeutics (*3*). To enable timely, reproducible, and high-throughput analyses of wastewater sequence data for SARS-CoV-2 monitoring, NWSS collaborated with CDC's Scientific Computing and Bioinformatics Services in the Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases, to develop the bioinformatics pipeline Aquascope, modeled after the CFSAN Wastewater Analysis Pipeline (C-WAP) (*5*). Aquascope will replace C-WAP on the CDC 1CDP platform, providing timely results to jurisdictions and the public. Aquascope is more robust than C-WAP; it includes quality metrics and logging features and can be deployed in high-performance computing and cloud platforms. Implemented in Nextflow, Aquascope uses open-source, containerized bioinformatic tools for quality control, variant identification, and estimation of lineage abundance from tiled-amplicon short-read and long-read wastewater sequence data.

Author affiliations: Centers for Disease Control and Prevention, Atlanta, Georgia, USA (D.J. Feistel, R. Welsh, J. Mercante, M. Mark-Carew, J. Caravas, M.H. Seabolt, D. Batra, S. Morrison, S. Jeon, D.M. Cornforth); Leidos Inc., Reston, Virginia, USA (A. Boddapati, S. Sevilla, S. Chavan); DCI Solutions, Aberdeen Proving Ground, Maryland, USA (J. Yoder); Palantir Technologies Inc., Denver, Colorado, USA (J. Yoder, H. Long, S. Mishra); Goldbelt Professional Services, Chesapeake, Virginia (B. Lorentz); ASRT Inc., Smyrna, Georgia, USA (A. Dhroso, I.V. Goraichuk)

DOI: https://doi.org/10.3201/eid3113.241411

Case Study

As a case study, we sought to retrospectively track the spread of the JN.1 variant of SARS-CoV-2, a closely related descendant of BA.2.86, first detected in clinical samples in early September 2023 (*2*). By early December 2023, JN.1 had become the predominant variant in the United States. Using Aquascope, we estimated the relative abundance of known SARS-CoV-2 lineages from wastewater sequence data collected from across the country. This activity has been reviewed by CDC and determined to be nonresearch public health surveillance that did not require review through the CDC Human Research Protection Office or Institutional Review Board.

We used a subset of wastewater surveillance data collected by Verily Life Sciences (https://verily.com) on behalf of NWSS (National Center for Biotechnology Information BioProject no. PRJNA1027353) to estimate variant relative abundances. By limiting analysis to data from this BioProject, we ensured consistency in laboratory methods across the data analyzed. We included in our analysis collection weeks with ≥10 samples, comprising 3,377 unique samples gathered from 130 sites across 87 counties in 32 US jurisdictions. The collection period was November 13, 2023–July 23, 2024. All samples were concentrated, DNAse treated, and reverse transcribed before amplification using the NEB Q5 High-Fidelity PCR kit with ARTIC version 5.3.2 primers (New England Biolabs, https://www.neb.com). Sequencing libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit and pair-end sequenced (2 × 300 bp) on the Illumina NextSeq 2000 (Illumina, https://www.illumina.com).

We processed raw sequencing data using Aquascope version 2.1.0, first performing quality checks, then removing adapters and low-quality regions. We aligned reads to the SARS-CoV-2 reference genome (GenBank accession no. MN908947.3) and trimmed primers used for amplification. We estimated the relative abundance of known SARS-CoV-2 lineages using Freyja (*6*) with SARS-CoV-2 UShER barcodes from July 26, 2024 (*7*). Full pipeline details are publicly available (https://github.com/CDCgov/aquascope). Lineage relative abundance estimates correspond to samples collected across jurisdictions within the same week after lineage aggregation and normalization; lineages representing ≤5% of the total being aggregated were categorized as Other. We tracked all lineage abundances, and we aggregated sublineages not

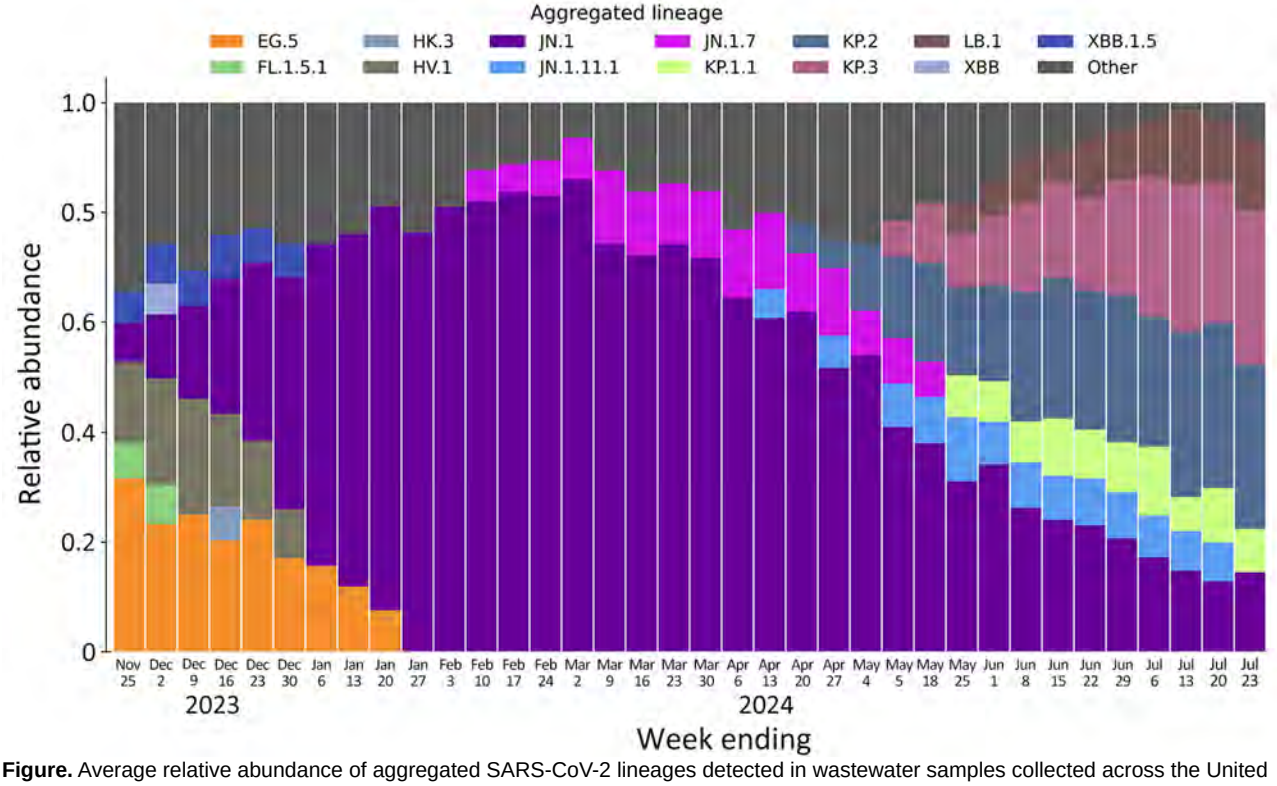


Figure. Average relative abundance of aggregated SARS-CoV-2 lineages detected in wastewater samples collected across the United States for the weeks ending November 25, 2023–July 23, 2024. The final time point shown (July 23, 2024) does not represent a full week of data.

enumerated with their parent lineages on the basis of Pango lineage definitions (*8*). We chose parent lineages to reflect those displayed on the CDC COVID Data Tracker (*9*) and NWSS dashboard (*10*).

Our analysis of wastewater sequence data revealed a distinct temporal trend in the emergence and spread of the JN.1 variant (Figure). JN.1 was first detected by the pipeline in a sample collected on November 15, 2023; however, because this week had <10 samples collected, the earliest displayed data are from samples collected the subsequent week. After initial detection, JN.1 increased in prevalence in early December 2023, peaked in early March 2024, and continued to decline through late July in the final displayed weeks. Results also showed other known lineages, such as the JN.1 sublineages JN.1.7 and JN.1.11.1, emerging sequentially and maintaining a significant presence. KP.2 and KP.3 lineages also appeared and grew to varying levels of prevalence. Our study of the JN.1 lineage with Aquascope demonstrates the ability of CDC's NWSS to monitor emerging SARS-CoV-2 variants.

Conclusions

Advances in wastewater bioinformatics pipelines, such as Aquascope, enhance our ability to track public health outbreaks by providing a relatively passive, low-cost, near–real-time surveillance approach that complements clinical genomic surveillance. Although JN.1 was first detected in a US clinical sample collected in late September 2023 (*11*), earlier than the first samples identified with Aquascope here, we note that the Bioproject analyzed in this study does not contain data preceding November 13, 2023. Still, trends in JN.1 proportions inferred in NWSS samples by Aquascope were similar to those in clinical sequence data, which first surpassed 0.1% prevalence in November 2023 and surpassed 50% prevalence from early January to the end of April 2024, similar to wastewater trends (Figure) (*11*).

Future work will cross-compare wastewater and clinical data using additional NWSS sites and account for differences in coverage, population normalization, and other analytical considerations. Although the pipeline we describe focuses on SARS-CoV-2 lineage abundance, Freyja's deconvolution algorithm (*6*) can use barcode libraries from additional pathogens to estimate their abundances in mixed wastewater samples, so that Aquascope can be adapted for broader pathogen detection. This pipeline relies on prior characterization of SARS-CoV-2 lineages; future advancements may enable identification of previously uncharacterized lineages.

One potential challenge for personal use of Aquascope is computational requirements robust enough to support large input datasets; it requires a high-performance computing environment with support for required dependencies. Aquascope will soon operate within a larger, scalable CDC computing platform freely available to public health partners for wastewater surveillance efforts. Continuous development of bioinformatics pipelines like Aquascope will broaden our capacity to monitor emerging infectious disease threats through wastewater surveillance.

About the Author

Mr. Feistel is a bioinformatician and microbiologist with the CDC National Wastewater Surveillance System, Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases. His research spans computational biology, microbial genomics, and infectious diseases, with a focus on developing advanced bioinformatics and computational methods to enhance disease detection and surveillance.

References

1. Kirby AE, Walters MS, Jennings WC, Fugitt R, LaCross N, Mattioli M, et al. Using wastewater surveillance data to support the COVID-19 response—United States, 2020–2021. MMWR Morb Mortal Wkly Rep. 2021;70:1242–4. https://doi.org/10.15585/mmwr.mm7036a2
2. Lambrou AS, South E, Ballou ES, Paden CR, Fuller JA, Bart SM, et al. Early detection and surveillance of the SARS-CoV-2 variant BA.2.86—worldwide, July–October 2023. MMWR Morb Mortal Wkly Rep. 2023;72:1162–7. https://doi.org/10.15585/mmwr.mm7243a2
3. Adams C, Bias M, Welsh RM, Webb J, Reese H, Delgado S, et al. The National Wastewater Surveillance System (NWSS): from inception to widespread coverage, 2020–2022, United States. Sci Total Environ. 2024; 924:171566. https://doi.org/10.1016/j.scitotenv.2024.171566
4. Kirby AE, Welsh RM, Marsh ZA, Yu AT, Vugia DJ, Boehm AB, et al.; New York City Department of Environmental Protection. Notes from the field: early evidence of the SARS-CoV-2 B.1.1.529 (Omicron) variant in community wastewater—United States, November–December 2021. MMWR Morb Mortal Wkly Rep. 2022;71:103–5. https://doi.org/10.15585/mmwr.mm7103a5
5. Kayikcioglu T, Amirzadegan J, Rand H, Tesfaldet B, Timme RE, Pettengill JB. Performance of methods for SARS-CoV-2 variant detection and abundance estimation within mixed population samples. PeerJ. 2023;11:e14596. https://doi.org/10.7717/peerj.14596
6. Karthikeyan S, Levy JI, De Hoff P, Humphrey G, Birmingham A, Jepsen K, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature. 2022;609:101–8. https://doi.org/10.1038/s41586-022-05049-6

7. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet. 2021;53:809–16. https://doi.org/10.1038/s41588-021-00862-7
8. Pango Network. Rules for the designation and naming of Pango lineages [cited 2024 Sep 3]. https://web.archive.org/web/20240116214031/https://www.pango.network/the-pango-nomenclature-system/statement-of-nomenclature-rules
9. Centers for Disease Control and Prevention. COVID data tracker [cited 2024 Sep 3]. https://covid.cdc.gov/covid-data-tracker/#variant-proportions
10. Centers for Disease Control and Prevention. COVID-19 variants in wastewater [cited 2024 Sep 3]. https://www.cdc.gov/nwss/rv/COVID19-variants.html
11. Ma KC, Castro J, Lambrou AS, Rose EB, Cook PW, Batra D, et al. Genomic surveillance for SARS-CoV-2 variants: circulation of Omicron XBB and JN.1 lineages — United States, May 2023–September 2024. MMWR Morb Mortal Wkly Rep. 2024;73:938–45. https://doi.org/10.15585/mmwr.mm7342a1

Address for correspondence: Daniel Cornforth, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H24-11, Atlanta, GA 30329-4018, USA; email: rio8@cdc.gov

SARS-CoV-2 Genomic Surveillance from Community-Distributed Rapid Antigen Tests, Wisconsin, USA

Isla E. Emmen, William C. Vuyk, Andrew J. Lail, Sydney Wolf, Eli J. O'Connor, Rhea Dalvie, Maansi Bhasin, Aanya Virdi, Caroline White, Nura R. Hassan, Alex Richardson, Grace VanSleet, Andrea Weiler, Savannah Rounds-Dunn, Kenneth Van Horn, Marc Gartler, Jane Jorgenson, Michael Spelman, Sean Ottosen, Nicholas R. Minor, Nancy Wilson, Thomas C. Friedrich, David H. O'Connor

In the United States, SARS-CoV-2 genomic surveillance initially relied almost entirely on residual diagnostic specimens from nucleic acid amplification–based tests. However, use of those tests waned after the end of the COVID-19 Public Health Emergency on May 11, 2023. In Dane County, Wisconsin, we partnered with local- and state-level public health agencies and the South Central Library System to continue genomic surveillance by obtaining SARS-CoV-2 genome sequences from freely available community rapid antigen tests (RATs). During August 15, 2023–February 29, 2024, we received 227 RAT samples, from which we generated 127 sequences with >10× depth of coverage for ≥90% of the SARS-CoV-2 genome. In a subset of tests, lower cycle threshold values correlated with sequence success. Our results demonstrated that collecting and sequencing results from RATs in partnership with community sites is a practical approach for sustaining SARS-CoV-2 genomic surveillance.

Genomic surveillance is a powerful tool that can inform public health responses to disease outbreaks (*1*). During the COVID-19 pandemic, genomic surveillance data were used to identify variants of concern, investigate patterns of transmission, and develop effective vaccines (*2*–*4*).

Genomic surveillance requires large, representative sets of samples. Initially, nucleic acid amplification tests (NAATs) were the standard for detecting SARS-CoV-2 infection (*5*). Laboratories received residual nasal swab samples leftover from NAAT testing for viral sequencing through contracts with companies and clinics performing NAATs. After the COVID-19 public health emergency ended on May 11, 2023, NAAT testing in clinical and public health facilities declined precipitously as government subsidies for performing NAATs ended (*6*). During 2020, an average of 587,975 NAATs were performed weekly in the United States. By 2023, that number decreased to ≈96,215 tests per week (*7*). Subsequently, the primary source of samples for genomic surveillance was greatly diminished.

The US Food and Drug Administration issued the first emergency use authorization for a COVID-19 rapid antigen test (RAT) in August of 2020 (*8*). At-home RAT usage increased significantly in 2021 during the rise of the Omicron lineage (*9*). RATs are less expensive than NAATs, provide faster results, and do not require trained personnel (*10*). RATs usually involve swabbing the insides of both nostrils, placing the swab into an inactivation buffer, and applying the buffer onto a lateral flow test strip. If SARS-CoV-2 antigen is present, a colorimetric test line will indicate positivity (*11*). By July 2024, the United States had 38 available over-the-counter SARS-CoV-2 RAT products authorized by the Food and Drug Administration and available to the public (*12*).

Author affiliations: University of Wisconsin–Madison School of Medicine and Public Health, Madison, Wisconsin, USA (I.E. Emmen, W.C. Vuyk, A.J. Lail, S. Wolf, E.J. O'Connor, R. Dalvie, M. Bhasin, A. Virdi, C. White, N.R. Hassan, N. Wilson, D.H. O'Connor); Madison West High School, Madison (E.J O'Connor); University of Wisconsin–Madison, Wisconsin National Primate Research Center, Madison (A. Richardson, G. VanSleet, A. Weiler, T.C. Friedrich); Public Health Madison Dane County, Madison (S. Rounds-Dunn, K. Van Horn); Madison Public Library, Madison (M. Gartler, J. Jorgensen, M. Spelman, S. Ottosen); University of Wisconsin–Madison School of Veterinary Medicine, Madison (T.C. Friedrich)

DOI: https://doi.org/10.3201/eid3113.241192

Multiple groups have investigated RATs as source material for SARS-CoV-2 genomic surveillance. SARS-CoV-2 RNA can be recovered from antigen tests and sequenced (*13*–*17*), enabling tracking of circulating lineages. We hypothesized that community members would be willing to send in SARS-CoV-2–positive RATs for genomic surveillance if the process were sufficiently simple. Thus, we partnered with local public libraries and a public health agency to create and assess a system for persons to anonymously submit positive RATs for viral RNA analysis and sequencing.

Materials and Methods

Collection of Rapid Antigen Tests

In Wisconsin, the South Central Library System and Public Health Madison Dane County (PHMDC) distributed RATs from the US national stockpile to the public free of charge. We partnered with 9 libraries and 2 sites through PHMDC. Six of the libraries were located in urban areas and 3 in rural areas (*18*) (Figure 1).

We designed a packet of materials to attach to each RAT to enable collection of positive tests (Figure 2). The packet included an instructional flyer affixed to the outside of a bubble mailer to which a business reply mail shipping label was affixed. Inside the bubble mailer, we included a zip-top bag with a unique quick response (QR) barcode for return of RAT tests. The flyer had instructions in both English and Spanish, describing the study and providing instructions on how to participate (Appendix Figure, https://wwwnc.cdc.gov/EID/article/31/13/24-1192-App1.pdf).

Participants could volunteer to submit their RAT to our program if they tested SARS-CoV-2 positive. Participants were directed to scan the unique QR code on an internet-connected device, seal the positive RAT strip inside the zip-top bag, place the bag inside the bubble mailer and seal it, then drop the sealed mailer in any post office mailbox. The inactivation buffer in a RAT inactivates SARS-CoV-2, rendering the tests nonbiohazardous and safe to send through the mail (*19*).

Upon arrival at our laboratory, we scanned the QR code to record the date of receipt and stored at −80°C until processing. Most RATs we received were BinaxNOW COVID-19 Antigen Self Tests (Abbott, https://www.abbott.com) or iHealth COVID-19 Antigen Rapid Tests (iHealth Labs Inc., https://ihealthlabs.com).

Ethics Statement

The University of Wisconsin institutional review board determined this project was human research exempt because participants were anonymous and self-identified. We created a secure website and database

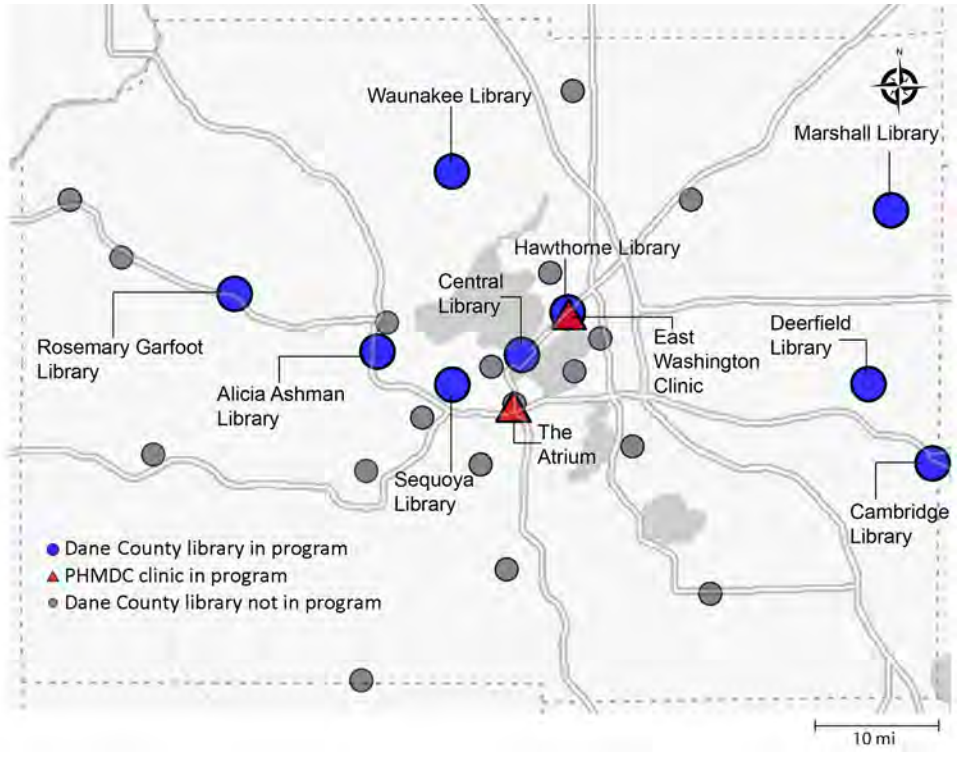


Figure 1. Locations for SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests, Wisconsin, USA. Nine of the South Central Library System libraries and 2 PHMDC sites distributed research packets and SARS-CoV-2 rapid antigen tests to patrons. Willing participants could send their positive tests to the AIDS Vaccine Research Laboratory, University of Wisconsin–Madison (Madison, WI, USA), for sequencing. PHMDC, Public Health Madison Dane County.

by using Node JS (https://nodejs.org) to collect the barcode, the date, and the location when a user scanned a randomly generated unique QR code. We assumed those data were a reasonable proxy for RAT date and location. The location of the scan was automatically converted to a census block group on the users' machines before submission to our database, so the actual location of each submission was not known to the study team. A census block group contains 250–550 housing units (*20*).

Nucleic Acid Extraction

We developed our approach to extract nucleic acids from used RATs per previously describe methods (*13*). We thawed and opened RATs to retrieve the testing strip, which we placed into a clean 5-mL freezer tube (Sarstedt, https://www.sarstedt.com). Some RATs also included a nasal swab, and we also placed those in the freezer tube.

We added 800 µL of Viral Transport Medium (Rocky Mountain Biologicals, LLC, https://rmbio.com), and incubated the tube at room temperature for 10 minutes on a Hulamixer (Thermo Fisher Scientific, https://www.thermofisher.com). We transferred 500 µL of that mixture to a clean 1.5-mL tube and added 5 µL of Dynabeads Wastewater Virus Enrichment Beads (Thermo Fisher Scientific). We incubated samples for 10 minutes on a Hulamixer, then placed on a magnetic rack for 3 minutes. Once clear, we discarded the supernatant and resuspended the beads in 500 µL of lysis buffer. We returned the tube to the magnet for 3 minutes and then transferred the clear supernatant to a clean tube. We isolated samples on a Kingfisher Apex instrument (Thermo Fisher Scientific) following the manufacturer's protocol (protocol no. MagMAX_Wastewater_DUO96.bdz).

After isolation, we treated the samples with Turbo DNase (Thermo Fisher Scientific), according to the manufacturer's protocol. After DNase treatment, we cleaned samples by using the RNA Clean and Concentrator-5 kit (Zymo Research, https://www.zymoresearch.com), following the manufacturer's protocol, but skipping the in-column DNase I Treatment.

Quantitative Reverse Transcription PCR and Sequencing

We selected a random subset of 75 samples to investigate trends between the quantitative reverse transcription PCR cycle threshold (Ct) and sequencing quality. We quantified SARS-CoV-2 RNA using the CDC N1 Taqman assay (*21*) (Appendix).

We generated PCR amplicons by using the QIAseq DIRECT SARS-CoV-2 Kit with Booster and Enhancer (QIAGEN, https://www.qiagen.com), according to the manufacturer's instructions. We normalized indexed samples to 4 nmol and pooled samples together. We diluted the pool to a concentration of 8 pmol and ran using 2 × 150 MiSeq Reagent Kits v2 on a MiSeq instrument (both Illumina, https://www.illumina.com).

Figure 2. Research packets distributed for SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests, Wisconsin, USA. A) Envelope and instructions; B) zip-top bag included in packet with quick response (QR) code; C) return address label. Research packets were attached to SARS-CoV-2 rapid antigen test boxes, enabling participants to send their positive tests to the laboratory through the US Postal Service. A folded flyer (A), attached to an envelope, explained in both English and Spanish the goal of the study and how to participate. Participants scanned the QR code inside the included zip-top bag (B) to document the date and location of their rapid antigen test, then sealed their test strip inside. The location of the scanned QR code was immediately converted to the census block group of the scan and stored in a secure database. Participants returned test strips in the provided envelope, which had a business-reply shipping label (C), enabling participants to mail to our laboratory from any post office drop box.

Sequencing Analysis

We quality-checked raw sequencing reads and aligned to the wild-type SARS-CoV-2 reference (GenBank accession no. MN908947.3), then variant-called by using

the open-source viralrecon pipeline from the nf-core project (*22*,*23*; B.E. Langer et al., unpub. data, https://doi.org/10.1101/2024.05.10.592912). We set the minimum frequency threshold for variant-calling to 0.01. Further details for how we ran viralrecon, alongside the custom R scripts we used to generate figures, are available in our GitHub repository (https://github.com/dholab/Library-Rapid-Antigen-Test-Manuscript).

Statistical Analysis

We used an unpaired 2-tailed *t*-test to compare the effect of Ct and length of transit time between samples that passed our sequencing quality threshold of ≥90% coverage at >10× depth and those that failed. We performed that analysis in Prism version 10.1.0 (GraphPad, https://www.graphpad.com).

We compared the identities of SARS-CoV-2 lineages detected in our RAT-derived sequences with surveillance data from the Wisconsin State Laboratory of Hygiene (WSLH) SARS-CoV-2 Genomic Dashboard (*24*). We analyzed data from August 28, 2023–February 25, 2024, dividing our passing sequences into 2-week intervals on the basis of test scan dates. We only included participant-scanned tests in that analysis. We assigned Pango lineages to our sequences by using Nextclade version 3.5.0 (*25*). For each 2-week period, we identified the 2 most prevalent lineage groups, which we based on Nextstrain clades, in the WSLH wastewater surveillance data and determined how often our RAT program also detected the same prevalent lineages.

Results

Test Collection

During August 15, 2023–February 29, 2024, we supplied 9 libraries and 2 public health clinics in Dane County with 7,775 research packets to attach to SARS-CoV-2 RATs distributed to patrons. Among distributed packets, 223 (2.9%) were mailed to our laboratory. Some packets contained multiple tests, resulting in 227 total tests for analysis. The return rates varied by month (Table 1), but the mean number received each month was 32 (SD 10).

Some tests arrived at the laboratory without the barcode or with a barcode that had never been scanned, resulting in loss of associated metadata. Of the 223 research packets received, 170 were properly associated with time and location metadata. Of those 170 samples, 1 was scanned in Sauk County, Wisconsin (adjacent to Dane County), and the rest were scanned in Dane County.

Sequencing Quality

We sequenced SARS-CoV-2 from all 227 RATs. We considered a sequence with genome coverage ≥90% at a depth of coverage >10× to be a passing sequence. Of the 227 RAT-derived sequences, 128 (56%) passed (Appendix Table 1).

Next, we evaluated whether SARS-CoV-2 viral RNA concentration or transit time correlated with successful sequencing. We randomly selected 75 samples for semiquantitative reverse transcription PCR. Of those samples, 15 had no detectable amplification of the N1 target. We obtained passing sequences for samples with Ct values up to 35.4. The mean Ct for samples that passed was 31.7 and the mean Ct for samples that failed was 35.3, a significant difference via unpaired, 2-tailed *t*-test ($p<0.0001$; degrees of freedom = 59) (Figure 3, panel A).

The time between a test being scanned by the participant and our receiving it (i.e., the transit time) ranged from 1 to 20 days, but transit time had little effect on sequencing success (Figure 3, panel B). The mean transit time for the passing samples was 6.3 days, compared with 6.6 days for failed samples ($p = 0.69$ by unpaired 2-tailed *t*-test).

Table 1. Monthly distribution and return of research packages in a study of SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests, Wisconsin, USA*

Collection month	Approximate no. packets supplied for RAT collection	No. packets with positive tests	No. tests that passed sequencing quality threshold†
2023			
Aug	100	13	5
Sep	1,470	33	14
Oct	1,390	37	18
Nov	2,405	33	21
Dec	300	46	28
2024			
Jan	1,160	28	20
Feb	950	33	21
Total no.	7,775	223	127

*RATs, rapid antigen tests.
†Quality threshold was >10× depth of coverage for ≥90% of the SARS-CoV-2 genome.

Tracking SARS-CoV-2 Lineages

We used Nextclade version 3.5.0 (*25*) to determine the Pango lineage of each successfully sequenced sample and tracked SARS-CoV-2 lineages detected by week on the basis of participant scan date (Figure 4). During August–November 2023, most detected lineages were assigned to the XBB clade. Beginning in December 2023, we observed a shift to the JN.1 lineage, which predominated in February 2024.

The identities of viral lineages in our RAT-derived sequences were concordant with statewide trends in lineages detected via wastewater surveillance, as summarized on the WSLH SARS-CoV-2 Wastewater Genomic Dashboard (*24*) (Table 2). Our program detected the dominant wastewater lineage in 12 of 13 two-week reporting periods and the second-most prevalent lineage in 7 of 13 periods. Concordance with wastewater surveillance data indicates that RAT-based surveillance can detect common circulating lineages. Moreover, RAT-based surveillance resulted in 6 of the earliest documented cases of a lineage in Wisconsin in GenBank and GISAID: JN.1.1, JN.1.2, XDD, XDA, XDP, and XDE (Table 3).

Discussion

Genomic surveillance has been crucial for tracking SARS-CoV-2 evolution during the COVID-19 pandemic (*26*). Because most persons now use RATs instead of NAATs to diagnose SARS-CoV-2 infection, we sought to evaluate a community genomic surveillance program predicated on voluntary mailing of positive RATs.

Despite the common narrative that the public is disinterested in COVID-19, we observed surprisingly strong participation. During August 12, 2023–February 24, 2024, Dane County's average COVID-19 test positivity rate, which we used to estimate the return rate on RATs, was 12.3% (range 7.8%–16%) (*27*). In an extreme case in which all the RATs distributed by our partners were used, we estimated that one quarter of all positive tests distributed with packets were returned to our laboratory for analysis. The true return rate is likely higher because some tests distributed with packets likely were not used.

The transit time during which RATs sat in uncontrolled (ambient) conditions had a negligible effect on overall sequencing success (Figure 3, panel B). Other studies have demonstrated that extraction of viral RNA is possible from RATs stored at room temperature for long periods (*14*,*16*); one study generated 75.2% genome coverage from a RAT stored at room temperature for 3 months. We obtained a sequence with >10× coverage for ≥90% of the SARS-CoV-2

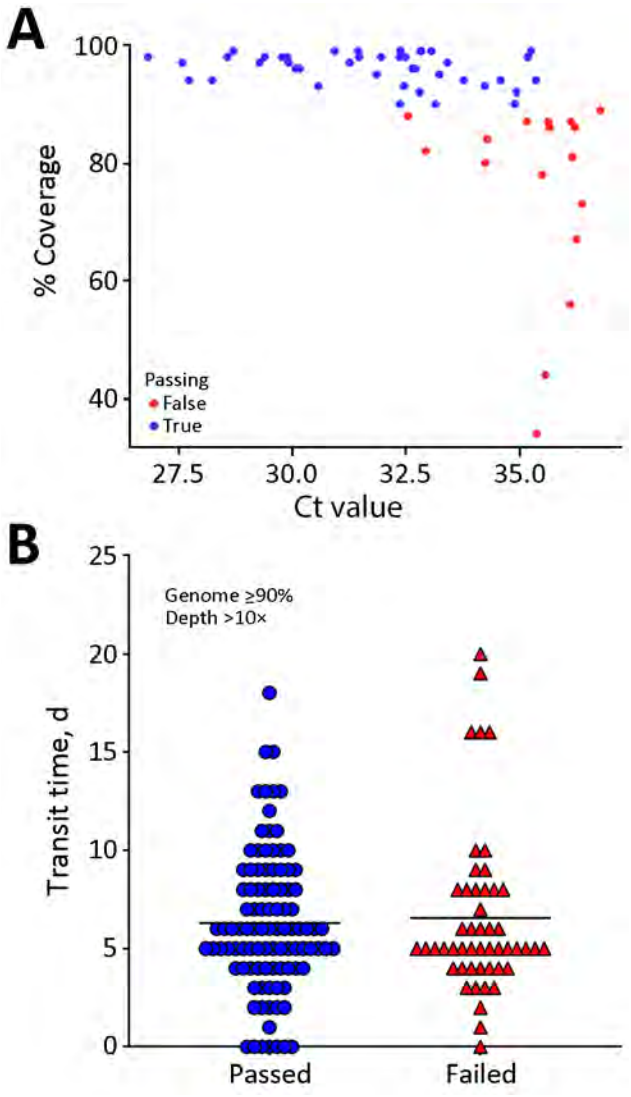


Figure 3. Comparison of passing and failing samples in a study of SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests, Wisconsin, USA. Scatterplots compare percentage coverage for Ct values (A) and transit times (B) for passing and failing RATs. Ct values were obtained through quantitative reverse transcription PCR. Sequences that passed the quality threshold had ≥90% coverage of the SARS-CoV-2 genome at >10× depth. The mean Ct for samples that passed was 31.7 and the mean Ct for those that failed was 35.3 ($p<0.0001$ by unpaired 2-tailed *t*-test; degrees of freedom = 59). Samples with lower Ct values correlated with higher SARS-CoV-2 coverage. Transit time refers to the number of days between a participant scanning the QR code provided with the RAT and receipt of positive RAT at our laboratory. The horizontal black line (B) is the mean value for each group. The mean transit time for passing samples was 6.3 (SD 3.6) days and the mean transit time for failing samples was 6.6 (SD 4.2) days. We noted no significant difference in transit times between passing and failing sequences ($p = 0.69$ by unpaired *t*-test). The amount of viral material present on RAT correlated with our ability to sequence samples, but time en route did not. Ct, cycle threshold; RAT, rapid antigen test.

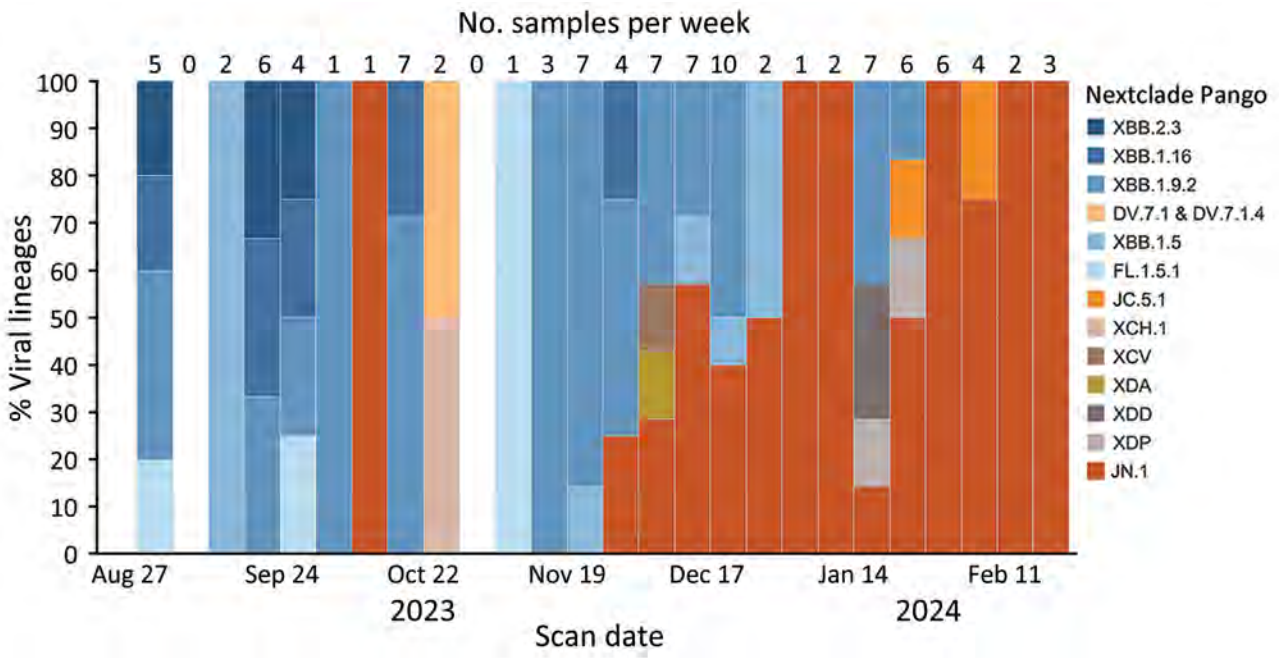


Figure 4. Number of samples collected per week and viral lineages detected in a study of SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests (RATs), Wisconsin, USA. The chart shows the percentage of SARS-CoV-2 lineages by week for samples that passed quality control thresholds of ≥90% of the SARS-CoV-2 genome at >10× depth. The date used reflected the date the participant scanned a provided QR code attached to a RAT. Unscanned RATs were excluded from the analysis. The number of samples included in each week's percentage is shown above the bar. We assigned Pango lineages by using Nextclade version 3.5.0 (*25*). From August to mid-November 2023, the most common lineages in our samples fell under XBB.1.5, XBB.1.9.2, XBB.1.16, and XBB.2.3. Beginning in early December 2023, we began to see an increase in the number of samples belonging to the lineage JN.1, which dominated RAT samples scanned in February 2024.

genome from a RAT that sat at uncontrolled temperatures for at least 17 days. Taken together, those results highlight that RATs stored at uncontrolled temperatures can be mailed from the point-of-testing to centralized laboratories for sequencing. Most (98%) of the US population is served by the United States Postal Service (*28*). Thus, the ability to self-collect samples for mail-in analysis could enable genomic surveillance even in settings that are typically underserved by academic and clinical research.

SARS-CoV-2 lineages identified by our RAT surveillance program were similarly prevalent in Wisconsin's statewide wastewater sequencing data (*24*). Of note, we also detected emerging lineages like JN.1 and rare variants like XDE, which was documented only 22 times in North America (*29*). Those findings demonstrate that RAT-based sequencing can effectively complement existing wastewater and NAAT surveillance methods.

One limitation of our study is the reliance on self-reported data, which is less precise than clinical specimen metadata. Our metadata depended on the participant's QR code scan to approximate the date and location of the test, which might have reduced data accuracy. Another limitation is that ≈25% of packets arrived unscanned or without a barcode; thus, we had no metadata for those samples. Our only communication with participants was through the flyer provided with each packet, and some participants might only skim the instructions and misinterpret the protocol. To reduce the frequency of unscanned tests, more simplified instructions that include visual cues could more clearly communicate the directions for returning RATs.

Census block groups of scanned tests showed a strong bias toward urban locations (*18*); only 1 test was scanned by a participant in a rural census block group. The 3 packet distribution sites in rural areas of Dane County received only 3% of the total packets we supplied, which might partially account for that low number of tests from rural areas. Rural and underrepresented areas might need stronger engagement efforts in future studies to achieve more representative genomic surveillance.

Our program relied on freely available RATs provided from the national government stockpile. The long-term sustainability of the programs that distribute those tests is unknown, which means this system

Table 2. Lineages detected during SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests, Wisconsin, USA*

Date†	Highest percentage‡		Second highest percentage§		No. RATs with passing sequence¶	Other lineages detected via RAT sequencing
	Wastewater lineage	Matching RAT lineages	Wastewater lineage	Matching RAT lineages		
2023 Aug 28	EG.5.1	EG.5.1.4	XBB.1.16	XBB.1.16.11	5	FL.1.5.1, XBB.2.3
2023 Sep 11	EG.5.1	EG.5.1.13	XBB.1.16	XBB.1.16	8	XBB.1.5.10, XBB.1.5, GE.1,HK.9, JF.1.1, HH.1
2023 Sep 25	EG.5.1	EG.5.1.4	XBB.1.16	XBB.1.16	4	GJ.1.2, FL.1.5.1
2023 Oct 9	EG.5.1	HV.1, HV.1.8	XBB.1.9	None	8	HK.29, JN.1.1, XBB.1.16.9
2023 Oct 23	EG.5.1	None	XBB.1.16	None	2	XCH.1, DV.7.1
2023 Nov 6	EG.5.1	EG.5.1.1, EG.5.1.6	XBB.1.16	None	6	FL.1.5.1, HK.13.2.1, HK.26
2023 No 20	EG.5.1	EG.5.1.1, EG.5.1.6, HV.1	XBB.1.16	XBB.1.16.6	10	HK.26, GK.1.1, JN.1.4.5, HK.3, JN.1
2023 Dec 4	EG.5.1	HV.1, HV.1.2, HV.1.6	BA.2.86	JN.1, JN.1.1, JN.1.38	14	XDA, XCV, GK.1.8
2023 Dec 18	BA.2.86	JN.1, JN.1.1, JN.1.4, JN.1.42	EG.5.1	EG.5.1, EG.5.1.8, HV.1	11	JG.3, GW.5.1.1, GK.1.6.1
2024 Jan 1	BA.2.86	JN.1, JN.1.39	EG.5.1	None	3	None
2024 Jan 15	BA.2.86	JN.1, JN.1.38, JN.1.4, JN.1.42	EG.5.1	HV.1	15	JG.3, XDD, XDP, JC.5.1, HK.3.2
2024 Jan 29	BA.2.86	JN.1, JN.1.1	EG.5.1	None	9	None
2024 Feb 12	BA.2.86	JN.1, JN.1.42	XBB.2.3	None	4	None

*Lineages represent successful rapid antigen test sequences that corresponded to the 2 most prevalent lineage groups in the wastewater signal in the state of Wisconsin for each 2-week period. RAT, rapid antigen test.
†Dates represent first day of each 2-week reporting period.
‡Lineage group comprising the first largest percentage of wastewater data.
§Lineage group comprising the second largest percentage of wastewater data.
¶Passing sequences had ≥90% genome coverage at >10× depth.

for collecting and sequencing RATs might not be sustainable long-term. A similar program could be established with RATs purchased by community members (e.g., by partnering with pharmacies to put them at point-of-sale), but that could greatly bias the results toward persons who have the resources and motivation to purchase costly tests. Providing free tests to members of the community gives them a valuable tool to minimize their risk for COVID-19 transmission while also potentially providing more inclusive, representative genomic surveillance.

In conclusion, the program described here could act as a framework for the creation of more expansive genomic surveillance programs. Regulators in some countries have approved at-home RATs for other respiratory viruses, including influenza A virus and respiratory syncytial virus (*30*–*32*), and those tests could be collected to set up surveillance programs for other viruses. Other studies have demonstrated the possibility of recovering various respiratory viruses from COVID-19 RATs (*15*,*33*). Thus, by collecting both positive and negative RATs from symptomatic persons, the prevalence of respiratory viruses circulating in communities could also be estimated, creating an innovative additional method for assessing the spread of respiratory viruses in communities.

Table 3. Lineages detected in a study of SARS-CoV-2 genomic surveillance from community-distributed rapid antigen tests, Wisconsin, USA*

GenBank accession no.	Scanned test date	Pango lineage
PP761647	2023 Oct 14	JN.1.1
PP747716	2023 Dec 4	XDA
PP747739	2023 Dec 21	JN.1.2
PP747779	2023 Dec 22	XDE
PP747696	2024 Jan 17	XDD
PP747750	2024 Jan 24	XDP

*These samples were the earliest recorded examples of respective Pango lineage in Wisconsin according to data submitted to GISAID (https://www.gisaid.org) and GenBank as of April 18, 2024.

This article was preprinted at https://www.medrxiv.org/content/10.1101/2024.08.12.24311680v1.

Acknowledgments

We thank the Wisconsin Department of Health Services for providing rapid antigen tests to libraries and the public during and after the COVID-19 public emergency. We thank the many patrons of the South Central Public Libraries and Public Health Madison Dane County, who were kind enough to take time while they were sick to scan and send us their tests, without which this study would not be possible. We also thank all the staff at the South Central Library System who helped implement this program. We specifically thank Leah Fritsche, Erick Plumb, Matt Rahner, Samantha Seeman, and Elizabeth Clauss, who agreed to distribute research packets at their libraries. We also thank Sam Petykowski, a high school student who volunteered his time to produce research packets for Alicia Ashman Library.

The sequencing data generated in this study are available in the National Center for Biotechnology Information Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject no. PRJNA1096364. The accession numbers to the sequences used in these analyses are available in Appendix Table 2. Analysis of these data was made possible by the Center for High Throughput Computing's High Performance Cluster at the University of Wisconsin–Madison.

This work was supported by the Wisconsin Department of Health Services (project no. 435100-A24-ELCProjE) and the Centers for Disease Control and Prevention (grant no. 75D30122C15355).

The chatbot Claude 3.5 Sonnet created by Anthropic (https://www.anthropic.com) was used to improve syntax and to make the text more concise.

About the Author

Ms. Emmen is a research specialist performing infectious disease research at the David O'Connor Laboratory at the University of Wisconsin–Madison in the School of Medicine and Public Health. Her research interests include infectious diseases of animals and humans, global public health, and ecology.

References

1. Ladner JT, Sahl JW. Towards a post-pandemic future for global pathogen genome sequencing. PLoS Biol. 2023;21:e3002225. https://doi.org/10.1371/journal.pbio.3002225
2. Rasmussen M, Møller FT, Gunalan V, Baig S, Bennedbæk M, Christiansen LE, et al. First cases of SARS-CoV-2 BA.2.86 in Denmark, 2023. Euro Surveill. 2023;28:36. https://doi.org/10.2807/1560-7917.ES.2023.28.36.2300460
3. Oliveira Roster KI, Kissler SM, Omoregie E, Wang JC, Amin H, Di Lonardo S, et al. Surveillance strategies for the detection of new pathogen variants across epidemiological contexts. PLOS Comput Biol. 2024;20:e1012416. https://doi.org/10.1371/journal.pcbi.1012416
4. Robishaw JD, Alter SM, Solano JJ, Shih RD, DeMets DL, Maki DG, et al. Genomic surveillance to combat COVID-19: challenges and opportunities. Lancet Microbe. 2021;2:e481–4. https://doi.org/10.1016/S2666-5247(21)00121-X
5. World Health Organization. Recommendations for national SARS-CoV-2 testing strategies and diagnostic capacities [cited 2024 Apr 29]. https://www.who.int/publications/i/item/WHO-2019-nCoV-lab-testing-2021.1-eng
6. Kates J, Cubanski J, Cox C, Published JT. Timeline of end dates for key health-related flexibilities provided through COVID-19 emergency declarations, legislation, and administrative actions [cited 2024 Nov 20]. https://www.kff.org/coronavirus-covid-19/issue-brief/timeline-of-end-dates-for-key-health-related-flexibilities-provided-through-covid-19-emergency-declarations-legislation-and-administrative-actions
7. Centers for Disease Control and Prevention. COVID data tracker [cited 2024 May 9]. https://covid.cdc.gov/covid-data-tracker
8. Centers for Disease Control and Prevention. COVID Museum COVID-19 timeline [cited 2025 Mar 6]. https://www.cdc.gov/museum/timeline/covid19.html
9. Rader B, Gertz A, Iuliano AD, Gilmer M, Wronski L, Astley CM, et al. Use of at-home COVID-19 tests—United States, August 23, 2021–March 12, 2022. MMWR Morb Mortal Wkly Rep. 2022;71:489–94.
10. Khalid MF, Selvam K, Jeffry AJN, Salmi MF, Najib MA, Norhayati MN, et al. Performance of rapid antigen tests for COVID-19 diagnosis: a systematic review and meta-analysis. Diagnostics (Basel). 2022;12:110. https://doi.org/10.3390/diagnostics12010110
11. American Society for Microbiology. How the SARS-CoV-2 EUA antigen tests work [cited 2024 Jul 25]. https://asm.org:443/Articles/2020/August/How-the-SARS-CoV-2-EUA-Antigen-Tests-Work
12. Food and Drug Administration. At-home OTC COVID-19 diagnostic tests [cited 2024 Jul 16]. https://www.fda.gov/medical-devices/coronavirus-covid-19-and-medical-devices/home-otc-covid-19-diagnostic-tests
13. Martin GE, Taiaroa G, Taouk ML, Savic I, O'Keefe J, Quach R, et al. Maintaining genomic surveillance using whole-genome sequencing of SARS-CoV-2 from rapid antigen test devices. Lancet Infect Dis. 2022;22:1417–8. https://doi.org/10.1016/S1473-3099(22)00512-6
14. Rector A, Bloemen M, Schiettekatte G, Maes P, Van Ranst M, Wollants E. Sequencing directly from antigen-detection rapid diagnostic tests in Belgium, 2022: a gamechanger in genomic surveillance? Euro Surveill. 2023;28:91. https://doi.org/10.2807/1560-7917.ES.2023.28.9.2200618
15. Paull JS, Petros BA, Brock-Fisher TM, Jalbert SA, Selser VM, Messer KS, et al. Optimisation and evaluation of viral genomic sequencing of SARS-CoV-2 rapid diagnostic tests: a laboratory and cohort-based study. Lancet Microbe. 2024;5:e468–77. https://doi.org/10.1016/S2666-5247(23)00399-3
16. Nguyen PV, Carmola LR, Wang E, Bassit L, Rao A, Greenleaf M, et al. SARS-CoV-2 molecular testing and whole genome sequencing following RNA recovery from used BinaxNOW COVID-19 antigen self tests. J Clin Virol. 2023;162:105426. https://doi.org/10.1016/j.jcv.2023.105426
17. Macori G, Russell T, Barry G, McCarthy SC, Koolman L, Wall P, et al. Inactivation and recovery of high quality RNA from positive SARS-CoV-2 rapid antigen tests suitable for whole virus genome sequencing. Front Public Health. 2022;10:863862. https://doi.org/10.3389/fpubh.2022.863862
18. Health Innovation Program. ZIP codes by rural and urban groupings: HIPxChange [cited 2025 Jan 3]. https://hipxchange.org/toolkit/ruralurbangroups
19. Coelho FF, da Silva MA, Lopes TB, Polatto JM, de Castro NS, Andrade LAF, et al. SARS-CoV-2 rapid antigen test based on a new anti-nucleocapsid protein monoclonal antibody: development and real-time validation. Microorganisms. 2023;11:2422. https://doi.org/10.3390/microorganisms11102422
20. US Census Bureau. Geographic areas reference manual. Washington: the Bureau; 1994.
21. Lu X, Wang L, Sakthivel SK, Whitaker B, Murray J, Kamili S, et al. US CDC real-time reverse transcription PCR panel for detection of severe acute respiratory syndrome coronavirus 2. Emerg Infect Dis. 2020;26:1654–65. https://doi.org/10.3201/eid2608.201246

22. Patel H, Monzón S, Varona S, Espinosa-Carrasco J, Garcia MU, Heuer ML, et al. nf-core/viralrecon: nf-core/viralrecon v2.6.0–rhodium raccoon [cited 2024 May 29]. https://zenodo.org/record/7764938
23. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–8. https://doi.org/10.1038/s41587-020-0439-x
24. Wisconsin State Laboratory of Hygiene. SARS-CoV-2 wastewater genomic dashboard [cited 2024 Jun 7]. https://dataportal.slh.wisc.edu/sc2-ww-dashboard
25. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6:3773. https://doi.org/10.21105/joss.03773
26. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global SARS-CoV-2 genomic surveillance: what we have learned (so far). Infect Genet Evol. 2023;108:105405. https://doi.org/10.1016/j.meegid.2023.105405
27. Public Health Madison & Dane County. Respiratory illness dashboard [cited 2024 Jun 7]. https://publichealthmdc.com/health-services/respiratory-illness/dashboard
28. US Postal Service. Postal Service delivery performance continues to average 2.6 days [cited 2024 Nov 13]. https://about.usps.com/newsroom/national-releases/2023/1222-usps-delivery-performance-continues-to-average-2-6-days.htm
29. Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics. 2022;38:1735–7. https://doi.org/10.1093/bioinformatics/btab856
30. Food and Drug Administration. Influenza diagnostic tests [cited 2024 Aug 2]. https://www.fda.gov/medical-devices/in-vitro-diagnostics/influenza-diagnostic-tests
31. Therapeutic Goods Administration. Respiratory combo panel RSV/SARS-CoV-2/Influenza A/B Rapid Antigen Test Kit RAT-19 (self-test) (nasal swab) (combination sel-tests) [cited 2024 Nov 14]. https://www.tga.gov.au/resources/covid-19-test-kits/respiratory-combo-panel-rsv-sars-cov-2-influenza-ab-rapid-antigen-test-kit-rat-19-self-test-nasal-swab-combination-self-tests
32. Therapeutic Goods Administration. COVID-19, Influenza A/B & RSV Antigen Nasal Test Kit for self-testing (Biolink Biopen) [cited 2024 Nov 14]. https://www.tga.gov.au/resources/covid-19-test-kits/covid-19-influenza-ab-rsv-antigen-nasal-test-kit-self-testing-biolink-biopen
33. Smith-Jeffcoat SE, Mellis AM, Grijalva CG, Talbot HK, Schmitz J, Lutrick K, et al.; RVTN-Sentinel Study Group. SARS-CoV-2 viral shedding and rapid antigen test performance — respiratory virus transmission network, November 2022–May 2023. MMWR Morb Mortal Wkly Rep. 2024;73:365–71. https://doi.org/10.15585/mmwr.mm7316a2

Address for correspondence: David H. O'Connor, AIDS Vaccine Research Laboratory, University of Wisconsin-Madison, 555 Science Dr, Madison, WI 53711, USA; email: dhoconno@wisc.edu

Establishing Methods to Monitor Influenza A(H5N1) Virus in Dairy Cattle Milk, Massachusetts, USA

Elyse Stachler, Andreas Gnirke, Kyle McMahon, Michael Gomez, Liam Stenson, Charelisse Guevara-Reyes, Hannah Knoll, Toni Hill, Sellers Hill, Katelyn S. Messer, Jon Arizti-Sanz, Fatinah Albeez, Elizabeth Curtis, Pedram Samani, Natalia Wewior, David H. O'Connor, William Vuyk, Sophia E. Khoury, Matthew K. Schnizlein, Nicole C. Rockey, Zachariah Broemmel, Michael Mina, Lawrence C. Madoff, Shirlee Wohl, Lorraine O'Connor, Catherine M. Brown, Al Ozonoff, Daniel J. Park, Bronwyn L. MacInnis,[1] Pardis C. Sabeti[1]

Highly pathogenic avian influenza A(H5N1) virus has caused a multistate outbreak among US dairy cattle, spreading across 16 states and infecting hundreds of herds since its onset. We rapidly developed and optimized PCR-based detection assays and sequencing protocols to support H5N1 molecular surveillance. Using 214 retail milk samples from 20 states for methods development, we found that H5N1 virus concentrations by digital PCR strongly correlated with quantitative PCR cycle threshold values; digital PCR exhibited greater sensitivity. Metagenomic sequencing after hybrid selection was best for higher concentration samples, whereas amplicon sequencing performed best for lower concentrations. By establishing these methods, we were able to support the creation of a statewide surveillance program to perform monthly testing of bulk milk samples from all dairy cattle farms in Massachusetts, USA, which remain negative to date. The methods, workflow, and recommendations described provide a framework for others aiming to conduct H5N1 surveillance efforts.

Highly pathogenic avian influenza A(H5N1) virus infection has caused large-scale outbreaks in wild and domestic birds, resulting in mass deaths, culling events, and economic losses (*1*). Viral spillover to mammals has become more frequent, including outbreaks involving mammal-to-mammal transmission and sporadic human infections (*2*). In March 2024, H5N1 clade 2.3.4.4b virus was found in unpasteurized milk produced by infected dairy cattle in the United States, the first confirmation of an outbreak that grew to span 927 herds in 16 states as of January 15, 2025 (*3*,*4*). The outbreak subsequently spread through interstate transport of cattle, milking practices, and shared milking machinery and farm equipment (*5*,*6*). Although confirmed human cases have thus far been sporadic and have primarily been associated with mild symptoms, the spread of H5N1 virus in cattle

Author affiliations: Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA (E. Stachler, A. Gnirke, K. McMahon, M. Gomez, L. Stenson, C. Guevara-Reyes, H. Knoll, T. Hill, S. Hill, K.S. Messer, J. Arizti-Sanz, F. Albeez, E. Curtis, P. Samani, N. Wewior, A. Ozonoff, D.J. Park, B.L. MacInnis, P.C. Sabeti); University of Puerto Rico-Rio Piedras, San Juan, Puerto Rico, USA (C. Guevara-Reyes); Harvard University, Cambridge (P. Samani, B.L. MacInnis, P.C. Sabeti); University College London, London, UK (P. Samani); University of Wisconsin-Madison, Madison, Wisconsin, USA (D.H. O'Connor, W. Vuyk); The University of Texas at Austin, Austin, Texas, USA (S.E. Khoury); Michigan State University, East Lansing, Michigan, USA (M.K. Schnizlein); Duke University, Durham, North Carolina, USA (N.C. Rockey, Z. Broemmel); Immune Observatory, Boston, Massachusetts, USA (M. Mina); University of Massachusetts Chan Medical School, Worcester, Massachusetts, USA (L.C. Madoff); Massachusetts Department of Public Health, Boston (L.C. Madoff, S. Wohl, C.M. Brown); Brigham and Women's Hospital, Boston (S. Wohl); Massachusetts Department of Agricultural Resources, Boston (L. O'Connor); Boston Children's Hospital, Boston (A. Ozonoff); Harvard Medical School, Boston (A. Ozonoff); Howard Hughes Medical Institute, Chevy Chase, Maryland, USA (P.C. Sabeti)

DOI: https://doi.org/10.3201/eid3113.250087

[1]These senior authors contributed equally to this article.

threatens the dairy industry and risks further adaptation to mammalian hosts, including humans.

This outbreak has highlighted the need for rapidly deployable H5N1 molecular surveillance capacity to detect infections, monitor viral spread and evolution, identify transmission routes, and target interventions to protect agricultural assets and food supply and prevent broader human transmission. Cow milk has emerged as an ideal sample source for H5N1 virus detection and surveillance during this outbreak; the virus is shed in high concentrations in milk, likely because of its affinity for infecting mammary gland epithelial cells (*7*). However, milk undergoes intense processing steps, including ultrapasteurization and homogenization, which have unknown effects on viral RNA quality.

We optimized methods for nucleic acid extraction, molecular detection, and sequencing of H5N1 virus in cow milk, first using synthetic nucleic acid material and subsequently validating those methods by using positive retail milk samples from affected states. By quickly establishing a robust workflow for detecting and sequencing H5N1 virus from milk as the outbreak emerged, we were positioned to support mandatory statewide surveillance for H5N1 virus in milk from dairy cattle farms across Massachusetts. This program, launched in August 2024, was implemented preemptively in the absence of H5N1 detection in the state and surrounding region to confirm the absence of H5N1 and to serve as an early warning system if a local outbreak occurs. State authorities worked with farms to collect samples from bulk milk tanks from all 95 dairy cattle farms across Massachusetts, initially within a 3-week period, followed by a rotating sampling schedule testing all farms monthly. On the basis of our workflow development and validation using retail milk samples (see next section), we extracted bulk milk samples using the MagMAX CORE extraction kit (Thermo Fisher Scientific, https://www.thermofisher.com) and performed digital PCR (dPCR) to detect H5N1 virus; we used the bovine RNaseP gene (RP_Bov) as a positive internal control. Although the surveillance program is ongoing, we have completed 4 rounds of statewide testing, and H5N1 has not been detected in the state. The RP_Bov–positive control has been routinely detected at similar levels to retail milk, providing confidence in the negative results obtained for H5N1 (Appendix Figure 1, https://wwwnc.cdc.gov/EID/article/31/13/25-0087-App1.pdf).

The sensitivity of our workflow allows for preemptive surveillance of H5N1 for the typical size of a Massachusetts dairy cattle farm (≈10,000 cows on 125 farms) (*8*). On the basis of our limit of detection (LOD) of 10^4 copies/mL milk, to detect 1 infected cow in a herd size of either 100 or 1,000 cows, the infected cow would have to be shedding 10^6 H5N1 copies/mL milk (for a herd of 100) or 10^7 H5N1 copies/mL milk (for a herd of 1,000). This level is within the concentration range of live virus shed by infected cattle (10^4–$10^{8.8}$ 50% tissue culture infectious dose/mL) (*7*). Despite the complexity of milk as a sample type, the robust detection of viral RNA in affected milk offers a unique surveillance mechanism to easily monitor lactating herds by testing pooled bulk milk tank samples, saving time and resources compared with the testing of individual cows.

Characteristics of the Validated Workflow

This article is meant to serve as a resource documenting how other laboratories can quickly validate and implement testing. The characteristics of the validated workflow are summarized next (Appendix). First, we tested performance of a previously published H5N1 assay targeting the H5 subtype of the hemagglutinin (HA) gene inclusive of the current virus outbreak strain (*9*) (H5_Taq) by both quantitative PCR (qPCR) and dPCR (Appendix). We optimized primer and probe concentrations using synthetic H5N1 RNA, selecting for optimal linearity, sensitivity, accuracy, precision, and qPCR efficiency (Appendix Figures 2, 3).

Overall, the H5N1 assay displayed robust performance on both platforms; dPCR outperformed qPCR in LOD and precision. The 90% LOD was 5 copies/µL by dPCR and 10 copies/µL by qPCR. In addition, although dPCR concentrations correlated well with qPCR cycle threshold (Ct) values (Figure 1, panel A), dPCR exhibited lower coefficients of variations, ranging from 10.5% to 26.4%, compared with 18.0% to 111.5% for qPCR (Figure 1, panel B). Both assays maintained linearity over their dynamic ranges (Figure 1, panels C, D).

As a positive internal control for nucleic acid extraction in cattle milk, we designed a PCR targeting the bovine Ribonuclease P gene (both DNA and RNA; RP_Bov). By dPCR, linearity was maintained across all dilutions tested (Figure 1, panel C) with a 90% LOD of 10 copies/µL. On the basis of the superior performance of dPCR for the H5N1 target virus, the RP_Bov assay was not evaluated as a qPCR. Overall, all PCRs performed well with minimal optimization.

We next evaluated preprocessing and extraction protocols to optimize sample preparation for subsequent H5N1 virus detection and sequencing. We tested 2 commercially available extraction kits, MagMAX Prime Viral/Pathogen (Prime) and

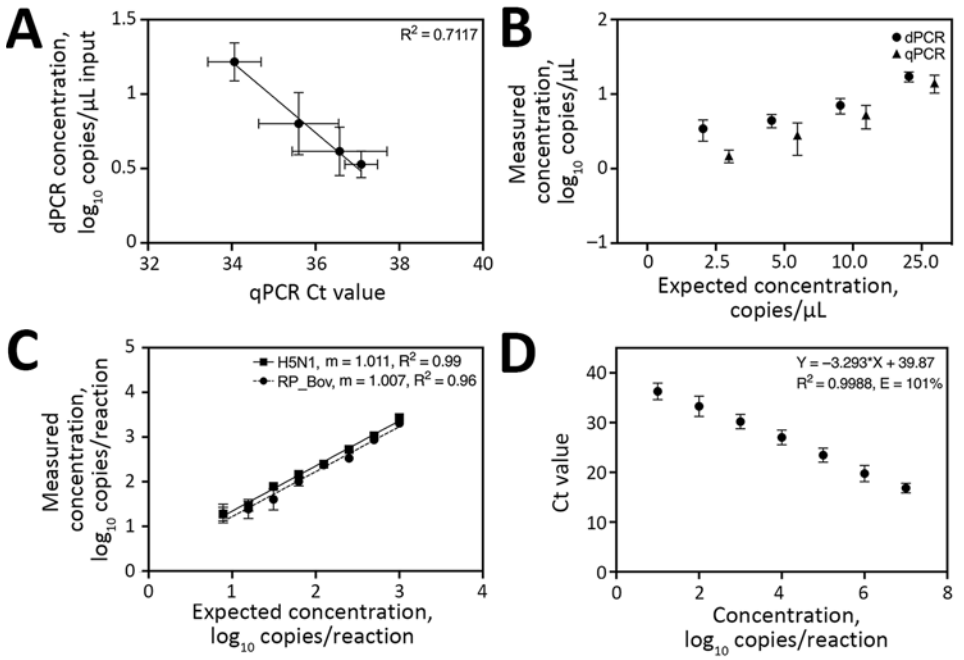


Figure 1. Validation and characterization of dPCR and qPCR on synthetic spike-in samples in study of methods to monitor influenza A(H5N1) virus in dairy cattle milk, Massachusetts, USA. A, B) Limit of detection analysis for correlation of dPCR concentrations with qPCR Ct values (A) and measured concentrations compared to expected concentrations for both qPCR and dPCR (B). C, D) Detection of dPCR (H5_Taq and RP_Bov) (C) and qPCR (H5_Taq) (D) assays using serial dilutions of synthetic H5N1 RNA standard material. For qPCR data, we combined and jointly analyzed all standard curve data from runs during retail milk testing. Fitted lines in panels A and D represent simple linear regression lines of best fit. Error bars indicate ±1 SD. Ct, cycle threshold; dPCR, digital PCR; qPCR, quantitative PCR; R^2, coefficient of determination.

MagMAX CORE (CORE) (both Thermo Fisher Scientific), by spiking serial dilutions of synthetic H5N1 nucleic acid into milk. We tested milk with various fat contents and examined the effect of pre-centrifugation (at either 1,200 × *g* or 12,000 × *g*) on outcomes. We also tested the MagMAX Wastewater kit (Wastewater) (Thermo Fisher Scientific) head-to-head with the CORE kit on a subset of 8 retail milk samples previously found to be H5N1 virus positive with CORE kit testing.

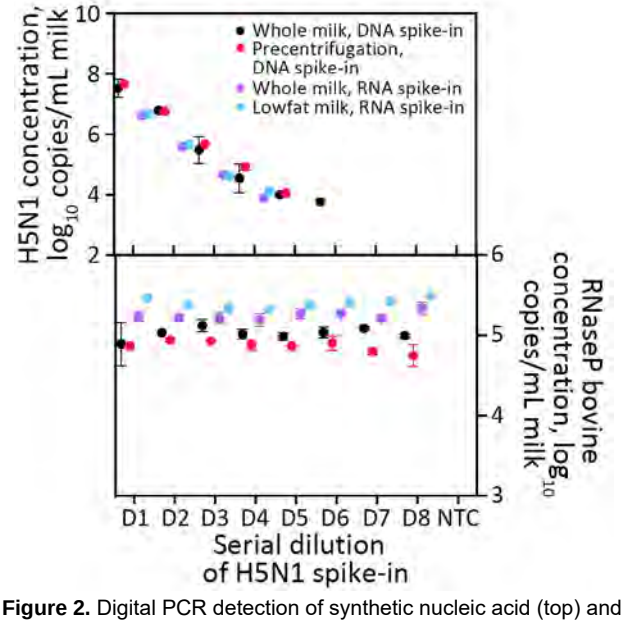


Figure 2. Digital PCR detection of synthetic nucleic acid (top) and RNaseP Bovine (bottom) in study of methods to monitor influenza A(H5N1) virus in dairy cattle milk, Massachusetts, USA. For direct extraction, we extracted 200 µL of milk spiked with serial dilutions of H5N1 synthetic gene fragments. For precentrifugation, we centrifuged samples for 12,000 × *g* for 10 minutes after spike-in, after which we extracted 200 µL. Extractions were performed using the MagMAX CORE extraction kit (Thermo Fisher Scientific, https://www.thermofisher.com).

All 3 extraction kits demonstrated similar recovery of H5N1 virus from milk; the CORE kit exhibited slightly better performance. The CORE (Figure 2) and Prime (Appendix Figure 4) kits showed comparable results in terms of total recovery (down to ≈10^4 H5N1 virus copies/mL milk) and linearity. Direct nucleic acid extraction from milk was efficient regardless of fat content, with pre-centrifugation offering no increase in viral RNA recovery, in accordance with previous findings (*10*; A. Lail et al., unpub. data, https://www.protocols.io/view/rna-extraction-from-milk-for-hpai-surveillance-n2bvjn6obgk5/v1). In addition, we found no significant difference in detection of H5N1 virus (p = 0.20) or RP_Bov (p = 0.17) using the Wastewater extraction kit on retail milk samples (Appendix Figure 5). We selected the CORE kit for ongoing testing given its low detection limit and slightly better detection of RP_Bov, as well as practical considerations, such as a manufacturer's protocol for processing milk and kit availability.

To validate protocols on in situ H5N1 virus in milk, we sourced 214 retail milk cartons with diverse characteristics, including fat content and

pasteurization processes, from 61 processing plants in 20 states (Table; Appendix Figure 6). Of those, 55 (26%) cartons tested positive for H5N1 RNA by dPCR, whereas 48 (22%) tested positive by qPCR. The platforms gave concordant positive/negative results for 95% (n = 203/214) of samples (Appendix, Figure 7). Nine samples were positive only by dPCR, which could be because of the slightly enhanced LOD of the dPCR assay. Conversely, 2 samples were positive only by qPCR, possibly because of the more stringent thresholding criteria for dPCR. Further, H5N1 RNA dPCR concentrations correlated strongly with qPCR Ct values (R^2 = 0.81; Figure 3), suggesting the assay is robust on either platform. However, we saw evidence of qPCR standard degradation throughout testing, highlighting the importance of standard material integrity for accurate qPCR quantification. Positive samples were from processing plants in 4 states with reported H5N1 outbreaks (Colorado, Idaho, Michigan, and Texas). We also detected 1 positive sample by both dPCR and qPCR that originated from a processing plant in Missouri, which has not reported H5N1 in cattle. Of note, the location of the processing plant reported on milk containers might or might not correspond to the state in which the milk was initially collected, and this linkage is not publicly available.

We used the RP_Bov assay as an internal sample process control to confirm sample integrity and ensure proper collection and extraction, especially useful to interpret negative H5N1 results. RP_Bov concentrations averaged 560 copies/µL extract (Figure 4); 98% of samples fell within 1 SD. Thus, detection of RP_Bov below ≈100 copies/µL could be effectively used as a measure of milk sample and process integrity.

We next sought to recover genomes from 23 H5N1 virus–positive retail milk samples, testing methods across a range of characteristics including virus concentration, milk type, and pasteurization process. To obtain higher H5N1 virus concentrations for library preparation, we first extracted, pooled, and concentrated 10 samples from each milk container. Ultrapasteurized samples exhibited significantly lower concentration factors than did pasteurized samples as measured by H5N1 copy number (p = 0.015; Appendix Figure 8). Despite being highly concentrated, samples showed no evidence of PCR inhibition by dPCR (p = 0.89; Appendix Figure 9). The recovered RNA content and quality from these samples spanned a wide range as determined by H5N1 copies, total RNA concentration, H5N1 copies per nanogram of RNA, and RNA integrity number score (Appendix Table 7).

We evaluated 3 library construction methods to assess their efficacy in producing genomes across the range of H5N1 virus concentrations and pasteurization

Table. Breakdown of milk samples tested and their results by processing plant state in study of methods to monitor influenza A(H5N1) virus in dairy cattle milk, Massachusetts, USA

Processing plant state	No. cartons tested	No. positive	Positivity rate, %
AZ	1		
CA	10		
CO*	59	33	56
CT	4		
IA*	9		
ID*	12	5	42
KS*	2		
KY	1		
MA	18		
ME	2		
MI*	14	5	36
MN*	9		
MO	3	1	33
NC*	7		
NH	6		
NY	2		
OH*	3		
TX*	42	13	31
UT	7		
VA	3		
Total	214	57	27

*Indicates state had reported cases of H5N1 in cattle at the time of testing.

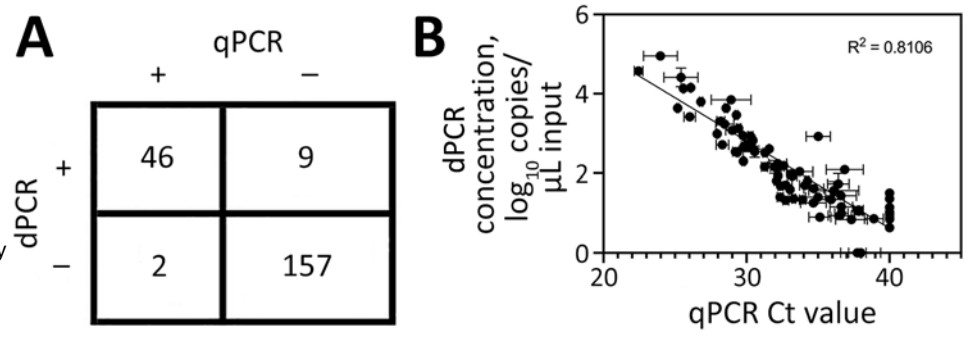


Figure 3. Comparison of dPCR and qPCR virus testing on retail milk samples in study of methods to monitor influenza A(H5N1) virus in dairy cattle milk, Massachusetts, USA. A) Agreement of positive and negative calls of milk samples between the 2 platforms; B) correlation of H5N1 measured by dPCR concentration compared with qPCR Ct value. For plotting purposes, samples not detected by dPCR were graphed with a dPCR concentration of 0 copies/µL, whereas samples not detected by qPCR were graphed with a Ct value of 40. Error bars indicate ±1 SD. Ct, cycle threshold; dPCR, digital PCR; qPCR, quantitative PCR; R^2, coefficient of determination.

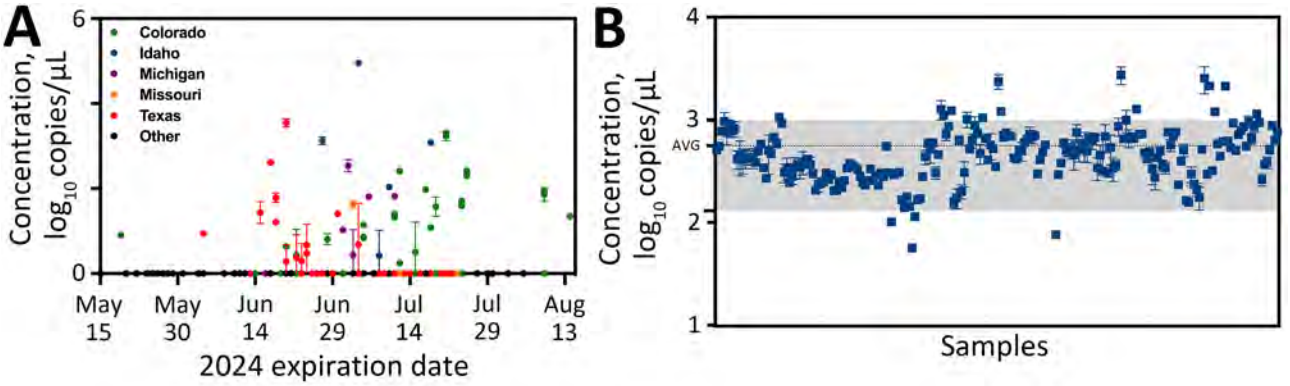


Figure 4. Virus and bovine ribonuclease P (RP_Bov) concentrations for all retail milk samples as measured by digital PCR in study of methods to monitor influenza A(H5N1) virus in dairy cattle milk, Massachusetts, USA. A) Concentration of H5N1 as a function of processing state and expiration date. B) RP_Bov data for all samples. The gray-shaded region corresponds to the average RP_Bov concentration of all data ±1 SD. Error bars indicate ±1 SD.

processes: untargeted metagenomic RNA sequencing (RNA-Seq), hybrid-selected RNA-Seq (hsRNA-Seq) enriched for human respiratory viruses including influenza A (albeit not explicitly H5N1) (*11*), and amplicon sequencing (Amp-Seq) of tiled 250-bp H5N1 PCR products (*12*). Despite intense milk preprocessing (such as ultrapasteurization), near-complete (>70% assembly) H5N1 virus genomes were readily recovered from all 23 samples, 12 by hsRNA-Seq (≥80%) and 11 by Amp-Seq (≥74%). Hybrid selection greatly increased the chances of genome recovery for higher concentration extracts (>500 copies/µL); hsRNA-Seq outperformed RNA-Seq for 11 of 12 samples. At lower concentrations, Amp-Seq resulted in the most complete genomes (Figure 5). Of note, we modified the PCR cycling conditions of a previously reported H5N1 Amp-Seq protocol (*12*), which resulted in improved amplicon generation and genome assemblies (Appendix Figure 10). However, PCR efficiency varied considerably across amplicons; a small fraction of amplicons produced most sequencing reads (Appendix Figure 11).

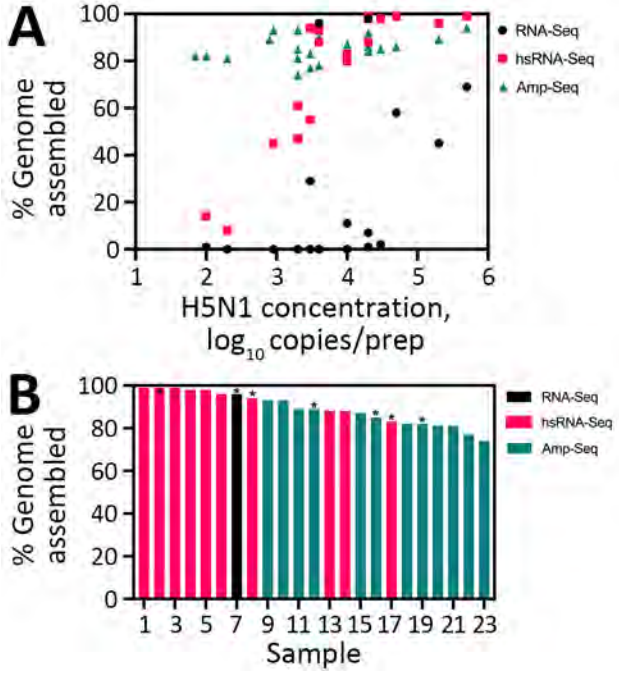


Figure 5. Virus genome assemblies from retail milk samples in study of methods to monitor influenza A(H5N1) virus in dairy cattle milk, Massachusetts, USA. A) Completeness of H5N1 genome assemblies generated by RNA-Seq, virus-enriched (hsRNA-Seq), and targeted H5N1 Amp-Seq as a function of H5N1 copies per milliliter of RNA. B) The most complete H5N1 assembly produced for each sample sorted by length and the underlying sequencing approach. Asterisks (*) above bars indicate ultrapasteurized samples. Amp-Seq, amplicon sequencing; hsRNA-Seq, hybrid-selected metagenomics; RNA-Seq, unbiased metagenomics.

Phylogenetic analysis showed geographic clustering with other publicly available H5N1 genomes associated with the dairy cattle outbreak (Appendix Figure 12), suggesting the origin of the viruses was consistent with the US state of the processing plant of the milk. Of note, the positive sample originating from Missouri (which has no reports of H5N1 in cattle) clustered with samples from Texas and Michigan, likely pointing to the farm location from which the milk originated, despite being processed in a Missouri plant.

Overall, this study contributes validated methods for the whole workflow from sample to analyzed data for rapid deployment for potential future epidemiologic studies and public health surveillance. On the basis of the methods testing and validation described, we have included a guide to establishing efficient, robust, and scalable H5N1 virus surveillance from bulk milk for implementation in molecular laboratory settings (Appendix). Enabling more laboratories to set up decentralized surveillance will enable us to stay ahead of current and future outbreaks of public concern. The

guidelines provided in this article are intended to serve as a blueprint for rapid validation of new molecular detection methods and establishment of surveillance systems for the current H5N1 outbreak and beyond.

This article was originally published as a preprint at https://www.medrxiv.org/content/10.1101/2024.12.04.24318491v1.

Acknowledgments

We thank the Boston Globe Media Partners, LLC, staff for sourcing and donating local New England milk samples; John Rinn, Noreen Beckie, Kristen Koneschik, and Michael Butts for collecting US-based milk samples; and other teammates/collaborators for helpful discussions.

This work was supported by funding from the Howard Hughes Medical Institute (HHMI) Investigator Program (to P.C.S.), the Centers for Disease Control and Prevention (BAA 75D30122C15113 to P.C.S and PGCoE NU50CK000629 to S.W., L.M., P.C.S, and B.L.M.), and the National Institutes of Health National Institute of Allergy and Infectious Diseases (GCID U19AI110818 to P.C.S. and D.J.P. and CREID U01AI151812 to P.C.S.). The diagnostic development work was made possible by support from Flu Lab and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. D.H.O.'s and W.V.'s work on this project was funded by Heart of Racing and the UW Institute for Clinical and Translational Research's Pilot Award program. This publication was supported by the Office of Advanced Molecular Detection, Centers for Disease Control and Prevention, through cooperative agreement no. CK22-2204. The content is solely the responsibility of the authors and does not necessarily represent the official views or policies of the Centers for Disease Control and Prevention or the US government. This study has been approved for public release; distribution is unlimited.

P.C.S. is cofounder and shareholder of Delve Bio. She was formerly cofounder and shareholder of Sherlock Biosciences and board member and shareholder of Danaher Corporation. D.H.O. is a cofounder and managing member of Pathogenuity LLC.

About the Author

Dr. Stachler is a research scientist in the Sabeti Lab Diagnostics Group at the Broad Institute of MIT and Harvard in Boston. Her research focuses on the intersection of environmental microbiology, infectious diseases, diagnostics, and public health.

References

1. Peacock T, Moncla L, Dudas G, VanInsberghe D, Sukhova K, Lloyd-Smith JO, et al. The global H5N1 influenza panzootic in mammals. Nature. 2024. https://doi.org/10.1038/s41586-024-08054-z
2. Graziosi G, Lupini C, Catelli E, Carnaccini S. Highly pathogenic avian influenza (HPAI) H5 clade 2.3.4.4b virus infection in birds and mammals. Animals (Basel). 2024;14:1372. https://doi.org/10.3390/ani14091372
3. World Health Organization. Avian influenza A(H5N1)—United States of America. 2024 April 9 [cited 2024 Oct 1]. https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON512
4. US Department of Agriculture Animal and Plant Health Inspection Service. HPAI confirmed cases in livestock. 2024 Jul 3 [cited 2024 Dec 3]. https://www.aphis.usda.gov/livestock-poultry-disease/avian/avian-influenza/hpai-detections/hpai-confirmed-cases-livestock
5. Halwe NJ, Cool K, Breithaupt A, Schön J, Trujillo JD, Nooruzzaman M, et al. H5N1 clade 2.3.4.4b dynamics in experimentally infected calves and cows. Nature. 2024. https://doi.org/10.1038/s41586-024-08063-y
6. Le Sage V, Campbell AJ, Reed DS, Duprex WP, Lakdawala SS. Persistence of influenza H5N1 and H1N1 viruses in unpasteurized milk on milking unit surfaces. Emerg Infect Dis. 2024;30:1721–3. https://doi.org/10.3201/eid3008.240775
7. Caserta LC, Frye EA, Butt SL, Laverack M, Nooruzzaman M, Covaleda LM, et al. Spillover of highly pathogenic avian influenza H5N1 virus to dairy cattle. Nature. 2024;634:669–76. https://doi.org/10.1038/s41586-024-07849-4
8. USDA National Agricultural Statistics Service. 2023 Agricultural Statistics Annual Bulletin New England [cited 2024 Sep 20]. https://www.nass.usda.gov/Statistics_by_State/New_England_includes/Publications/Annual_Statistical_Bulletin/2023/2023_NewEngland_Annual_Bulletin.pdf
9. Wolfe MK, Duong D, Shelden B, Chan EMG, Chan-Herur V, Hilton S, et al. Detection of hemagglutinin H5 influenza A virus sequence in municipal wastewater solids at wastewater treatment plants with increases in influenza A in spring, 2024. Environ Sci Technol Lett. 2024;11:526–532. https://doi.org/10.1021/acs.estlett.4c00331
10. Minsky BB, Kadam A, Tanner NA, Cantor E, Patton GC. Facilitating purification and detection of viral nucleic acids from milk. New England Biolabs [cited 2024 Jun 7]. https://www.neb.com/en-us/-/media/nebus/files/application-notes/appnote_facilitating_purification_and_detection_of_viral_nucleicacids_from_milk.pdf
11. Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, Brehio P, et al.; Viral Hemorrhagic Fever Consortium. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. Nat Biotechnol. 2019;37:160–8. https://doi.org/10.1038/s41587-018-0006-x
12. Vuyk WC, Lail A, Emmen I, Hassa N, Tiburcio PB, Newman C, et al. Whole genome sequencing of H5N1 from dairy products with tiled 250bp amplicons [cited 2024 Jun 5]. https://www.protocols.io/view/whole-genome-sequencing-of-h5n1-from-dairy-product-kqdg322kpv25/v1

Address for correspondence: Elyse Stachler, Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142, USA; email: estachle@broadinstitute.org

Real-Time Use of Monkeypox Virus Genomic Surveillance, King County, Washington, USA, 2022–2024

Kathryn M. Lau, Michaela Banks, Kaila Bryant, Joanie D. Lambert, Laura Marcela Torres, Stephanie M. Lunn, Cory Yun, Pavitra Roychoudhury, B. Ethan Nunley, Jaydee Sereewit, Alexander L. Greninger, Allison Black, Vance Kawakami, Sargis Pogosjans, Elysia Gonzales, Eric J. Chow

A monkeypox virus genomic surveillance pilot began in King County, Washington, USA, during the 2022 outbreak. Genomic surveillance proved critical in determining local versus international exposure of a case where no known exposures were identified by interview, illustrating the value of genomics in case investigation and public health practice.

Mpox, an infectious disease caused by monkeypox virus (MPXV), emerged as a virus largely endemic to western and central Africa (*1*). When local transmission began to occur in many additional countries in 2022 (*1*), whole-genome sequencing (WGS) of MPXV offered a potential new tool to complement traditional case investigations and identify epidemiologic linkages. Public Health–Seattle & King County (PHSKC), Washington State Department of Health (WADOH), and the University of Washington Virology Laboratory (UW Virology) collaborated to pilot a retrospective genomic surveillance program investigating mpox in King County, Washington, USA, in September 2022. PHSKC subsequently used retrospective data from the pilot and real-time WGS to support investigation of a case of mpox with unknown exposure.

Data were collected as part of routine public health surveillance and are considered nonresearch. Patient consent was not required, but verbal consent was obtained from the patient whose case is described, and all identifying details of the patient have been removed in accordance with the institutional policy of PHSKC.

Methods

In September 2022, WADOH and PHSKC retrospectively linked WGS and epidemiologic data for cases of mpox occurring since May 2022 in King County. PHSKC collected epidemiologic data using case interviews and chart reviews for every reported case of mpox. UW Virology performed WGS on MPXV-positive residual diagnostic specimens using a hybridization probe-capture–based approach with probes designed using the MPXV 2022/MA001 strain (*2*). Laboratory staff generated consensus genomes using a custom Nextflow pipeline (https://github.com/greninger-lab/nf_mpxv_f13l). WADOH built a phylogenetic tree of all local cases and contextual sequences from other regions (*3*), and PHSKC annotated the tree with epidemiologic data. During September 2022–July 2024, PHSKC actively pursued WGS of all new mpox cases, linked WGS results to cases, and analyzed phylogenetic relatedness using Nextstrain and Nextclade (*4*,*5*).

Results

The retrospective analysis linked WGS results to 126 mpox cases that occurred during May–September 2022 (29.6% of 426 total cases during the pilot period). All isolates belonged to clade IIb, lineage B.1—the lineage identified in most sequences from the 2022–2023 global outbreak (*1*,*6*).

In fall 2023, an adult man residing in King County, Washington, had mpox symptoms develop 2 days after a 5-day trip to Kenya with an overnight layover

Author affiliations: Public Health–Seattle & King County, Seattle, Washington, USA (K.M. Lau, M. Banks, K. Bryant, J.D. Lambert, V. Kawakami, S. Pogosjans, E. Gonzales, E.J. Chow); Washington State Department of Health, Olympia, Washington, USA (L.M. Torres, S.M. Lunn, C. Yun, A. Black); University of Washington School of Medicine, Seattle (P. Roychoudhury, B.E. Nunley, J. Sereewit, A.L. Greninger, E.J. Chow); Fred Hutchinson Cancer Research Center, Seattle (P. Roychoudhury, A.L. Greninger); University of Washington School of Public Health, Seattle (E.J. Chow)

DOI: http://doi.org/10.3201/eid3113.241242

in United Arab Emirates (UAE). Symptoms began with small, blister-like penile lesions. Five days after lesions appeared, the man developed headache, groin pain, fatigue, and subjective fever. Eight days after symptom onset, healthcare professionals collected a lesion specimen, which tested positive for mpox by real-time PCR. The man completed a prescribed course of oral tecovirimat and reported fever and pain resolution 18 days after symptom onset and complete scab resolution 24 days after symptom onset. On the basis of symptom onset date and a typical incubation period of 3–17 days, the exposure period spanned dates when the man was in King County, Kenya, and the UAE (*7*). During case interviews, he reported having no exposure to anyone with known or suspected mpox and no sexual activity or close physical contact with anyone 3 weeks before symptom onset. According to interview alone, investigators remained uncertain where and when the man was likely exposed.

During September 2022–fall 2023, 71 cases from Washington were sequenced in real time (37.4% of 190 total cases), and all real time sequenced cases were lineage B.1. WGS results for this case were shared with PHSKC 15 days after the case report, and the

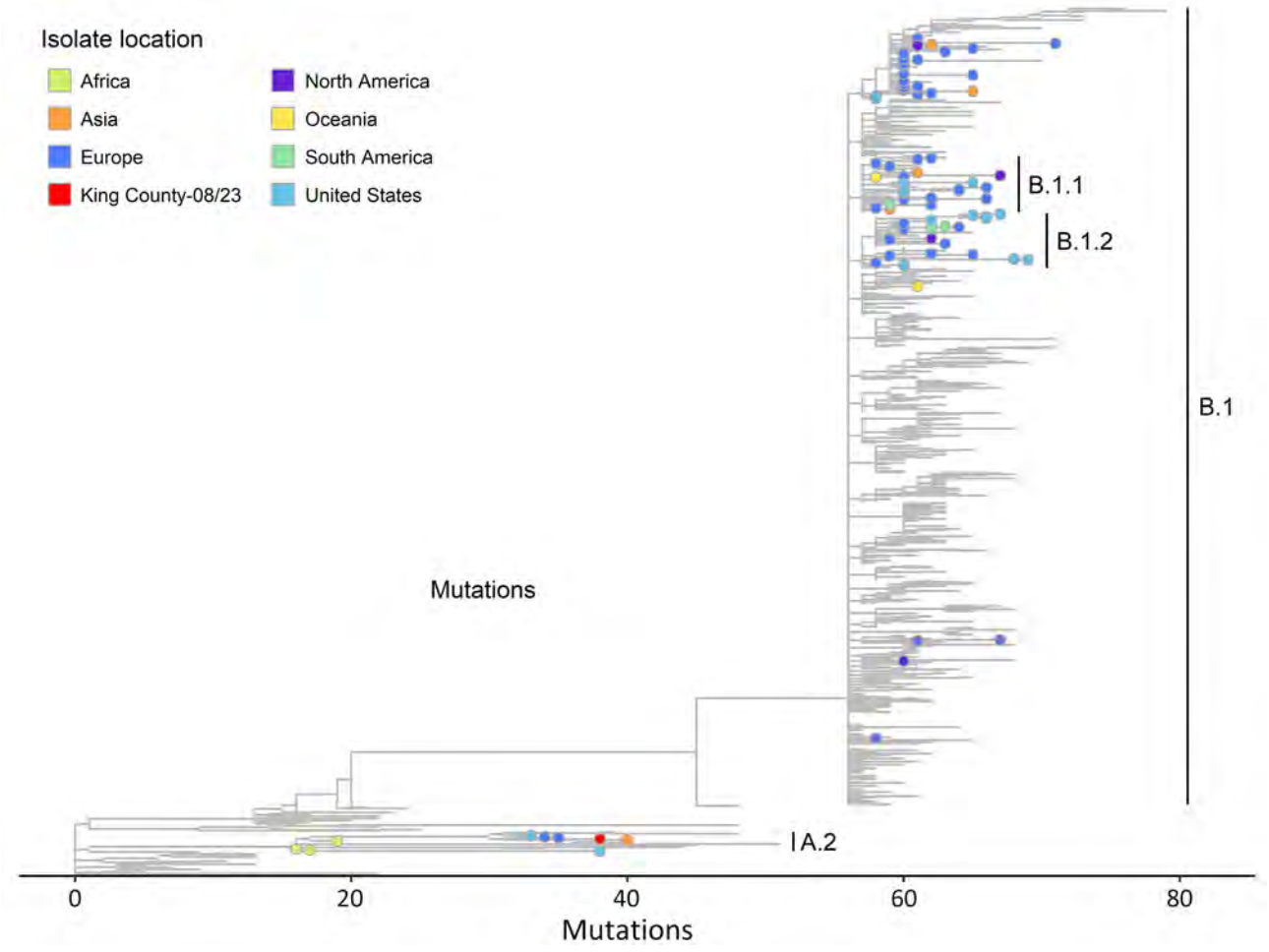


Figure. Phylogenetic reference tree of clade IIb monkeypox virus sequences as of December 2023 from study of real-time use of monkeypox virus genomic surveillance, King County, Washington, USA, 2022–2024. The tree shows the viral strain of sublineage A.2.1 identified from a King County resident in fall 2023 (in red) is highly diverged from the B.1 lineage. Colors correspond to the location of the case from which a viral isolate was sampled and are either a region, the United States, or King County for the specific isolate of interest in this case report. Phylogenetic tree generated using Nextclade dataset for "Mpox virus (All Clades)" in Auspice v2.61.1 (https://github.com/nextstrain/auspice.us). The reference sequence used in the tree is the clade IIb Genbank reference sequence (accession no. NC_063383.1). Data sourced from GenBank on December 8, 2023 (*13*). Sequences were downsampled by the Nextstrain team to ≈500 sequences with the goal of capturing monkeypox virus diversity across geography, collection dates, and lineages. The dataset is archived by Nextclade (https://github.com/nextstrain/Nextclade_data/tree/master/data_output/nextstrain/mpox/all-clades/2024-01-16--20-31-02Z). The figure is filtered to the clade IIb branch within the larger dataset. Branches are shown for all lineages within clade IIb, and specific nodes are shown for selected lineages A.2, A.2.1, B.1.1, B.1.2, and B.1.3. A table of the nodes with metadata is included in the Appendix (https://wwwnc.cdc.gov/EID/article/31/13/24-1242-App1.pdf).

sequence was identified as Clade IIb, sublineage A.2.1. The sequence was highly divergent from lineage B.1, showing ≈94 nt variations (Figure). Although lineage B.1 was identified in most sequences in the 2022–2023 outbreak, including in local transmission in the United States, lineage A.2 had been detected infrequently outside of endemic areas (*1*,*6*). A published analysis of the 3 previous mpox cases in the United States with lineage A.2 concluded each was likely a separate introduction, with suspected exposure during travel to the Middle East or West Africa (*6*). The high divergence from other Washington cases and the rarity of reported A.2 lineages in the United States strongly suggested that this case-patient's exposure occurred during international travel.

As of fall 2023, there were zero cases of mpox reported in Kenya and 16 in the UAE; lineage A.2 specifically was reported among at least nine travelers returning from the UAE in 2022 (*8*–*11*), strongly suggesting exposure in the UAE. During fall 2023–July 2024, 37 cases of mpox were reported in King County, and 76% of the cases had samples that were sequenced (n = 28). All sequences were lineage B.1, suggesting no onward transmission of lineage A.2 locally.

Discussion

This case illustrates the value of genomic surveillance in mpox public health response. No known exposures could be identified during case investigation through patient interview, but sequencing helped public health staff determine that this case was unlikely representative of undetected local spread, which reduced concern regarding additional, unreported cases. Applying WGS in practice required close collaboration across agencies to proactively establish local genomic diversity and conduct active genomic surveillance. The 2022–2023 mpox outbreak disproportionately affected men who have sex with men, and interview participation could have been limited by concerns about stigma (*12*). Genomic surveillance has potential to answer broad questions of concern for public health, like whether local transmission is occurring, without requiring detailed exposure information. As use of pathogen sequencing expands, additional work should be conducted to anticipate potential ethical concerns and establish practices that protect privacy while advancing infectious disease prevention and response.

Acknowledgments

For sharing MPXV sequencing data from around the world through GISAID (https://www.gisaid.org), we gratefully acknowledge all data contributors, including the authors and their originating laboratories responsible for obtaining the specimens, as well as their submitting laboratories, for generating the genetic sequence and metadata (Appendix, https://wwwnc.cdc.gov/EID/article/31/13/24-1242-App1.xlsx). We also gratefully acknowledge Michelle Perry for her collaboration with the case investigation. The sequence for this isolate has been deposited into GenBank (accession no. OR455100.1) (*13*).

This work was supported in part by the Centers for Disease Control and Prevention (grant no. NU50CK000630).

P.R. reports honoraria from the Bill and Melinda Gates Foundation, Association for Molecular Pathology, and Kentucky AIDS Education & Training Center and travel support to attend meetings from the Association of Molecular Pathology, outside the described work. A.L.G. reports contract testing from Abbott, Cepheid, Novavax, Pfizer, Janssen, and Hologic and research support from Gilead, outside the described work. E.J.C. reports honoraria from Providence Regional Medical Center for presentations on COVID-19, travel support from IDSA to attend IDWeek 2022, and travel support from the Northwest Healthcare Response Network to attend a Common Health Coalition workshop, all outside the described work. All other authors report no potential conflicts.

About the Author

Ms. Lau works as an epidemiologist at Public Health–Seattle & King County in the Communicable Disease Epidemiology and Immunizations section and participates in the Northwest Pathogen Genomics Center of Excellence. Her work focuses on implementing pathogen genomics in local public health practice.

References

1. Laurenson-Schafer H, Sklenovská N, Hoxha A, Kerr SM, Ndumbi P, Fitzner J, et al.; WHO mpox Surveillance and Analytics team. Description of the first global outbreak of mpox: an analysis of global surveillance data. Lancet Glob Health. 2023;11:e1012–23. https://doi.org/10.1016/S2214-109X(23)00198-5
2. Roychoudhury P, Sereewit J, Xie H, Nunley E, Bakhash SM, Lieberman NAP, et al. Genomic analysis of early monkeypox virus outbreak strains, Washington, USA. Emerg Infect Dis. 2023;29:644–6. https://doi.org/10.3201/eid2903.221446
3. Genomic epidemiology of monkeypox virus – Washington state focused build. [cited 2024 Nov 19] https://nextstrain.org/groups/waphl/wa/hmpxv1
4. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3. https://doi.org/10.1093/bioinformatics/bty407
5. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6:3773. https://doi.org/10.21105/joss.03773

6. Gigante CM, Korber B, Seabolt MH, Wilkins K, Davidson W, Rao AK, et al. Multiple lineages of monkeypox virus detected in the United States, 2021-2022. Science. 2022;378:560–5. https://doi.org/10.1126/science.add4153
7. Centers for Disease Control and Prevention. Mpox in the U.S. 2024. [cited 2024 Nov 19] https://www.cdc.gov/poxvirus/mpox/symptoms/index.html.
8. Gehre F, Lagu HI, Achol E, Omari N, Ochido G, Shand K, et al. The East African Community mobile laboratory network prepares for monkeypox outbreaks. J Public Health Africa. 2023;14:2309. https://doi.org/10.4081/jphia.2023.2309
9. Hermez J, El Helou R, Sawaya T, Sader G, Jamil MS, Alaama AS, et al. Emergence of mpox in the Eastern Mediterranean Region: Data assessment and implications for a public health response. J Infect Public Health. 2024;17:102565. https://doi.org/10.1016/j.jiph.2024.102565
10. Dung NT, Hung LM, Hoa HTT, Nga LH, Hong NTT, Thuong TC, et al. Monkeypox virus infection in 2 female travelers returning to Vietnam from Dubai, United Arab Emirates, 2022. Emerg Infect Dis. 2023;29:778–81. https://doi.org/10.3201/eid2904.221835
11. Shete AM, Yadav PD, Kumar A, Patil S, Patil DY, Joshi Y, et al. Genome characterization of monkeypox cases detected in India: Identification of three sub clusters among A.2 lineage. J Infect. 2023;86:66–117. https://doi.org/10.1016/j.jinf.2022.09.024
12. El Dine FB, Gebreal A, Samhouri D, Estifanos H, Kourampi I, Abdelrhem H, et al. Ethical considerations during Mpox Outbreak: a scoping review. BMC Med Ethics. 2024;25:79. https://doi.org/10.1186/s12910-024-01078-0
13. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, et al. GenBank 2024 Update. Nucleic Acids Res. 2024;52(D1):D134–7. https://doi.org/10.1093/nar/gkad903

Address for correspondence: Kathryn M. Lau, Communicable Disease Epidemiology and Immunizations, 401 5th Ave, Ste 1250, Seattle, WA 98104, USA; email: katlau@kingcounty.gov

Nationwide Implementation of HIV Molecular Cluster Detection by Centers for Disease Control and Prevention and State and Local Health Departments, United States

Anne Marie France,[1] Camden J. Hallmark,[1] Nivedha Panneer, Rachael Billock, Olivia O. Russell, Mary Plaster, Jessica Alberti, Fathima Nuthan, Neeraja Saduvala, David Philpott, M. Cheryl Bañez Ocfemia, Scott Cope, Angela L. Hernandez, Sergei L. Kosakovsky Pond, Joel O. Wertheim, Steven Weaver, Saja Khader, Kevin Johnson, Alexandra M. Oster

Detecting and responding to clusters of rapid HIV transmission is a core HIV prevention strategy in the United States, guiding public health interventions and identifying gaps in prevention and care services. In 2016, the Centers for Disease Control and Prevention (CDC) initiated molecular cluster detection using data from 27 jurisdictions. During 2016–2023, CDC expanded sequence reporting nationwide and deployed Secure HIV-TRACE, an application supporting health department (HD) molecular cluster detection. CDC conducts molecular cluster detection quarterly; state and local HDs analyze local data monthly. HDs began routinely reporting clusters to CDC by using cluster report forms in 2020. During 2018–2023, CDC identified 404 molecular clusters of rapid HIV transmission; 325 (80%) involved multiple jurisdictions. During 2020–2023, HDs reported 298 molecular clusters to CDC; 249 were first detected by HDs. Expanding molecular cluster detection has provided a foundation for improving service delivery to networks experiencing rapid HIV transmission.

HIV clusters or outbreaks are defined as rapid HIV transmission among persons in a sex or drug-using network (*1*); network refers to persons in an HIV cluster and those with whom they have sex or share drugs, who might or might not have HIV. Identifying rapid HIV transmission through cluster detection guides public health efforts designed to identify and address gaps in care and prevention services that are not effectively reaching HIV transmission networks (*2*). Before 2016, HIV clusters in the United States were detected sporadically, typically by astute medical providers or partner services and frontline staff (*2*).

In 2016, after responding to a large outbreak of HIV among persons who inject drugs in Scott County, Indiana, USA (*3*), the Centers for Disease Control and Prevention (CDC) initiated proactive cluster detection through routine analysis of CDC's National HIV Surveillance System (NHSS) data. HIV is a nationally notifiable disease condition, and state, tribal, local, and territorial (STLT) health departments (HDs) collect demographic, transmission risk, and clinical information and report deidentified data to CDC. NHSS data are routinely used at federal and STLT levels to monitor HIV distribution and transmission, plan and evaluate prevention and care programs, allocate resources, inform policy development, and identify and respond to rapid transmission across the United States (*4*). Laboratory reporting, including HIV

Author affiliations: US Public Health Service Commissioned Corps, Atlanta, Georgia, USA (A.M. France, A.M. Oster); Centers for Disease Control and Prevention, Atlanta (A.M. France, C.J. Hallmark, N. Panneer, R. Billock, O.O. Russell, D. Philpott, M.C.B. Ocfemia, S. Cope, A.L. Hernandez, A.M. Oster); DLH Corporation, Atlanta (M. Plaster, J. Alberti, F. Nuthan, S. Khader); SeKON Enterprise Inc., Atlanta (N. Saduvala); Temple University, Philadelphia, Pennsylvania, USA (S.L. Kosakovsky Pond, S. Weaver); University of California San Diego, La Jolla, California, USA (J.O. Wertheim, S. Weaver); Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA (K. Johnson)

DOI: https://doi.org/10.3201/eid3113.241143

[1]These first authors contributed equally to this article.

molecular viral sequence data generated through routine clinical HIV drug-resistance testing, is an integrated component of NHSS.

Surveillance-based cluster detection methods include approaches to identify geospatial increases in HIV diagnoses (*5*), referred to as time-space analysis, as well as molecular cluster detection, in which clusters are recognized through analysis of HIV sequence data. HIV mutates rapidly, and molecular cluster detection methods assess differences between virus nucleotide sequences (genetic distance) to distinguish infections that are more closely related in transmission networks from those more distantly related. Time-space analysis and molecular cluster detection, in addition to cluster detection by providers or partner services and frontline staff, are complementary (*1*).

Many approaches are available for HIV sequence analysis, but not all approaches focus on clusters with the highest transmission rates, which contribute disproportionately to current and future transmission (*6*,*7*). CDC developed an approach focusing on clusters representing rapid transmission (*8*), which yields clusters having transmission rates >8 times the overall national transmission estimate among all persons living with HIV (*8*) and some clusters with rates >33 times the national rate (*9*).

Molecular cluster detection at both local and national levels is critical. Local HIV surveillance data are available during collection, but analyses by HDs are limited to persons diagnosed or living within the HD's administrative boundary who have had data reported to the local surveillance system. However, populations are mobile (*10*), and HIV transmission can be geographically dispersed (*11*). National HIV surveillance data are not as timely but can be analyzed by CDC to identify clusters that cross jurisdictional boundaries (*8*,*12*).

CDC initiated routine analysis of HIV surveillance data in 2016 to identify clusters at the national level, including clusters spanning multiple jurisdictions. Those analyses included sequence data reported to CDC by 27 STLT HDs to assess drug resistance and general transmission patterns (*13*). The HIV TRAnsmission Cluster Engine (HIV-TRACE) (*14*), a tool developed to assess global HIV transmission patterns (*15*), was adapted to characterize transmission patterns in a local installation within CDC that adhered to stringent HIV data protections (*13*). However, secure local installation and implementation of HIV-TRACE was technically complex and not feasible for most STLT HDs. In 2018, cluster detection and response (CDR) was expanded nationwide (*16*).

To promote analysis of molecular HIV data to elucidate transmission patterns at STLT levels, CDC initiated the development of Secure HIV-TRACE in 2015 through funding from CDC's Advanced Molecular Detection program, part of the National Center for Emerging and Zoonotic Infectious Diseases. It was essential to ensure that this web-based tool met stringent HIV data protections. This secure, web-based bioinformatics application incorporates sequence analysis methods similar to HIV-TRACE. Secure HIV-TRACE was first released for HD use in 2017 for characterizing general transmission patterns. After CDR expansion nationwide in 2018 (*16*), Secure HIV-TRACE was refined to focus on detecting and monitoring clusters of rapid HIV transmission. Secure HIV-TRACE is computationally efficient, scales to accommodate large datasets and was designed for use by public health staff who might lack bioinformatics expertise.

With expanded cluster detection capabilities both at national and STLT levels and the recognition of the importance of cluster response to reduce HIV incidence, CDR was included as 1 of 4 pillars of the federal Ending the HIV Epidemic Initiative launched in 2019 (*17*). We describe the implementation of molecular HIV cluster detection at both the national and STLT levels and assess the contributions of each to overall cluster detection in the United States.

Materials and Methods

Data Collection

HDs report HIV surveillance data to CDC according to STLT laws and regulations. HIV surveillance programs collect demographic, clinical, laboratory, vital statistics, and behavioral data. NHSS data collection includes HIV laboratory results indicative of HIV infection, such as CD4+ T lymphocyte numbers, viral load test results, and HIV sequence data. HIV sequences are collected from genotypic resistance tests performed as part of routine clinical care at commercial, private, and public health laboratories and reported electronically to STLT HDs (*4*).

During 1997–2012, sequences were collected through supplemental surveillance projects focused on drug resistance and virus diversity, expanding from 4 jurisdictions in 1997 to 17 in 2012 (*18*). During 2013–2017, reporting of HIV sequence data expanded to 27 jurisdictions with the additional aim to assess transmission patterns and, in 2018, expanded to 59 HD HIV surveillance programs with the charge to use those data for HIV CDR (*16*).

National Cluster Detection

A CDC-provided software application (Enhanced HIV/AIDS Reporting System [eHARS]) is installed at

HDs for entry, storage, management, and reporting of HIV data conducted by surveillance programs; the application is secured behind each HD's firewall and accessible by designated HD staff only. HDs transfer monthly deidentified HIV surveillance data to CDC. Each quarter, CDC produces national-level datasets that deduplicate and reconcile data for persons reported by all jurisdictions into a single, unified dataset for reporting, analyses, and evaluation (*4*). Data are protected by a CDC Assurance of Confidentiality (*1*), and release is governed according to the data rerelease agreement with each HD. For annual HIV surveillance reports, data are considered preliminary until a 12-month reporting delay has elapsed. After 12 months, data are considered provisional and are subject to change as additional data are reported (*4*).

National cluster detection is conducted after each quarterly preliminary national-level dataset is created by using all available HIV sequences of suitable quality, which can include multiple sequences per person. Cluster detection is conducted using preliminary datasets to expedite timely detection of clusters. At each quarterly interval, clusters of rapid transmission among persons with HIV diagnosed in the previous 3 years are identified by using a secure local installation of HIV-TRACE (*14*); this transmission network analysis includes a 1,497-nt segment of the HIV protease and reverse transcriptase genomic region. Sequences are aligned with the HIV-1 *pol* gene in reference strain HXB2, pairwise genetic distances are calculated, and clusters are defined when the pairwise genetic distance is ≤0.005 substitutions/site (i.e., meeting a 0.5% genetic distance threshold) (*8*). HIV clusters meeting CDC's national priority definition are then identified.

CDC defines national priority clusters to focus on clusters having evidence of rapid transmission by using the 0.5% genetic distance threshold and data for persons with an HIV diagnosis within the previous 3 years. Such clusters with ≥3 HIV diagnoses in the previous 12 months have high transmission rates (*8*). If that definition was used nationwide, many jurisdictions would have more priority clusters than they could feasibly conduct response activities for; thus, that definition is applied to low-burden jurisdictions (those with <200 diagnoses annually), whereas a higher threshold of ≥5 diagnoses is used for most jurisdictions nationally. All clusters meeting those definitions are considered national priority clusters.

Each quarter, new national priority clusters are identified and growth in previously identified clusters is monitored. Clusters that continue to meet or newly meet national priority criteria are flagged for review. Because of disruptions related to the COVID-19 pandemic, national cluster detection was not conducted for the March 2020 quarter.

Many national priority clusters cross jurisdictional boundaries. The primary jurisdiction for each cluster is defined as the jurisdiction where >50% of cluster members resided at the time of HIV diagnosis. In 2022, CDC began systematically identifying jurisdictions with substantial involvement in national priority clusters; substantial involvement is defined as ≥3 diagnoses (resident at diagnosis or current resident) in the previous 12 months. Multiple jurisdictions can be substantially involved in a cluster at the same time.

Each quarter, summary reports of new and previously identified national priority clusters are generated for jurisdictions. Those reports include current case counts, jurisdictions involved, and current priority status. In addition, line lists are generated for newly detected clusters and clusters that continue to meet priority criteria and have grown. Those data are securely transmitted to HDs. HDs are asked to review the information in conjunction with results of their local molecular cluster analysis. CDC epidemiologists meet with HD personnel, as needed, to discuss findings and offer technical assistance to respond to clusters.

State and Local Cluster Detection

Beginning in 2018, CDC expanded support to all HDs to collect HIV sequence data and conduct molecular cluster detection monthly (*16*). Methods for state and local molecular cluster detection parallel those used for national cluster detection; however, several critical differences exist (Table 1). To provide an accessible tool for HDs to conduct molecular analysis, CDC contracted with the University of California San Diego (La Jolla, CA, USA) and Temple University (Philadelphia, PA, USA) to develop Secure HIV-TRACE. Although the original intent was to characterize transmission patterns, the development and intended use evolved to focus on detecting and monitoring HIV molecular clusters. Funding for Secure HIV-TRACE development was obtained through CDC's Advanced Molecular Detection program (fiscal years 2015–2017) and the Division of HIV Prevention, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, since that time.

To conduct molecular cluster detection, HDs export data from their local eHARS, process the data by using a CDC-supplied SAS software program (SAS Institute Inc., https://www.sas.com), and securely upload data without personal identifiers to Secure HIV-TRACE. Similar to the national analysis, Secure HIV-TRACE aligns the sequences to the HIV-1 *pol* gene from reference strain HXB2, computes pairwise genetic distances, and defines clusters at the 0.5% and

Table 1. Comparison of HIV cluster detection by CDC and state and local health departments, United States*

Comparisons	CDC	State/local
Interval	Quarterly	At least monthly
Data	National, deduplicated	State or local
Ability to detect multijurisdictional clusters	Yes	Not at the time of writing
Analytic tool	Local installation of HIV-TRACE	Secure HIV-TRACE
Identifiable information	No personal identifiers	Linked to personal identifiers†
Initial notification to health departments	CDC securely transmits notification of priority clusters to state/local health departments	Secure HIV-TRACE automatically flags priority clusters identified in each analysis
Reporting from health departments to CDC	Response activities reported to CDC via cluster report forms	Response activities reported to CDC via cluster report forms

*CDC, Centers for Disease Control and Prevention; HIV-TRACE, HIV TRAnsmission Cluster Engine.
†Identifiers are not uploaded to Secure HIV-TRACE or transmitted to CDC.

1.5% genetic distance thresholds. Secure HIV-TRACE users can visualize data and review summary statistics, epidemiologic curves, and line-listed information that can be exported for further analysis.

When first released, Secure HIV-TRACE focused broadly on transmission networks, using a 1.5% genetic distance threshold to define clusters. Ten additional releases since 2017 have expanded application functionality; cluster detection has been aligned with CDC's focus on clusters of rapid transmission by using the national priority cluster definition with a 0.5% genetic distance threshold. In May 2023, Secure HIV-TRACE was enhanced to automatically identify and monitor growth in clusters meeting CDC's national priority criteria. Before that enhancement, a CDC-supplied SAS program was used to identify national priority clusters according to Secure HIV-TRACE output. Users can now also elect to monitor other clusters of interest on the basis of locally defined priority criteria. Additional enhancements have included a data quality screen, HIV drug resistance and subtype analysis, improved visualization, and optional data visualizations using external platforms such as Power BI (Microsoft, https://www.microsoft.com) (Figure 1).

Figure 1. Timeline of nationwide implementation of HIV cluster detection and response by CDC and state and local HDs in the United States. CDC, Centers for Disease Control and Prevention; HD, health department; Secure HIV-TRACE, Secure HIV TRAnsmission Cluster Engine.

Secure HIV-TRACE hosting was also modernized by a move to the CDC cloud platform with access through CDC's Secure Access Management Services.

Cluster Report Forms

Since 2020, CDC-funded STLT HDs have been required to submit HIV cluster report forms each quarter to CDC to report clusters of public health concern, including molecular clusters meeting CDC's national priority cluster criteria and other clusters of concern detected through both molecular and other approaches. Cluster report forms promote communication between HDs and CDC and have information about methods of cluster detection, key cluster attributes, and cluster response activities (*19*). Using a secure REDCap (https://www.project-redcap.org) platform, HDs submit initial cluster report forms, which have information on how the cluster was first detected, cluster size at detection, and other characteristics. Follow-up forms are submitted during the response to report ongoing response activities and findings, and annual/closeout forms are submitted at the end of the response (or annually for ongoing responses) to provide additional activity and outcome information.

When cluster report forms are submitted, jurisdictions are also asked to enter cluster-related variables in eHARS for persons with HIV who are known to be part of the cluster. Cluster report forms and cluster-related variables do not include personal identifiers. The variables include cluster identification numbers assigned locally and national cluster identification numbers for those clusters also identified through national cluster detection. That information enables HDs and CDC to elucidate characteristics and care status of persons with HIV who are part of reported clusters and help guide response activities.

Analysis

We assessed the number of diagnoses for which HIV sequence data were reported and described characteristics of clusters detected through national molecular analysis during 2018–2023. We defined sequence completeness as the percentage of persons with a reported HIV diagnosis for which a sequence of ≥100 bp was reported. To characterize clusters identified through state/local cluster detection, we analyzed data reported on cluster report forms from 2020–2023.

Results

Sequence Reporting

Using data reported to CDC's NHSS until December 2023, we determined sequences were available for 52% of HIV diagnoses during 2021–2023, an increase

Table 2. Number of clusters detected each year through nationwide analysis of National HIV Surveillance System data by Centers for Disease Control and Prevention and state and local health departments, United States, 2018–2023*

Characteristics	2018	2019	2020†	2021	2022	2023	Total no.
National priority clusters detected‡	69	74	48	61	77	75	404
Clusters meeting national priority criteria as of December 2023‡	2 (3)	8 (11)	5 (10)	9 (15)	10 (13)	40 (53)	74 (18)
Median cluster size when first detected (range)	7 (3–24)	8 (3–19)	8 (3–16)	7 (4–18)	7 (3–14)	7 (3–23)	7 (3–24)§
Median cluster size as of December 2023 (range)	17 (3–193)	17.5 (3–90)	20 (3–49)	16 (6–72)	11 (3–47)	9 (3–26)	13 (3–193)§
Clusters involving >1 jurisdiction as of December 2023	58 (84)	62 (84)	43 (90)	51 (84)	56 (73)	55 (73)	325 (80)
Clusters with no primary jurisdiction at detection	6 (9)	9 (12)	4 (8)	5 (8)	5 (6)	6 (8)	35 (9)
Jurisdictions with ≥1 cluster as the primary jurisdiction at detection¶	25	25	25	20	26	28	43#
Jurisdictions with ≥1 cluster with substantial involvement**	NA	NA	NA	NA	27	31	NA
Clusters with no jurisdiction ever substantially involved	NA	NA	NA	NA	2 (3)	5 (7)	NA
Clusters with only 1 jurisdiction ever substantially involved	NA	NA	NA	NA	68 (88)	67(89)	NA
Clusters with ≥2 jurisdictions ever substantially involved	NA	NA	NA	NA	7 (9)	3 (4)	NA

*Values are no. (%) except as indicated. NA, not applicable.
†National cluster detection was not conducted for the March 2020 quarter because of disruptions related to the COVID-19 pandemic.
‡National priority clusters are defined as clusters at the 0.5% genetic distance threshold with ≥5 diagnoses in the previous 12 months or ≥3 diagnoses in the previous 12 months in low-burden jurisdictions (those having <200 reported HIV diagnoses annually). National priority criteria is assessed on an ongoing basis.
§Overall median and range.
¶Primary jurisdiction is defined as the jurisdiction in which >50% of cluster members resided at the time of HIV diagnosis. If there is no single jurisdiction in which >50% of cluster members reside at diagnosis, no primary jurisdiction is assigned.
#Total number of unique jurisdictions.
**Substantial involvement in national priority clusters is defined as ≥3 diagnoses in a given jurisdiction (resident at diagnosis or current resident) in the previous 12 months. Multiple jurisdictions can be substantially involved in a cluster at the same time.

from 41% sequence completeness observed for the 3-year period of 2016–2018 (as reported by December 2018). During 2023, the median timeframe from collection of the sample to receipt of the sequence at the HD was 34 (interquartile range 20–50) days.

National Molecular Analysis

During 2018–2023, CDC detected 404 national priority clusters (Table 2). An average of 67 (range 48–77) new clusters were detected each year; the fewest were detected during the COVID-19 pandemic peak in 2020, when only 3 quarterly analyses were conducted. The median number of persons in each cluster at the time of detection was 7 (range 3–24). Clusters had grown to a median size of 13 (range 3–193) by December 2023; clusters detected in earlier years (and therefore with more time for potential growth) had higher median sizes as of December 2023. Of the 404 clusters, 74 (18%) continued to meet national priority criteria in December 2023, indicating that the cluster had ≥5 diagnoses during 2023 or ≥3 diagnoses in low-burden jurisdictions. Of 59 CDC-funded HDs, 43 (73%) were the primary jurisdiction for ≥1 cluster.

Most clusters spanned multiple jurisdictions; 325 (80%) clusters involved >1 jurisdiction, including clusters with ≥1 diagnosis outside the primary jurisdiction and those with no primary jurisdiction. Among 152 clusters identified since substantial involvement was first systematically assessed beginning in 2022, 7 (5%) had no jurisdiction, 135 (89%) had only 1 jurisdiction, and 10 (7%) had ≥2 jurisdictions susbtantially involved.

State and Local Cluster Detection

Secure HIV-TRACE was first released in August 2017 as an optional tool for jurisdictions collecting sequence data at that time. In January 2018, the application became available to all jurisdictions as HIV CDR expanded. Use of Secure HIV-TRACE expanded over time. In January 2018, 30 HDs had been using Secure HIV-TRACE. By April 2024, 148 HD users representing 54 HDs had been using Secure HIV-TRACE and had uploaded 792,360 sequences to the application for cluster detection analysis.

During 2020–2023, a total of 403 (90–116/year) clusters newly detected through any method were reported by HDs to CDC via cluster report forms (Figure 2). Of those, 298 (74%) were first detected through national or state/local molecular analysis; 248 (83%) of 298 were first detected through state/local analysis. Clusters were most often (48%) reported by jurisdictions in the South of the United States (Figure 3). The median size of molecular clusters first detected through state/local analysis was 7 (range 2–85) at detection, and 224 (90%) of 248 met national priority criteria when first reported. Of 59 CDC-funded jurisdictions, 37 (62%) reported ≥1 cluster first detected through state/local molecular analysis.

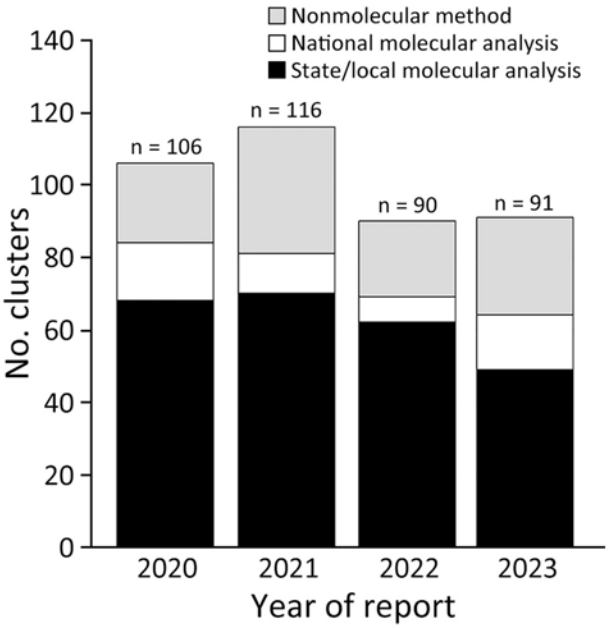


Figure 2. Clusters of HIV newly reported to the Centers for Disease Control and Prevention by state and local health departments, United States, during 2020–2023. Clusters were reported to Centers for Disease Control and Prevention through cluster report forms. Methods by which clusters were first detected are indicated; nonmolecular cluster detection methods include time-space cluster detection, partner services, and provider notification. Numbers on top of bars indicate exact number of HIV clusters reported each year.

Discussion

CDC and state and local HDs have successfully implemented routine detection and monitoring of clusters of rapid HIV transmission through analysis of molecular sequence data. Clusters of rapid HIV transmission are now frequently detected in the United States; both national and state/local molecular cluster analyses are essential for HIV cluster detection. Most jurisdictions have had ≥1 national priority cluster identified through CDC analysis, and most jurisdictions have reported a molecular cluster to CDC that was first identified through state/local molecular analysis. The higher frequency of reported clusters in the South is consistent with the greater burden of new diagnoses in that region (*20*). Multijurisdictional clusters are commonly identified, and clusters often exhibit ongoing growth many years after detection.

National and state/local cluster detection are complementary, both contributing to comprehensive and timely cluster identification. Because NHSS

datasets are only available quarterly, monthly state/local cluster detection is essential for more timely cluster detection. Most molecular clusters reported to CDC by HDs have first been detected through state/local analysis, promoting a timely response. In addition, state and local analyses enable flexibility to detect clusters not yet meeting priority criteria but still of concern because of local epidemiology or other factors. Secure HIV-TRACE is the primary tool for state and local cluster detection. Although the tool used for cluster detection is not explicitly reported by HDs on cluster report forms, nearly all reported clusters detected through state/local molecular analysis have likely been detected by using Secure HIV-TRACE. That secure, accessible tool has been essential for implementing cluster detection, and tool enhancements have improved identification and monitoring of clusters meeting national or local priority.

State and local HIV surveillance systems in the United States are decentralized, and state/local cluster detection analyses by HDs are limited to data from their own jurisdictions. Therefore, national-level cluster detection conducted by CDC is essential for detection of multijurisdictional clusters; 20% of nationally identified clusters are missed by local cluster detection (*21*). HIV is a chronic infection, and persons living with HIV in the United States are more mobile than the general US population (*10*). The frequent identification of multijurisdictional clusters is consistent with findings that transmission networks are often geographically dispersed (*11*). Multijurisdictional involvement can range from a single diagnosis in an otherwise geographically focused cluster to clusters with no geographic focus. The substantial involvement definition captures evidence of rapid transmission occurring in multiple jurisdictions, suggesting the need for meaningful response engagement and coordination from the jurisdictions involved.

Traditional approaches to identifying rapid transmission, including time-space analysis, rely on observing increases in diagnoses in a population or area. However, those approaches are limited by factors such as a median HIV diagnosis delay of 3 years in the United States (*22*) and difficulty detecting increases in diagnoses in areas with higher baseline HIV incidence. In addition, populations are mobile (*10*), and detecting geographically dispersed rapid transmission is difficult using traditional approaches (*11*). Molecular cluster detection can detect rapid transmission that is geographically dispersed or in areas with a high baseline HIV incidence. Robust HIV surveillance systems are critical for effective cluster detection.

For cluster detection to have the intended effect, response is essential. Rapid transmission occurs

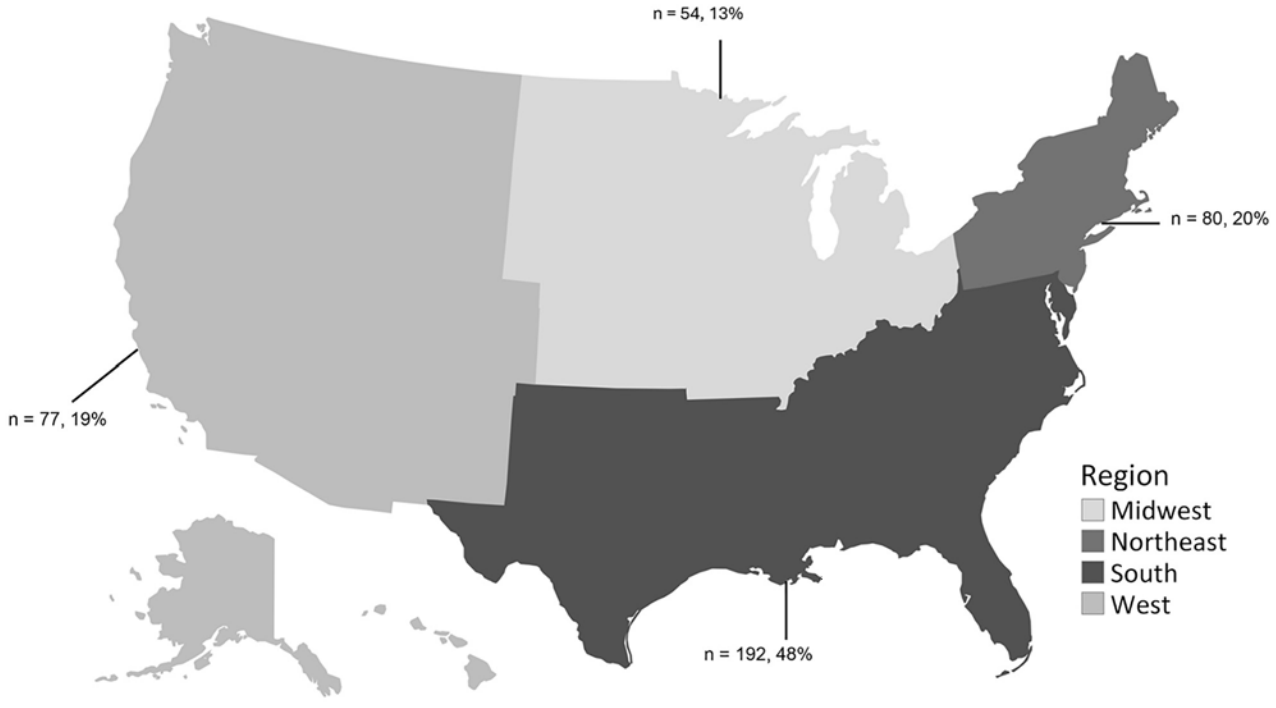


Figure 3. Region of reporting health department for clusters of HIV newly reported to the Centers for Disease Control and Prevention by state and local health departments, United States, during 2020–2023. Clusters were reported to Centers for Disease Control and Prevention through cluster report forms. Numbers and percentages of clusters are indicated for each region.

because affected communities are not adequately reached by existing services (*1*). Clusters affect various populations and can grow rapidly (*23*,*24*); most clusters are associated with sexual transmission (*24*). Responses to clusters can vary depending on local priorities, resources, and needs identified through cluster investigation and might include individual, network, and structure-level interventions to ensure that testing, care, and prevention programs are reaching persons and places that can most benefit. Response activities and outcomes have been previously described (*2*,*19*).

Evidence that HIV CDR strengthens HIV prevention and care services has been observed in a growing body of field investigations, as well as analytic and modeling studies (*2*,*6*,*7*,*25*). Priority clusters correspond to underlying networks that are 3–9 times the size of the detected cluster; those networks have high transmission rates and disproportionate numbers of persons with undiagnosed HIV infections, indicating opportunities for public health intervention (*25*). Clusters can grow rapidly (*23*) and contribute disproportionately to future infections (*6*,*7*).

Community engagement is essential to implement HIV CDR. Although analysis of molecular data provides a unique and powerful tool to identify communities affected by rapid HIV transmission, some advocates and community members have raised concerns about collection and analysis of molecular sequence data, including concerns about the potential use of molecular HIV data in criminal transmission cases (*26*,*27*). CDC has held numerous discussions with community members, community-based organizations, advocates, and other key partners to inform responsible establishment of CDR activities and guide the use of sequence data (*27*,*28*). CDC has strong security measures to ensure privacy and confidentiality of persons with HIV, requires health departments to comply with data security and confidentiality guidelines, and has further strengthened guidance for protection of sequence data (*29*,*30*). Secure HIV-TRACE is not an open database. The deidentified individual-level information submitted to NHSS is not publicly available. Given the potential harms of disclosure, CDC's data use agreement prohibits release without individual-level consent. CDC also requires HDs to communicate and collaborate with community members and partners for input on CDR activities and to design responses to specific clusters and outbreaks, including, where needed, enhancing or improving processes or procedures for protecting privacy and confidentiality (*1*,*30*,*31*). That engagement helps HDs address community concerns, provides a foundation for trust and collaboration, and supports collaboration with community partners for cluster response.

We are in an era where we have the tools to successfully treat and prevent HIV. However, those tools often do not reach the persons who need them most. Molecular cluster detection enables us to identify rapid HIV transmission that would have previously been unrecognized. That detection creates new opportunities to identify and close gaps in HIV prevention and care services and advances efforts to end the HIV epidemic in the United States.

About the Author

Dr. France leads the Cluster Detection and Molecular Epidemiology Team in the Detection and Response Branch, Division of HIV Prevention, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, at CDC. Her work focuses on developing, implementing and assessing approaches to identify and monitor clusters and outbreaks of rapid transmission of HIV.

References

1. Centers for Disease Control and Prevention. Cluster detection and response guidance for health departments. 2024 [cited 2025 Apr 30]. https://www.cdc.gov/hivpartners/php/cdr/health-department-guidance.html
2. Oster AM, Lyss SB, McClung RP, Watson M, Panneer N, Hernandez AL, et al. HIV cluster and outbreak detection and response: the science and experience. Am J Prev Med. 2021;61:S130–42. https://doi.org/10.1016/j.amepre.2021.05.029
3. Peters PJ, Pontones P, Hoover KW, Patel MR, Galang RR, Shields J, et al.; Indiana HIV Outbreak Investigation Team. HIV infection linked to injection use of oxymorphone in Indiana, 2014–2015. N Engl J Med. 2016;375:229–39. https://doi.org/10.1056/NEJMoa1515195
4. Satcher Johnson A, Peruski A, Oster AM, Balaji A, Siddiqi AE, Sweeney P, et al. Enhancements to the National HIV Surveillance System, United States, 2013–2023. Public Health Rep. 2024;139:654–61. https://doi.org/10.1177/00333549241253092
5. Fitzmaurice AG, Linley L, Zhang C, Watson M, France AM, Oster AM. Novel method for rapid detection of spatiotemporal HIV clusters potentially warranting intervention. Emerg Infect Dis. 2019;25:988–91. https://doi.org/10.3201/eid2505.180776
6. Panneer N, Schlanger K, Billock RM, Farnham PG, Oster AM, Islam H, et al. Clusters contribute disproportionately to future HIV transmission in the United States. In: Abstracts of the 25th International AIDS Conference; Munich, Germany; 2024 Jul 22–26. Abstract WEPEC191. Geneva: International AIDS Society; 2024.
7. Billock RM, France AM, Saduvala N, Panneer N, Hallmark CJ, Wertheim JO, et al. Contribution of HIV transmission bursts to future HIV infections, United States. AIDS. 2024 Dec 24 [online ahead of print]. https://doi.org/10.1097/QAD.0000000000004101
8. Oster AM, France AM, Panneer N, Bañez Ocfemia MC, Campbell E, Dasgupta S, et al. Identifying clusters of recent and rapid HIV transmission through analysis of molecular

surveillance data. J Acquir Immune Defic Syndr. 2018;79:543–50. https://doi.org/10.1097/QAI.0000000000001856
9. McClung RP, Atkins AD, Kilkenny M, Bernstein KT, Willenburg KS, Weimer M, et al.; 2019 Cabell County HIV Outbreak Response Team. Response to a large HIV outbreak, Cabell County, West Virginia, 2018–2019. Am J Prev Med. 2021;61:S143–50. https://doi.org/10.1016/j.amepre.2021.05.039
10. Okello A, Song R, Hall HI, Dailey A, Johnson AS. Interstate mobility of people with diagnosed HIV in the United States, 2011–2019. Public Health Rep. 2024;139:467–75. https://doi.org/10.1177/00333549231208488
11. Board AR, Oster AM, Song R, Gant Z, Linley L, Watson M, et al. Geographic distribution of HIV transmission networks in the United States. J Acquir Immune Defic Syndr. 2020;85:e32–40. https://doi.org/10.1097/QAI.0000000000002448
12. Oster AM, France AM, Mermin J. Molecular epidemiology and the transformation of HIV prevention. JAMA. 2018;319:1657–8. https://doi.org/10.1001/jama.2018.1513
13. Oster AM, Wertheim JO, Hernandez AL, Bañez Ocfemia MC, Saduvala N, Hall HI. Using molecular HIV surveillance data to understand transmission between subpopulations in the United States. J Acquir Immune Defic Syndr. 2015;70:444–51. https://doi.org/10.1097/QAI.0000000000000809
14. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (TRAnsmission cluster engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Mol Biol Evol. 2018;35:1812–9. https://doi.org/10.1093/molbev/msy016
15. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The global transmission network of HIV-1. J Infect Dis. 2014;209:304–13. https://doi.org/10.1093/infdis/jit524
16. Centers for Disease Control and Prevention. Funding opportunity announcement (FOA) PS18–1802: integrated human immunodeficiency virus (HIV) surveillance and prevention programs for health departments. 2019 [cited 2025 Apr 21]. https://web.archive.org/web/20241129190907/https://www.cdc.gov/hiv/funding/announcements/ps18-1802/index.html
17. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic: a plan for the United States. JAMA. 2019;321:844–5. https://doi.org/10.1001/jama.2019.1343
18. Wheeler WH, Ziebell RA, Zabina H, Pieniazek D, Prejean J, Bodnar UR, et al.; Variant, Atypical, and Resistant HIV Surveillance Group. Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, U.S.—2006. AIDS. 2010;24:1203–12. https://doi.org/10.1097/QAD.0b013e3283388742
19. Philpot DC, Curran KG, Russell OO, McClung RP, Hallmark CJ, Pieczykolan LL, et al. HIV clusters reported by state and local health departments in the United States, 2020–2022. J Acquir Immune Defic Syndr. 2025 Mar 6 [online ahead of print]. PubMed https://doi.org/10.1097/QAI.0000000000003658
20. Centers for Disease Control and Prevention. Issue brief: HIV in the southern United States. 2019 [cited 2025 Apr 21]. https://stacks.cdc.gov/view/cdc/92279
21. Johnson K, Billock R, Hallmark C, France AM, Plaster M, Panneer N. Added value of national-level HIV molecular cluster detection compared to local cluster detection. In: Abstracts of the 2024 CSTE Annual Conference; Pittsburgh, PA, USA; 2024 Jun 9–13. Abstract 20375. Atlanta: Council of State and Territorial Epidemiologists; 2024.
22. Dailey AF, Hoots BE, Hall HI, Song R, Hayes D, Fulton P Jr, et al. Vital signs: human immunodeficiency virus testing and diagnosis delays—United States. MMWR Morb Mortal Wkly Rep. 2017;66:1300–6. https://doi.org/10.15585/mmwr.mm6647e1
23. Perez SM, Panneer N, France AM, Carnes N, Curran KG, Denson DJ, et al. Clusters of rapid HIV transmission among gay, bisexual, and other men who have sex with men—United States, 2018–2021. MMWR Morb Mortal Wkly Rep. 2022;71:1201–6. https://doi.org/10.15585/mmwr.mm7138a1
24. Oster AM, Panneer N, Lyss SB, McClung RP, Watson M, Saduvala N, et al. Increasing capacity to detect clusters of rapid HIV transmission in varied populations—United States. Viruses. 2021;13:577. https://doi.org/10.3390/v13040577
25. France AM, Panneer N, Farnham PG, Oster AM, Viguerie A, Gopalappa C. Simulation of full HIV cluster networks in a nationally representative model indicates intervention opportunities. J Acquir Immune Defic Syndr. 2024;95:355–61. PubMed https://doi.org/10.1097/QAI.0000000000003367
26. Watson M, Sweeney P. Furthering discussion of ethical implementation of HIV cluster detection and response. Am J Bioeth. 2020;20:24–6. https://doi.org/10.1080/15265161.2020.1806398
27. Centers for Disease Control and Prevention. Meeting summary: responsible use of HIV cluster data for public health action: amplifying benefits, minimizing harms. 2020 [cited 2025 Apr 21]. https://web.archive.org/web/20241127181524/https://www.cdc.gov/hiv/programresources/guidance/cluster-outbreak/responsible-use.html
28. Oster A, Hallmark C, Macomber K, Brandt MG, Robilotto S. HIV Cluster Detection and Response Institute 101: connecting data, partners, and programs to close gaps. In: 2022 National Ryan White Conference on HIV Care and Treatment; virtual meeting; 2022 Aug 23–26. Abstract 21040. Rockville (MD): Health Resources and Services Administration; 2022.
29. Sweeney P, Gardner LI, Buchacz K, Garland PM, Mugavero MJ, Bosshart JT, et al. Shifting the paradigm: using HIV surveillance data as a foundation for improving HIV care and preventing HIV infection. Milbank Q. 2013;91:558–603. https://doi.org/10.1111/milq.12018
30. Centers for Disease Control and Prevention. Additional implementation guidance for PS18–1802 strategy 3: cluster detection and response. 2018 [cited 2025 Apr 21]. https://web.archive.org/web/20241221052552/https://www.cdc.gov/hiv/pdf/funding/announcements/ps18-1802/cdc-hiv-additional-cluster-implementation-guidance.pdf
31. Centers for Disease Control and Prevention. Notice of funding opportunity PS24–0047: high-impact HIV prevention and surveillance programs for health departments. 2024 [cited 2025 Apr 21]. https://web.archive.org/web/20241128223748/https://www.cdc.gov/hiv/funding/announcements/ps24-0047/index.html

Address for correspondence: Anne Marie France, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H24-5, Atlanta, GA 30329-4018, USA; email: afrance@cdc.gov

Effects of Decentralized Sequencing on National *Listeria monocytogenes* Genomic Surveillance, Australia, 2016–2023

Patiyan Andersson, Sally Dougall, Karolina Mercoulia, Kristy A. Horan, Torsten Seemann, Jake A. Lacey, Tuyet Hoang, Lex E.X. Leong, David Speers, Louise Cooley, Karina Kennedy, Rob Baird, Rikki Graham, Qinning Wang, Avram Levy, Dimitrios Menouhos, Norelle L. Sherry, Susan A. Ballard, Vitali Sintchenko, Amy V. Jennison, Benjamin P. Howden

We assessed turnaround times in the national *Listeria monocytogenes* genomic surveillance system in Australia before and after decentralized sequencing. Using 1,204 samples collected during 2016–2023, we observed statistically significant reductions in median time from sample collection to issuance of national genomic surveillance report to 26 days, despite sample numbers doubling in 2022 and 2023. During 2016–2018, all jurisdictions referred samples to the National Listeria Reference Laboratory for sequencing and analysis, but as jurisdictional sequencing capacity increased, 4 jurisdictions transitioned to sequencing their own samples and referring sequence data to the national laboratory. One jurisdiction had well-established genomics capacity, transitioned without noticeable disruption, and continued to improve. Another 3 jurisdictions initially had increased turnaround times, highlighting the need for defined sequence referral mechanisms. Overall, timeliness and throughput improved, and sequencing decentralization strengthened Australia's genomic surveillance system while maintaining timeliness. The practices described could be beneficial and achievable in other countries.

Listeria monocytogenes is an etiologic agent for gastroenteritis but can also cause serious invasive disease (*1*). The incidence of invasive listeriosis is relatively low, but the case-fatality rate is one of the highest among foodborne infections (*2*,*3*). The severity of *L. monocytogenes* infection, along with its ubiquitous environmental presence and frequent outbreaks from commercially manufactured foods, results in major social and economic impacts (*4*–*6*). Collecting detailed information for both the case and the pathogen enhances the success of public health investigations.

Whole-genome sequencing (WGS) provides high-resolution characterization of pathogens and has been shown to be critical in identifying outbreak clusters, separating outbreaks from endemic cases, and linking food and environmental samples to human cases with greater confidence (*7*,*8*). Consequently, WGS has been implemented for routine surveillance of *L. monocytogenes* in several countries, including Australia (*9*–*16*).

In Australia, invasive listeriosis has been a notifiable disease since 1991 and is recorded in the National Notifiable Diseases Surveillance System (NNDSS) (*17*).

Author affiliations: The University of Melbourne, Melbourne, Victoria, Australia (P. Andersson, S. Dougall, K. Mercoulia, K.A. Horan, T. Seeman, J.A. Lacey, T. Hoang, N.L. Sherry, S.A. Ballard, B.P. Howden); SA Pathology, Adelaide, South Australia, Australia (L.E.X. Leong); Queen Elizabeth II Medical Centre, Perth, Western Australia, Australia (D. Speers, A. Levy); Royal Hobart Hospital, Hobart, Tasmania, Australia (L. Cooley); Canberra Health Services, Australian National University Medical School, Canberra, Australian Capital Territory, Australia (K. Kennedy); Territory Pathology, Royal Darwin Hospital, Darwin, Northern Territory, Australia (R. Baird, D. Menouhos); Queensland Public Health and Scientific Services, Queensland Health, Brisbane, Queensland, Australia (R. Graham, A.V. Jennison); Institute of Clinical Pathology and Medical Research, NSW Health Pathology, Sydney, New South Wales, Australia (Q. Wang, V. Sintchenko); Austin Health, Heidelberg, Victoria, Australia (N.L. Sherry, B.P. Howden); Sydney Institute for Infectious Diseases, The University of Sydney, Sydney (V. Sintchenko)

DOI: https://doi.org/10.3201/eid3113.241357

Public health monitoring and action is managed by OzFoodNet, the national foodborne disease surveillance network. In 2010, the National Enhanced Listeriosis Surveillance System (NELSS) was established to collate both enhanced epidemiologic data from cases and molecular laboratory data from isolates (*18*,*19*). Once NELSS was established, the National Listeria Reference Laboratory (NLRL), based at the Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL) in the state of Victoria, was tasked with providing national molecular characterization of all referred *L. monocytogenes* samples, including typing with pulsed-field gel electrophoresis (PFGE). In July 2015, the NLRL commenced routine WGS for all referred samples and, after a 12-month trial of parallel use with PFGE, WGS became the preferred typing method (*20*). The NLRL also conducts centralized genomic analysis and issues a national genomic surveillance report.

As a federation, Australia's 8 jurisdictions are independently responsible for their public health activities, including pathogen genomics. Since 2016, genomic sequencing capacity has expanded in Australia; MDU PHL continued WGS for Victoria, and 4 additional jurisdictions successively became responsible for their own *L. monocytogenes* WGS during 2018–2023. The other 3 jurisdictions still refer samples to the NLRL for WGS. We investigated the timeliness and continued evolution of national *L. monocytogenes* surveillance from the perspective of the transition to a decentralized sequencing model.

Materials and Methods

Setting

Australia is a federated nation composed of 8 jurisdictions and had a combined estimated residential population of 27,000,000 in 2023 (*21*). We obtained annual listeriosis incidence rates from the NNDSS dashboard (*17*).

In Australia, samples from listeriosis notifications and relevant positive food and environmental samples are forwarded to public health laboratories (PHL) in each jurisdiction for confirmation, and PHLs subsequently refer sequences or isolates to the NLRL for national genomic analysis (Figure 1). Sequencing and bioinformatic analysis of *L. monocytogenes* at the NLRL are to ISO 17025 and ISO 15189 standards and accredited by the Australian National Association of Testing Authorities (https://www.nata.com.au).

Study Sample Dataset

This study included all *L. monocytogenes* samples referred to the NLRL for sequencing and all *L. monocytogenes* sequences referred by PHLs during 2016–2023, representing the first complete year of WGS to the most recent full year of data. Sample metadata included referring laboratory, residential jurisdiction of the case, and sample source categorized as human, food, or environmental. All sequence data were generated on Illumina (https://www.illumina.com) platforms, and the NLRL analysis workflow applied quality control thresholds of ≥40× coverage, *L. monocytogenes* species detected, and genome size within 10% of expected maximum genome size. We assessed timeliness of the genomic surveillance system by using temporal data, including date sample was collected, date sample was received at the jurisdictional PHL, date sample was sequenced, date NLRL received sample or sequence data, and date NLRL issued national genomics report.

Statistical Analysis

We used a Shapiro-Wilk test to assess for normality in the processing times at each stage, for each year, and for each jurisdiction. We excluded years with <3 observations from the normality testing. Because most of the dataset was not normally distributed, we used a nonparametric Kruskal-Wallis test to assess differences in processing times across years and, where statistically significant ($p \leq 0.05$), performed a Dunn's posthoc test with Bonferroni correction on the pairwise comparison of years for each jurisdiction.

Results

Notifications and Study Sample Set

Listeriosis became a notifiable disease in Australia in 1991. The average number of notifications recorded in the NNDSS was 65.2 (range 35–93) cases/year. The annual incidence rate of listeriosis remained relatively stable since 1991, ranging from 0.2 to 0.4/100,000 population.

During 2016–2023, Australia had 545 notified listeriosis cases, and yearly case numbers ranged from 43 to 89. We included a total of 543 sequences from 508 individual cases in the study, representing sequences from an average of 93.2% (range 87.3%–100%) of cases per year. We also included an additional 418 sequences of *L. monocytogenes* cultured from food samples and 243 sequences from environmental samples, bringing the complete dataset to 1,204 samples.

Samples from the 2 most populous jurisdictions, New South Wales (NSW) and Victoria (VIC), made up 65% of the dataset, and Queensland (QLD) and South Australia (SA) comprised another 22.5% (Figure 2). All jurisdictions submitted isolates or sequences from human, food, and environmental sources, except

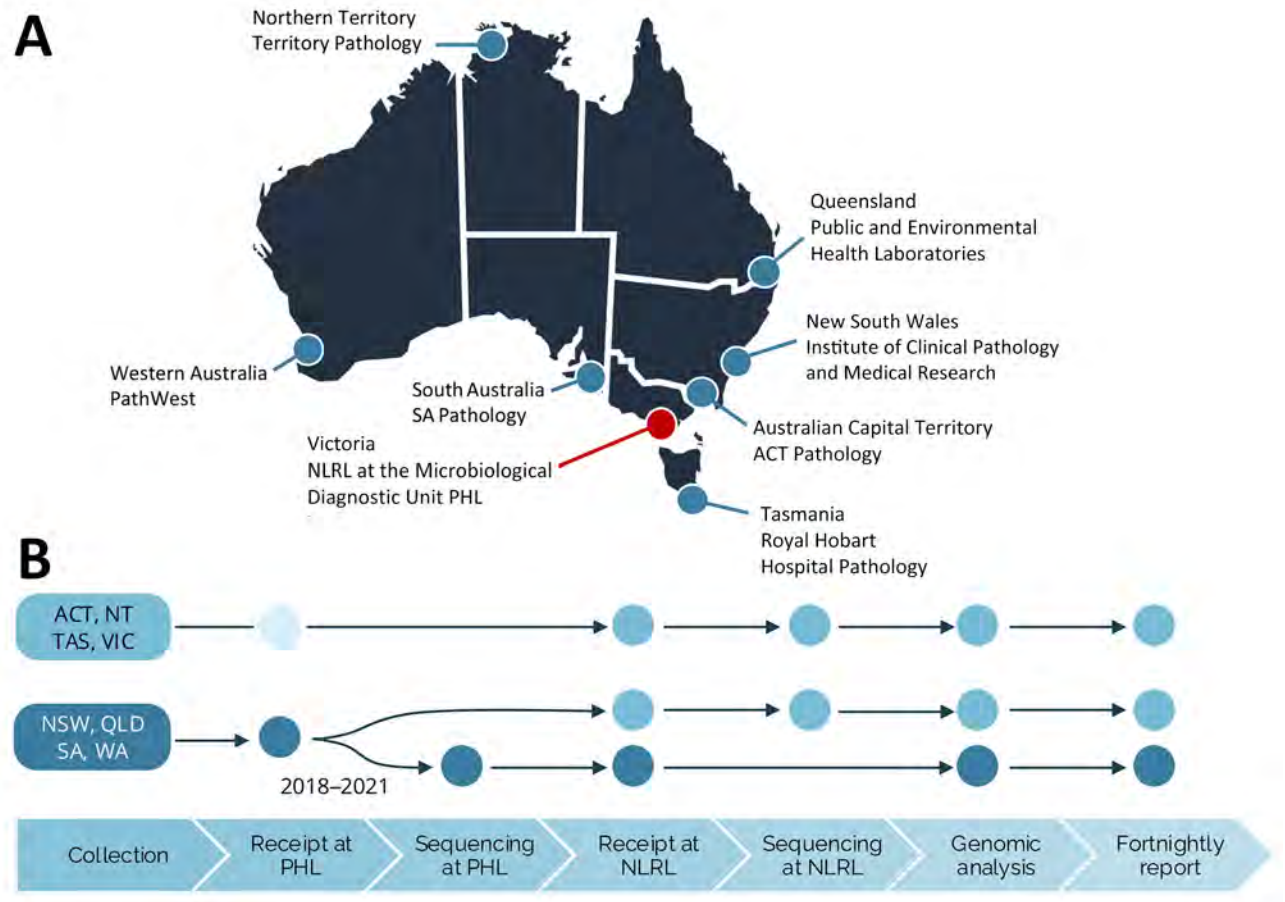


Figure 1. National decentralized sequencing system for *Listeria monocytogenes* genomic surveillance, Australia. A) Eight jurisdictional public health laboratories contributing to genomic surveillance. The NLRL is based at the Microbiological Diagnostic Unit (MDU) PHL in the state of Victoria. B) Overview of the steps in the national genomic surveillance system; dots indicate where sample processing occurs. The process is the same for human and nonhuman samples. For jurisdictions ACT, NT, TAS, and VIC, sequencing is performed by the NLRL at the Microbiological Diagnostic Unit PHL. The unfilled circle indicates that some samples are referred directly from the primary pathology laboratory to NLRL. For jurisdictions NSW, QLD, SA, and WA, the referral pathway transitioned during 2018–2021 from sequencing performed by the NLRL to jurisdictional sequencing and referral of sequences for genomic analysis. ACT, Australian Capital Territory; NLRL, National Listeria Reference Laboratory; NSW, New South Wales; NT, Northern Territory; PHL, public health laboratory; QLD, Queensland; SA, South Australia; TAS, Tasmania; VIC, Victoria; WA, Western Australia.

Northern Territory (NT) and Western Australia (WA). However, the distribution was uneven; VIC had a high (49%) percentage of food samples and NSW had a high (43%) percentage of environmental samples, demonstrating some differences in investigation practices. We noted a marked increase in the number of submissions from 2021 onward, mainly driven by increases from food and environmental sources.

Sample Referral Pathways

Samples from notified cases and food and environmental sources from all jurisdictions are referred to the NLRL for inclusion in genomic analysis (Figure 1, panel B). Before 2018, all jurisdictions referred either primary samples (food or environmental) or cultured isolates to the NLRL where, after *L. monocytogenes* culture, if required, isolates were subject to WGS and bioinformatic analysis. PHLs gradually transitioned to performing sequencing locally and referring *L. monocytogenes* genome sequences to the NLRL for inclusion in the national analysis. By 2023, four jurisdictions had transitioned to local sequencing: QLD in 2018, NSW in 2019, SA in 2020, and WA in 2021. The NLRL processed samples from VIC and continued to support NT, Australian Capital Territory (ACT), and Tasmania (TAS) with WGS services.

Genomic Surveillance Reporting

The genomic surveillance report includes samples collected within a 24-month rolling window and provides phylogenetic and clustering data, including historical data for context when relevant. Single-linkage clustering

is performed and reported as highly related if the pairwise difference is ≤5 single-nucleotide polymorphisms, and possibly related if the pairwise difference is 6–20 single-nucleotide polymorphisms. Since genomic surveillance began in 2015, NLRL has issued >200 formal national *L. monocytogenes* genomic surveillance reports.

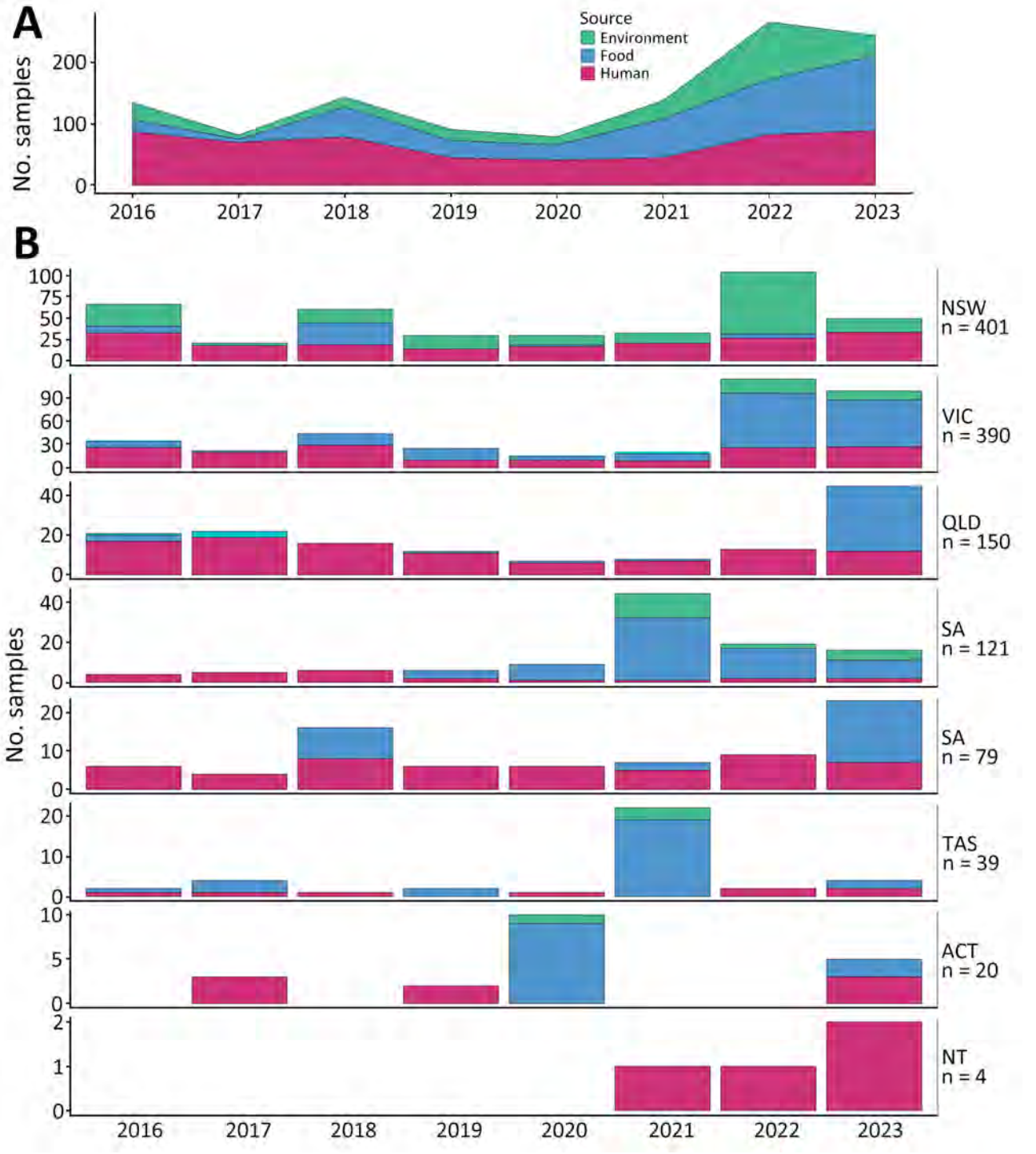


Figure 2. Summary of *Listeria monocytogenes* samples included in a study of effects of decentralized sequencing on national *L. monocytogenes* genomic surveillance, Australia, 2016–2023. A) Number of samples per year by source; B) number of samples per jurisdiction per year and source. Total number of samples per jurisdiction are provided; note varying scales of the y-axes. A notable increase in samples from food and environmental sources has occurred since 2021. ACT, Australian Capital Territory; NSW, New South Wales; NT, Northern Territory; QLD, Queensland; SA, South Australia; TAS, Tasmania; VIC, Victoria; WA, Western Australia.

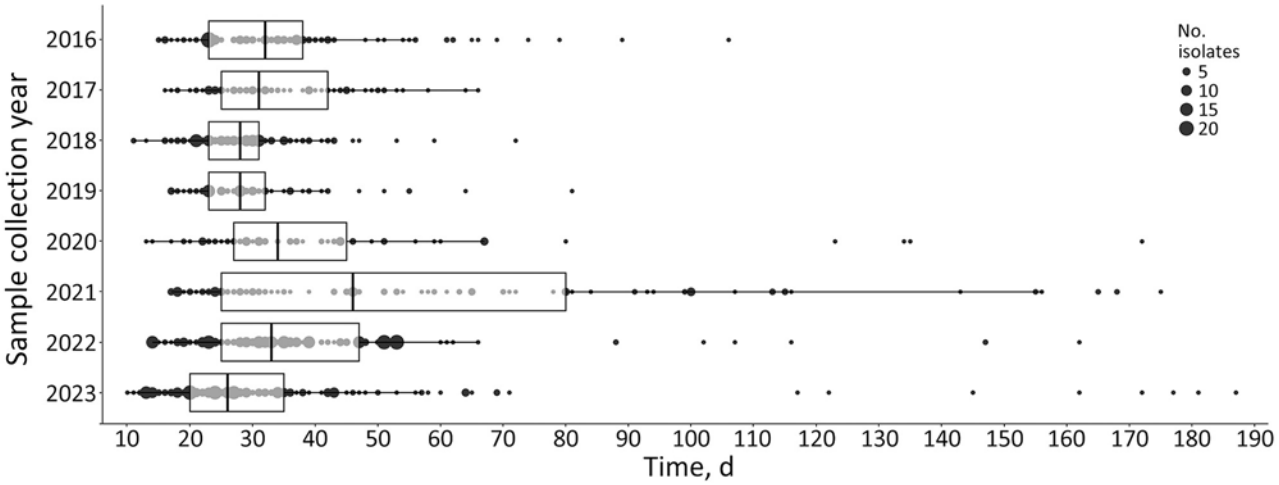


Figure 3. Box and whisker representation of end-to-end timeliness in a study of effects of decentralized sequencing on national *Listeria monocytogenes* genomic surveillance, Australia, 2016–2023. Time represents date of sample collection to date genomic surveillance report was issued, by year. Boxplots show medians (vertical lines within boxes), First and third quartiles (box left and right edges), and 1.5× interquartile range from each quartile (whiskers); points outside that range are considered outliers. Underlying data are shown as dots, and dot size corresponds to the number of samples at each timepoint. Statistically significant differences ($p<0.001$) between the years were calculated by using Kruskal-Wallis χ^2 test. Dunn's posthoc test showed statistically significant ($p<0.001$) increases in times for years 2020 and 2021, compared with previous years ($p = 0.029$), and subsequent significant decreases in times in 2022 and 2023 (adjusted $p<0.001$). Median time in 2023 was 26 days, compared with a median of 32 days in 2016.

Those reports are distributed to the referring PHLs, and to the national coordinating OzFoodNet epidemiologists and jurisdictional OzFoodNet epidemiologists. Reports are issued every 2 weeks, but genomic analysis is conducted weekly at a minimum and more frequently during outbreak investigations (Appendix 1, https://wwwnc.cdc.gov/EID/article/31/13/24-1357-App1.pdf). Critical findings from analyses are communicated immediately to epidemiologists via phone or email.

National Genomic Surveillance System Timeliness

Overall System Timeliness

In the dataset of 1,204 samples, we excluded 39 historical samples because those samples were not sequenced in real time. Thus, the timeliness analysis included 1,165 samples. We calculated the end-to-end turnaround times of the surveillance system, from date of sample collection to issuance of the national genomic surveillance report (Figure 3). We observed a pattern of pre–COVID-19 pandemic improvements but a statistically significant increase in turnaround times in 2020 and 2021 compared with previous years (pairwise comparisons years 2016 to 2021, adjusted $p<0.001$ to $p = 0.029$), and subsequent time reductions in 2022 and 2023 (adjusted $p<0.001$). We also observed a statistically significant improvement in timeliness between 2016 (median 32 days) and 2023 (median 26 days) (adjusted $p<0.001$).

We noted differing patterns of timeliness across the jurisdictions (Appendix 2 Figure 1, https://wwwnc.cdc.gov/EID/article/31/13/24-1357-App2.pdf), and all jurisdictions, except NT, showed statistically significant changes over the years ($p<0.001$–0.007). NT had too few observations for analysis. VIC was the most stable, consistently maintaining median turnaround times of 24–27 days, although that range increased in the 2 most recent years, and VIC had a significantly higher median in 2022 ($p = 0.032$) compared with other all years ($p<0.001$). Of the jurisdictions that have transitioned from referring samples to referring sequences, NSW remained consistent over time but had larger variations in median turnaround times, 24–42 days, and higher upper limits, 60–70 days, not considering outliers (Figure 3). NSW and QLD demonstrated statistically significant improvements in 2023 compared with 2016 (Table). However, QLD, SA, and WA all showed increases in median and range of turnaround times immediately after transitioning to referring sequence data instead of isolates.

When comparing the 2016 and 2023 median turnaround times for only human samples, we noted most jurisdictions improved timeliness (Appendix 2 Figure 2). We noted statistically significant variations only in NSW ($p<0.001$) and WA ($p<0.02$), despite the appearance of large variations in the SA and QLD data for human sequences.

We also assessed time before and after transitioning to sequence referrals for the 4 relevant states (Appendix 2 Figure 3). Although NSW did not show any difference before and after transition, the

3 other states had significant differences in median turnaround times (p<0.001). Only the 2 earliest transitioning states, QLD (adjusted p<0.001) and NSW (adjusted p = 0.003), achieved medians in 2023 that were significantly lower than those of 2016.

Primary Referral Time

The time for collection of human samples to referral of samples to the jurisdictional PHLs was consistent over time; median times were ≈5 days and upper limits <10 days for most jurisdictions except NSW (Table; Appendix 2 Figure 4). In VIC, SA, TAS, and particularly WA, referral of food samples had higher upper time limits. However, VIC had a reduced median because a large number were referred in ≤1 day. We also observed that trend in NSW and QLD.

Sequencing Times

Median processing times, from sample receipt at the PHL to date sequencing preformed, were relatively consistent (10 days) across jurisdictions, with some minor improvements between 2016 and 2023 (Table; Appendix 2 Figure 5). The sequence processing times include potential culturing of samples received as

Table. Sample processing times used in a study of effects of decentralized sequencing on national *Listeria monocytogenes* genomic surveillance, Australia, 2016–2023*

	Median time, d (range)							
Collection year	ACT	NSW	NT	QLD	SA	TAS	VIC	WA
Overall referral time†								
All sample sources								
2016	ND	29 (15–106)	ND	36 (30–55)	37 (33–61)	34 (29–39)	25 (15–37)	44 (36–65)
2023	36 (36–43)	24 (10–65)	75 (28–122)	20 (12–43)	45 (14–71)	30 (18–32)	26 (12–56)	50 (26–187)
p value	ND	0.003	ND	<0.001	NS	NS	NS	NS
Primary referral time‡								
Human samples								
2016	ND	5 (1–14)	ND	4 (3–7)	2 (2–3)	4 (4)	5 (2–10)	6 (2–10)
2023	13 (12–13)	6 (0–20)	3 (3)	5 (2–6)	3 (1–5)	9 (8–9)	3 (1–7)	4 (0–10)
p value	ND	NS	ND	NS	NS	NS	<0.001	NS
Food samples								
2016	ND	1 (1)	ND	1 (0–11)	ND	19 (19)	4 (0–14)	ND
2023	15 (15)	ND	ND	0 (0–2)	9 (6–17)	9 (9)	4 (0–25)	32 (0–129)
p value	NS	ND	ND	NS	ND	0.026	NS	ND
Environmental samples								
2016	ND	1 (0–2)	ND	1 (1)	ND	ND	0	ND
2023	ND	1 (1)	ND	ND	29 (29)	ND	1 (0–6)	ND
p value	ND	NS	ND	ND	ND	ND	NS	ND
Sequencing time§								
Human samples								
2016	ND	10 (3–18)	ND	10 (8–14)	10 (5–13)	11 (11)	10 (6–18)	12 (10–20)
2023	13	5 (1–12)	11 (10–11)	8 (2–12)	3 (2–5)	8 (6–11)	8 (5–11)	6 (1–12)
p value	ND	<0.001	ND	0.014	NS	NS	0.003	NS
Food samples								
2016	ND	3 (3–10)	ND	14 (10–18)	ND	16 (16)	20 (15–20)	ND
2023	9 (9)	15 (5–15)	ND	7 (6–14)	9 (6–14)	8 (7–9)	11 (6–25)	7 (2–8)
p value	ND	<0.001	ND	NS	ND	NS	0.002	ND
Environmental samples								
2016	ND	10 (3–14)	ND	14 (14)	ND	ND	17 (17)	ND
2023	ND	5 (2–5)	ND	ND	6 (6)	ND	16 (9–21)	ND
p value	ND	<0.001	ND	ND	ND	ND	NS	ND
Sequence referral time§								
All sample sources								
2016	ND	20 (9–96)	ND	27 (13–41)	25 (17–52)	ND	ND	34 (21–48)
2023	ND	12 (4–54)	ND	12 (3–19)	26 (7–31)	ND	ND	12 (7–26)
p value	ND	<0.001	ND	<0.001	NS	ND	ND	0.007
Genomic analysis time#								
2016	ND	11 (0–14)	ND	8 (1–11)	10 (4–17)	9 (4–14)	8 (3–17)	9 (2–11)
2023	19 (15–19)	8 (1–13)	7 (2–12)	1 (1–19)	9 (3–14)	12 (4–15)	9 (1–21)	14 (3–18)
p value	ND	0.025	ND	0.002	NS	NS	NS	<0.001

*Times are shown as median (range) in days for each jurisdiction for years 2016 and 2023, and adjusted p values from Dunn's post-hoc tests of pairwise comparisons. Range defined as data points within 1.5 from each quartile, with points outside interquartile range considered outliers. ACT, Australian Capital Territory; ND, no data available; NS, not statistically significant; NSW, New South Wales; NT, Northern Territory; QLD, Queensland; SA, South Australia; TAS, Tasmania; VIC, Victoria; WA, Western Australia.
†Date of collection to date genomic surveillance report issued.
‡Date of collection to date sample received at jurisdictional public health laboratory.
§Date received at jurisdictional public health laboratory to date sequenced.
¶Date received at jurisdictional PHL to date sequence available for bioinformatic analysis at the National Listeria Reference Laboratory.
#Date received at national *Listeria* reference laboratory to date national genomic report issued.

primary specimen. Processing times for human samples improved considerably between 2016 and 2023 in NSW, SA, VIC, and WA. The sequence processing times for food samples were similar to those of human samples in all jurisdictions except VIC. The difference for VIC can be explained by the referral workflow, in which the NLRL in VIC would predominantly receive cultured food isolates from other jurisdictions for sequencing, but NLRL received local VIC samples as primary specimens that require culture and isolation before sequencing. The considerable improvements we noted in VIC for the 2 most recent years can partially be attributed to a larger number of local food samples received as cultured isolates.

Effects of Transition to Sequence Referral

To compare the effect of transitioning to sequence referral, we considered the entire process from sample collection to bioinformatic analysis, thereby accounting for processing times at each phase, including culturing and isolation, sequencing, and sample or sequence referral, either at the jurisdictional PHL or the NLRL (Appendix 2 Figure 6). Median processing time at each of the 4 jurisdictions before and after transition showed significant reductions associated with referral of sequences for NSW (adjusted $p<0.001$) and QLD (adjusted $p<0.001$), but a significant increase in processing times for referred sequences in SA (adjusted $p = 0.015$). We found no overall significant differences for WA. We noted considerable variation in processing times between years for QLD, SA, and WA, but NSW was more stable. Comparing the extreme timepoints of 2016 and 2023, we observed statistically significant reductions for NSW (adjusted $p<0.001$), QLD (adjusted $p<0.001$), and WA (adjusted $p = 0.007$).

Genomic Analysis Times

We calculated the genomic analysis times on the basis of the date a sequence was available (either sequencing completed at the NLRL or PHL sequence received by the NLRL) and the date the fortnightly genomic surveillance report was issued (Appendix 2 Figure 7). Thus, times varied depending on when in the reporting cycle the sequence became available. Genomic analysis and reporting were consistent and timely across the study period; 84.4% (983/1,165) of samples were reported within a 14-day reporting cycle and 99.4% (1,158/1,165) of samples reported within 2 reporting cycles.

Given the potential effect of the fortnightly reporting cycle, we analyzed the time from sample collection to a sequence being available for analysis and reporting. The pattern for all jurisdictions remained unchanged, but the shortest times in 2023 were just 5–8 days for NSW, QLD, VIC, and SA (data not shown).

Discussion

We describe the maturation of a multijurisdictional genomic surveillance system for *L. monocytogenes* in Australia and the effects of transitions to decentralized sequencing of isolates as local genomic capacity improved. The overall median time for genomic data to be available was 32 days in 2016, but 2023, the last year in the review, demonstrated the lowest recorded median at 26 days. That difference is a marked improvement when compared with the predecessor analysis method of PFGE, in which the median time from notification to data availability to NELSS during 2010–2013 was 50 days (*18*). We believe the reductions we report are associated with use and increased capacity of automated robotics workflows for WGS; accelerated establishment of strong WGS capacity during the COVID-19 pandemic; and replacement of a physical sample transport step, and potential batching of samples for courier transport, with electronic data transfer. Compared with the first full year of genomic data from the system in 2016, the overall median end-to-end processing time was lower in 2022 and 2023 despite a substantial increase in the number of samples, mainly food and environmental samples. The shift to decentralized sequencing in some jurisdictions might contribute to the ability of the system to manage increased sample volumes without detrimental effects on the timeliness.

Of the 4 jurisdictions that transitioned to sequence referral to NLRL, 3 had considerable increases in overall turnaround times after shifting to PHL sequencing for *L. monocytogenes* but then resumed a downward trajectory in turnaround times. Delays associated with sample batching resulting from limited throughput could be expected during the early stages of establishing sequencing capacity but were not evident from the sequencing times we observed. Instead, we mainly observed delays in referral of sequences to the NLRL. In part, those transitions coincided with the COVID-19 pandemic, during which all PHLs were managing an unprecedented additional workload from real-time SARS-CoV-2 sequencing. Delays might also have been associated with a lack of a national protocol for inclusion of nonhuman samples in the national genomic analyses and variability in the software solutions and processes for sequence referrals. Those observations highlight the need to define and adequately resource sequence referral mechanisms during implementation of local sequencing to ensure optimal turnaround times.

Sequencing capacity was strengthened across all jurisdictions in Australia during the COVID-19 pandemic. After the initial years of the pandemic, sequencing priorities were realigned, which enabled PHLs to apply the enhanced WGS capacity to other pathogens. The capacity for continued improvement speaks well for the future national capability of managing increased volumes of nonhuman samples to improve source identification for *L. monocytogenes*. The benefits of integration of cross-sectoral samples (food and environment) were immediately apparent after the implementation of genomic analysis in Australia, and the sequence from an unresolved case was linked to stone fruits imported from the United States when the sequence was deposited in GenomeTrakr (*22*). By late 2024, Australia had contributed 770 sequences to GenomeTrakr, and that network and the Pathogen Detection Portal continue to be a highly valuable resources for monitoring potential common outbreak sources with international data (*9*,*16*).

The fact that listeriosis notification rates remained stable in Australia over the study period is an indication that public health management of listeriosis remains complex. Although WGS has greatly enhanced the capacity to detect and characterize outbreaks, its effectiveness in reducing overall case numbers is contingent on rapid and comprehensive public health actions, effective control of persistent contamination sources, and improvements in food safety protocols and compliance. Large-scale analysis of international data has shown numerous multinational clusters and emphasized the power of genomics to manage the challenges of persistent environmental contamination and highly interconnected food supply networks (*15*,*23*). The observation of long-term clusters and limited initial epidemiologic signals is further echoed in descriptions from other national genomic surveillance programs (*10*,*11*,*13*,*14*,*24*–*27*). Those findings make a strong argument for coordinated monitoring of *L. monocytogenes* at the global level through consistent and timely data sharing from national surveillance efforts.

Here, we have shown the evolution of timeliness in a longstanding national genomic surveillance system for *L. monocytogenes* and an immediate 30% reduction in median processing time compared with PFGE, then a further 20% reduction to 26 days from sample to notification report observed in 2023. We also demonstrated that surveillance processes can be disrupted and result in delays in data availability during the establishment of decentralized sequencing processes but that those disruptions can be resolved as the capacity matures. Of note, we found that when genomic capacity was already strong in the referring jurisdiction, the transition was managed without noticeable detrimental effects, even in the extraordinary circumstances of the COVID-19 pandemic response. That finding should be considered when making process changes for pathogens that have time-sensitive surveillance objectives.

In summary, we report the overall picture for decentralized sequencing of *L. monocytogenes* in Australia as one of reduced turnaround times and continued improvement. Decentralization of sequencing strengthened the genomic surveillance system in the country through increased throughput while maintaining timeliness. Such practices could be beneficial and achievable in other countries with sequencing capacity.

Acknowledgments

We recognize the tireless work of all staff in the primary pathology laboratories, jurisdictional public health laboratories, OzFoodNet units in the jurisdictional departments of health and at the Australian Government Department of Health and Aged Care, and their ongoing contributions to the surveillance of *Listeria monocytogenes* in Australia. We thank all the members the Communicable Diseases Genomics Network of Australia for conducting this work and for their continued contributions to strengthening pathogen genomics in Australia.

This work was supported by Australian National Health and Medical Research Council, Medical Research Futures Fund (Australian Pathogen Genomics program, grant no. FSPGN00049), and Investigator Grant (no. GNT1196103) to B.P.H.

About the Author

Dr. Andersson is a senior research fellow and genomic epidemiologist at the University of Melbourne, Australia, with a background in molecular biology, genomics, and field epidemiology. He works with translational research and application of pathogen genomics in public health at state, national and regional levels.

References

1. Charlier C, Perrodeau É, Leclercq A, Cazenave B, Pilmis B, Henry B, et al.; MONALISA Study Group. Clinical features and prognostic factors of listeriosis: the MONALISA national prospective cohort study. Lancet Infect Dis. 2017; 17:510–9. https://doi.org/10.1016/S1473-3099(16)30521-7
2. European Food Safety Authority (EFSA); European Centre for Disease Prevention and Control (ECDC). The European Union One Health 2022 zoonoses report. EFSA J. 2023;21:e8442. https://doi.org/10.2903/j.efsa.2023.8442

3. OzFoodNet Working Group. Monitoring the incidence and causes of disease potentially transmitted by food in Australia: annual report of the OzFoodNet network, 2017. Commun Dis Intell (2018). 2022:46. https://doi.org/10.33321/cdi.2022.46.59
4. Glass K, McLure A, Bourke S, Cribb DM, Kirk MD, March J, et al. The cost of foodborne illness and its sequelae in Australia circa 2019. Foodborne Pathog Dis. 2023;20:419–26. https://doi.org/10.1089/fpd.2023.0015
5. Desai AN, Anyoha A, Madoff LC, Lassmann B. Changing epidemiology of *Listeria monocytogenes* outbreaks, sporadic cases, and recalls globally: a review of ProMED reports from 1996 to 2018. Int J Infect Dis. 2019;84:48–53. https://doi.org/10.1016/j.ijid.2019.04.021
6. Quereda JJ, Morón-García A, Palacios-Gorba C, Dessaux C, García-Del Portillo F, Pucciarelli MG, et al. Pathogenicity and virulence of *Listeria monocytogenes*: a trip from environmental to medical microbiology. Virulence. 2021;12:2509–45. https://doi.org/10.1080/21505594.2021.1975526
7. Besser JM, Carleton HA, Trees E, Stroika SG, Hise K, Wise M, et al. Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. Foodborne Pathog Dis. 2019;16:504–12. https://doi.org/10.1089/fpd.2019.2650
8. Cooper AL, Wong A, Tamber S, Blais BW, Carrillo CD. Analysis of antimicrobial resistance in bacterial pathogens recovered from food and human sources: insights from 639,087 bacterial whole-genome sequences in the NCBI Pathogen Detection database. Microorganisms. 2024;12:709. https://doi.org/10.3390/microorganisms12040709
9. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol. 2016;54:1975–83. https://doi.org/10.1128/JCM.00081-16
10. Coipan CE, Friesema IHM, van Hoek AHAM, van den Bosch T, van den Beld M, Kuiling S, et al. New insights into the epidemiology of *Listeria monocytogenes*—a cross-sectoral retrospective genomic analysis in the Netherlands (2010–2020). Front Microbiol. 2023;14:1147137. https://doi.org/10.3389/fmicb.2023.1147137
11. Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen N, et al. Real-time whole-genome sequencing for surveillance of *Listeria monocytogenes*, France. Emerg Infect Dis. 2017;23:1462–70. https://doi.org/10.3201/eid2309.170336
12. Rivas L, Paine S, Dupont PY, Tiong A, Horn B, Moura A, et al. Genome typing and epidemiology of human listeriosis in New Zealand, 1999 to 2018. J Clin Microbiol. 2021;59:e0084921. https://doi.org/10.1128/JCM.00849-21
13. Tsai YH, Moura A, Gu ZQ, Chang JH, Liao YS, Teng RH, et al. Genomic surveillance of *Listeria monocytogenes* in Taiwan, 2014 to 2019. Microbiol Spectr. 2022;10:e0182522. https://doi.org/10.1128/spectrum.01825-22
14. Zhang H, Chen W, Wang J, Xu B, Liu H, Dong Q, et al. 10-year molecular surveillance of *Listeria monocytogenes* using whole-genome sequencing in Shanghai, China, 2009–2019. Front Microbiol. 2020;11:551020. https://doi.org/10.3389/fmicb.2020.551020
15. Stevens EL, Carleton HA, Beal J, Tillman GE, Lindsey RL, Lauer AC, et al. Use of whole genome sequencing by the federal interagency collaboration for genomics for food and feed safety in the United States. J Food Prot. 2022;85:755–72. https://doi.org/10.4315/JFP-21-437
16. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis. 2016;63:380–6. https://doi.org/10.1093/cid/ciw242
17. Popovic I, Heron B, Covacin C. *Listeria*: an Australian perspective (2001–2010). Foodborne Pathog Dis. 2014;11:425–32. https://doi.org/10.1089/fpd.2013.1697
18. Polkinghorne B, Draper A, Harlock M, Leader R. OzFoodNet into the future: the rapid evolution of foodborne disease surveillance in Australia. Microbiol Aust. 2017;38:179–83. https://doi.org/10.1071/MA17063
19. OzFoodNet Working Group. Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: annual report of the OzFoodNet network, 2010. Commun Dis Intell Q Rep. 2012;36:E213–41.
20. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol. 2016;54:333–42. https://doi.org/10.1128/JCM.02344-15
21. Australian Bureau of Statistics. National, state and territory population, Dec 2023 [cited 2024 Aug 5]. https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/dec-2023
22. Kwong JC, Stafford R, Strain E, Stinear TP, Seemann T, Howden BP. Sharing is caring: international sharing of data enhances genomic surveillance of *Listeria monocytogenes*. Clin Infect Dis. 2016;63:846–8. https://doi.org/10.1093/cid/ciw359
23. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol. 2016;2:16185. https://doi.org/10.1038/nmicrobiol.2016.185
24. Daza Prieto B, Pietzka A, Martinovic A, Ruppitsch W, Zuber Bogdanovic I. Surveillance and genetic characterization of *Listeria monocytogenes* in the food chain in Montenegro during the period 2014–2022. Front Microbiol. 2024;15:1418333. https://doi.org/10.3389/fmicb.2024.1418333
25. Morton V, Kandar R, Kearney A, Hamel M, Nadon C. Transition to whole genome sequencing surveillance: the impact on national outbreak detection and response for *Listeria monocytogenes*, *Salmonella*, Shiga toxin–producing *Escherichia coli*, and *Shigella* clusters in Canada, 2015–2021. Foodborne Pathog Dis. 2024;21:689–97. https://doi.org/10.1089/fpd.2024.0041
26. Paduro C, Montero DA, Chamorro N, Carreño LJ, Vidal M, Vidal R. Ten years of molecular epidemiology surveillance of *Listeria monocytogenes* in Chile 2008–2017. Food Microbiol. 2020;85:103280. https://doi.org/10.1016/j.fm.2019.103280
27. Pietzka A, Allerberger F, Murer A, Lennkh A, Stöger A, Cabal Rosel A, et al. Whole genome sequencing based surveillance of *L. monocytogenes* for early detection and investigations of listeriosis outbreaks. Front Public Health. 2019;7:139. https://doi.org/10.3389/fpubh.2019.00139

Address for correspondence: Benjamin P. Howden, University of Melbourne, Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, 792 Elizabeth St, Melbourne, VIC 3000, Australia; email: bhowden@unimelb.edu.au

Genomic Modeling of an Outbreak of Multidrug-Resistant *Shigella sonnei*, California, USA, 2023–2024

Tyler Lloyd,[1] Sana M. Khan,[1] Dustin Heaton, Munira Shemsu, Vici Varghese, Jay Graham, Misha Gregory, Penny Dorfman, Megan Talton, Jessica DeVol, Nicola F. Müller,[2] Kavita K. Trivedi[2]

We report the detection of a *Shigella sonnei* outbreak from a small investigation in the San Francisco Bay area, California, USA, in 2024. By combining outbreak investigation with genomic sequencing, we show the utility of phylodynamics to aid outbreak investigations of bacterial pathogens by state or local public health departments.

In January 2024, a board and care facility (facility A) in the San Francisco Bay area, California, USA, reported 4 cases of *Shigella sonnei* infection to the Alameda County Public Health Department (ACPHD; San Leandro, California, USA). Shigellemia was confirmed in 3 patients. In February 2024, an independent living center (facility B) reported 3 cases of *Shigella* infection. *Shigella* bacteremia was confirmed in 2 patients (Figures 1,2). *Shigella* bacteremia (shigellemia) is rare but associated with immature immune responses or immunocompromised adults (*1*). We performed a 10-year retrospective review of *Shigella* cases in Alameda County and found 0.7% of cases had positive blood samples reported, consistent with other reviews on *Shigella* bacteremia (*2*). The 7 cases from 2 facilities prompted patient investigations at facilities A and B, and investigations into other *S. sonnei* patients in Alameda County during December 2023–February 2024.

Methods and Materials

Case investigations were limited but included symptom onset, severity, housing status, and other attainable risk factors. The outbreak investigation linked patients from facility B and 2 unhoused community members to a third location (facility C) where marginally housed community members gather. No clear transmission pattern was determined through epidemiologic investigation. We identified 19 genotypically identical *S. sonnei* isolates during December 2, 2023–February 26, 2024, among all cases in facilities A, B, and C.

Of the 19 patients, 13 (68%) were male and 6 (32%) female; median age was 59 years, and 9 (47%) were White and 10 (52%) non-Hispanic. Case investigations were completed on 16 of the 19 patients; 5 (26%) were experiencing homelessness, 4 (21%) were associated with facility A, 3 (15%) were associated with facility B, 4 (21%) had stable housing, and 3 (15%) had unknown housing. Drug use history was known in 3 patients. Of the 5 patients with shigellemia, 1 reported drug use. Sexual contact was unknown or denied during the incubation period for all patients. All treatment regimens where data were available were appropriate for the antimicrobial drug susceptibility data; all patients recovered.

Of the 3 patients associated with facility C, 2 were experiencing homelessness and 1 volunteered as a food handler at facility C while ill. The third patient from facility B visited facility C and had symptoms develop 14 days after exposure to the food handler at facility C. No other epidemiologic links were established among the 19 cases. No comorbidities were found in electronic medical records. However, determining precise risk factors in patients experiencing homelessness, such as where they sheltered during their infectious period, contact with each other, using the same resources, public restrooms or transportation, was not possible.

Author affiliations: Alameda County Public Health Department, San Leandro, California, USA (T. Lloyd, S.M. Khan, D. Heaton, M. Shemsu, V. Varghese, M. Gregory, P. Dorfman, M. Talton, J. DeVol, K.K. Trivedi); University of California, Berkeley, California, USA (T. Lloyd, J. Graham); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (S.M. Khan); University of California, San Francisco, California, USA (N.F. Müller)

DOI: https://doi.org/10.3201/eid3113.241307

[1]These first authors contributed equally to this article.

[2]These senior authors contributed equally to this article.

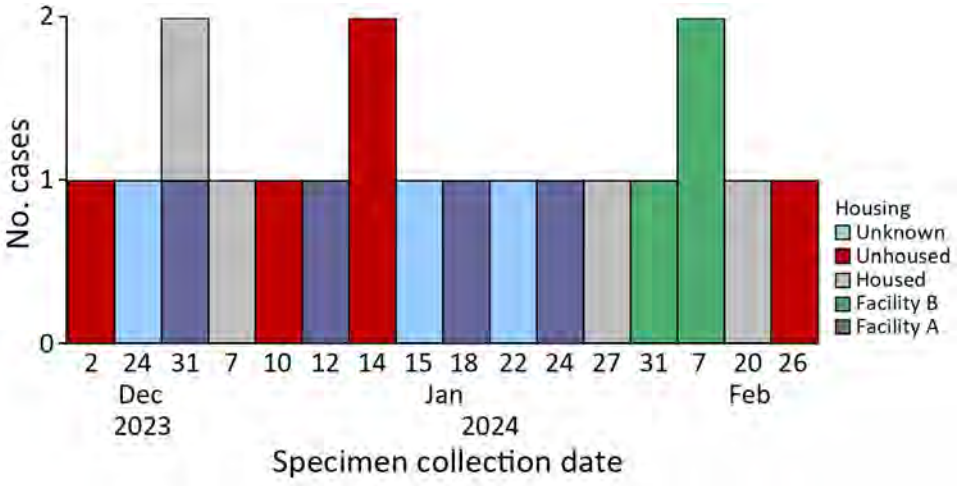


Figure 1. Epidemiologic curve showing specimen collection dates of shigellemia cases and housing status in *Shigella sonnei* case investigation with known linkages, California, USA, 2023–2024.

Results

Whole-genome sequencing (WGS) of *Shigella* isolates revealed highly similar sequences, suggesting an epidemiologic link. The time from notification of a potential outbreak in facility A to WGS confirmation was 8 days. We genotyped the isolates, which belonged to genotype 3.7.26, as previously described (*3*). This method removes repetitive regions with higher rates of potentially erroneous single-nucleotide polymorphisms (SNPs). The method simplified interpretation by providing a numerical genotype, making it easy to determine close ancestry and it was part of our routine bioinformatics workflow (*4*). References for genotype 3.7.26 are from the United Kingdom (2013) and France (2014). We confirmed phenotypic multidrug resistance by using antimicrobial drug resistance gene detection (Appendix, https://wwwnc.cdc.gov/EID/article/31/13/24-1307-App1.pdf).

During the retrospective sequencing of *S. sonnei* from patients treated in Alameda County, we identified patients with highly similar isolates in neighboring counties. The lack of specific links in the investigation and detection of cases from neighboring counties prompted ACPHD to notify the California Department of Public Health in March 2024, which led to a prioritization of *Shigella* isolates for sequencing. A total of 75 genetically related isolates were identified by California Department of Public Health by using PulseNet whole-genome multilocus sequence typing (MLST) (*5*), which showed relatedness but did not incorporate metadata. To reconstruct the spatial transmission dynamics of the outbreak, we performed a time-resolved, phylogeographic method known as the marginal approximation of the structured coalescent (MASCOT-skyline) in collaboration with the University of California, San Francisco. This approach uses Bayesian inference to reconstruct spatiotemporal transmission of pathogens and is implemented in the open-source program BEAST2 (*6*). MASCOT-skyline incorporates sampling time and sampling location of

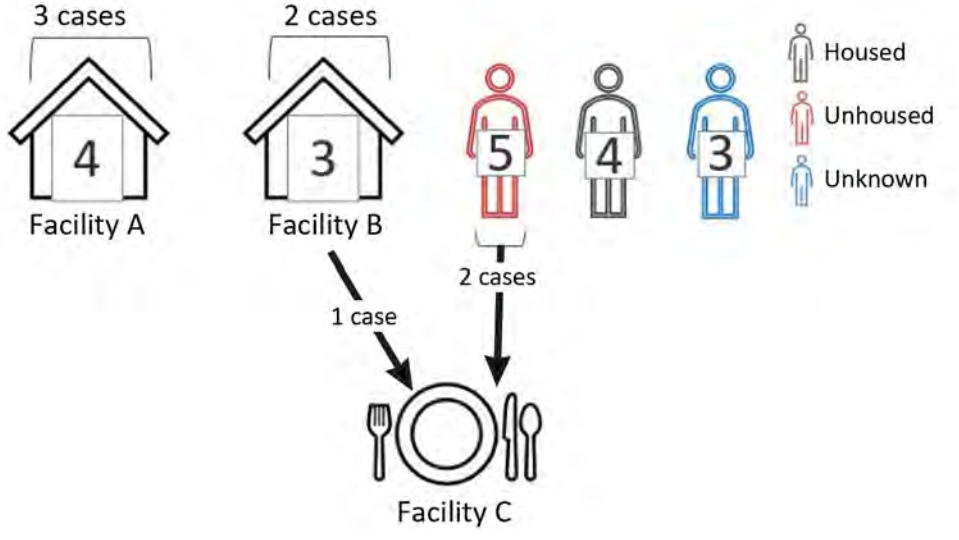


Figure 2. Diagram showing the number of cases from linked facilities, the housing status of patients, and the linkage between cases in *Shigella sonnei* case investigation with known linkages, California, USA, 2023–2024.

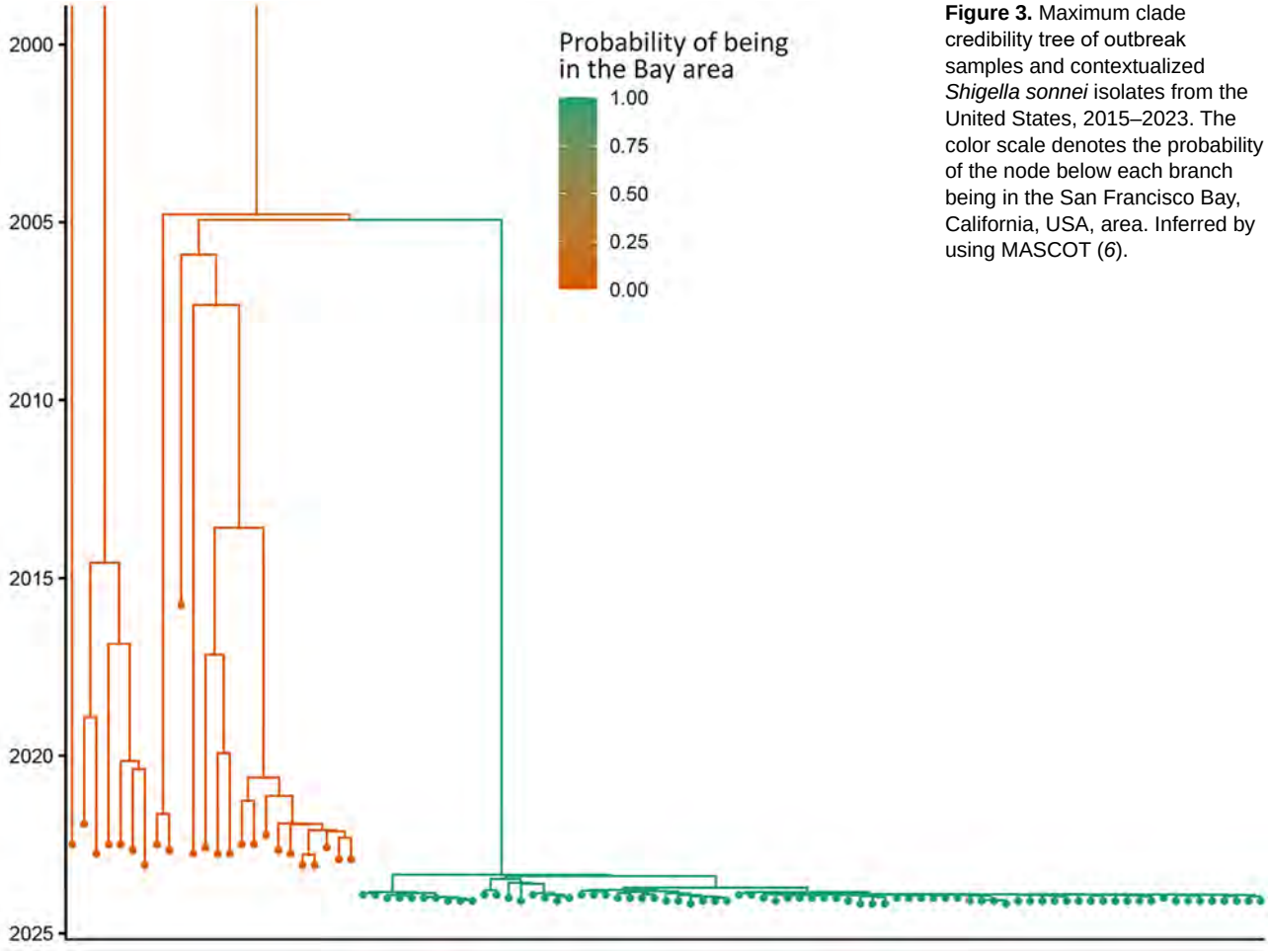


Figure 3. Maximum clade credibility tree of outbreak samples and contextualized *Shigella sonnei* isolates from the United States, 2015–2023. The color scale denotes the probability of the node below each branch being in the San Francisco Bay, California, USA, area. Inferred by using MASCOT (*6*).

isolates. MASCOT-skyline then infers a posterior estimate of where the bacterial lineage was in the past. From this result, we inferred that all isolates were the result of a single introduction into the area (*7*; N.F. Müller, et al., unpub. data, https://pmc.ncbi.nlm.nih.gov/articles/PMC10942421). We obtained the molecular clock rate by contextualizing outbreak samples with 24 *S. sonnei* Pulsenet sequences from 2015–2023 to ensure an appropriate number of samples and the time span to effectively estimate the clock rate. The mean clock rate for the core-genome SNP alignment (length 1491 bp) was 3.341×10^{-3} substitutions per site per year. The sequences formed a distinct cluster from all other historical sequences (Figure 3). We inferred time to most recent common ancestor was most likely June 2023 (95% CI November 2022–August 2023), providing an upper bound on the time of introduction (Figure 4). We analyzed the outbreak at a more granular scale by using the same regional alignment and fixing clock rate while removing contextual sequences. We found the samples were geographically clustered by county within the outbreak (Figure 5).

One patient was a food handler at facility C, but no evidence of foodborne transmission was found. The shigellemia cases prompted us to investigate this cluster; however, patients were immunocompetent, and virulence markers were identical to nonbloodstream infections. Host factors such as intravenous drug use and sexual contact were incomplete and remain possible factors for shigellemia.

Discussion

The advent of phylodynamic approaches and genomic epidemiology has provided public health with additional insight into the spread of diseases, transmission chains, and mutations when using laboratory data paired with epidemiologic information. In this article, we demonstrate the use of phylodynamic modeling alongside a traditional case investigation to better determine outbreak dynamics and inform public health actions. Bacterial genomic epidemiology has historically relied on MLST, SNPs, whole-genome MLST, or a combination of techno-

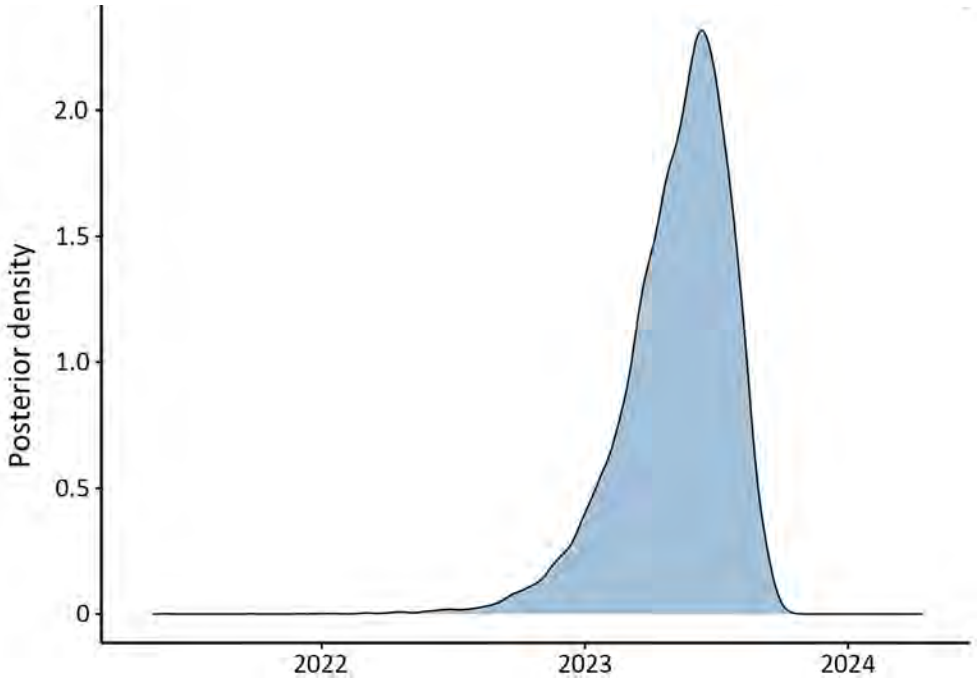


Figure 4. Distribution curve of the predicted dates of the most recent common ancestor of the *Shigella sonnei* outbreak isolates, California, USA, 2023–2024. The plot shows the posterior density for the common ancestor times of the *Shigella* sequences collected in the San Francisco Bay, California, USA, area. For a single introduction, the common ancestor time provides a lower bound on the timing of the introduction into the San Francisco Bay area.

logic tools. However, those tools do not enable us to characterize the direction and timing of disease spread. SNP cutoff levels have shown variable specificity and sensitivity in identifying closely related bacterial isolates (*8*).

We also describe the role of the local public health laboratory to initiate enhanced WGS of *S. sonnei* to discover unlinked cases and identify a regional outbreak. We describe the timeline of the outbreak identification, notification of the state public health department, and phylodynamic methods to provide evidence of a single introduction and incorporate metadata into bacterial genomic epidemiology. However, those models do not guarantee complete

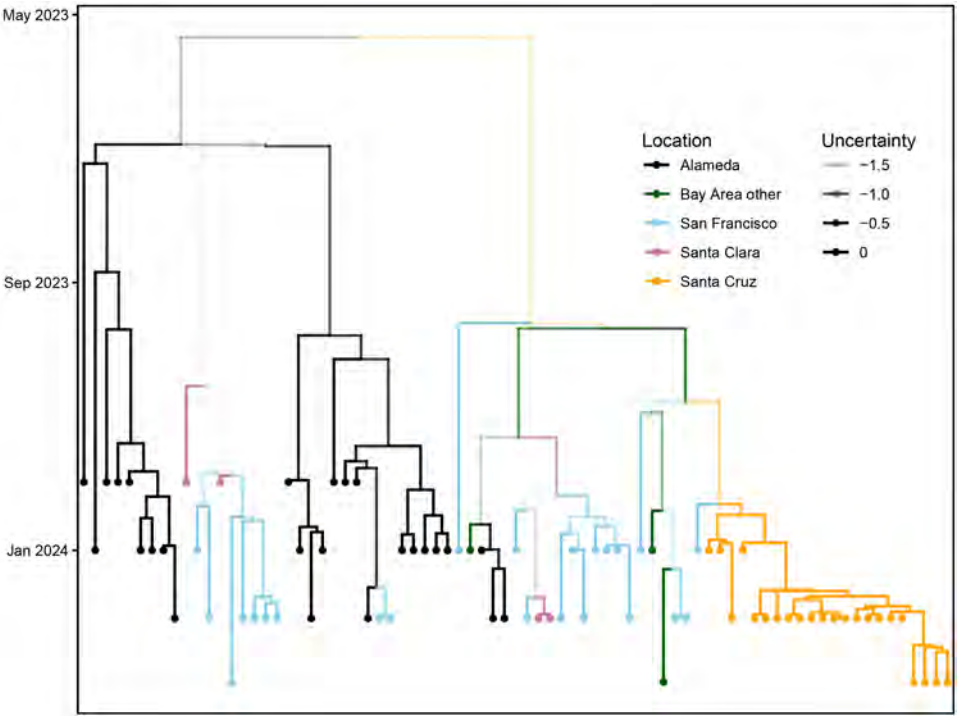


Figure 5. Phylogenetic tree of *Shigella sonne*i outbreak isolates in the San Francisco Bay, California, USA, area with spatiotemporal metadata and tree uncertainty, 2023–2024. Branches are colored according to location. The opacity of the branches is equal to the uncertainty of the placement of each branch. Phylodynamic methods are incorporated into phylogenetic trees with time and location.

ascertainment of transmission, and the inability to gather complete data on risk factors to link specific case manifestations, symptoms, or other factors associated with shigellemia or its mode of transmission is a limitation of our study. Ideally, genomic sequencing paired with epidemiologic information gathered, such as case manifestation, risk factors identified, and symptoms, can provide improved insights into the drivers of transmission. This information can be particularly helpful when investigating outbreaks in communities such as persons experiencing homelessness, when epidemiologic information may be limited. We recommend public health prevention measures focus on the proper maintenance, routine disinfection, and cleaning of public restroom facilities and handwashing stations, particularly in places that are frequented by persons experiencing homelessness.

N.F.M. is supported in part by a Noyce initiative award and the CDC CFA grant 1NU38FT00007.

This activity was reviewed by CDC, deemed not research, and was conducted consistent with applicable federal law and CDC policy.

About the Author

Mr. Lloyd is a public health microbiologist with the Alameda County Department of Public Health and a PhD student at the University of California, Berkeley. His primary research interests are antimicrobial resistance and bacterial evolution through a public health lens. Dr. Khan is an Epidemic Intelligence Service Officer with the Centers for Disease Control and Prevention. She is stationed in Alameda County, California.

References

1. Rotramel HE, Zamir HS. Shigella bacteremia in an immunocompetent patient. Cureus. 2021;13:e19778.
2. Stefanovic A, Matic N, Ritchie G, Lowe CF, Leung V, Hull M, et al. Multidrug-resistant *Shigella sonnei* bacteremia among persons experiencing homelessness, Vancouver, British Columbia, Canada. Emerg Infect Dis. 2023;29:1668–71. https://doi.org/10.3201/eid2908.230323
3. Hawkey J, Paranagama K, Baker KS, Bengtsson RJ, Weill FX, Thomson NR, et al. Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, *Shigella sonnei*. Nat Commun. 2021;12:2684. https://doi.org/10.1038/s41467-021-22700-4
4. Libuit KG, Doughty EL, Otieno JR, Ambrosio F, Kapsak CJ, Smith EA, et al. Accelerating bioinformatics implementation in public health. Microb Genom. 2023;9:mgen001051. https://doi.org/10.1099/mgen.0.001051
5. Ribot EM, Freeman M, Hise KB, Gerner-Smidt P. PulseNet: entering the age of next-generation sequencing. Foodborne Pathog Dis. 2019;16:451–6. https://doi.org/10.1089/fpd.2019.2634
6. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLOS Comput Biol. 2019;15:e1006650. https://doi.org/10.1371/journal.pcbi.1006650
7. Müller NF, Rasmussen D, Stadler T. MASCOT: parameter and state inference under the marginal structured coalescent approximation. Bioinformatics. 2018;34:3843–8. https://doi.org/10.1093/bioinformatics/bty406
8. Mitchell HD, Mikhail AFW, Painset A, Dallman TJ, Jenkins C, Thomson NR, et al. Use of whole-genome sequencing to identify clusters of *Shigella flexneri* associated with sexual transmission in men who have sex with men in England: a validation study using linked behavioural data. Microb Genom. 2019;5:e000311. https://doi.org/10.1099/mgen.0.000311

Address for correspondence: Tyler Lloyd or Kavita R. Trivedi, Alameda County Public Health Department, 1100 San Leandro Blvd, San Leandro, CA 94577, USA; email: tyler_lloyd@berkeley.edu or kavita.trivedi@acgov.org

Successful Transition to Whole-Genome Sequencing and Bioinformatics to Identify Invasive *Streptococcus* spp. Drug Resistance, Alaska, USA

Karen M. Miernyk, Sopio Chochua, Ben Metcalf, Alisa Reasonover, Brenna Simons-Petrusa

The Centers for Disease Control and Prevention's Arctic Investigations Program evaluated whole-genome sequencing (WGS) workflows and bioinformatics pipelines developed by the Centers' *Streptococcus* Laboratory. We compared WGS-based antimicrobial drug resistance predictions with phenotypic testing for group B (n = 130) and group A (n = 217) *Streptococcus* and *Streptococcus pneumoniae* (n = 293). Isolates were collected in Alaska during January 2019–February 2021. We also included a historical phenotypically nonsusceptible subset. Concordances between phenotypic testing and WGS predictions were 99.9% (895/896) for group B *Streptococcus*, 100% (1,298/1,298) for group A *Streptococcus*, and 99.98% (3,516/3,517) for *S. pneumoniae*. Common resistance determinants were *ermTR*, *ermB*, and *mef* for macrolides, *tetM* for tetracyclines, and g*yrA* and *parC* for levofloxacin. *S. pneumoniae* trimethoprim/sulfamethoxazole nonsusceptibility was associated with *folP* gene insertions and *folA* mutations. In 2022, the Arctic Investigations Program transitioned *Streptococcus* spp. workflows to WGS, enabling more rapid monitoring and prevention of invasive disease.

The Arctic Investigations Program (AIP), the Centers for Disease Control and Prevention (CDC) infectious disease field station in Alaska, USA, began surveillance of invasive pneumococcal disease (IPD) in 1986 (*1*) and added invasive cases of *Streptococcus agalactiae* (group B *Streptococcus*; GBS) and *S. pyogenes* (group A *Streptococcus*; GAS) in 2000 (*2*,*3*). Clinical laboratories across Alaska send case isolates to AIP for species confirmation and strain characterization. Data from AIP's Invasive Bacterial Disease Surveillance (IBDS) are critical for understanding disease patterns. Alaska Native persons have higher rates of S*treptococcus* spp. infections than non–Alaska Native persons (*3*,*4*). In 2015, CDC's Office of Advanced Molecular Detection, National Center for Emerging and Zoonotic Infectious Diseases, granted AIP funds to develop a technical and bioinformatics infrastructure and workforce to operationalize microbial genomics and enhance AIP's ability to detect outbreaks, provide information about genetic lineage, and identify genetic determinants associated with antimicrobial drug resistance and virulence.

CDC's *Streptococcus* Laboratory, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, transitioned its Active Bacterial Core Surveillance workflows to whole-genome sequencing (WGS) in 2015 (*5*–*7*). WGS analyses enable antimicrobial drug susceptibility predictions and strain typing, establish mechanisms of resistance, identify genotypes, and characterize surface protein antigens without requiring the labor and specialized skills needed for conventional phenotypic characterization. WGS also contributes to outbreak and disease transmission investigations (*8*–*10*). After a 2016–2017 GAS *emm* type 26.3 outbreak investigation (*8*), AIP began discussing WGS technology transfer with the *Streptococcus* Laboratory. We describe WGS validation, antimicrobial susceptibility mechanisms, and next steps for *Streptococcus* spp. WGS at AIP. This work was reviewed by a CDC research review board, which deemed it not research.

Author affiliations: Centers for Disease Control and Prevention, Anchorage, Alaska, USA (K.M. Miernyk, A. Reasonover, B. Simons-Petrusa); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (S. Chochua, B. Metcalf)

DOI: https://doi.org/10.3201/eid3113.241828

Materials and Methods

We validated isolates collected during 2019 and during January–February 2021. We also included a subset of isolates from 1986–2018 (*S. pneumoniae*) and 2000–2018 (GAS and GBS) because phenotypic testing showed those isolates were nonsusceptible to ≥1 antimicrobial drug. AIP had determined antimicrobial drug MICs, *S. pneumoniae* serotypes, and GAS *emm* types previously by using phenotypic microbiologic methods. We cultured isolates according to a previously established protocol (Appendix, https://wwwnc.cdc.gov/EID/article/31/13/24-1828-App1.pdf). We extracted genomic DNA and prepared DNA libraries by using Nextera DNA Flex with 96 dual indices (Illumina, https://www.illumina.com). We pooled libraries and performed WGS by using a MiSeq instrument and MiSeq v2 500 cycle reagent kit (Illumina). We used bioinformatic pipelines developed and validated by the CDC *Streptococcus* Laboratory (https://github.com/BenJamesMetcalf) (Appendix) (*5*–*7*).

We compared MICs from phenotypic testing with MICs predicted by WGS. We considered results to be preliminarily discordant when MICs between phenotypic and WGS methods differed by ≥2 dilutions. We retested isolates with discordant results by using Etest (bioMérieux, https://www.biomerieux.com), BD BBL Sensi-Disc for the D-Zone (BD, https://www.bd.com), or other disk diffusion methods (BD). We confirmed results were discordant when the original phenotypic result agreed with follow-up testing but continued to differ from the WGS prediction by ≥2 dilutions.

For GAS isolates, we compared *emm* type results from Sanger sequencing with WGS assignments. We extracted DNA again from isolates that had preliminary discordant results and tested all 3 extracts (original extract, extract for WGS, and reextraction) by using Sanger sequencing. We confirmed those results were discordant if the Sanger sequencing results continued to differ from the WGS assignment.

For *S. pneumoniae* isolates, we compared phenotypic serotype results with WGS assignments. We tested those isolates with preliminary discordant results by using the Immulex Pneumotest (SSI Diagnostica A/S, https://ssidiagnostica.com) with Quellung reaction confirmation. We confirmed results were discordant when the follow-up testing agreed with the original serotype result and continued to differ from the WGS assignment.

Results

We tested 130 GBS (82 from 2019/2021, 48 historical), 217 GAS (169 from 2019/2021, 48 historical), and 293 *S. pneumoniae* (203 from 2019/2021, 90 historical) isolates. Initial comparisons showed a preliminary concordance of 99.4% (891/896) for GBS, 99.2% (1,288/1,298) for GAS, and 98.7% (3,470/3,517) for *S. pneumoniae*. Follow-up testing confirmed the WGS prediction for 4/5 (80%) GBS isolates, 10/10 (100%) GAS isolates, and 46/47 (97.9%) *S. pneumoniae* isolates. Final concordance was 99.9% (895/896) for GBS, 100% (1,298/1,298) for GAS, and 99.98% (3,516/3,517) for *S. pneumoniae*.

Initial phenotypic analysis showed 192 nonsusceptible results for GBS isolates (Table 1). Two GBS isolates exhibiting high (MIC ≥8 μg/mL) levofloxacin resistance contained amino acid substitutions in the GyrA subunit of DNA gyrase (S81L) and the ParC subunit of topoisomerase IV, of which 1 had S79F and 1 had S79Y. Tetracycline nonsusceptibility was most often caused by the *tetM* determinant (n = 37); the *tetO* determinant accounted for the other 7 instances. The 5 preliminary discordant results for GBS occurred with erythromycin (n = 2) and clindamycin (n = 3) (Table 2); follow-up testing confirmed the WGS prediction for 4/5 (80%) isolates. Most combined erythromycin and clindamycin nonsusceptibility was associated with the presence of the 23S rRNA methylase genes, *ermTR* (n = 42) or *ermB* (n = 19). All *ermB*-positive isolates were constitutively clindamycin resistant, whereas 12/42 (28.6%) *ermTR*-positive isolates were

Table 1. Initial phenotypic testing results for isolates included in whole-genome sequencing workflow validation to identify invasive *Streptococcus* spp. drug resistance in Alaska, USA, during 2000–2021*

	Streptococcus agalactiae isolates			*S. pyogenes* isolates		
Antimicrobial drug	Total no.	No. susceptible	No. nonsusceptible	Total no.	No. susceptible	No. nonsusceptible
Ampicillin	112	112	0	165	165	0
Cefotaxime	45	45	0	34	34	0
Clindamycin	118	51	67	180	108	72
Erythromycin	119	40	79	180	99	81
Levofloxacin	118	116	2	181	177	4
Linezolid	103	103	0	176	176	0
Penicillin	115	115	0	166	166	0
Tetracycline	47	3	44	35	2	33
Vancomycin	119	119	0	181	181	0

*130 *S. agalactiae* and 217 *S. pyogenes* isolates were analyzed.

Table 2. Discordance between phenotypic testing and whole-genome sequence predictions along with resistance determinants for *Streptococcus* spp. in study to identify invasive *Streptococcus* spp. drug resistance, Alaska, USA, 1986–2021*

Organism resistance mechanism	Antimicrobial drug	WGS predicted MIC, μg/mL	Initial phenotypic MIC, μg/mL	Follow-up testing for ≥2 dilution discrepancy		
				MIC, μg/mL	Agrees with WGS, no.	True discrepancy, no.
Streptococcus agalactiae						
lsaC gene	Clindamycin	≥1, R	0.25, S	0.25, S	0	1
negative	Clindamycin	≤0.25, S	2, R	0.064, S	1	0
	Clindamycin	≤0.25, S	2, R	0.047, S	1	0
	Erythromycin	≤0.25, S	0.5, I	0.047, S	1	0
	Erythromycin	≤0.25, S	8, R	0.064, S	1	0
S. pyogenes						
ermTR gene	Clindamycin†	≥1, R	≤0.12, S	All D-Zone +	8	0
Negative	Tetracycline	≤2, S	4, I	0.125, S	1	0
	Erythromycin	≤0.25, S	2, R	0.064, S	1	0
S. pneumoniae						
mef gene	Erythromycin	8, R	0.12, S	4, R	1	0
negative	Chloramphenicol	≤2, S	8, R	2, S	1	0
	Levofloxacin	≤2, S	4, I	0.5, S	1	0
	Erythromycin	0.06, S	1, R	0.06, S	1	0
	Tetracycline	≤0.25, S	4, I	0.125, S	1	0
	Tetracycline	≤0.25, S	>16, R	0.25, S	1	0
	Tetracycline	≤0.25, S	>16, R	0.125, S	1	0
	Quinupristin, dalfopristin	≤1, S	>4, R	Sensitive‡	13	0
	Rifampin	≤1, S	>4, R	≤0.064, S	24	0

*I, intermediate resistance; R, resistant; S, susceptible; WGS, whole-genome sequencing; +, positive.
†Follow-up testing for clindamycin used BD BBL Sensi-Disc for the D-Zone test (BD, https://www.bd.com).
‡Determined by disk diffusion.

inducibly clindamycin resistant. Erythromycin resistance and clindamycin susceptibility (M phenotype) were detected in 13 isolates with the *mef*-positive genotype. Two isolates contained the *lsaC* gene, which confers resistance to clindamycin. One GBS isolate with a discordant result contained the *lsaC* gene but was phenotypically sensitive to clindamycin. Three isolates constitutively resistant to erythromycin and clindamycin contained multiple resistance mechanisms, 1 each of *ermB* plus *lsaC*, *ermB* plus *mef*, and *lsaC* plus *mef*.

Initial phenotypic analysis showed 190 nonsusceptible results for GAS isolates (Table 1). Three isolates with the S81F substitution in ParC had intermediate levofloxacin resistance; 1 isolate with high (MIC ≥8 μg/mL) levofloxacin resistance contained amino acid substitutions in both GyrA (S81Y) and ParC (D85N) protein subunits. We identified preliminary discordant GAS results when comparing tetracycline (n = 1), clindamycin (n = 8), and erythromycin (n = 1) (Table 2); follow-up testing confirmed the WGS predictions for all 10 of those isolates. All 32 isolates nonsusceptible to tetracycline contained the *tetM* determinant. Most combined erythromycin and clindamycin nonsusceptibility was associated with the presence of *ermTR* (n = 71), *ermB* (n = 5), or *ermT* (n = 1) gene determinants. All *ermB*-positive and *ermT*-positive isolates were constitutively clindamycin resistant, whereas 29/71 (40.8%) *ermTR*-positive isolates were inducibly clindamycin resistant. Three isolates constitutively resistant to erythromycin and clindamycin contained >1 resistance mechanism, *ermT* plus *ermTR* (n = 2) and *ermTR* plus *lsaC* (n = 1).

We completed *emm* typing by Sanger sequencing for 201 GAS isolates, identifying 37 *emm* types. The initial comparison between Sanger sequencing and WGS showed 199/201 (99.0%) *emm* type concordance. Sanger sequencing of a third DNA extraction from both isolates confirmed 100% WGS concordance.

Initial phenotypic analysis showed 467 nonsusceptible results for *S. pneumoniae* (Table 3). At initial comparison, 13 isolates were nonsusceptible to quinupristin/dalfopristin and 24 isolates were nonsusceptible to rifampin; follow-up testing confirmed the WGS prediction for all 37 isolates (Table 2). One isolate exhibited high (MIC ≥8 μg/mL) fluoroquinolone resistance and had amino acid substitutions in both GyrA (S81F) and ParC (S79F) subunits. All 15 isolates containing the chloramphenicol acetyl transferase (*cat*) gene were resistant to chloramphenicol, and all 35 isolates nonsusceptible to tetracycline contained the *tetM* determinant. Three isolates without a WGS-identified tetracycline resistance mechanism were phenotypically nonsusceptible to tetracycline at initial testing; follow-up testing confirmed the WGS prediction for all 3 isolates.

We detected nonsusceptibility to trimethoprim/sulfamethoxazole in 111 *S. pneumoniae* isolates; 62 of those had an insertion of 1 or 2 codons within the *folP* gene, conferring intermediate drug resistance. The

Table 3. Initial phenotypic testing results for 293 *Streptococcus pneumoniae* isolates used for whole-genome sequencing workflow validation in Alaska in study to identify invasive *Streptococcus* spp. drug resistance, 1986–2021

Antimicrobial drugs	Total no. isolates	No. susceptible isolates	No. nonsusceptible isolates
Cefotaxime	75	38*	37*
Ceftriaxone	278	271*	26*
Chloramphenicol	293	278	15
Clindamycin	274	248	26
Erythromycin	291	221	70
Levofloxacin	277	276	1
Linezolid	271	271	0
Meropenem	278	247	31
Penicillin	293	218*	75*
Rifampin	246	222	24
Quinupristin/dalfopristin	275	262	13
Tetracycline	87	49	38
Trimethoprim/sulfamethoxazole	286	175	111
Vancomycin	293	293	0

*Determined by using the Clinical and Laboratory Standards Institute (https://www.clsi.org) breakpoint for meningitis cases.

remaining 49 isolates were double mutants, leading to full resistance consisting of a *folA* gene mutation (I100L amino acid substitution) and insertion of 1 or 2 codons within *folP*. Most combined erythromycin and clindamycin nonsusceptibility was associated with the presence of *ermB* (n = 16); all of those were constitutively clindamycin resistant. An additional 3 isolates were *ermB* positive, phenotypically erythromycin nonsusceptible, but were not tested phenotypically for clindamycin nonsusceptibility. Ten isolates constitutively resistant to erythromycin and clindamycin were *ermB* and *mef* positive; 41 isolates with the M phenotype were *mef* positive. Two isolates were initially discordant for erythromycin sensitivity; 1 was *mef* positive but phenotypically sensitive, and 1 did not contain a resistance mechanism but was phenotypically nonsusceptible. Follow-up testing confirmed the WGS predication for both isolates.

Most *S. pneumoniae* isolates had MIC predictions related to penicillin-binding protein (PBP) gene types that indicated sensitivity to ceftriaxone (n = 271), meropenem (n = 247), penicillin (n = 218), and cefotaxime (n = 38) (Table 3). Penicillin and cefotaxime MIC predictions were concordant with phenotypic testing for all *S. pneumoniae* isolates. One initial discordant result for meropenem and 3 initial discordant results for ceftriaxone were observed; all 4 isolates were phenotypically sensitive, but WGS predicted nonsusceptibility (Table 4). Follow-up testing confirmed the WGS prediction for 3/4 (75%) isolates; 1 isolate was predicted to be nonsusceptible to ceftriaxone but was sensitive according to both the initial and additional phenotypic testing.

Quellung serotyping was completed for 258 *S. pneumoniae* isolates, which comprised 30 serotypes. During the initial comparison, 2 discordant results were observed, likely related to an isolate mixup during phenotypic testing; additional testing confirmed the WGS-assigned result for both, indicating 100% concordance.

Discussion

Despite data published by the CDC's *Streptococcus* Laboratory supporting the accuracy of WGS-based antimicrobial drug susceptibility predictions (*5*–*7*), some collaborators have not pursued WGS workflows because they are uncertain those predictions are accurate (K.M. Miernyk, unpub. data). The antimicrobial susceptibility data shown here provide further evidence to address those concerns. For the 3 *Streptococcus* spp., we found 99.96% (5,709/5,711) concordance between phenotypic testing and genomic predictions. In addition, the WGS predictions were more accurate than phenotypic testing. Of 62 initial discordant results, WGS was confirmed to be correct for 60 (97%) of those.

Antimicrobial drug susceptibility data are needed to inform patient treatment and develop population treatment guidelines. IPD disproportionately impacts Alaska Native persons. AIP's IBDS indicated the IPD

Table 4. Discordant *Streptococcus pneumoniae* β-lactam resistance predicted by penicillin-binding protein sequences compared with phenotypic testing results, Alaska, 1986–2021*

	WGS predicted	Initial phenotypic	Follow-up testing for ≥2 dilution discrepancy		
Antimicrobial drug	MIC, μg/mL	MIC, μg/mL	MIC, μg/mL	Agrees with WGS, no.	True discrepancy, no.
Meropenem	1, R	0.25, S	1, R	1	0
Ceftriaxone†	1, I/S	<0.5, S/S	1, I/S	2	0
	1, I/S	<0.5, S/S	0.019, S/S	0	1

*I, intermediate resistance; R, resistant; S, susceptible; WGS, whole-genome sequencing.
†Determined by using the Clinical and Laboratory Standards Institute (https://www.clsi.org) breakpoints for meningitis/nonmeningitis cases.

rates associated with serotypes not targeted by the licensed 13-valent pneumococcal conjugate vaccine (PCV13) were 27.2/100,000 Alaska Native persons and 6.7/100,000 non–Alaska Native persons during April 2010–December 2013 (*4*). Similarly, the percentage of persons carrying non-PCV13 serotype *S. pneumoniae* in their nasopharynx that was nonsusceptible to erythromycin or penicillin increased significantly from 16.8% (n = 709) during 2008–2011 to 26.5% (n = 1,466) during 2012–2015 (*11*). This finding suggests Alaska Native persons might be at increased risk for IPD caused by *S. pneumoniae* that are resistant to commonly used antimicrobial drugs. Capsular *S. pneumoniae* serotyping data are necessary to evaluate vaccines for IPD prevention. Finally, GAS *emm* typing data predict the M protein serotype (*12*), a potential target for vaccines to prevent invasive GAS infections. Traditional phenotypic methods are time consuming and require many different specialized technical skills, reagents, and consumables. The WGS workflow described here provides those data for 48 samples simultaneously by using 1 method. In addition, as the COVID-19 pandemic revealed, supply chains can be unreliable. WGS provides data in a single workflow, enabling a more streamlined process with fewer consumables and reagents.

AIP has not previously had the capacity to consistently characterize resistance mechanisms in any *Streptococcus* spp. bacteria. To better elucidate changes in macrolide resistance after PCV13 introduction, we briefly used PCR to characterize *ermB* and *mef* macrolide resistance mechanisms in *S. pneumoniae* (*13*). However, we have not investigated those mechanisms in GAS or GBS collected from persons in Alaska, which could be of particular importance for GAS. Candidate GAS vaccines target the M protein (*14*), and >275 known M types exist. Therefore, vaccine pressure on targeted M types could affect circulating strains of GAS, which has been observed for *S. pneumoniae* after PCV13 introduction (*15*). Macrolide nonsusceptibility is not uncommon for GAS, and it will be critical to understand whether changes in nonsusceptibility after vaccine introduction are caused by expansion of existing strains, by introduction of new strains, or by some other mechanism.

In conclusion, our antimicrobial susceptibility, serotype, and *emm* type validation data confirm the accuracy of WGS-based predictions for GAS, GBS, and *S. pneumoniae* when performed at AIP. The single WGS workflow is more efficient than multiple workflows needed for phenotypic testing. WGS pipelines identify previously unknown genotypic mechanisms for nonsusceptibility of *Streptococcus* spp. isolates collected in Alaska and provide additional data, such as GBS serotypes and multilocus sequence types. WGS also provides a genetic sequence for every isolate, which is available for future investigations. In 2022, AIP transitioned all IBDS workflows for *Streptococcus* spp. to WGS, and we continue to improve those processes. We have decreased cost by sequencing more extracts on a flow cell and have been increasing local analysis capabilities and data storage for more rapid and local pathogen detection. We also perform biannual phenotypic testing on a subset of isolates to monitor for new resistance genes. Future work includes validating a WGS workflow for GAS that can be used in remote field settings, enabling AIP to provide more rapid outbreak response.

Acknowledgments

We thank clinical microbiology laboratory personnel from across Alaska who submitted isolates to the Alaska IBDS program and the AIP laboratory team members for processing and phenotypically testing the isolates chosen for validation.

This work was funded, in part, by the CDC's Advanced Molecular Detection initiative.

According to agreements with Alaska Native Tribal Health leaders, genomic sequences generated in this work were not submitted into public-facing, open access databases.

About the Author

Ms. Miernyk is a biologist and laboratory manager in the Arctic Investigations Program, Division of Infectious Disease Readiness and Innovation, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, in Anchorage, Alaska, USA. Her research interests focus on preventing diseases caused by *Streptococcus* spp., *Haemophilus influenz*ae, *Neisseria meningitidis*, *Helicobacter pylori*, respiratory syncytial virus, influenza viruses, SARS-CoV-2, and other respiratory pathogens.

References

1. Davidson M, Parkinson AJ, Bulkow LR, Fitzgerald MA, Peters HV, Parks DJ. The epidemiology of invasive pneumococcal disease in Alaska, 1986–1990—ethnic differences and opportunities for prevention. J Infect Dis. 1994;170:368–76. https://doi.org/10.1093/infdis/170.2.368
2. Castrodale L, Gessner B, Hammitt L, Chimonas MA, Hennessy T. Invasive early-onset neonatal group B streptococcal cases—Alaska, 2000–2004. Matern Child Health J. 2007;11:91–5. https://doi.org/10.1007/s10995-006-0144-5
3. Rudolph K, Bruce MG, Bruden D, Zulz T, Reasonover A, Hurlburt D, et al. Epidemiology of invasive group A streptococcal disease in Alaska, 2001 to 2013. J Clin Microbiol. 2016;54:134–41. https://doi.org/10.1128/JCM.02122-15

4. Bruce MG, Singleton R, Bulkow L, Rudolph K, Zulz T, Gounder P, et al. Impact of the 13-valent pneumococcal conjugate vaccine (PCV13) on invasive pneumococcal disease and carriage in Alaska. Vaccine. 2015;33:4813–9. https://doi.org/10.1016/j.vaccine.2015.07.080
5. Chochua S, Metcalf BJ, Li Z, Rivers J, Mathis S, Jackson D, et al. Population and whole genome sequence based characterization of invasive group A streptococci recovered in the United States during 2015. mBio. 2017;8:e01422-17. https://doi.org/10.1128/mBio.01422-17
6. Metcalf BJ, Chochua S, Gertz RE Jr, Hawkins PA, Ricaldi J, Li Z, et al. Short-read whole genome sequencing for determination of antimicrobial resistance mechanisms and capsular serotypes of current invasive *Streptococcus agalactiae* recovered in the USA. Clin Microbiol Infect. 2017;23:574.e7–14. PubMed https://doi.org/10.1016/j.cmi.2017.02.021
7. Metcalf BJ, Chochua S, Gertz RE Jr., Li Z, Walker H, Tran T, et al. Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. Clin Microbiol Infect. 2016;22:1002.e1–8. PubMed https://doi.org/10.1016/j.cmi.2016.08.001
8. Mosites E, Frick A, Gounder P, Castrodale L, Li Y, Rudolph K, et al. Outbreak of invasive infections from subtype emm26.3 group A *Streptococcus* among homeless adults—Anchorage, Alaska, 2016–2017. Clin Infect Dis. 2018;66:1068–74. https://doi.org/10.1093/cid/cix921
9. Nolen LD, DeByle C, Topaz N, Simons BC, Tiffany A, Reasonover A, et al. Genomic diversity of *Haemophilus influenzae* serotype a in an outbreak community—Alaska, 2018. J Infect Dis. 2022;225:520–4. https://doi.org/10.1093/infdis/jiab376
10. Nolen LD, Topaz N, Miernyk K, Bressler S, Massay SC, Geist M, et al. Evaluating a cluster and the overall trend of invasive *Haemophilus influenzae* serotype b in Alaska 2005–2019. Pediatr Infect Dis J. 2022;41:e120–5. https://doi.org/10.1097/INF.0000000000003470
11. Plumb ID, Gounder PP, Bruden DJT, Bulkow LR, Rudolph KM, Singleton RJ, et al. Increasing non-susceptibility to antibiotics within carried pneumococcal serotypes—Alaska, 2008–2015. Vaccine. 2020;38:4273–80. https://doi.org/10.1016/j.vaccine.2020.04.048
12. Beall B, Facklam R, Thompson T. Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci. J Clin Microbiol. 1996;34:953–8. https://doi.org/10.1128/jcm.34.4.953-958.1996
13. Rudolph K, Bulkow L, Bruce M, Zulz T, Reasonover A, Harker-Jones M, et al. Molecular resistance mechanisms of macrolide-resistant invasive *Streptococcus pneumoniae* isolates from Alaska, 1986 to 2010. Antimicrob Agents Chemother. 2013;57:5415–22. https://doi.org/10.1128/AAC.00319-13
14. Fan J, Toth I, Stephenson RJ. Recent scientific advancements towards a vaccine against group A *Streptococcus*. Vaccines (Basel). 2024;12:272. https://doi.org/10.3390/vaccines12030272
15. Singleton RJ, Hennessy TW, Bulkow LR, Hammitt LL, Zulz T, Hurlburt DA, et al. Invasive pneumococcal disease caused by nonvaccine serotypes among Alaska native children with high levels of 7-valent pneumococcal conjugate vaccine coverage. JAMA. 2007;297:1784–92. https://doi.org/10.1001/jama.297.16.1784

Address for correspondence: Karen Miernyk, Centers for Disease Control and Prevention, 4055 Tudor Centre Dr, Anchorage, AK 99508, USA; email: kmiernyk@cdc.gov

Genomic Characterization of *Escherichia coli* O157:H7 Associated with Multiple Sources, United States

Joseph S. Wirth, Molly M. Leeper, Peyton A. Smith, Michael Vasser, Lee S. Katz, Eshaw Vidyaprakash, Heather A. Carleton, Jessica C. Chen

In the United States, Shiga toxin–producing *Escherichia coli* (STEC) outbreaks cause >265,000 infections and cost $280 million annually. We investigated REPEXH01, a persistent strain of STEC O157:H7 associated with multiple sources, including romaine lettuce and recreational water, that has caused multiple outbreaks since emerging in late 2015. By comparing the genomes of 729 REPEXH01 isolates with those of 2,027 other STEC O157:H7 isolates, we identified a highly conserved, single base pair deletion in *espW* that was strongly linked to REPEXH01 membership. The biological consequence of that deletion remains unclear; further studies are needed to elucidate its role in REPEXH01. Additional analyses revealed that REPEXH01 isolates belonged to Manning clade 8; possessed the toxins *stx2a*, *stx2c*, or both; were predicted to be resistant to several antimicrobial compounds; and possessed a diverse set of plasmids. Those factors underscore the need to continue monitoring REPEXH01 and clarify aspects contributing to its emergence and persistence.

Shiga toxin–producing *Escherichia coli* (STEC) outbreaks associated with produce were first identified in 1991, and the trend of produce-associated STEC outbreaks remains prevalent, among which romaine lettuce is the most common leafy green outbreak vehicle (*1*–*4*). Each year in the United States, >265,000 STEC infections occur, costing $280 million and resulting in ≈3,600 hospitalizations and ≈30 deaths (*4*,*5*). *E. coli* O157:H7, a specific serotype of STEC, causes ≈25% of those infections and ≈67% of all STEC deaths (*5*). STEC O157:H7 infections often induce abdominal cramps, vomiting, and bloody diarrhea. In particularly severe cases, a rare condition known as hemolytic uremic syndrome (HUS) develops, which can cause anemia, acute renal failure, and death (*6*). STEC O157:H7 outbreaks are commonly linked to consumption of leafy greens or beef. Although nearly 60% of STEC O157:H7 infections have been attributed to vegetable row crops, a category that includes leafy greens, ruminants, especially cattle, are the suspected primary STEC O157:H7 reservoir (*7*–*9*). During 2009–2018, 32 STEC O157:H7 outbreaks in the United States and Canada were linked to contaminated leafy greens (*4*).

Since April 2017, nine separate outbreaks of the same strain of STEC O157:H7, hereafter referred to as REPEXH01, have occurred (Table 1). A large REPEXH01 outbreak affecting 37 states occurred in 2018, from which 238 STEC O157:H7 infections, 104 hospitalizations, 28 cases of HUS, and 5 deaths were reported (*3*). Most (85%) interviewed patients reported consuming romaine lettuce, and a subsequent investigation linked those infections to romaine lettuce grown in the Yuma, Arizona, region of the United States (*3*). By March 29, 2024, the United States reported 762 persons in 46 states infected with the REPEXH01 strain, and new infections continue to be identified. In this study, we compared whole-genome sequences of 729 REPEXH01 isolates with 2,027 other STEC O157:H7 isolates to examine genomic factors in REPEXH01 that might have contributed to the emergence and public health impacts of that strain.

Author affiliations: Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA (J.S. Wirth); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (J.S. Wirth, M.M. Leeper, P.A. Smith, M. Vasser, L.S. Katz, E. Vidyaprakash, H.A. Carleton, J.C. Chen)

DOI: https://doi.org/10.3201/eid3113.240686

Table 1. Outbreaks caused by reoccurring STEC O157:H7 strain REPEXH01 in a genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States*

Outbreak	Timeframe	Source†	Origin‡	No. reported illnesses	No. states	No. HUS cases	No. deaths	No. sequences	Allele differences (range)§
A	Apr–May 2017	Unknown	Unknown	9	5	0	0	7	3 (0–8)
B	Jul–Sep 2017	Recreational water¶	California	10	1	4	0	13	0 (0)
C	Mar–Jun 2018	Romaine lettuce¶	Arizona	238	37	28	5	238	4 (0–12)
D	Aug–Oct 2018	Ground beef#	Unknown	12	4	1	0	4	7 (4–10)
E	Oct–Dec 2018	Leafy greens#	Unknown	25	10	4	0	8	7 (1–11)
F	May–Oct 2019	Ground beef#	Unknown	44	12	4	0	44	0 (0–5)
G	Nov 2019	Unknown	Unknown	8	1	0	0	8	0 (0–1)
H	Dec 2020–Mar 2021	Unknown	Unknown	22	7	3	1	22	0 (0–0)
I	Apr–May 2021	Unknown	Unknown	5	3	0	0	5	0 (0–1)

*Outbreak dates are based on reported or estimated illness onset dates. HUS, hemolytic uremic syndrome; REPHEXHO1, recurring strain of STEC O157:H7; STEC, Shiga toxin–producing *Escherichia coli*.
†Confirmed sources were implicated by epidemiologic plus traceback or laboratory data. Suspected sources were implicated by epidemiologic data only (https://www.cdc.gov/foodsafety/outbreaks/multistate-outbreaks/annual-summaries.html).
‡Geographic origin of a confirmed outbreak source might not always be known, which can happen when a food containing multiple ingredients (e.g., bagged salad blend) is confirmed as the source, but the evidence cannot implicate a specific ingredient, or when evidence confirms an outbreak source but traceback cannot pinpoint the exact geographic origin of the source.
§ Values indicate the median allele differences between the isolates of each respective outbreak. Values in parentheses indicate the range of minimum and maximum allele differences between the isolates of each respective outbreak.
¶Confirmed source.
#Suspected source.

Methods

Sequence Selection and Retrieval

We used sequences from 729 REPEXH01 isolates and 598 closely related isolates previously classified as REPEXH01 for this study. All isolates were in PulseNet (https://www.cdc.gov/pulsenet/index.html) and had whole-genome sequences available in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) (Appendix 1, https://wwwnc.cdc.gov/EID/article/31/13/24-0686-App1.pdf). To compare a diverse collection of STEC O157:H7, we randomly selected 1,429 non-REPEXH01 STEC O157:H7 isolates, for a total of 2,756 genomes analyzed. That total accounts for roughly 20% of all 13,778 STEC O157:H7 isolates within PulseNet that had whole-genome sequences available in NCBI as of September 5, 2023. We downloaded whole-genome sequences from GenBank and assemblies and raw reads from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) during May 23–August 1, 2023 (Appendix 2 Table 1, https://wwwnc.cdc.gov/EID/article/31/13/24-0686-App1.xlxs). We used Genbank annotated genomes when available and used Prokka version 1.14.5 (*10*) to annotate SRA genomes that did not have annotations.

Identification of Genomic Features

We used Roary version 3.11.2 (*11*) to perform pangenome analysis on Prokka-annotated genomes, then screened pangenomes for linkage to REPEXH01 isolates by using Scoary version 1.6.16 (*12*). Because those steps are computationally intensive, we used a subset of genomes comprising 181 current and 103 former REPEXH01 isolates and 2 closely related non-REPEXH01 isolates. We identified multiple alleles of espW, a known virulence gene, in that initial dataset and subsequently profiled the expanded dataset (n = 2,756) for those alleles and their association with REPEHX01 (*13,14*) (Appendix 2). We screened assemblies for antimicrobial resistance determinants, plasmid determinants, antimicrobial resistance determinant–associated point mutations, membership in O157 clades (hereafter referred to as Manning clades), and stx subtypes (Appendix 1).

Phylogenetic Reconstruction

From the subset of genomes profiled for pangenome analysis, we constructed a single-nucleotide polymorphism (SNP) analysis by using Lyve-SET version 1.1.4f (https://github.com/lskatz/lyve-SET) (*15*) and presets for *Escherichia* using the single chromosomal contig of 2018C-3602 (BioSample accession no. SAMN08964444) as the reference. We used Gubbins version 3.0.0 (Sanger, https://sanger-pathogens.github.io/gubbins) to generate a recombination-free SNP alignment from the Lyve-SET core alignment (*15,16*). We then generated a time-scaled phylogenetic tree from the SNP alignment for a subset of 286 isolates in BEAST2 version 2.6.3 (*17*), accounting for constant sites and using bModelTest version 1.2.1 (*18*) to average across appropriate substitution models. We used BioNumerics version 7.6.3 (Applied Maths, http://www.applied-maths.com) to construct an

allele-based dendrogram for 2,754 isolates by using UPGMA as the clustering technique. We excluded 2 isolates from the dendrogram because the submitting state agencies had requested those isolates be removed from PulseNet.

Prophage Detection

We detected prophage sequences in the reference genome and categorized their genes by using the PHASTER online phage search tool (*19*,*20*). We used BLASTn version 2.14.0 (https://blast.ncbi.nlm.nih.gov) to search all espW-containing contigs for prophages (Appendix 1).

Obtaining and Visualizing Isolate Metadata

Unless otherwise specified, we obtained all metadata associated with isolates in this study from the System for Enteric Disease Response, Investigation, and Coordination (SEDRIC) (https://www.cdc.gov/foodsafety/outbreaks/tools/sedric.html) or the PulseNet national database (*21*). We visualized data alongside phylogenies by using the Interactive Tree of Life version 5 webtool (https://itol.embl.de) (*22*).

Results

Epidemiology of REPEXH01

All REPEXH01 isolates belonged to Manning clade 8, the clade most strongly correlated with patients developing HUS (*23*,*24*). In fact, nearly every outbreak associated with the REPEXH01 strain included cases of HUS, and an average of 11% (median 9%) of reported illnesses displayed HUS (Table 1). Of the 729 REPEXH01 isolates, all possessed *stx2a*, *stx2c*, or both: 699 (96%) isolates possessed *stx2a*, 574 (79%) possessed *stx2c*, and 544 (75%) possessed both *stx2a* and *stx2c* (Appendix 2 Table 3). Because all REPEXH01 isolates belonged to Manning clade 8, those isolates likely all possessed stx2a, and the absence of *stx2a* in 4% of isolates was likely an artifact of the genome assemblies (*23*,*24*).

Relationship between *espW* and REPEXH01

We performed a preliminary Roary/Scoary pangenome analysis on a subset of 264 isolates, which indicated that the presence of *espW* was linked to membership in REPEXH01, but that same linkage was absent when analyzing the 286 isolates in the time-scaled tree (Figure 1). Closer inspection revealed that espW was in all isolates but often possessed a conserved single base pair deletion, and that deletion appeared to be linked to REPEXH01. We confirmed that hypothesis by analyzing the *espW* alleles in 2,756 isolates, 729 of which were REPEXH01, 598 were former REPEXH01 isolates, and the other 1,429 were a random sampling of all other STEC O157:H7 isolates in the PulseNet database that had publicly available genomes in NCBI (Table 2, Figure 2; Appendix 2 Table 2). We used a χ^2 statistical test, ignoring ambiguous data, to examine the relationship between *espW* alleles and REPEXH01 membership and found the association between those

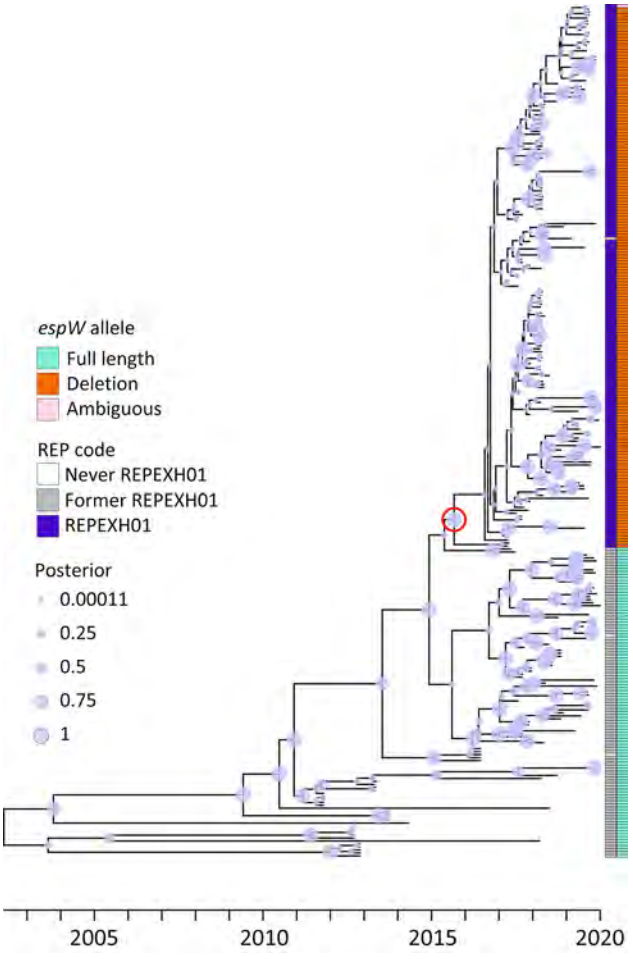


Figure 1. Time-calibrated tree for 286 former and current REPEXH01 isolates with associated metadata used for genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States. The tree was constructed using BEAST2 (https://beast.community) on an alignment of high-quality single-nucleotide polymorphisms. The red circle indicates the most common recent ancestor of the REPEXH01 isolates and corresponds to December 2015. On the right side, the first column indicates the current REPEXH01 isolates (purple), former REPEXH01 isolates (gray), or isolates that were never part of the REPEXH01 definition (white). The second column indicates the *espW* allele: teal indicates the full-length allele, orange indicates the presence of a single base pair deletion; and pink indicates that *espW* is present but the allele could not be determined due to inadequate sequencing data. Circles on the branches indicate the posterior probability. REP, reocurring, emerging, and persistent; REPHEXH01, recurring strain of Shiga toxin–producing *E. coli* O157:H7.

Table 2. Distribution of *espW* alleles in the 2,756 STEC O157:H7 isolates included in a genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States*

REP code	*espW* allele	Count
REPEXH01	Full length	0
	Deletion	727
	Insertion	0
	Ambiguous	2
	Absent	0
non-REPEXH01	Full length	1,892
	Deletion	77
	Insertion	22
	Ambiguous	6
	Absent	30

*Values represent 20% of all STEC O157:H7 isolates in the PulseNet (https://www.cdc.gov/pulsenet/index.html) database with whole-genome sequences available in the National Center for Biotechnology Information Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra); all 729 REPEXH01 isolates with whole-genome sequences are included, alongside 2,027 non-REPEXH01 STEC O157:H7 isolates. The *espW* allele was statistically associated with REPEXH01 membership. REPHEXH01, recurring strain of STEC O157:H7; STEC, Shiga toxin–producing *Escherichia coli*.

variables was significant (p<0.0001). REPEXH01 isolates were more likely to have the deletion than other STEC O157:H7 isolates.

The deletion in *espW* consisted of the loss of a single adenine residue, converting a homopolymer within codons 174–176 from 8 adenine to 7 adenine residues. That deletion introduced a frameshift that ultimately resulted in an early termination codon. We observed insertion of an adenine residue, from 8 to 9 residues, within the same locus in 22 isolates. That insertion also introduced an early termination codon.

REPEXH01 Emergence

Analysis of 286 current and former REPEXH01 isolates revealed that the strain emerged around December 23, 2015 (95% highest posterior density interval March 5, 2015–September 4, 2016), before it was detected in clinical cases in April 2017 (Figure 1). That phylogeny appeared to suggest that members of REPEXH01 shared a common ancestor and the single base pair deletion in espW associated with REPEXH01 appeared to coincide with the emergence of the REPEXH01 strain in late 2015 (Figure 1).

espW Association with STEC O157:H7 Prophages

Examining the gene synteny surrounding *espW* in the reference sequence for REPEXH01 (BioSample accession no. SAMN0896444) showed that many neighboring genes appeared to be of phage origin. Analyzing that genome using PHASTER *(20)* revealed that *espW* was contained within a putative prophage that was most closely related to *Escherichia* phage 500465-1 (GenBank accession no. NC_049342.1) (Figure 3). We examined the genomic regions containing *espW*, and most isolates possessed *espW* within the same putative prophage (Appendix 2 Table 2). Although we detected additional loci in ≈43 isolates, most were of phage origin. Of the 2,626 isolates with assembled contigs that contained *espW*, 87% (n = 2,292) possessed *espW* in or near a putative prophage region (Appendix 2 Table 2). One isolate (SRA accession no. SRR93211959) possessed *espW* directly adjacent to a prophage in what appeared to be an effector exchange locus (13). Another isolate (SRA accession no. SRR6870099) contained espW in a nonprophage region. In the other 332 (13%) isolates, presence of *espW* in a phage-associated region was ambiguous.

Additional REPEXH01 Genomic Features

We evaluated antimicrobial resistance determinants in REPEXH01 (Table 3; Appendix 2 Table 3). REPEXH01 is known to be resistant to several antimicrobial drugs and our dataset confirmed that resistance (*25*). Of note, our results predicted that >99% of REPEXH01 isolates would be resistant to aminoglycosides, folate pathway inhibitors, phenicols, quaternary ammonium compounds, sulfonamides, and tetracylines. However, data predict few isolates would be resistant to cephalosporins (<2%), fluoroquinolones (<1%), or penicillins (<1%).

We also investigated REPEXH01 plasmids (Table 4; Appendix 2 Table 3) and detected ≥1 plasmid replicons in >95% of isolates. Most isolates possessed the IncFIB replicon, IncFIA replicon, or both replicons, but other replicons were not as prevalent. Approximately 9% of isolates contained IncFII replicons, but IncI1-Iγ, IncI2, IncB/O/K/Z, Col, IncX4, and pEC4115 were detected in <5% of isolates.

Discussion

A key finding in this study was identification of a SNP mutation in the espW gene that is largely characteristic of the REPEXH01 strain. The EspW protein has been shown to be secreted by a type III secretion system (T3SS) in E. coli O157:H7 and was previously observed within effector exchange locus (13). Once secreted into the host intestinal epithelial cell, EspW reorganizes host-cell actin in a Rac1-dependent manner to enable extracellular attachment (14). A Pseudomonas syringae homolog of that protein, HopW1, has been shown to solubilize cytosolic actin when injected into plant cells by a T3SS, which disrupts normal localization of proteins and might interfere with the plant immune response (26). The T3SS and secretory proteins such as EspA have been shown to play integral roles in the colonization of the surface of leaves and deeper tissues of the phyllosphere in spinach and

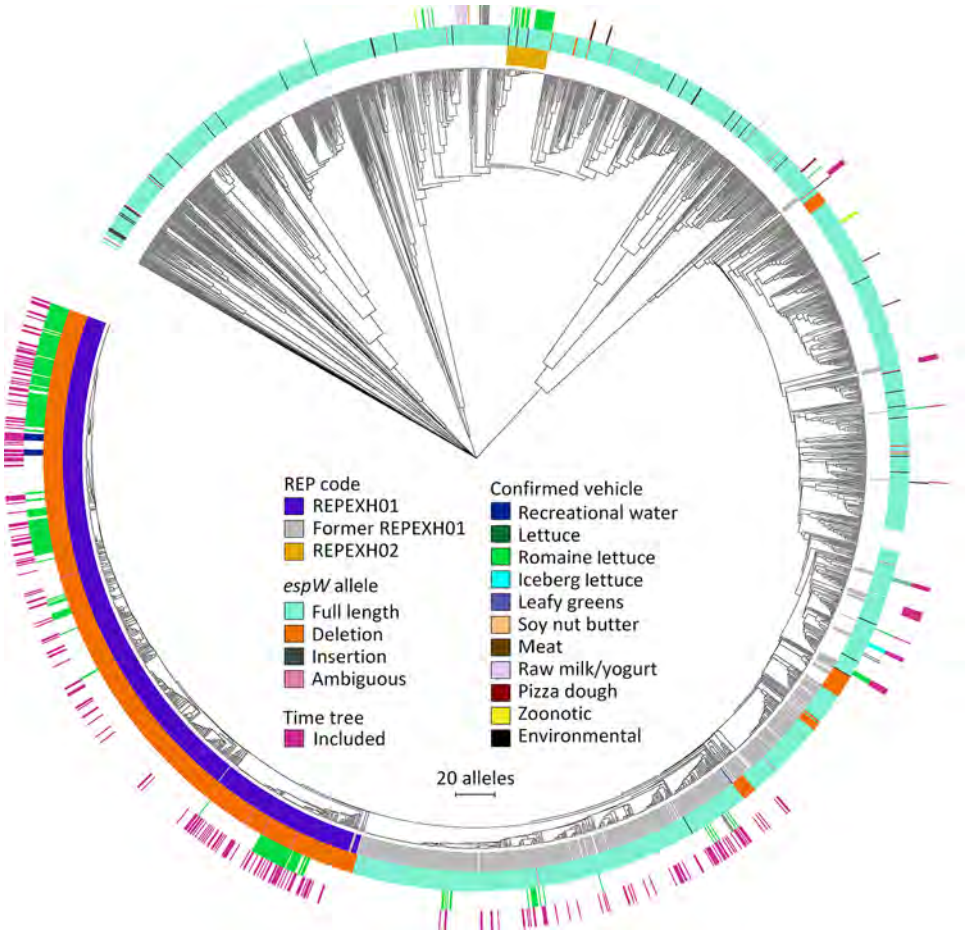


Figure 2. Allele-based core genome multilocus sequence typing dendrogram for 2,752 STEC O157:H7 isolates and associated metadata used for genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States. The dendrogram was constructed in BioNumerics version 7.6.3 (Applied Maths, http://www.applied-maths.com) by using UPGMA as the clustering technique. The innermost track indicates whether an isolate was included in a STEC O157:H7 REP strain. The second innermost track indicates the *espW* allele showing the full-length allele, single base pair deletions or insertion, where *espW* allele could not be determined due to inadequate sequencing data, and where *espW* was not detected in the isolate. The second outermost track indicates vehicles that have been confirmed via epidemiologic investigations of coded outbreaks. The outermost track indicates whether an isolate was included in the time-calibrated tree (Figure 1). An interactive version of this tree is available at https://itol.embl.de/tree/158111236232136169169200 3. The following isolates were omitted from this tree: National Center for Biotechnology Information Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) accession nos. SRR7094189, SRR8956189, SRR7540755, and SRR5588761. The scale bar indicates the number of allele differences between isolates. REP, reocurring, emerging, and persistent; REPHEX01, recurring strain of STEC HO1:H7; STEC, Shiga toxin–producing *Escherichia coli*.

lettuce, where STEC O157:H7 can continue to grow under favorable conditions (27,28).

However, the biological significance of the single base pair deletion in *espW* remains unclear. That deletion could be an example of a gene truncation; another study observed truncations of *espW* in other pathogenic strains of *E. coli* (*14*). Alternatively, the resulting frame shift might silence expression

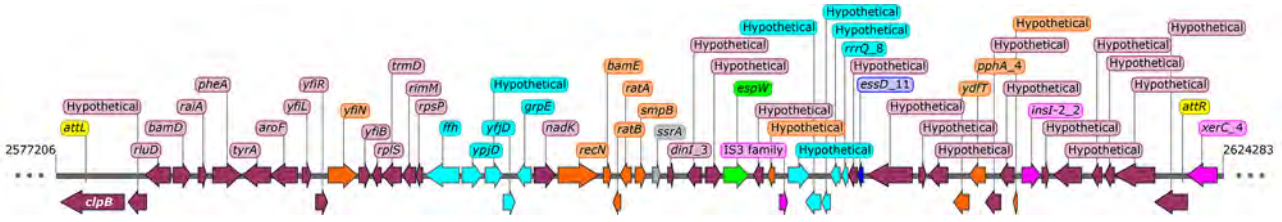


Figure 3. Prophage architecture for most *espW*-containing isolates from a genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States. This locus was annotated using PROKKA version 1.14.5 (https://github.com/tseemann/prokka) and represents a 47,078-bp sequence of Shiga toxin–producing *E. coli* O157:H7 PNUSAE013304 (BioSample accession no. SAMN0896444; SRA accession no. SRR7050023) from positions 2577206–2624283. Arrows indicate genes and their direction indicates the DNA strand. *att* sites were annotated by PHASTER (https://phaster.ca) and are shown in yellow. Genes are colored on the basis of the category assigned by PHASTER: teal indicates phage tail; orange indicates phage-like proteins; pink indicates transposases, integrases, or recombinases; blue indicates phage lysis proteins; gray indicates tRNA; and maroon indicates genes that were not categorized by PHASTER. *espW* is highlighted in green. IS, insertion.

Table 3. Antimicrobial resistance determinants in the 729 REPEXH01 isolates in a genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States*

Antimicrobial class	% Resistant isolates
Aminoglycosides†	99.6
Folate pathway inhibitors‡	99.6
Phenicols§	99.6
Sulfonamides¶	99.6
Quaternary ammonium compounds#	99.6
Tetracyclines**	99.5
Cephalosporins††	1.9
Fluoroquinolones‡‡	0.3
Penicillins§§	0.3

*Antimicrobial resistance determinants were determined by using ResFinder (https://cge.cbs.dtu.dk/services/ResFinder). Resistance was defined by the presence of one or more determinants. REPHEXH01, recurring strain of Shiga toxin–producing *Escherichia coli* O157:H7.
†*aadA1*, *aph(3")-Ib*, and *aph (6)-Id*.
‡*dfrA1* and *dfrA8*.
§*floR*.
¶*sul1* and *sul2*.
#*qacE*.
***tet(A)* and *tet(B)*.
††*bla*CMY-2 and *bla*$_{CTX-M-27}$.
‡‡*qnrB19*.
§§*bla*$_{TEM-1B}$.

of *espW*, or *espW* might be regulated by a homopolymeric tract mechanism where slippage of RNA polymerase could produce heterogenous transcripts, some of which could encode the in-frame functional gene product (*29*). In each of those scenarios, reduced EspW could promote colonization of romaine lettuce through several mechanisms. For example, EspW might elicit an immune response from an infected plant, causing stomata to close, thus restricting access to the interior of leaves by colonizing STEC. Alternatively, EspW could function like HopW1 and cause a more severe infection in plant tissues, lowering the likelihood that the infected leaves are harvested and consumed. Further experiments are required to elucidate the role of that single base pair deletion in REPEXH01 isolates.

Table 4. Plasmid content in the 729 REPEXH01 isolates from a genomic characterization of *Escherichia coli* O157:H7 associated with multiple sources, United States*

Plasmid replicon	% Isolates
IncFIB	92.7
IncFIA	92.3
IncFII†	8.8
IncI1-I(gamma)	4.9
IncI2‡	3.6
IncB/O/K/Z	1.5
Col§	0.8
IncX4	0.7
pEC4115	0.7
IncFIC(FII)	0.3

*Presence of plasmid replicons was determined by using PlasmidFinder (Center for Genomic Epidemiology, https://genomicepidemiology.org). REPHEXH01, recurring strain of Shiga toxin–producing *Escherichia coli* O157:H7.
†IncFII, IncFII(29), IncFII(pCoo), IncFII(pHN7A8), and IncFII(pSE11).
‡IncI2 and IncI2(delta).
§ColE1, ColpVC, Col(KPHS6), Col(MG828), and Col(pHAD28).

In this study, we performed key molecular profiling to provide information on molecular attributes of REPEXH01. Certain *stx* subtypes are associated with more severe disease, and the prevalence of *stx2a* in REPEXH01 highlights the need for surveillance of this strain (*30*). All isolates of this strain belonged to Manning clade 8, the clade most strongly correlated with poor disease outcomes (*23*,*24*). Nearly all REPEXH01 isolates possessed antimicrobial resistance determinants, but that finding does not have direct clinical significance because antimicrobial drugs are not indicated for treating STEC infections because those drugs can increase toxin concentrations in the patient (*25*). However, the plasmids observed in REPEXH01 isolates have been implicated in horizontal gene transfer, and those plasmids were in >95% of REPEXH01 isolates (Appendix 2 Table 3) (*31*). Taken together, those findings suggests that although the presence of antimicrobial resistance determinants has minimal effects on clinical outcomes of STEC infections, and REPEXH01 isolates could still serve as a reservoir of antimicrobial resistance.

Among the limitations of this study, although we included all current and former REPEXH01 isolates in this study, we only screened 20% of the total STEC O157:H7 isolates to decrease the computational demand of the analyses. That subsampling has the potential to bias the data, but the random selection of non-REPEXH01 STEC O157:H7 genomes might alleviate that bias. The genomes used in this study were primarily derived from short-read sequencing technology, and most were at the draft level, indicating that the replicons had not been fully assembled. Although use of draft genomes could result in *espW* being erroneously called absent, steps such as read recruitment using ARIBA (https://github.com/sanger-pathogens/ariba) helped mitigate those potential errors.

REPEXHO1 is a persistent strain of STEC O157:H7 that we estimate emerged in late 2015, before the detection of clinical cases beginning in April 2017. We detected a single base pair deletion in the *espW* virulence gene in >99% of REPEXH01 isolates but in only a few (<4%) non-REPEXH01 STEC O157:H7 isolates (Table 2). That deletion can be useful as a genomic signature of this strain for molecular surveillance and as a subject of future research to clarify the strain's evolution. Additional research addressing the role of the single base pair mutation in this strain's colonization and survival on leafy vegetables could yield valuable insights.

In summary, REPEHX01 belongs to *E. coli* O157:H7 Manning Clade 8, and most isolates possess

stx2a, both factors that are associated with severe clinical outcomes. Those factors, along with its harboring of multiple resistance determinants, underscore the continued need to monitor REPEXH01 and understand factors contributing to its emergence and persistence.

Acknowledgments

We thank Kaitlin Tagg for providing subject matter expertise on plasmid classification. We thank both Kaitlin Tagg and Hattie Webb for their helpful discussions and insightful comments.

This work was made possible by support from the Office of Advanced Molecular Detection at the Centers for Disease Control and Prevention and is covered by activities approved by the Centers for Disease Control and Prevention Institutional Review Board (approval no. 7172).

About the Author

Dr. Wirth is bioinformatician on the Molecular Virology Team in the Viral Vaccine-Preventable Disease Branch, Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. His research interests include the application of computational techniques to microbiological problems, especially those involving the evolution and physiology of human pathogens.

References

1. Rangel JM, Sparling PH, Crowe C, Griffin PM, Swerdlow DL. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982–2002. Emerg Infect Dis. 2005;11:603–9. https://doi.org/10.3201/eid1104.040739
2. Dewey-Mattia D, Manikonda K, Hall AJ, Wise ME, Crowe SJ. Surveillance for foodborne disease outbreaks—United States, 2009–2015. MMWR Surveill Summ. 2018;67:1–11. https://doi.org/10.15585/mmwr.ss6710a1
3. Bottichio L, Keaton A, Thomas D, Fulton T, Tiffany A, Frick A, et al. Shiga toxin–producing *Escherichia coli* infections associated with romaine lettuce—United States, 2018. Clin Infect Dis. 2020;71:e323–30. https://doi.org/10.1093/cid/ciz1182
4. Marshall KE, Hexemer A, Seelman SL, Fatica MK, Blessington T, Hajmeer M, et al. Lessons learned from a decade of investigations of Shiga toxin–producing *Escherichia coli* outbreaks linked to leafy greens, United States and Canada. Emerg Infect Dis. 2020;26:2319–28. https://doi.org/10.3201/eid2610.191418
5. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M-A, Roy SL, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis. 2011;17:7–15. https://doi.org/10.3201/eid1701.P11101
6. Mead PS, Griffin PM. *Escherichia coli* O157:H7. Lancet. 1998; 352:1207–12. https://doi.org/10.1016/S0140-6736(98)01267-7
7. Interagency Food Safety Analytics Collaboration. Foodborne illness source attribution estimates for 2020 for *Salmonella, Escherichia coli* O157, and *Listeria monocytogenes* using multi-year outbreak surveillance data, United States. US Department of Health and Human Services, Centers for Disease Control and Prevention, Food and Drug Administration, US Department of Agriculture Food Safety and Inspection Service, editors. Atlanta and Washington; The Departments; 2020.
8. Chen JC, Patel K, Smith PA, Vidyaprakash E, Snyder C, Tagg KA, et al. Reocurring *Escherichia coli* O157:H7 strain linked to leafy greens–associated outbreaks, 2016–2019. Emerg Infect Dis. 2023;29:1895–9. https://doi.org/10.3201/eid2909.230069
9. Bielaszewska M, Schmidt H, Liesegang A, Prager R, Rabsch W, Tschäpe H, et al. Cattle can be a reservoir of sorbitol-fermenting Shiga toxin–producing *Escherichia coli* O157:H(–) strains and a source of human diseases. J Clin Microbiol. 2000;38:3470–3. https://doi.org/10.1128/JCM.38.9.3470-3473.2000
10. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153
11. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3. https://doi.org/10.1093/bioinformatics/btv421
12. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17:1. https://doi.org/10.1186/s13059-016-1108-8
13. Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, Fivian A, et al. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. P Proc Natl Acad Sci U S A. 2006;103:14941–6.
14. Sandu P, Crepin VF, Drechsler H, McAinsh AD, Frankel G, Berger CN. The enterohemorrhagic *Escherichia coli* effector EspW triggers actin remodeling in a Rac1-dependent manner. Infect Immun. 2017;85:e00244-17. https://doi.org/10.1128/IAI.00244-17
15. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, et al. A comparative analysis of the lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. Front Microbiol. 2017;8:375. https://doi.org/10.3389/fmicb.2017.00375
16. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15. https://doi.org/10.1093/nar/gku1196
17. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLOS Comput Biol. 2014;10:e1003537. https://doi.org/10.1371/journal.pcbi.1003537
18. Bouckaert RR, Drummond AJ. bModelTest: Bayesian phylogenetic site model averaging and model comparison. BMC Evol Biol. 2017;17:42. https://doi.org/10.1186/s12862-017-0890-6
19. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39):W347-52. https://doi.org/10.1093/nar/gkr485
20. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44:W16–21. https://doi.org/10.1093/nar/gkw387

21. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV; CDC PulseNet Task Force. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis. 2001;7:382–9. https://doi.org/10.3201/eid0703.017303
22. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293–6. https://doi.org/10.1093/nar/gkab301
23. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, et al. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. Proc Natl Acad Sci U S A. 2008;105:4868–73. https://doi.org/10.1073/pnas.0710834105
24. Iyoda S, Manning SD, Seto K, Kimata K, Isobe J, Etoh Y, et al. Phylogenetic clades 6 and 8 of enterohemorrhagic *Escherichia coli* O157:H7 with particular *stx* subtypes are more frequently found in isolates from hemolytic uremic syndrome patients than from asymptomatic carriers. Open Forum Infect Dis. 2014;1:ofu061. https://doi.org/10.1093/ofid/ofu061
25. Centers for Disease Control and Prevention. Persistent strain of *E. coli* O157:H7 (REPEXH01) linked to multiple sources [cited 2024 Mar 7]. https://www.cdc.gov/ecoli/php/data-research/repexh01-e-coli-o157h7.html
26. Kang Y, Jelenska J, Cecchini NM, Li Y, Lee MW, Kovar DR, et al. HopW1 from *Pseudomonas syringae* disrupts the actin cytoskeleton to promote virulence in Arabidopsis. PLoS Pathog. 2014;10:e1004232. https://doi.org/10.1371/journal.ppat.1004232
27. Xicohtencatl-Cortes J, Sánchez Chacón E, Saldaña Z, Freer E, Girón JA. Interaction of *Escherichia coli* O157:H7 with leafy green produce. J Food Prot. 2009;72:1531–7. https://doi.org/10.4315/0362-028X-72.7.1531
28. Saldaña Z, Sánchez E, Xicohtencatl-Cortes J, Puente JL, Girón JA. Surface structures involved in plant stomata and leaf colonization by Shiga-toxigenic *Escherichia coli* O157:H7. Front Microbiol. 2011;2:119. https://doi.org/10.3389/fmicb.2011.00119
29. Orsi RH, Bowen BM, Wiedmann M. Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC Genomics. 2010;11:102. https://doi.org/10.1186/1471-2164-11-102
30. Byrne L, Adams N, Jenkins C. Association between Shiga toxin-producing *Escherichia coli* O157:H7 *stx* gene subtype and disease severity, England, 2009–2019. Emerg Infect Dis. 2020;26:2394–400. https://doi.org/10.3201/eid2610.200319
31. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun. 2020;11:3602. https://doi.org/10.1038/s41467-020-17278-2

Address for correspondence: Joseph S. Wirth, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H18-5, Atlanta, GA 30329-4018, USA; email: jwirth@cdc.gov

etymologia revisited

Scrapie
[skra'pe]

Scrapie is a fatal neurodegenerative disease of sheep and goats that was the first of a group of spongiform encephalopathies to be reported (1732 in England) and the first whose transmissibility was demonstrated by Cuille and Chelle in 1936. The name resulted because most affected sheep develop pruritis and compulsively scratch their hides against fixed objects. Like other transmissible spongiform encephalopathies, scrapie is associated with an alteration in conformation of a normal neural cell glycoprotein, the prion protein. The scrapie agent was first described as a prion (and the term coined) by Stanley Prusiner in 1982, work for which he received the Nobel Prize in 1997.

Originally published in June 2020

References:

1. Brown P, Bradley R. 1755 and all that: a historical primer of transmissible spongiform encephalopathy. BMJ. 1998;317:1688–92.
2. Cuillé J, Chelle PL. The so-called "trembling" disease of sheep: is it inoculable? [in French]. Comptes Rendus de l'Académie Sciences. 1936;203:1552.
3. Laplanche J-L, HunterN, Shinagawa M, Williams E. Scrapie, chronic wasting disease, and transmissible mink encephalopathy. In: Prusiner SB, editor. Prion biology and diseases. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1999. p. 393–429.
4. Prusiner SB. Novel proteinaceous infectious particles cause scrapie. Science. 1982;216:136–44.

https//wwwnc.cdc.gov/eid/article/26/6/et-2606_article

Lessons from 5 Years of Routine Whole-Genome Sequencing for Epidemiologic Surveillance of Shiga Toxin–Producing *Escherichia coli*, France, 2018–2022

Gabrielle Jones, Carolina Silva Nodari, Laëtitia Fabre, Henriette de Valk, Harold Noel, Aurélie Cointe, Stéphane Bonacorsi, François-Xavier Weill, Yann Le Strat

Whole-genome sequencing (WGS) is routine for surveillance of Shiga toxin–producing *Escherichia coli* human isolates in France. Protocols use EnteroBase hierarchical clustering at ≤5 allelic differences (HC5) as screening for cluster detection. We assessed current implementation after 5 years for 1,002 sequenced isolates. From genomic distances of serotypes O26:H11, O157:H7, O80:H2, and O103:H2, we determined statistical thresholds for cluster determination and compared those with HC5 clusters. Thresholds varied by serotype, 5–16 allelic distances and 15–20 single-nucleotide polymorphisms, showing limits of a single-threshold approach. We confirmed validity of HC5 screening for 3 serotypes because statistical thresholds had limited effect on isolate clustering (high sensitivity and specificity). For O80:H2, results suggest that HC5 is less reliable, and other approaches should be explored. Public health officials should regularly assess WGS used for Shiga toxin–producing *E. coli* surveillance to account for serotype and genomic evolution and to interpret WGS-linked isolates in light of epidemiologic data.

Shiga toxin–producing *Escherichia coli* (STEC) are responsible for a spectrum of disease that ranges from self-resolving diarrhea to bloody diarrhea and severe complications, including hemolytic uremic syndrome (HUS). STEC continues to be a public health risk, and although infections are largely sporadic, STEC has substantial outbreak potential (*1*–*3*). Therefore, surveillance and outbreak detection remain public health priorities (*4*). Advances in STEC detection and typing methods over the past decade, including the expansion of culture-independent diagnostic tests and whole-genome sequencing (WGS), have affected diagnostic approaches, expanded knowledge of pathogenicity, informed source attribution, improved outbreak detection capacities, and guided surveillance protocols (*5*–*9*).

Author affiliations: Santé publique France, Saint-Maurice, France (G. Jones, H. de Valk, H. Noel, Y. Le Strat); Institut Pasteur, Université Paris Cité, Paris, France (C. Silva Nodari, L. Fabre, F.-X. Weill); Centre hospitalier universitaire Robert Debré, Assistance Publique–Hôpitaux de Paris, Paris (A. Cointe, S. Bonacorsi)

DOI: https://doi.org/10.3201/eid3113.241950

Advantages of implementing WGS for epidemiologic surveillance are widely documented. WGS is the primary method of foodborne pathogen surveillance and outbreak detection in numerous countries in Europe and North America (*5*,*10*–*12*). Diverse studies have confirmed superiority of WGS for cluster determination, shown validation of thresholds used for cluster detection in surveillance protocols, and described WGS-linked isolates in light of epidemiologic data (*6*,*11*,*13*–*19*). WGS improves outbreak detection and investigation capacity by providing more timely cluster detection and discriminatory case definitions and detecting geographically and temporally diffuse clusters. Such studies are essential for guiding the international adoption of widespread use of WGS for disease surveillance and outbreak detection. However, surveillance systems and epidemiologic context differ between countries, and multiple WGS approaches are possible for isolate comparison (*6*,*9*,*15*,*20*). Therefore, assessing the implementation of WGS for epidemiologic surveillance specific to a given pathogen and country is vital.

WGS was implemented in France for STEC surveillance in early 2017 (*3*). Surveillance uses the EnteroBase (https://enterobase.warwick.ac.uk) core-genome multilocus sequence typing (cgMLST) hierarchical clustering method (HierCC) for *E. coli* as an initial screening step for cluster detection at ≤5 allelic differences (HC5) (*21*–*23*). HC5 clusters are confirmed by core-genome single-nucleotide polymorphism (SNP) tree analysis.

On the basis of 5 years (2018–2022) of retrospective data available from STEC surveillance, this study aimed to assess implementation of WGS for cluster detection protocols in France. The first objective was to apply statistical approaches to pairwise allelic distance (AD) and SNP distance data to evaluate whether thresholds could be determined to define genomic proximity. The second objective was to assess the performance of those statistical thresholds compared with HC5. Finally, we described genomic distance data by considering HC5 and associated epidemiologic data.

Methods

STEC Surveillance and Cluster Detection in France

STEC surveillance and outbreak detection in France rely on 2 previously described parallel voluntary systems: epidemiologic surveillance of HUS in children <15 years of age, coordinated at the national level by the food and waterborne disease surveillance and outbreak investigation unit at Santé publique France (French public health agency, https://www.santepubliquefrance.fr); and microbiological surveillance coordinated by the National Reference Center for *E. coli*, *Shigella*, *Salmonella* (NRC-ESS) and its associated NRC at Robert Debré hospital, Paris (NRC-RD) (*1*,*3*). Epidemiologists at regional offices of Santé Publique France can also contribute to investigations but are not dedicated to foodborne disease surveillance. Santé publique France links epidemiologic data from pediatric STEC-HUS surveillance and epidemiologic investigations to WGS data, generating a consolidated anonymous dataset for annual surveillance reports (*3*).

A cluster is typically defined as cases grouped in space, time, or both. An outbreak defines cases for which an epidemiologic link is identified. A microbiological cluster defines isolates grouped on the basis of an established typing method: phenotypic serogroup and serotype or genomic typing using cgMLST or SNP analysis. Cluster detection in France relies on pediatric HUS notifications and microbiological data (Appendix 1, https://wwwnc.cdc.gov/EID/article/31/13/24-1950-App1.pdf).

In current WGS protocols, STEC genomic data are submitted to EnteroBase with limited metadata (isolate source, e.g., human, food; sampling year; and country). The cgMLST and HierCC schemes implemented in that platform assist in identification of genomic clusters (Appendix 1) (*21*). The platform uses multilevel, static, cluster assignments of bacterial genomes to describe genetic diversity (*23*). At the French NRC, the HC5 level of the HierCC scheme is used for screening of genomic relatedness for epidemiologic purposes. If necessary, particularly for HC5s that persist over time, an additional SNP analysis using the EnteroBase pipeline serves as a confirmatory step. Epidemiologists assess cluster characteristics (size, space-time distribution, clinical severity, case-patient characteristics) to decide whether investigations should be initiated. Decisions to investigate small (<5 isolates) or temporally dispersed WGS clusters also depend on availability of human resources.

Study Data

We included STEC isolates sequenced at the NRC-ESS and uploaded to EnteroBase as part of routine WGS data analysis from January 1, 2018–December 31, 2022. We considered isolates from the same patient as duplicates and excluded those if sampling dates were ≤2 weeks apart and WGS identified the same strain. We restricted analyses to 4 serotypes with sufficient historical data: O26:H11 (n = 478), O80:H2 (n = 226), O157:H7 (n = 223), and O103:H2 (n = 75). We conducted all data management and statistical analyses in R version 4.2 (The R Project for Statistical Computing, https://www.r-project.org).

The assembled short-read data for the list of genomes are available from EnteroBase (https://enterobase.warwick.ac.uk/species/ecoli/search_strains?query=workspace:127168) (Appendix 2 Table 1, https://wwwnc.cdc.gov/EID/article/31/13/24-1950-App1.xlsx). Short-read sequences are available at the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home; project no. PRJEB50273.

Allelic and SNP Distance Distributions

We generated pairwise allele and SNP distance matrices for each serotype. We extracted the cgMLST allelic profiles from EnteroBase and excluded alleles if they were missing from >5% of isolates within a given serotype (2 excluded from O157:H7 AD matrices). We calculated AD from allelic profiles on the basis of the number of mismatched loci and determined SNP distances on a recombination-free multisequence alignment of the core genome of each studied serotype (Appendix 1).

We merged isolate characteristics (sampling date, HC2–HC50, epidemiologic data) from consolidated surveillance datasets with the AD and SNP matrices by using a unique anonymous identifier from the NRC-ESS. For each serotype, we plotted overall distribution of pairwise AD and SNP. We censured data at 50 AD and SNP distance for statistical analysis and primary graphical representations.

Determination of Statistical Genomic Distance Thresholds

We applied a mixture of distributions approach to test whether statistical thresholds to describe genomic proximity of isolates could be determined. Mixture of distributions is a classic statistical approach for determining thresholds from continuous data distributions, such as seroprevalence data (*24*). We used the mixR package in R, which determines the best fit to continuous data from several distribution families and selects the optimal number of components for the mixture model on the basis of the lowest value of the Bayesian information criterion (*25*). The underlying hypothesis was that outbreak-related isolates have smaller pairwise AD and SNP distance. For each pair of isolates, the probability of belonging to the first distribution (comprising the smallest genomic distances) is calculated and plotted according to AD and SNP distance. We set a threshold as the AD or SNP distance at which the probability of belonging to the first distribution was ≥50%.

Comparison of Genomic Distributions and Statistical Thresholds to HC5 Clusters

Although isolates are assigned to HC on the basis of AD, HC does not strictly translate to AD because of the multilevel clustering approach, which defines that when AD at a given level is equal, the genome is assigned to the oldest HC value (*23*). For example, isolates assigned to a given HC5 or HC10 cluster are not all within 5 or 10 AD of each other. Therefore, assessing the observed genomic distance distributions and performance of statistically defined thresholds in relation to HC5 clusters is necessary. We calculated sensitivity and specificity of statistically determined thresholds compared with HC5.

We also assessed the relationship between time and genomic distances within HC5 clusters. We studied time in categories constituted on equal distribution of isolates and as a continuous variable (days) by using a multivariable fractional polynomial (MFP) linear regression. To assess concordance between HC5 and SNP analysis as a confirmatory step for cluster determination, we visualized HC5 clusters (≥4 isolates) and year (all isolates) into generated SNP-based maximum-likelihood trees by using iTOL (https://itol.embl.de) (*26*) (Appendix 1).

Genomic and Epidemiologic Characteristics of HC5 Clusters

We assessed characteristics for each HC5 cluster, including genomic distance range, number of isolates, temporal distribution, geographic distribution (same administrative department or region, multiple regions), and epidemiologic link. Epidemiologic links included clusters of household transmission and single patients (isolates sampled >15 days apart), isolates with a confirmed or suspected outbreak link, and isolates for which the link was unable to be determined from investigations.

Results

Pairwise Distance Distributions

Genomic distance distributions varied by serotype (Appendix 1 Figure 1, panels A, B). For O26:H11 and O157:H7, we observed a peak at 0–5 AD (Figure 1, panel A). Conversely, fewer O80:H2 isolate pairs had shorter AD, and we observed no similar peak but noted a normal distribution. Few O103:H2 isolate pairs had AD <10. The O26:H11 SNP distance distribution showed a plateau from 1 to 20 SNPs (Figure 1, panel B). For O157:H7, we observed a peak of 0–20 for pairwise SNP distances. The SNP distance distribution for O80:H2 showed a sloping increase, and few isolate pairs had <10 SNP distance. The O103:H2 SNP difference distribution was sparse, limiting description of specific characteristics.

Determination of Statistical Thresholds

The mixture of distributions model retained the gamma distribution for determination of both AD and SNP distance thresholds. The number of components fitting the genomic distance distributions in the final model varied by serotype (Figure 2, panel A; Figure 3, panel A). The AD statistical thresholds were ≤8 AD for O26:H11, ≤16 AD for O157:H7, ≤9 AD for O80:H2, and ≤5 AD for O103:H2 (Figure 2, panel B). The SNP distance statistical thresholds were ≤15 SNP for O26:H11, ≤20 SNP for O157:H7, ≤17 SNP for O80:H2, and ≤15 SNP for O103:H2 (Figure 3, panel B). For O157:H7 SNP distances, we determined the threshold from the probability of belonging to the second distribution, because the first distribution was at 0, with mean and SD close to 0. Although we determined a threshold for O103:H2, the result was less robust because of the small number of pairwise isolates, particularly at shorter genomic distances.

Genomic Distance Distributions within HC5

The number of HC5 clusters increased with serotype frequency: 6 for O103:H2, 19 for O157:H7, 23 for O80:H2, and 39 for O26:H11. The AD and SNP distance distributions observed in HC5 clusters varied within and between serotypes (Appendix 1 Figure 2). Applying statistically determined thresholds, all HC5 cluster isolates were under the AD threshold for serotypes O103:H2 and O157:H7. Only O103:H2 HC5 cluster isolates were under the SNP distance threshold (Appendix 1 Figure 2). A greater number of O26:H11 and O80:H2 HC5 clusters contained isolate pairs surpassing statistical thresholds.

Sensitivity and specificity of the statistical thresholds compared with HC5 clusters varied between serotypes (Table). For O157:H7 and O103:H2, the statistical thresholds had high sensitivity (≥99%) and specificity (83%–100%). For O26:H11, sensitivity was close to 100%, and specificity was 73% for AD threshold and 88% for SNP threshold. Finally, for O80:H2, although the mixture of distributions determined a statistical threshold,

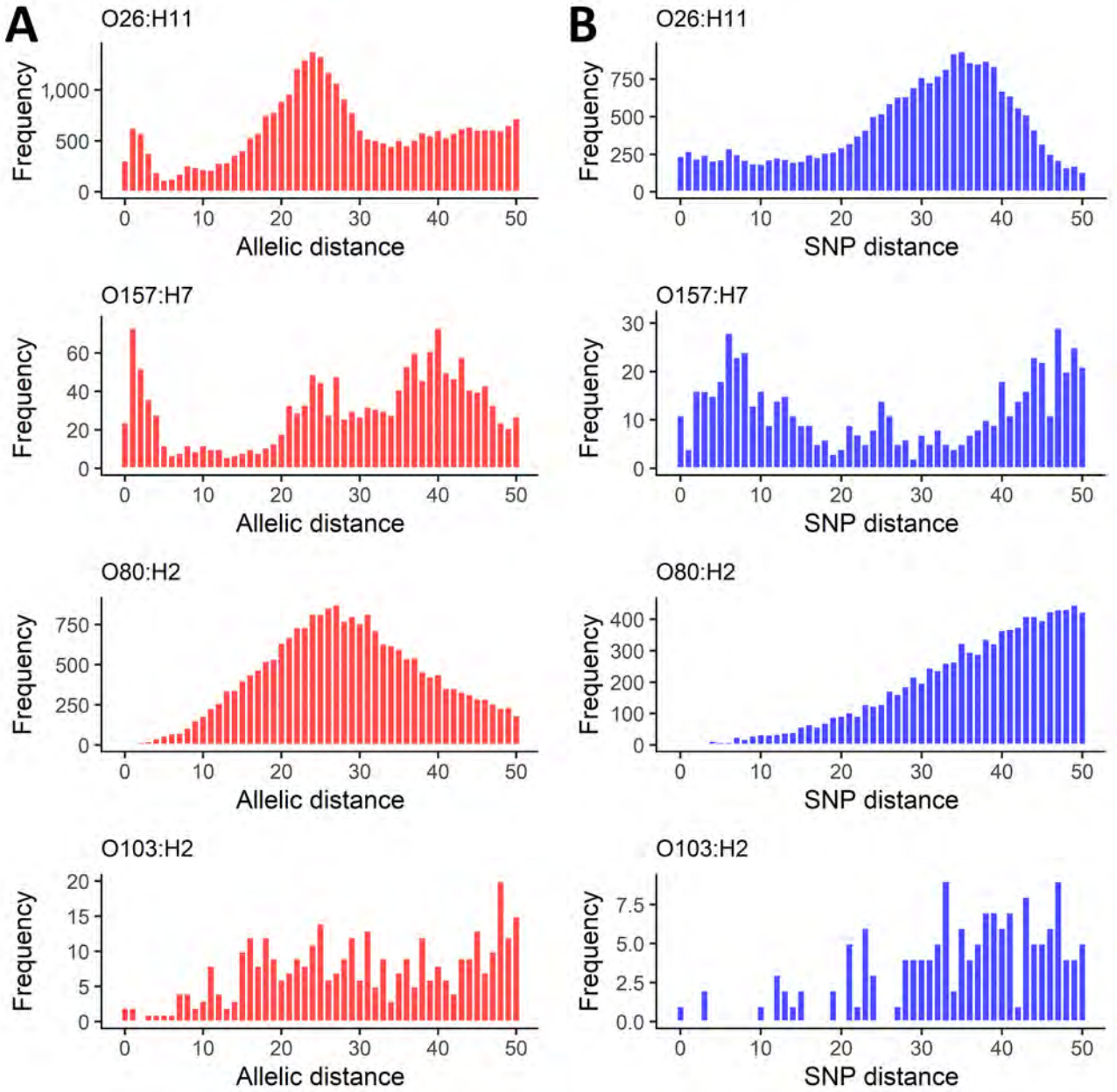


Figure 1. Characteristics from 5 years of routine whole-genome sequencing for epidemiologic surveillance of Shiga toxin–producing *Escherichia coli*, France, 2018–2022. A) Distribution of pairwise allelic distances; B) SNP distances, censured at 50. Shiga toxin–producing *Escherichia coli* serotypes are shown for each panel. SNP, single-nucleotide polymorphism.

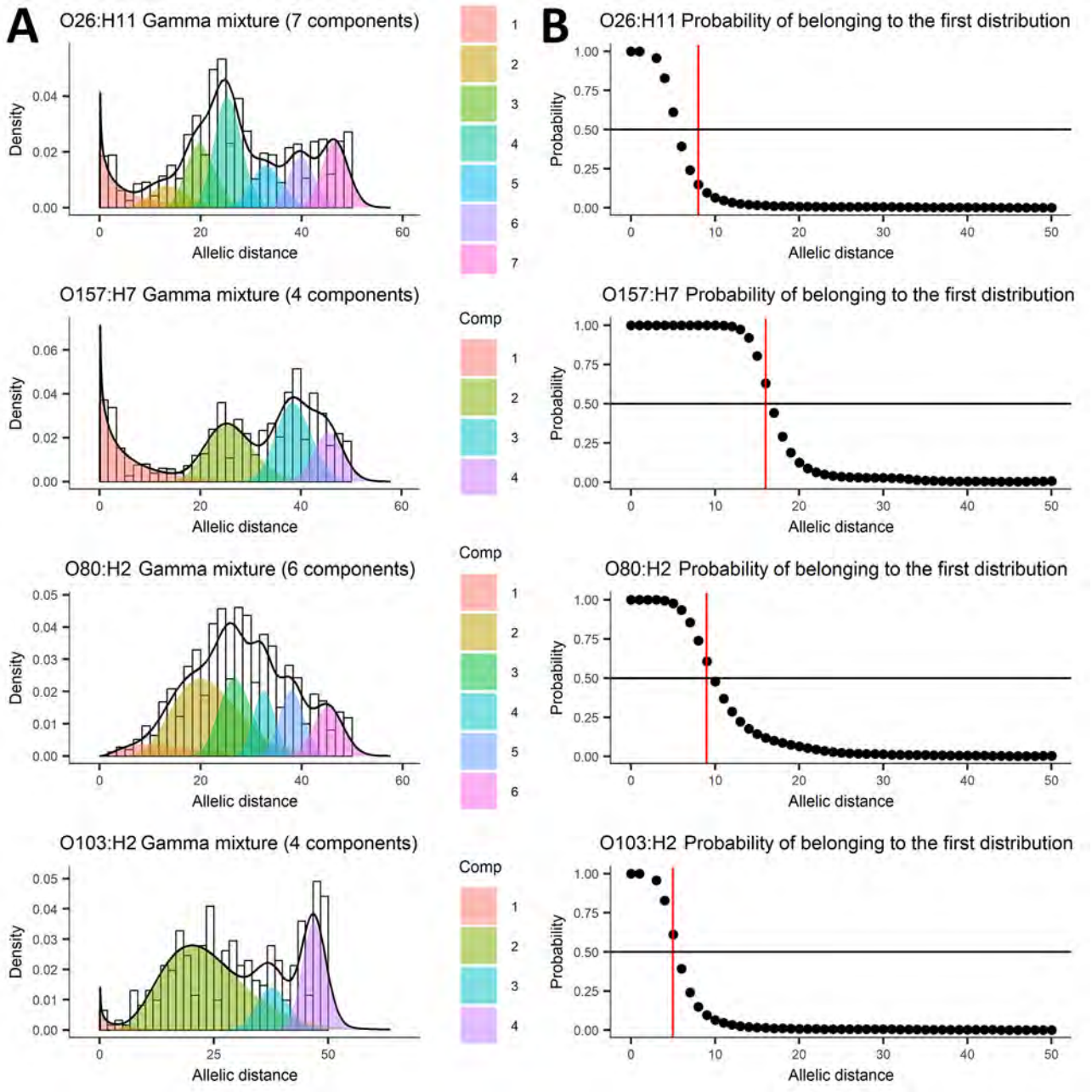


Figure 2. Mixture of distributions model applied to allelic distance data from 5 years of routine whole-genome sequencing for epidemiologic surveillance of Shiga toxin–producing *Escherichia coli*, France, 2018–2022. A) Number of components fit to the data distribution; B) threshold represented as the probability of belonging to the first distribution. Shiga toxin–producing *Escherichia coli* serotypes are shown for each panel. Black line indicates global estimated density; black circles, probability of belonging to first distribution for each observed allelic or single-nucleotide polymorphism distance; red line, largest allelic or single-nucleotide polymorphism distance that has a 50% probability of belonging to the first distribution. Comp, component.

specificity was poor for both AD (34%) and SNP (35%) thresholds.

Genomic Distance Distributions within HC5 Clusters as a Function of Time

With time represented in classes, we observed a slight positive association between AD and HC5 (Appendix 1 Figure 3). MFP regression integrated time as a continuous variable and confirmed a linear relationship with AD for all serotypes, but with varied strength of association (Figure 4, panel A). Of note, we found a negative association between AD and time observed for O26:H11 and O157:H7 at the smallest temporal distances (≤5 days) and then a positive linear

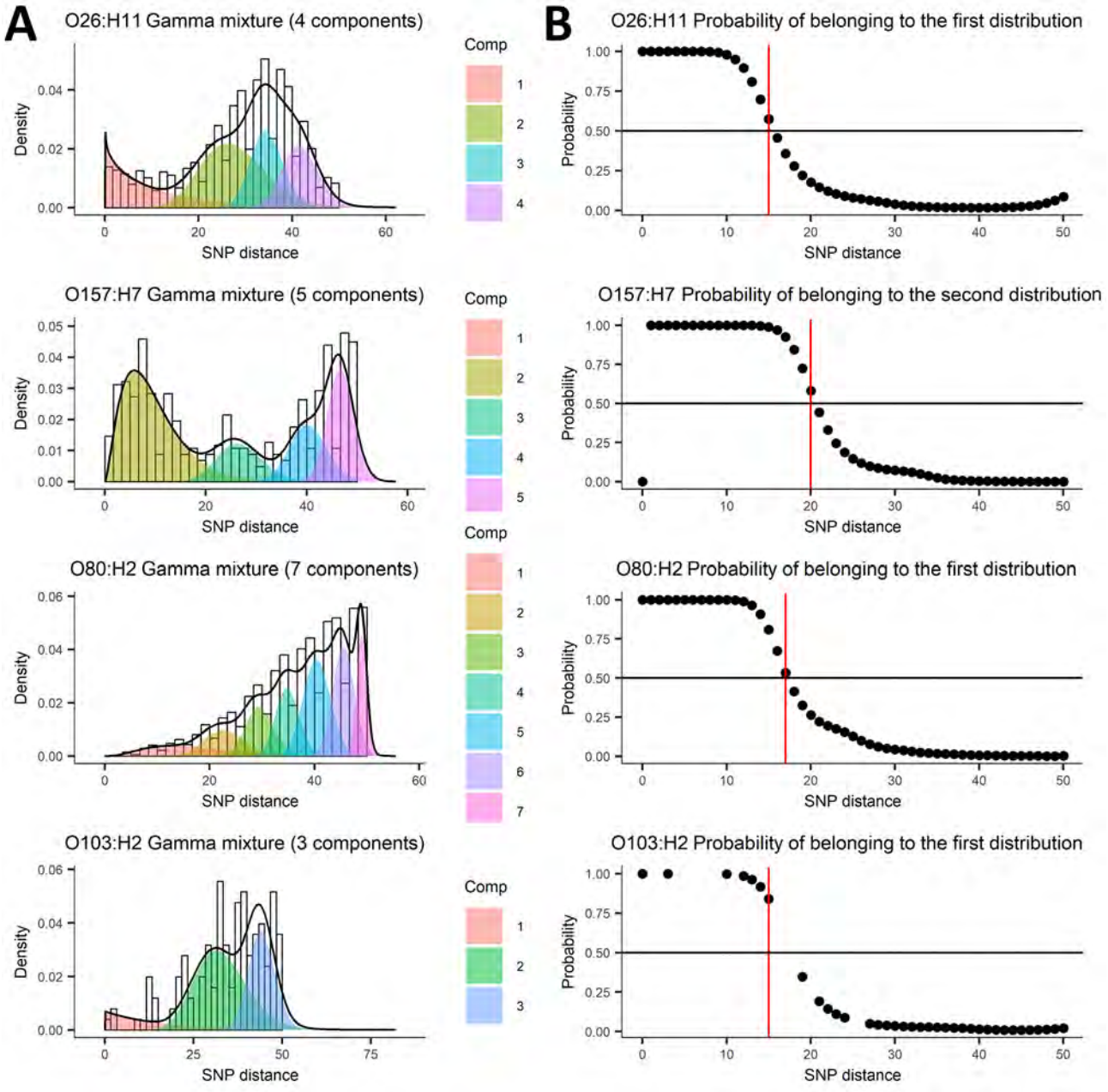


Figure 3. Mixture of distributions model applied to SNP distance data from 5 years of routine whole-genome sequencing for epidemiologic surveillance of Shiga toxin–producing *Escherichia coli*, France, 2018–2022. A) Number of components fit to the data distribution; B) threshold represented as a probability of belonging to the first or second distribution. Shiga toxin–producing *Escherichia coli* serotypes are shown for each panel. Comp, component; SNP, single-nucleotide polymorphism.

relationship as temporal distance increased. For O103:H2, the relationship was linear, but the number of HC5 clusters was small, and the maximum temporal distance was comparatively short (≈100 days). Analysis with SNP distance yielded similar results as AD, with 1 distinct difference: MFP regression did not identify the same negative association at small temporal distances for O26:H11 and O157:H7 (Figure 4, panel B; Appendix 1 Figure 4).

Concordance between HC5 and SNP

SNP analysis generally confirmed HC5 clusters for all serotypes except O80:H2 (Appendix 1 Figures 5–7). For O80:H2, although SNP distance confirmed clustering for some HC5s, for others, such as HC5_35789 and HC5_80832, HC5 was not predictive of SNP clustering because genomes belonging to the same HC5 were dispersed in the phylogenetic tree (Figure 5).

Genomic Distance and Epidemiologic Characteristics of HC5 Clusters

Because HC5 informed cluster detection and guided epidemiologic investigations during the study period, data are not independent. However, examining differences in genomic distance in light of epidemiologic characteristics of HC5 clusters is of interest.

We identified 87 HC5 clusters (≥2 isolates) comprising 449 isolates over the study period. Most (81/87; 93%) clusters comprised 2–10 isolates; 80% (70/87) of the HC5 clusters comprised 2–4 isolates, and 13% (11/87) comprised 5–10 isolates (Appendix 2 Table 2).

For the 81 clusters with 2–10 isolates, 58 (72%) comprised isolates with sampling dates within 1 year of each other. Twenty (25%) clusters had a duration of 1–2 years, and 4 (5%) clusters had a duration ≥3 years. Of the 6 HC5 clusters with >10 isolates, 4 lasted ≥3 years and 2 had isolates sampled within 3-month periods. Geographic distribution expanded with cluster

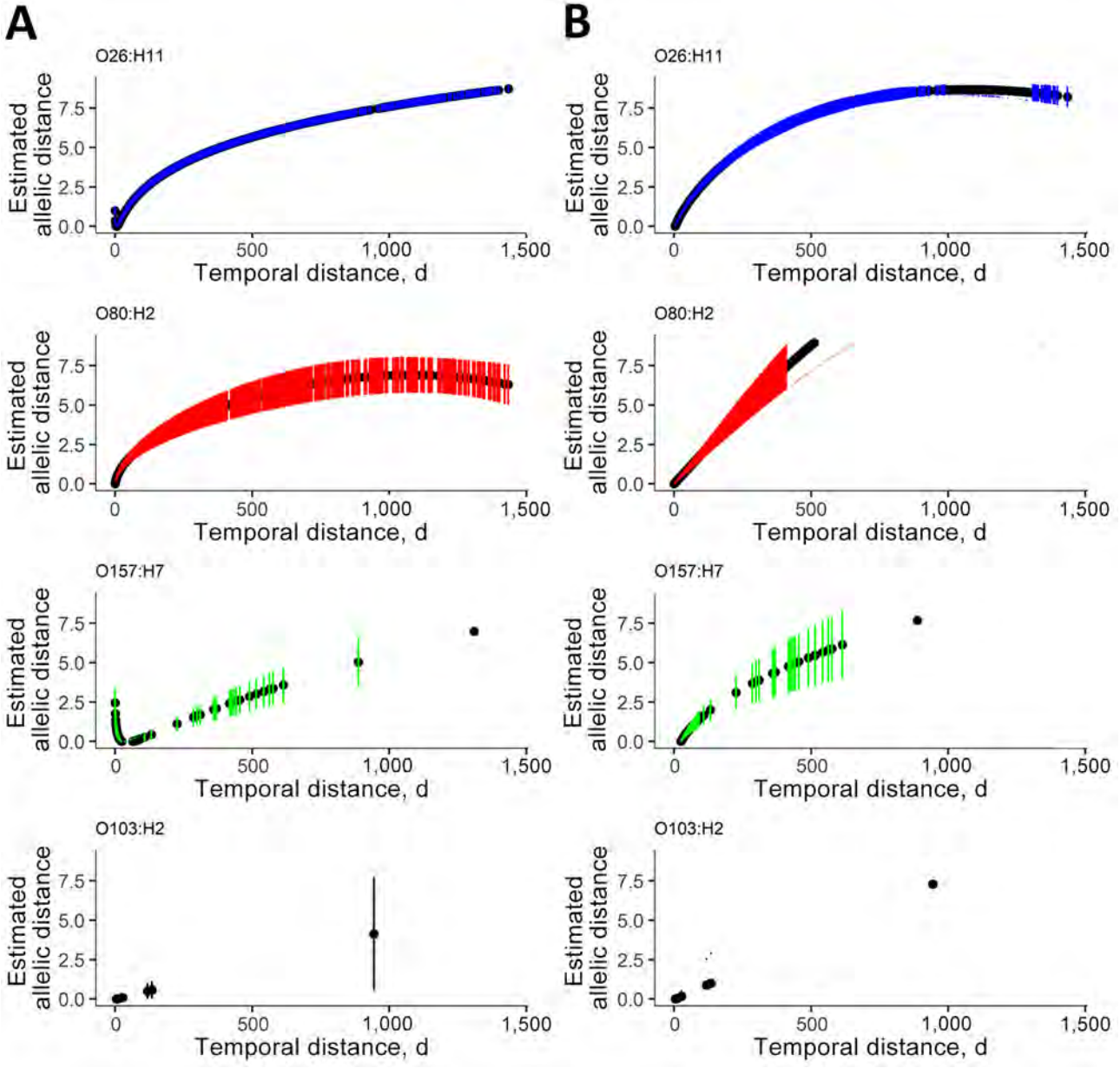


Figure 4. Regression from hierarchical clustering at a threshold of 5 allelic differences from 5 years of routine whole-genome sequencing for epidemiologic surveillance of Shiga toxin–producing *Escherichia coli*, France, 2018–2022. A) Allelic distance; B) SNP distance. Distances calculated as a function of time in days by multivariable fractional polynomial linear regression. Black circles indicate estimated allelic or SNP distance for each observed temporal distance in days; blue, red, green, and black vertical lines, 95% CIs of the estimated genomic distances for each observed temporal distance in days. SNP, single-nucleotide polymorphism.

Table. Sensitivity and specificity of statistically determined allelic and SNP distance thresholds from 5 years of routine whole-genome sequencing for epidemiologic surveillance of Shiga toxin–producing *Escherichia coli*, France, 2018–2022*

	Allelic distance			SNP distance		
Serotype	Threshold, no. alleles	Sensitivity, %	Specificity, %	Threshold, no. SNPs	Sensitivity, %	Specificity, %
O26:H11	≤8	99.9	73.1	≤15	99.6	87.8
O157:H7	≤16	99.7	99.6	≤20	99.9	96.7
O80:H2	≤9	99.8	33.6	≤17	99.6	35.1
O103:H2	≤5	100	83.3	≤15	99.0	100

*SNP, single-nucleotide polymorphism.

size. All HC5 clusters within the same administrative department had <5 isolates, and all clusters within the same region had <10 isolates.

For clusters of 2–10 isolates, median AD ranged from 0–5 (O103:H2) to 0–15 (O26:H11), and median SNP distance ranged from 0–14 (O103:H2) to 5–28 (O157:H7). For the 6 larger clusters (>10 isolates), 2 were point-source outbreaks (O157:H7 HC5_116498 [suspected] and O26:H11 HC5_190514 [confirmed]), with reasonably small median genomic distances: median AD = 1 for both and median SNP distance = 6 for O157:H7 HC5_116498 and 4 for O26:H11 HC5_190514. The 4 other large clusters with isolates sampled over 3–5 years had median AD of 2–10 and median SNP distance of 8–21. We observed the highest median and maximum genomic distances for O80:H2.

We linked 6 HC5 clusters (all <5 isolates) exclusively to household transmission, and we linked 1 cluster to 1 patient. Those median genomic distances were small. Of the additional 27 HC5 clusters that led to epidemiologic investigations of all or some cases (depending on space-time distribution), we identified a confirmed or suspected epidemiologic link for 20 (74%) clusters, corresponding to 146 isolates (15% of the study population) (Appendix 2 Table 2) (*27*–*29*). Those links included 2 persistent O26:H11 clusters (HC5_65006 and HC5_75047) comprising isolates associated with multiple point-source outbreaks and sporadic isolates with no identified epidemiologic link to each other or with previous outbreak sources (*28*,*29*). Within O26:H11 clusters that comprised isolates with documented epidemiologic links to several different point-source outbreaks, the median genomic distances of epidemiologically linked isolates were smaller than those of the overall cluster (Appendix 2 Table 2).

Discussion

The results of this study describe advantages and challenges of WGS for epidemiologic surveillance of STEC and inform potential adaptations in surveillance protocols in France. In this study, we used pairwise genomic distances to explore the robustness of using WGS-based clustering, particularly the HC5 level of EnteroBase's HierCC scheme, as a screening threshold for outbreak detection in STEC surveillance in France after 5 years of routine use. We first determined statistical thresholds to define genomic proximity. The heterogeneity of the thresholds across serotypes showed the necessity of verifying the suitability of a given approach strictly on the basis of genomic distance thresholds to all serotypes. Except for O80:H2, we confirmed the validity of using HC5 for a screening step for microbiological cluster determination; applying the statistical thresholds had a limited effect on how isolates grouped compared with HC5.

The O80:H2 genomic distance distributions were visually distinct, with near normal distributions versus multimodal distributions. SNP analysis for O80:H2 showed limited concordance with specific HC5 clusters compared with the other serotypes. Factors influencing genomic diversity, including mutation rate, reservoir, and transmission pathways, may differ for O80:H2 and explain its limited genomic diversity (*30*). The lack of concordance between cgMLST, including HC5, and epidemiologically relevant clusters has also been observed for another pathogenic clone of *E. coli* that exhibits limited genomic diversity, the human-restricted enteric pathogen *Shigella sonnei*, leading to a reliance on high-resolution techniques for surveillance (*31*). That observation suggests O80:H2 cluster determination should rely on SNP-based phylogenies. Such approaches require selection of an appropriate reference isolate and continuous integration of emerging strains into the analysis. Those approaches do not confer the same advantages of cgMLST and the EnteroBase's HierCC scheme, such as ease of comparing isolates with standardized methodology and nomenclature. Although O80:H2 is in the top 3 serotypes isolated in France since 2015, it is an uncommon hybrid pathotype (STEC/ExPEC [extraintestinal pathogenic *E. coli*]) that emerged in the early 2010s, and its reservoirs remain unclear (*1*,*30*). Indeed, a case–case study comparing characteristics and reported risk factors of *E. coli* O80–infected children with HUS with those infected by other STEC serogroups in France concluded that epidemiologic characteristics of O80:H2-infected pediatric HUS cases differed from O157:H7 and other

serotypes (*32*). Also, although O80:H2 was isolated in healthy cattle in France in 2023 and diarrheic calves in Belgium, no outbreaks have been documented in France after epidemiologic investigations (*33*,*34*). Improving cluster discrimination could increase the likelihood of resolving epidemiologic investigations and advancing knowledge on potential sources of contamination and reservoirs.

This study had several limitations related to data availability. Of note, analyses depended on the number of isolates available in surveillance data for France and pertained to 4 primary serotypes. The results suggest that conclusions may differ for other serotypes, and when sufficient isolates are available, expanding the study will be pertinent. Because STEC surveillance in France is voluntary, isolate data are not representative of all STEC in France. Pediatric HUS surveillance data are considered representative (*3*). However, that is not the case for other clinical isolates because patients with more severe illness are more likely to have consultations or be hospitalized and have biological sampling (*35*). Few environmental, food, and animal isolate data are available, and no routine sequencing has been implemented in France thus far. Therefore, this analysis was limited to clinical isolates. Increasing the number of nonclinical isolates and associated metadata would provide greater insight into the genomic diversity of circulating STEC

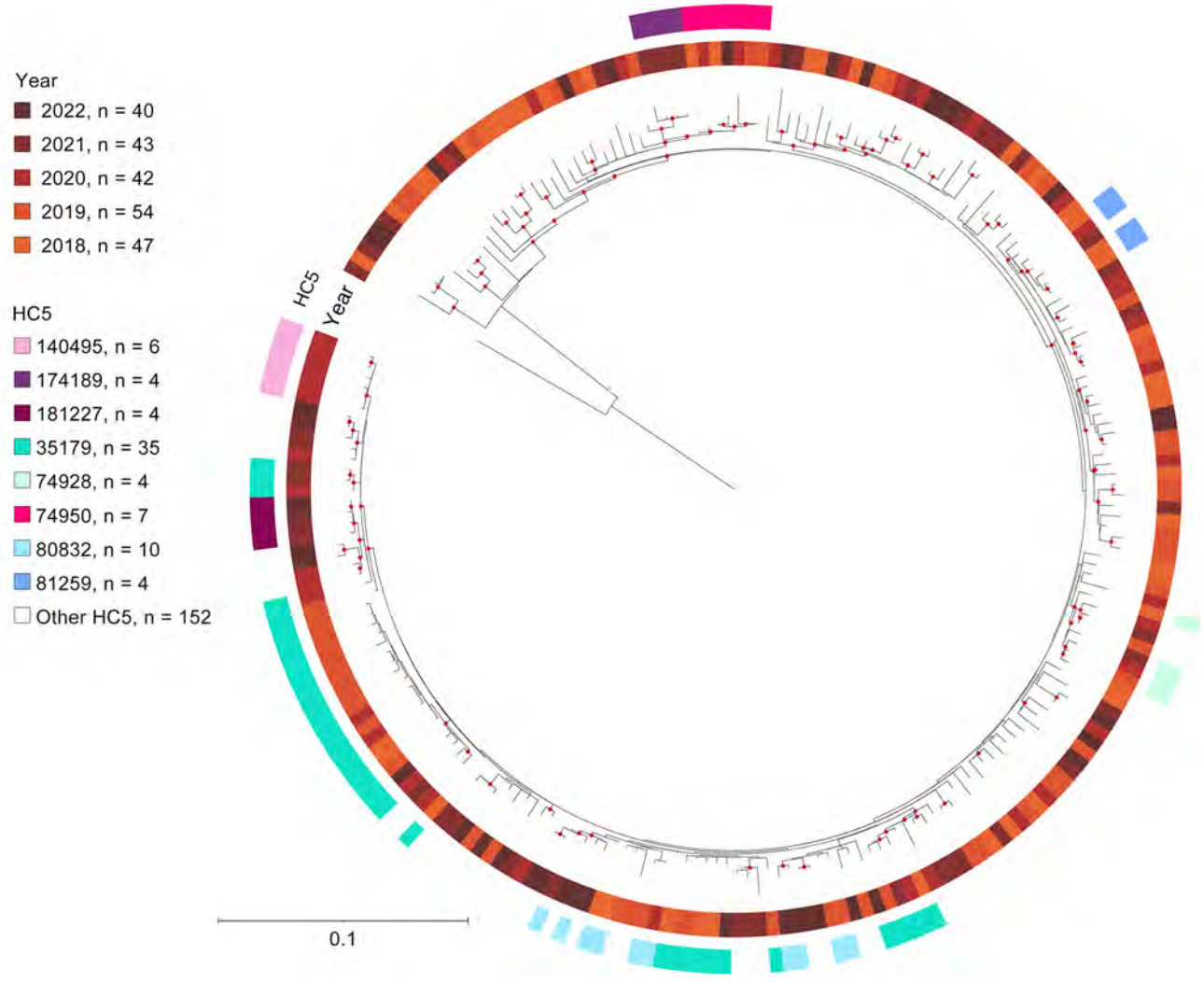


Figure 5. Single-nucleotide polymorphism–based maximum likelihood phylogenetic tree of 226 O80:H2 isolates from 5 years of routine whole-genome sequencing for epidemiologic surveillance of Shiga toxin–producing *Escherichia coli*, France, 2018–2022. Tree was built based on the sequence alignment of 3,949 single-nucleotide variant sites of the recombination-free core genome of *E. coli* strain MOD1-EC6881 (GenBank accession no. GCF_002520045.1). Tree was midpoint-rooted and visualized with iTOL (https://itol.embl.de). Bootstrap support values ≥90% are indicated with red dots on the branches. Branch lengths and corresponding scale bar indicate numbers of single-nucleotide polymorphisms per base of the final alignment. HC5, hierarchical clustering at a threshold of 5 allelic differences.

isolates in France and enable exploration of potential transmission chains and links with clinical isolates. Those links will be particularly relevant because certain geographic zones have shown greater risk for sporadic pediatric HUS, including WGS clusters with no identified epidemiologic link (*3*).

Although WGS provides a major advance for foodborne pathogen surveillance, epidemiologic data remain essential for confirming a common source for WGS-linked isolates (*36*). This study provides insight into the diversity of situations faced by epidemiologists after introduction of WGS. Indeed, a prior study described the complexity of interpreting WGS data in light of the effects of pathogen interactions with host and reservoir and the multiple transmission mechanisms involved in STEC circulation and contamination (*36*). Within HC5 clusters, the AD and SNP distributions were variable for a single serotype and between serotypes. Although some HC5 clusters linked to point-source outbreaks had low genomic diversity, others did not, particularly O157:H7, which was historically the predominant serotype in France before 2015 (*1*). The relationship between genomic and temporal distances within HC5 clusters also illustrates that variability. Although we observed an overall positive association, we noted a negative relationship between AD and temporal distance ≤5 days for O26:H11 and O157:H7. That relationship could be because of a limited number of point-source outbreaks linked to a diversity of food vehicles (vegetables, raw milk cheeses, industrial frozen pizzas) and caused by strains that accumulated greater genomic diversity before the outbreak (e.g., in reservoirs, in the manufacturing ingredients or environment). Different manufacturing processes for primary and final ingredients may also contribute (initial inoculum, bottlenecks, duration of processing or aging, temperature, stress) (*37*). Periodically assessing methods of WGS cluster determination, particularly HC5, used in surveillance approaches to ensure their continued validity will be needed.

During 2018–2022, epidemiologists in France regularly investigated WGS-linked isolates with case-patients closely related in space or time, but with no common source suspected despite extensive case interviews. Although we know of inherent limitations to epidemiologic investigations (interview based, memory bias), such clusters are necessary for documenting experiences with WGS in STEC surveillance and outbreak investigations. Similar to findings reported previously, most of the HC5 clusters from France are small (<5 isolates) (*2*). Limited public health resources are directed toward investigation of larger clusters or those including severe clinical manifestations such as HUS. However, even when very small clusters are investigated, identifying a common source of contamination can be challenging because of limited epidemiologic or traceback data. Moving toward systematic documentation of epidemiologic information for all WGS-linked isolates would provide more complete data to explore and interpret relatedness but would require evolutions in prioritization of activities or additional human resources. Finally, the numerous HC5 clusters comprising isolates over several years show that, as time progresses, genomic proximity evolves to different degrees, reinforcing that a SNP-based analysis remains an essential confirmatory step for cluster determination. Threshold-based approaches, although appropriate for screening in some serotypes, may therefore not be universally applicable for a given pathogen (*12*,*38*). Overall, public health professionals should strike a balance between consideration of serotype-related limits and the advantages conferred through more standardized genomic approaches. STEC surveillance protocols on the basis of WGS data should integrate regular assessment to ensure continued validity of genomic approaches.

In summary, after 5 years of implementation of WGS for STEC surveillance, our results validate the current approach of using cgMLST HC5 as a screening step for cluster detection for 3 major serotypes in France. For the fourth major serotype, O80:H2, our results indicate that HC5 is less reliable. Regular assessment of WGS-based STEC surveillance protocols to document the effects of serotype and time (genomic evolution) is appropriate. Exploring possibilities for routinely collecting epidemiologic data for WGS clusters could enrich the capacity to describe the relationship between WGS-linked isolates and epidemiologic links.

Data anonymization and storage authorizations for STEC surveillance at Santé publique France were previously described (*3*). Because the study used the existing consolidated and anonymous surveillance datasets and anonymous sequence data extracted from EnteroBase, no additional ethics approval was required.

About the Author

Ms. Jones is an epidemiologist for the French national public health agency (Santé publique France) working in foodborne disease surveillance and outbreak investigation. Her primary research interests include surveillance of Shiga toxin–producing *Escherichia coli* infections, Shiga toxin–producing *Escherichia coli*–associated hemolytic uremic syndrome, and viral gastroenteritis.

References

1. Bruyand M, Mariani-Kurkdjian P, Le Hello S, King LA, Van Cauteren D, Lefevre S, et al.; Réseau français hospitalier de surveillance du SHU pédiatrique. Paediatric haemolytic uraemic syndrome related to Shiga toxin–producing *Escherichia coli*, an overview of 10 years of surveillance in France, 2007 to 2016. Euro Surveill. 2019 24:1800068. https://doi.org/10.2807/1560-7917.ES.2019.24.8.1800068
2. Lipman DJ, Cherry JL, Strain E, Agarwala R, Musser SM. Genomic perspectives on foodborne illness. Proc Natl Acad Sci U S A. 2024;121:e2411894121. https://doi.org/10.1073/pnas.2411894121
3. Jones G, Mariani-Kurkdjian P, Cointe A, Bonacorsi S, Lefèvre S, Weill FX, et al. Sporadic Shiga toxin–producing *Escherichia coli*–associated pediatric hemolytic uremic syndrome, France, 2012–2021. Emerg Infect Dis. 2023; 29:2054–64. https://doi.org/10.3201/eid2910.230382
4. Joseph A, Cointe A, Mariani Kurkdjian P, Rafat C, Hertig A. Shiga toxin–associated hemolytic uremic syndrome: a narrative review. Toxins (Basel). 2020;12:67. https://doi.org/10.3390/toxins12020067
5. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect. 2018;24:335–41. https://doi.org/10.1016/j.cmi.2017.10.013
6. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-genome sequencing for national surveillance of Shiga toxin–producing *Escherichia coli* O157. Clin Infect Dis. 2015;61:305–12. https://doi.org/10.1093/cid/civ318
7. Tolar B, Joseph LA, Schroeder MN, Stroika S, Ribot EM, Hise KB, et al. An overview of PulseNet USA databases. Foodborne Pathog Dis. 2019;16:457–62. https://doi.org/10.1089/fpd.2019.2637
8. Gerner-Smidt P, Besser J, Concepción-Acevedo J, Folster JP, Huffman J, Joseph LA, et al. Whole genome sequencing: bridging One-Health surveillance of foodborne diseases. Front Public Health. 2019;7:172. https://doi.org/10.3389/fpubh.2019.00172
9. Paranthaman K, Mook P, Curtis D, Evans EW, Crawley-Boevey E, Dabke G, et al. Development and evaluation of an outbreak surveillance system integrating whole genome sequencing data for non-typhoidal *Salmonella* in London and South East of England, 2016–17. Epidemiol Infect. 2021;149:e164. https://doi.org/10.1017/S0950268821001400
10. Nouws S, Verhaegen B, Denayer S, Crombé F, Piérard D, Bogaerts B, et al. Transforming Shiga toxin–producing *Escherichia coli* surveillance through whole genome sequencing in food safety practices. Front Microbiol. 2023; 14:1204630. https://doi.org/10.3389/fmicb.2023.1204630
11. Fruth A, Lang C, Größl T, Garn T, Flieger A. Genomic surveillance of STEC/EHEC infections in Germany 2020 to 2022 permits insight into virulence gene profiles and novel O-antigen gene clusters. Int J Med Microbiol. 2024;314:151610. https://doi.org/10.1016/j.ijmm.2024.151610
12. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. Front Microbiol. 2018;9:1482. https://doi.org/10.3389/fmicb.2018.01482
13. Berenger BM, Berry C, Peterson T, Fach P, Delannoy S, Li V, et al. The utility of multiple molecular methods including whole genome sequencing as tools to differentiate *Escherichia coli* O157:H7 outbreaks. Euro Surveill. 2015;20. https://doi.org/10.2807/1560-7917.ES.2015.20.47.30073
14. Joensen KG, Kiil K, Gantzhorn MR, Nauerby B, Engberg J, Holt HM, et al. Whole-genome sequencing to detect numerous *Campylobacter jejuni* outbreaks and match patient isolates to sources, Denmark, 2015–2017. Emerg Infect Dis. 2020;26:523–32. https://doi.org/10.3201/eid2603.190947
15. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol. 2016;2:16185. https://doi.org/10.1038/nmicrobiol.2016.185
16. Pijnacker R, van den Beld M, van der Zwaluw K, Verbruggen A, Coipan C, Segura AH, et al. Comparing multiple locus variable-number tandem repeat analyses with whole-genome sequencing as typing method for *Salmonella* enteritidis surveillance in The Netherlands, January 2019 to March 2020. Microbiol Spectr. 2022;10:e0137522. https://doi.org/10.1128/spectrum.01375-22
17. Rumore J, Tschetter L, Kearney A, Kandar R, McCormick R, Walker M, et al. Evaluation of whole-genome sequencing for outbreak detection of verotoxigenic *Escherichia coli* O157:H7 from the Canadian perspective. BMC Genomics. 2018;19:870. https://doi.org/10.1186/s12864-018-5243-3
18. Coipan CE, Dallman TJ, Brown D, Hartman H, van der Voort M, van den Berg RR, et al. Concordance of SNP- and allele-based typing workflows in the context of a large-scale international *Salmonella* enteritidis outbreak investigation. Microb Genom. 2020;6:e000318. https://doi.org/10.1099/mgen.0.000318
19. Radomski N, Cadel-Six S, Cherchame E, Felten A, Barbet P, Palma F, et al. A simple and robust statistical method to define genetic relatedness of samples related to outbreaks at the genomic scale—application to retrospective *Salmonella* foodborne outbreak investigations. Front Microbiol. 2019;10:2413. https://doi.org/10.3389/fmicb.2019.02413
20. Ribot EM, Freeman M, Hise KB, Gerner-Smidt P. PulseNet: entering the age of next-generation sequencing. Foodborne Pathog Dis. 2019;16:451–6. https://doi.org/10.1089/fpd.2019.2634
21. Achtman M, Zhou Z, Charlesworth J, Baxter L. EnteroBase: hierarchical clustering of 100,000s of bacterial genomes into species/subspecies and populations. Philos Trans R Soc Lond B Biol Sci. 2022;377:20210240. https://doi.org/10.1098/rstb.2021.0240
22. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M; Agama Study Group. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. Genome Res. 2020;30:138–52. https://doi.org/10.1101/gr.251678.119
23. Zhou Z, Charlesworth J, Achtman M. HierCC: a multi-level clustering scheme for population assignments based on core genome MLST. Bioinformatics. 2021;37:3645–6. https://doi.org/10.1093/bioinformatics/btab234
24. Hens NS, Aerts M, Faes C, Van Damme P, Beutels P. Modeling the prevalence and the force of infection directly from antibody levels. In: Gail M, Krickeberg K, Sarnet J, Tsiatis A, Wong W, editors. Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective. Amsterdam: Springer; 2012. p. 167–83.
25. Yu Y. mixR: an R package for finite mixture modeling for both raw and binned data. J Open Source Softw. 2022;7:4031. https://doi.org/10.21105/joss.04031

26. Letunic I, Bork P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49:W293–6. https://doi.org/10.1093/nar/gkab301
27. Haeghebaert S, Devien L, Lefèvre S, Mariani-Kurkdjian F, Sergentet D, Jones G. Multi-site collective food-borne illness with Shiga toxin–producing *E. coli* O157 (STEC), linked to consumption of raw cucumbers, Hauts-de-France, September 2021 [in French]. Med Infect Dis Edu. 2023;2:S132–3. https://doi.org/10.1016/j.mmifmc.2023.03.309
28. Jones G, Lefèvre S, Donguy MP, Nisavanh A, Terpant G, Fougère E, et al. Outbreak of Shiga toxin–producing *Escherichia coli* (STEC) O26 paediatric haemolytic uraemic syndrome (HUS) cases associated with the consumption of soft raw cow's milk cheeses, France, March to May 2019. Euro Surveill. 2019;24:1900305. https://doi.org/10.2807/1560-7917.ES.2019.24.22.1900305
29. Jones G, de Valk, H. Outbreak of Shiga toxin–producing *Escherichia coli* O26 infections linked to the consumption of raw milk reblochon cheese. France, March–May 2018 [in French] [cited 2024 Nov 29]. https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-infectieuses-d-origine-alimentaire/syndrome-hemolytique-et-uremique-pediatrique/documents/rapport-synthese/epidemie-d-infections-a-escherichia-coli-o26-producteur-de-shiga-toxines-liees-a-la-consommation-de-reblochon-au-lait-cru.-france-mars-mai-2018.
30. Cointe A, Birgy A, Mariani-Kurkdjian P, Liguori S, Courroux C, Blanco J, et al. Emerging multidrug-resistant hybrid pathotype Shiga toxin–producing *Escherichia coli* O80 and related strains of clonal complex 165, Europe. Emerg Infect Dis. 2018;24:2262–9. https://doi.org/10.3201/eid2412.180272
31. Hawkey J, Paranagama K, Baker KS, Bengtsson RJ, Weill FX, Thomson NR, et al. Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, *Shigella sonnei*. Nat Commun. 2021;12:2684. https://doi.org/10.1038/s41467-021-22700-4
32. Ingelbeen B, Bruyand M, Mariani-Kurkjian P, Le Hello S, Danis K, Sommen C, et al. Emerging Shiga-toxin–producing *Escherichia coli* serogroup O80 associated hemolytic and uremic syndrome in France, 2013–2016: differences with other serogroups. PLoS One. 2018;13:e0207492. https://doi.org/10.1371/journal.pone.0207492
33. Habets A, Crombé F, Nakamura K, Guérin V, De Rauw K, Piérard D, et al. Genetic characterization of shigatoxigenic and enteropathogenic *Escherichia coli* O80:H2 from diarrhoeic and septicaemic calves and relatedness to human shigatoxigenic *E. coli* O80:H2. J Appl Microbiol. 2021;130:258–64. https://doi.org/10.1111/jam.14759
34. Soleau N, Ganet S, Werlen S, Collignon L, Cointe A, Bonacorsi S, et al. First isolation of the heteropathotype Shiga toxin–producing and extra-intestinal pathogenic (STEC-ExPEC) *E. coli* O80:H2 in French healthy cattle: genomic characterization and phylogenetic position. Int J Mol Sci. 2024;25:5428. https://doi.org/10.3390/ijms25105428
35. Van Cauteren D, De Valk H, Vaux S, Le Strat Y, Vaillant V. Burden of acute gastroenteritis and healthcare-seeking behaviour in France: a population-based study. Epidemiol Infect. 2012;140:697–705. https://doi.org/10.1017/S0950268811000999
36. Besser JM, Carleton HA, Trees E, Stroika SG, Hise K, Wise M, et al. Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. Foodborne Pathog Dis. 2019;16:504–12. https://doi.org/10.1089/fpd.2019.2650
37. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, et al. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci U S A. 2012;109:3065–70. https://doi.org/10.1073/pnas.1121491109
38. Duval A, Opatowski L, Brisse S. Defining genomic epidemiology thresholds for common-source bacterial outbreaks: a modelling study. Lancet Microbe. 2023;4:e349–57. https://doi.org/10.1016/S2666-5247(22)00380-9

Address for correspondence: Gabrielle Jones, Santé Publique France, Direction des Maladies Infectieuses, 12 rue du Val d'Osne, 94415 Saint-Maurice CEDEX, France; email: gabrielle.jones@santepubliquefrance.fr

16S Ribosomal RNA Gene PCR and Sequencing for Pediatric Infection Diagnosis, United States, 2020–2023

Guyu Li, Christopher A. Reis, Rebecca M. Kruc, Ziyuan Zhang, Nicholas T. Streck, Elizabeth H. Ristagno, Jay Mandrekar, Matthew J. Wolf, James T. Gaensbauer, Robin Patel

Gene PCR and sequencing using 16S ribosomal RNA (rRNA) can help diagnose challenging bacterial infections. Data on the optimal clinical settings for this type of testing are limited. We performed a retrospective study at Mayo Clinic, Rochester, Minnesota, USA, with typically sterile specimens from children that underwent 16S rRNA PCR testing during September 2020–December 2023. Of 162 tests performed on 124 patients, 20% were positive; 58% of positive samples were from culture-negative specimens. Fluid specimens were >3 times as likely to test positive as tissue specimens (odds ratio 3.07 [95% CI 1.32–7.11]; p = 0.007), and pleural fluid demonstrated the highest positivity rate (50%). Of 33 positive results, 4 (12%) specimens qualified for reporting to the state health department for communicable diseases. Those single-laboratory findings demonstrate that the highest positivity rate of 16S rRNA PCR and sequencing is pleural fluid, although many specimen types tested positive.

Gene PCR using 16S ribosomal RNA (rRNA) followed by sequencing can identify bacteria in normally sterile body tissues and fluids (*1*,*2*). This method may serve as a diagnostic tool in complex bacterial infections, particularly when conventional tests fail to identify pathogens (*3*,*4*). The clinical use of 16S rRNA PCR and sequencing has been shown to yield concordant results with bacterial cultures (when positive), to enhance detection of fastidious bacteria, and to assist in antimicrobial drug stewardship (*4*–*8*). However, the diagnostic yield of 16S rRNA PCR and sequencing from various specimen sources has been variable in published studies (*4*,*6*,*9*–*11*); diagnostic yield may vary on the basis of patient and specimen characteristics. Data on optimal clinical settings and specimen selection for this testing are limited, particularly in pediatrics (*9*,*12*).

Mayo Clinic (Rochester, MN, USA) began offering 16S rRNA PCR and sequencing clinically in 2017; the sequencing initially involving Sanger sequencing alone (*4*). Then, in 2019, to increase positivity rates and to decatenate mixed sequences because of 16S rRNA gene copy variants or polymicrobial infections, next-generation sequencing (NGS) was substituted for or added to Sanger sequencing of the PCR-amplified 16S rRNA gene when needed (*13*). This study reviews Mayo Clinic's clinical experience with 16S rRNA PCR and sequencing of specimens from children to identify clinical syndromes where this testing is useful and to optimize specimen choice.

Methods

Study Design

We performed a retrospective study involving specimens collected from Mayo Clinic patients 0–18 years of age whose normally sterile tissue or fluid specimens underwent 16S rRNA PCR and sequencing during September 2020–December 2023. We identified patients and 16S rRNA PCR and sequencing results by using the clinical microbiology laboratory database and collected demographic, clinical, and microbiologic data from the electronic medical record. If a patient had specimens collected from the same source during different encounters, we included only specimens from the first encounter. In routine clinical practice, holding a specimen in the clinical microbiology

Author affiliations: Mayo Clinic, Rochester, Minnesota, USA (G. Li, C.A. Reis, R.M. Kruc, N.T. Streck, E.H. Ristagno, J. Mandrekar, M.J. Wolf, J.T. Gaensbauer, R. Patel); Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA (Z. Zhang)

DOI: https://doi.org/10.3201/eid3113.241101

laboratory for 14 days for potential 16S rRNA PCR and sequencing, if clinically needed, was offered as an option. This study was approved by the Mayo Clinic Institutional Review Board (protocol no. 20–012373).

Definitions

Immunocompromised hosts included patients with malignancies undergoing chemotherapy, those who had undergone solid organ or hematopoietic stem cell transplantation, and those receiving high-dose steroids (pulse dose steroids 20 mg/d for ≥14 days, or dexamethasone for ≥10 days) or other immunosuppressive agents. We defined intensive care unit (ICU) admission as receiving medical care in the neonatal, pediatric, or cardiovascular ICU at the time of specimen collection.

We categorized cerebrospinal fluid, ovarian fluid, pericardial fluid, peritoneal fluid, pleural fluid, subdural fluid, synovial fluid, and vitreous fluid as fluid specimens and other specimens (e.g., bone) as tissue specimens. We collected the results of conventional testing, which included Gram stain, bacterial culture, BioFire Meningitis and Encephalitis (ME) panel (bioMérieux, https://www.biomerieux.com), and *Kingella kingae* PCR if clinically performed on specimens collected at the same time and from the same site as specimens for 16S rRNA PCR and sequencing. We calculated the turnaround time as the interval from specimen collection to result finalization. We defined prior antibacterial therapy as any antimicrobial drugs administered within 24 hours before the test order for 16S rRNA PCR and sequencing.

Specimen Processing

We performed specimen processing and bacterial culture in the Clinical Bacteriology Laboratories of the Division of Clinical Microbiology at Mayo Clinic. We identified isolated bacteria by using conventional biochemical methods or matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Details of the 16S rRNA PCR and sequencing procedure have been described previously (*13*). In brief, the assay involved an up-front real-time PCR assay, reported as negative or submitted to Sanger or NGS on the basis of cycle threshold (Ct) value. Specimens with Ct values <32 cycles underwent bidirectional Sanger sequencing by using an Applied Biosystems 3500xL Genetic Analyzer (Thermo Fisher Scientific, https://www.thermofisher.com). We sent specimens with Ct values of 32–34 or <32 with Sanger sequencing that yielded an uninterpretable result to NGS by using an Illumina MiSeq System (Illumina, https://www.illumina.com) with a 500-cycle (2 × 250 paired-end read) v2 nano kit. We reported specimens with Ct values >34 as negative, except if we observed a well-defined melting temperature peak (≥0.4), in which case we sent them to NGS. We used Pathogenomix (https://www.pathogenomix.com) for quality control processes and the Pathogenomix PRIME database for sequence analysis. The Pathogenomix Prime database contains 48,139 curated 16S rRNA gene sequences. The processor filters low-quality reads (Q<30) and clusters sequences on the basis of ≥210-bp length, ≥100 copies, and 0% variation.

Statistical Analysis

We compared characteristics between positive and negative tests by using a 2-sample *t*-test for continuous variables. For categorical variables with ≥5 observations, we calculated odds ratios (ORs) and 95% CIs by using unconditional maximum likelihood estimation; we obtained p values by using χ^2 tests. For categorical variables with <5 observations, we calculated ORs and 95% CIs by using conditional maximum likelihood estimation and obtained p values were by using Fisher exact tests. We considered a 2-tailed p value <0.05 statistically significant.

Results

Patients

A total of 124 pediatric patients with 162 tests from typically sterile sources were included (Figure). At sampling, 20% (n = 24) of patients were identified as immunocompromised hosts, and 37% (n = 46) of patients were in ICUs (Table 1). The most common suspected clinical manifestations were meningoencephalitis, musculoskeletal infection, and pleural effusion.

16S rRNA PCR and Sequencing Results

The mean turnaround time for positive 16S rRNA PCR and sequencing tests was 8 days (3.2–12.8 days), whereas for negative tests it was 3 days (0–6.7 days) (Table 2). A total of 84 (50%) specimens were collected from patients who received antimicrobial drugs within 24 hours before sampling, which was associated with a higher likelihood of positive results (p = 0.001).

The overall 16S rRNA PCR and sequencing positivity rate was 20% among all 162 specimens collected from 124 patients (Figure). Fluid specimens were 3-fold more likely to test positive compared with tissue specimens (OR 3.07 [95% CI 1.32–7.11]; p = 0.007) (Table 2). The most frequent specimen sources were cerebrospinal fluid, bone tissue, deep soft tissue, synovial fluid, and pleural fluid. Among those, specimens

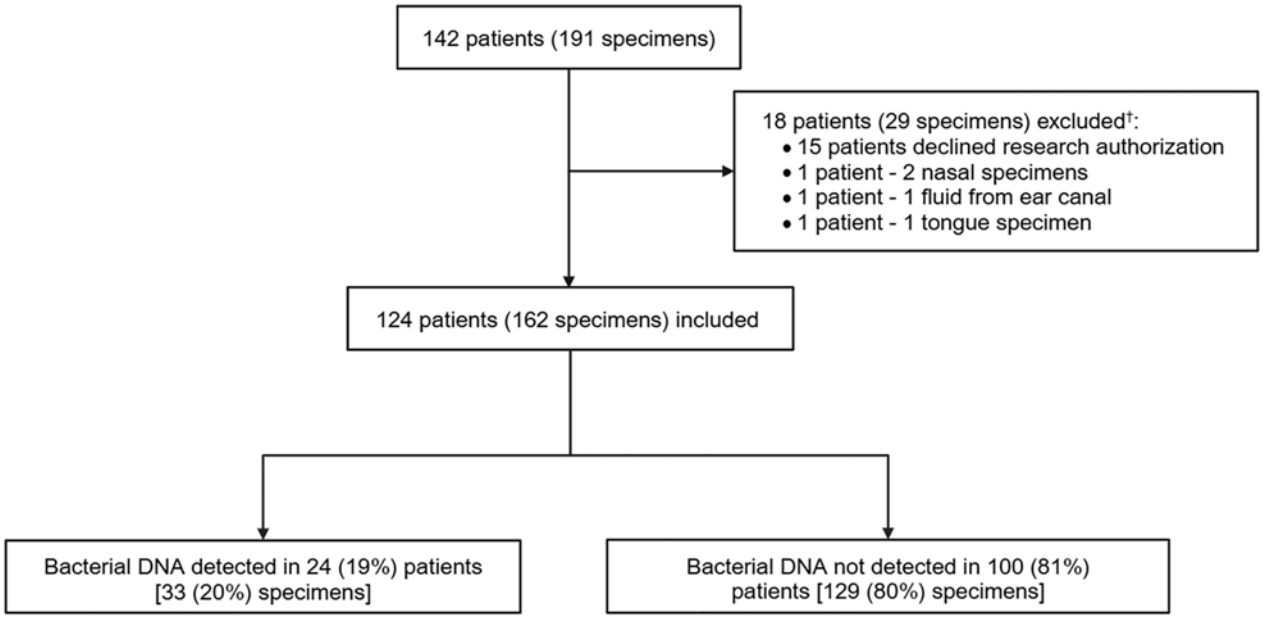


Figure. Specimen flowchart from a study of 16S ribosomal RNA gene PCR and sequencing for pediatric infection diagnosis, Mayo Clinic, Rochester, Minnesota, USA, 2020–2023. Specimens from tongue, ear canal, and nose were excluded.

with high positivity rates included pleural fluid (50%, n = 5) and synovial fluid (43%, n = 9); there were no positive results from deep soft tissue specimens.

Among the 33 positive tests, 12 (36%) tests were polymicrobial detections. The most common single bacteria identified was *Staphylococcus aureus* complex in 4 (12%) positive tests, followed by *Kingella kingae* in 3 (9%) positive tests (all synovial fluid); other bacteria each accounted for 3%–9% of positive tests from various sources (Table 3). We recorded details of test results and clinical diagnoses for 24 patients with positive 16S rRNA PCR and sequencing results (Appendix Table, https://wwwnc.cdc.gov/EID/article/31/13/24-1101-App1.pdf).

Table 1. Patient characteristics from study of 16S ribosomal RNA gene PCR and sequencing for pediatric infection diagnosis, Mayo Clinic, Rochester, Minnesota, USA, 2020–2023*

Patient characteristics	Value, n = 124 patients
Median age, y (IQR)	9.6 (2.2–15.0)
Sex	
F	58 (47)
M	66 (53)
Immunocompromised host†	24 (20)
Intensive care unit admission	46 (37)
Suspected clinical syndrome	
Meningoencephalitis	38 (31)
Musculoskeletal infection: septic arthritis, osteomyelitis	33 (27)
Pleural effusion	10 (8)
Surgical wound infection, including hardware infection	10 (8)
Lymphadenopathy	5 (4)
Bone mass	4 (3)
Intracranial abscess/fluid collection	2 (2)
Pericardial effusion	4 (3)
Pulmonary nodules	4 (3)
Traumatic wound infection	4 (3)
Intraabdominal abscess/fluid collection	4 (3)
Endocarditis	2 (2)
Splenic mass	1 (1)
Infected pseudoaneurysms	1 (1)
Mediastinitis	1 (1)
Retinal detachment	1 (1)
Suggested by pediatric infectious diseases team	56 (45)

*Values are no. (%) patients except as indicated. IQR, interquartile range.
†Immunocompromised hosts include those with history of solid organ transplant, history of hematopoietic stem cell transplant, active chemotherapy, or receiving an immunosuppressive agent.

Table 2. Specimen characteristics from study of 16S ribosomal RNA gene PCR and sequencing for pediatric infection diagnosis, Mayo Clinic, Rochester, Minnesota, USA, 2020–2023*

Specimen characteristics	Positive 16S rRNA PCR and sequencing, n = 33	Negative 16S rRNA PCR and sequencing, n = 129	OR (95% CI)	p value
Turnaround time, d, mean ± SD	8.0 ± 4.8	3.0 ± 3.7	NA	<0.0001
Received antimicrobial drugs within 24 h	28 (85)	56 (43)	7.30 (2.65–20.11)	0.001
Specimen hold ordered before testing	3 (9)	14 (11)	0.82 (0.22–3.04)	1.000
Specimen type				
Fluid	24 (73)	60 (47)	3.07 (1.32–7.11)	0.007
Tissue	9 (27)	69 (54)	0.33 (0.14–0.76)	0.007
Specimen source†				
Cerebrospinal fluid	5 (15)	33 (26)	NA	NA
Bone tissue	4 (12)	24 (19)	NA	NA
Deep soft tissue	0	22 (17)	NA	NA
Synovial fluid	9 (27)	12 (9)	NA	NA
Pleural fluid	5 (15)	5 (4)	NA	NA
Synovial tissue	2 (6)	7 (5)	NA	NA
Lymph node	0	5 (4)	NA	NA
Subdural fluid	3 (9)	1 (1)	NA	NA
Pericardial fluid	0	4 (3)	NA	NA
Peri-implant tissue	2 (6)	2 (2)	NA	NA
Peritoneal fluid	1 (3)	2 (2)	NA	NA
Lung parenchyma	0	3 (2)	NA	NA
Pacemaker pocket tissue	0	3 (2)	NA	NA
Vitreous fluid	1 (3)	1 (1)	NA	NA
Brain tissue	0	2 (2)	NA	NA
Ovarian fluid	0	1 (1)	NA	NA
Heart valve tissue	1 (3)	0	NA	NA
Spleen tissue	0	1 (1)	NA	NA
Vascular tissue	0	1 (1)	NA	NA

*Values are no. (%) tests except as indicated. NA, not applicable; OR, odds ratio.
†Statistical analysis was not performed because of the limited sample size.

Comparison to Conventional Tests

Among 152 specimens tested with both Gram stain and 16S rRNA PCR and sequencing, 21% (n = 7) of positive specimens had a corresponding positive Gram stain, whereas none of the negative tests were associated with a positive Gram stain. Patients with positive Gram stains had a higher likelihood of positive 16S rRNA PCR and sequencing results compared with patients with negative Gram stains (p<0.0001) (Table 4).

Of the 161 specimens tested with both bacterial cultures and 16S rRNA PCR and sequencing, 133 (83%) specimens demonstrated concordant results between the 2 methods: 14 (9%) specimens were positive after both tests and 119 (74%) specimens were negative after both tests. In addition, 19 specimens with negative bacterial cultures were positive by 16S rRNA PCR and sequencing: polymicrobial infections (n = 9), *K. kingae* (n = 3), *Fusobacterium naviforme/nucleatum* (n = 2), *Streptococcus mitis* group (n = 2), *Cardiobacterium hominis* (n = 1), *Pseudomonas aeruginosa* (n = 1), and *Streptococcus pyogenes* (n = 1).

Table 3. Microorganisms detected by 16S rRNA PCR and sequencing and associated specimen sources from a study of 16S ribosomal RNA gene PCR and sequencing for pediatric infection diagnosis, Mayo Clinic, Rochester, Minnesota, USA, 2020–2023

Identified bacteria	No. (%) positive results, n = 33	Specimen source (no. tests)
Polymicrobial	12 (36)	Bone tissue (4*), subdural fluid (3*), cerebrospinal fluid (1), peri-implant tissue (1), vitreous fluid (1*), synovial tissue (1), peritoneal fluid (1*)
Staphylococcus aureus complex	4 (12)	Synovial fluid (4)
Kingella kingae	3 (9)	Synovial fluid (3*)
Streptococcus mitis group	3 (9)	Pleural fluid (1), pleural fluid (2*)
Fusobacterium naviforme/nucleatum	2 (6)	Synovial fluid (1*), CSF (1*)
Streptococcus intermedius	2 (6)	Cerebrospinal fluid (2)
Cardiobacterium hominis	1 (3)	Peri-implant tissue (1*)
Enterococcus faecalis	1 (3)	Heart valve tissue (1)
Fusobacterium necrophorum	1 (3)	Pleural fluid (1)
Pseudomonas aeruginosa	1 (3)	Synovial tissue (1*)
Staphylococcus epidermidis	1 (3)	Cerebrospinal fluid (1)
Streptococcus dysgalactiae	1 (3)	Synovial fluid (1*)
Streptococcus pyogenes	1 (3)	Pleural fluid (1*)

*Specimens with negative bacterial cultures from the same specimen source.

Table 4. Comparison of conventional tests and 16S rRNA PCR and sequencing results from study of 16S ribosomal RNA gene PCR and sequencing for pediatric infection diagnosis, Mayo Clinic, Rochester, Minnesota, USA, 2020–2023*

Conventional tests characteristics	16S rRNA PCR and sequencing results		p value
	Positive	Negative	
Gram stain			
Positive	7	0	<0.0001
Negative	26	119	
Bacterial culture			
Positive	14	9	<0.0001
Negative	19	119	
BioFire Meningitis/Encephalitis panel, cerebrospinal fluid only			
Positive	0	2†	1.000
Negative	2	19	
Synovial fluid *Kingella kingae* PCR, when clinically ordered			
Positive	3	0	0.143
Negative	1	3	

*Specimens were collected from same specimen source.
†One panel was positive for parechovirus and another was positive for human herpesvirus 6.

Nine 16S rRNA PCR and sequencing tests were negative despite positive cultures: 4 positive results for *Cutibacterium acnes* from bacterial cultures in periimplant and bone tissue, 1 positive result for *Staphylococcus capitis* from bacterial culture in a bone tissue specimen, and 5 cases of suspected culture contamination. The contamination cases involved isolations of *Staphylococcus epidermidis* from pleural fluid (n = 1), deep soft tissue (n = 1), and lymph node tissue (n = 1) and *Niallia circulans* group from bone (n = 1) and deep soft tissue (n = 1).

Of the 23 specimens tested with both the BioFire ME panel and 16S rRNA PCR and sequencing, 2 were negative by the panel with positive 16S rRNA PCR and sequencing results (*S. epidermidis* and *F. naviforme/nucleatum*). The *S. epidermidis* case was considered a contaminant. No bacterial pathogens were identified by the BioFire ME panel that were not also detected by 16S rRNA PCR and sequencing.

Of the 7 synovial fluid specimens tested with both *K. kingae* PCR and 16S rRNA PCR and sequencing, 86% (n = 6) of specimens showed concordant positive or negative results. 16S rRNA PCR and sequencing detected *K. kingae* in 1 synovial fluid specimen that tested negative with synovial fluid *K. kingae* PCR.

Multiple Tests on the Same Specimen Type

At least 2 16S rRNA PCR and sequencing tests were ordered for 22 patients on the same specimen source during the same procedure (Table 5), mostly bone tissue, deep soft tissue, and synovial fluid. All tests yielded concordant results, either negative or positive.

Specimen Hold Strategy

Clinicians placed a request to hold a specimen for potential future 16S rRNA PCR and sequencing testing on 17 specimens (Table 2). Over the ensuing clinical course, because of positive Gram stains and negative bacterial cultures after 24–48 hours of incubation, 16S rRNA PCR and sequencing tests were performed on the saved specimens. Of those, 3 tests had positive 16S rRNA PCR and sequencing results, including identification of *S. mitis* group in 2 pleural fluid specimens and *S. dysgalactiae* in 1 synovial fluid specimen.

Discussion

In this study, we conducted a 3-year retrospective evaluation of the diagnostic yield of 16S rRNA PCR and sequencing in children by using various specimen types. We were unable to find many other

Table 5. Characteristics of multiple 16S rRNA PCR and sequencing tests ordered on the same patient from the same specimen source during the same procedure from study of 16S ribosomal RNA gene PCR and sequencing for pediatric infection diagnosis, Mayo Clinic, Rochester, Minnesota, USA, 2020–2023*

Characteristics	Bone tissue, n = 7	Deep soft tissue, n = 6	Synovial fluid, n = 4	Cerebrospinal fluid, n = 2	Brain tissue, n = 1	Pacemaker pocket tissue, n = 1	Subdural fluid, n = 1
Frequency of tests ordered							
Two	4 (57)	3 (50)	3 (75)	2 (100)	1 (100)	1 (100)	0
Three	2 (29)	3 (50)	1 (25)	0	0	0	1 (100)
Four	1 (14)	0	0	0	0	0	0
Assay results							
Concordant negative	6 (86)	6 (100)	3 (75)	1 (50)	1 (100)	1 (100)	0
Concordant positive	1 (14)	0	1 (25)	1 (50)	0	0	1 (100)
Discordant	0	0	0	0	0	0	0

*Values are no. (%) patients.

published studies exploring the application of 16S rRNA PCR and sequencing in pediatric patients. The overall test positivity rate we found was 20%, consistent with previous studies in pediatric patients, which reported positivity rates ranging from 14% to 23% (*6*,*9*,*10*,*14*). Initiation of empiric therapy within 24 hours before sampling did not negatively affect positivity rates, consistent with findings from studies in adults and children (*4*,*9*,*11*).

Subgroup analysis revealed that 16S rRNA PCR and sequencing had a higher positivity rate in fluid compared with tissue specimens, especially in pleural fluid, which provided additional diagnostic value for pathogens such as *S. mitis* group and *S. pyogenes*. Despite the limited pediatric sample size, our findings are consistent with prior studies indicating that pleural fluid yields a high positivity rate (*10*,*14*,*15*).

A potential limitation of our study is that bronchoalveolar lavage (BAL) fluid was not tested; in prior studies, BAL fluid has been reported as a common specimen source for 16S rRNA PCR and sequencing testing. However, despite high positivity rates in BAL fluids, the clinical relevance of those findings has been questionable (*6*,*9*,*16*), possibly because BAL fluid is not sterile. In contrast, sample dilution during bronchoscopy can increase the likelihood of false-negative results.

The yield of 16S rRNA PCR and sequencing in bone and joint infection has varied in previous research, ranging from 21% to 32% (*17*–*19*). Bone tissue and synovial fluids or tissues were the most common sources in this study. Compared with the single *K. kingae* PCR test on synovial fluid used at the Mayo Clinic, the 16S rRNA PCR and sequencing offered additional diagnostic value in only 1 of 7 cases. Given the shorter turnaround time of the *K. kingae* PCR test, a single PCR test remains the optimal first-line test for suspected bone and joint infections in toddlers. This target is also available on the BioFire joint infection (JI) panel (*20*). The BioFire JI panel has been used for rapid diagnosis of pediatric septic arthritis, offering a fast turnaround, and sensitive and specific detection of on-panel microorganisms and select antimicrobial resistance genes (*20*,*21*). Compared with the BioFire JI panel, 16S rRNA PCR and sequencing demonstrated higher sensitivity in periprosthetic JI (PJI) because the BioFire JI panel does not include *S. epidermidis*, a common cause of PJI (*22*–*24*).

We found discrepancies in *C. acnes* testing, in which cultures were positive but 16S rRNA PCR and sequencing was negative (4 peri-implant and bone tissue specimens). Those discrepancies are likely because of the limited ability to report low abundance *C. acnes* from 16S rRNA PCR and sequencing because of its frequent presence in background sequences, as published previously (*25*,*26*).

Of the 23 BioFire ME panels performed, 19 had concordant negative results by 16S rRNA PCR and sequencing, in keeping with other studies' findings (*4*,*27*). 16S rRNA PCR and sequencing uniquely identified *F. naviforme/nucleatum*, which is not included in the ME panel (*28*). Two cases were negative by 16S rRNA PCR and sequencing but positive for viruses by the BioFire ME panel; this is expected because 16S rRNA PCR and sequencing targets bacterial DNA, while the BioFire ME panel includes viral targets. Turnaround time is a key factor to consider. Our findings underscore the value of using the BioFire ME panel ahead of 16S rRNA PCR and sequencing, proceeding to 16S rRNA PCR and sequencing when the BioFire ME panel is negative (*29*).

In this study, multiple 16S rRNA PCR and sequencing tests performed on specimens from the same specimen source collected during the same procedure resulted in no discordant results. Assessment of the clinical value of performing multiple 16S rRNA PCR and sequencing tests has been limited. A multicenter study on adult PJI showed that collecting 5 perioperative samples per patient for culture and 16S rRNA PCR and sequencing showed a lack of sensitivty of the latter in the diagnosis of PJI (*30*). Another report indicated that testing multiple samples per patient may help rule out potential contaminating microorganisms (*31*). Our findings indicate a single 16S rRNA PCR and sequencing test on 1 specimen, collected along with at $\geq$2 specimens for bacterial culture during the same procedure, may be adequate.

This study also explored the role of collecting and holding a specimen for future testing if clinically indicated. Positive detections were found in 3 cases managed with this strategy. We conceive that use of this diagnostic pathway could optimize testing resource use. Further research with larger sample sizes is necessary to determine the clinical syndromes and specimen sources that would benefit from delayed or reflexive testing.

The use of 16S rRNA PCR and sequencing in clinical practice has implications for public health, including enhanced detection of bacteria that may be notifiable infectious diseases. Clinical laboratories should establish protocols for reporting detected pathogens to public health authorities, and public health laboratories should define which molecularly detected species are reportable from which specimen types. As demonstrated in this study, *K. kingae*, often missed by conventional cultures, is readily detected by 16S

rRNA PCR and sequencing. Clinical use of this assay can provide data useful for identifying outbreaks and informing timely public health interventions (*32*).

The first limitation of this study is that the small sample size limits statistical power. Second, the study was conducted at a single institution, limiting generalizability. Finally, subgroup analysis of suspected clinical syndromes and outcomes was not performed. Future studies with larger sample sizes, specimens collected from multiple sites, comprehensive clinical outcomes recorded, and adjustments for potential confounders are warranted.

In conclusion, this study demonstrates that 16S rRNA PCR and sequencing yields the highest positivity rate in fluid specimens, particularly pleural and synovial fluids from children. A strategy of collecting specimens for future testing, if clinically indicated, is described as a diagnostic stewardship tool. Further research should focus on optimizing use of the described testing use in conjunction with other testing, while considering overall turnaround time. Implementation research is needed to evaluate the effect of 16S rRNA PCR and sequencing on patient outcomes.

Author contributions: G.L., C.A.R., and J.T.G. designed the study; G.L., C.A.R., R.M.K., and N.T.S. collected the data; G.L., Z.Z., and J.M. analyzed the data; M.J.W. and R.P. developed the assay; G.L., Z.Z., and R.M.K. wrote the manuscript; C.A.R., R.M.K., N.T.S., E.H.R., J.M., J.T.G., and R.P. revised the manuscript.

About the Author

Dr. Li is a physician and assistant professor of pediatrics in the Division of Pediatric Infectious Diseases, Department of Pediatrics, at the Mayo Clinic. His primary research interests include novel molecular microbiological diagnostics in children and using phage therapy to treat multidrug-resistant bacteria.

References

1. Drevinek P, Hollweck R, Lorenz MG, Lustig M, Bjarnsholt T. Direct 16S/18S rRNA PCR followed by Sanger sequencing as a clinical diagnostic tool for detection of bacterial and fungal infections: a systematic review and meta-analysis. J Clin Microbiol. 2023;61:e0033823. https://doi.org/10.1128/jcm.00338-23
2. Church DL, Cerutti L, Gürtler A, Griener T, Zelazny A, Emler S. Performance and application of 16S rRNA gene cycle sequencing for routine identification of bacteria in the clinical microbiology laboratory. Clin Microbiol Rev. 2020;33:e00053–19. https://doi.org/10.1128/CMR.00053-19
3. Rampini SK, Bloemberg GV, Keller PM, Büchler AC, Dollenmaier G, Speck RF, et al. Broad-range 16S rRNA gene polymerase chain reaction for diagnosis of culture-negative bacterial infections. Clin Infect Dis. 2011;53:1245–51. https://doi.org/10.1093/cid/cir692
4. Fida M, Khalil S, Abu Saleh O, Challener DW, Sohail MR, Yang JN, et al. Diagnostic value of 16S ribosomal RNA gene polymerase chain reaction/Sanger sequencing in clinical practice. Clin Infect Dis. 2021;73:961–8. https://doi.org/10.1093/cid/ciab167
5. Akram A, Maley M, Gosbell I, Nguyen T, Chavada R. Utility of 16S rRNA PCR performed on clinical specimens in patient management. Int J Infect Dis. 2017;57:144–9. https://doi.org/10.1016/j.ijid.2017.02.006
6. Lucas EJ, Leber A, Ardura MI. Broad-range PCR application in a large academic pediatric center: clinical value and challenges in diagnosis of infectious diseases. Pediatr Infect Dis J. 2019;38:786–90. https://doi.org/10.1097/INF.0000000000002308
7. Clarridge JE III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004;17:840–62. https://doi.org/10.1128/CMR.17.4.840-862.2004
8. Ursenbach A, Schramm F, Séverac F, Hansmann Y, Lefebvre N, Ruch Y, et al. Revised version (INFD-D-20-00242): impact of 16S rDNA sequencing on clinical treatment decisions: a single center retrospective study. BMC Infect Dis. 2021;21:190. https://doi.org/10.1186/s12879-021-05892-4
9. Naureckas Li C, Nakamura MM. Utility of broad-range PCR sequencing for infectious diseases clinical decision making: a pediatric center experience. J Clin Microbiol. 2022;60:e0243721. https://doi.org/10.1128/jcm.02437-21
10. Lim PPC, Stempak LM, Malay S, Moore LN, Cherian SSS, Desai AP. Determining the clinical utility of 16S rRNA sequencing in the management of culture-negative pediatric infections. Antimicrobial drugs. Basel. 2022;11:159. https://doi.org/10.3390/antimicrobial drugs11020159
11. Eamsakulrat P, Santanirand P, Phuphuakrat A. Diagnostic yield and impact on antimicrobial management of 16S rRNA testing of clinical specimens. Microbiol Spectr. 2022;10:e0209422. https://doi.org/10.1128/spectrum.02094-22
12. Basein T, Gardiner BJ, Andujar Vazquez GM, Joel Chandranesan AS, Rabson AR, Doron S, et al. Microbial identification using DNA target amplification and sequencing: clinical utility and impact on patient management. Open Forum Infect Dis. 2018;5:ofy257. https://doi.org/10.1093/ofid/ofy257
13. Flurin L, Wolf MJ, Mutchler MM, Daniels ML, Wengenack NL, Patel R. Targeted metagenomic sequencing-based approach applied to 2146 tissue and body fluid samples in routine clinical practice. Clin Infect Dis. 2022;75:1800–8. https://doi.org/10.1093/cid/ciac247
14. Mongkolrattanothai K, Dien Bard J. The utility of direct specimen detection by Sanger sequencing in hospitalized pediatric patients. Diagn Microbiol Infect Dis. 2017;87:100–2. https://doi.org/10.1016/j.diagmicrobio.2016.10.024
15. Kerkhoff AD, Rutishauser RL, Miller S, Babik JM. Clinical utility of universal broad-range polymerase chain reaction amplicon sequencing for pathogen identification: a retrospective cohort study. Clin Infect Dis. 2020;71:1554–7. https://doi.org/10.1093/cid/ciz1245
16. Zachariah P, Ryan C, Nadimpalli S, Coscia G, Kolb M, Smith H, et al. Culture-independent analysis of pediatric bronchoalveolar lavage specimens. Ann Am Thorac Soc. 2018;15:1047–56. https://doi.org/10.1513/AnnalsATS.201802-146OC
17. Jensen KH, Dargis R, Christensen JJ, Kemp M. Ribosomal PCR and DNA sequencing for detection and identification

of bacteria: experience from 6 years of routine analyses of patient samples. APMIS. 2014;122:248–55. https://doi.org/10.1111/apm.12139
18. Grif K, Heller I, Prodinger WM, Lechleitner K, Lass-Flörl C, Orth D. Improvement of detection of bacterial pathogens in normally sterile body sites with a focus on orthopedic samples by use of a commercial 16S rRNA broad-range PCR and sequence analysis. J Clin Microbiol. 2012;50:2250–4. https://doi.org/10.1128/JCM.00362-12
19. Alraddadi B, Al-Azri S, Forward K. Influence of 16S ribosomal RNA gene polymerase chain reaction and sequencing on antimicrobial drug management of bone and joint infections. Can J Infect Dis Med Microbiol. 2013;24:85–8. https://doi.org/10.1155/2013/747145
20. Esteban J, Salar-Vidal L, Schmitt BH, Waggoner A, Laurent F, Abad L, et al. Multicenter evaluation of the BIOFIRE joint infection panel for the detection of bacteria, yeast, and AMR genes in synovial fluid samples. J Clin Microbiol. 2023;61:e0035723. https://doi.org/10.1128/jcm.00357-23
21. Gaillard T, Dupieux-Chabert C, Roux AL, Tessier E, Boutet-Dubois A, Courboulès C, et al. A prospective multicentre evaluation of BioFire® joint infection panel for the rapid microbiological documentation of acute arthritis. Clin Microbiol Infect. 2024;30:905–10. https://doi.org/10.1016/j.cmi.2024.03.022
22. Azad MA, Wolf MJ, Strasburg AP, Daniels ML, Starkey JC, Donadio AD, et al. Comparison of the BioFire joint infection panel to 16S ribosomal RNA gene-based targeted metagenomic sequencing for testing synovial fluid from patients with knee arthroplasty failure. J Clin Microbiol. 2022;60:e0112622. https://doi.org/10.1128/jcm.01126-22
23. Tai DBG, Patel R, Abdel MP, Berbari EF, Tande AJ. Microbiology of hip and knee periprosthetic joint infections: a database study. Clin Microbiol Infect. 2022;28:255–9. https://doi.org/10.1016/j.cmi.2021.06.006
24. Zeller V, Kerroumi Y, Meyssonnier V, Heym B, Metten MA, Desplaces N, et al. Analysis of postoperative and hematogenous prosthetic joint-infection microbiological patterns in a large cohort. J Infect. 2018;76:328–34. https://doi.org/10.1016/j.jinf.2017.12.016
25. Namdari S, Nicholson T, Abboud J, Lazarus M, Ramsey ML, Williams G, et al. *Cutibacterium acnes* is less commonly identified by next-generation sequencing than culture in primary shoulder surgery. Shoulder Elbow. 2020;12:170–7. https://doi.org/10.1177/1758573219842160
26. Dyrhovden R, Rippin M, Øvrebø KK, Nygaard RM, Ulvestad E, Kommedal Ø. Managing contamination and diverse bacterial loads in 16S rRNA deep sequencing of clinical samples: Implications of the law of small numbers. MBio. 2021;12:e0059821. https://doi.org/10.1128/mBio.00598-21
27. Esparcia O, Montemayor M, Ginovart G, Pomar V, Soriano G, Pericas R, et al. Diagnostic accuracy of a 16S ribosomal DNA gene-based molecular technique (RT-PCR, microarray, and sequencing) for bacterial meningitis, early-onset neonatal sepsis, and spontaneous bacterial peritonitis. Diagn Microbiol Infect Dis. 2011;69:153–60. https://doi.org/10.1016/j.diagmicrobio.2010.10.022
28. Posnakoglou L, Siahanidou T, Syriopoulou V, Michos A. Impact of cerebrospinal fluid syndromic testing in the management of children with suspected central nervous system infection. Eur J Clin Microbiol Infect Dis. 2020; 39:2379–86. https://doi.org/10.1007/s10096-020-03986-6
29. Leber AL, Everhart K, Balada-Llasat JM, Cullison J, Daly J, Holt S, et al. Multicenter evaluation of BioFire FilmArray meningitis/encephalitis panel for detection of bacteria, viruses, and yeast in cerebrospinal fluid specimens. J Clin Microbiol. 2016;54:2251–61. https://doi.org/10.1128/JCM.00730-16
30. Bémer P, Plouzeau C, Tande D, Léger J, Giraudeau B, Valentin AS, et al.; Centre de Référence des Infections Ostéo-articulaires du Grand Ouest Study Team. Evaluation of 16S rRNA PCR sensitivity and specificity for diagnosis of prosthetic joint infection: a prospective multicenter cross-sectional study. J Clin Microbiol. 2014;52:3583–9. https://doi.org/10.1128/JCM.01459-14
31. Wallander K, Vondracek M, Giske CG. Evaluation of multi-sample 16S ribosomal DNA sequencing for the diagnosis of postoperative bone and joint infections during antimicrobial treatment. BMC Res Notes. 2022;15:113. https://doi.org/10.1186/s13104-022-05992-7
32. Sacchi CT, Whitney AM, Mayer LW, Morey R, Steigerwalt A, Boras A, et al. Sequencing of 16S rRNA gene: a rapid tool for identification of *Bacillus anthracis*. Emerg Infect Dis. 2002;8:1117–23. https://doi.org/10.3201/eid0810.020391

Address for correspondence: Guyu Li, Mayo Clinic, 200 1st St SW, Rochester, MN 55905, USA; email: li.guyu@mayo.edu

Vincent van Gogh (1853–1890), *The Starry Night* (1889). Oil on canvas, 29 in × 36¼ in/73.7 cm × 92.1 cm. Museum of Modern Art, New York, New York, USA. Public domain image from Google Art Project.

Beyond the Brushstrokes—Illuminating Patterns and Interactions to Find Order in Complex Systems

Duncan MacCannell, Bronwyn MacInnis, Scott Santibanez

Vincent van Gogh's *The Starry Night* is widely considered a postimpressionist masterpiece and is one of the most recognizable pieces of art in modern history. It also presents a metaphor for public health innovation, particularly the advances in pathogen genomics and genomic epidemiology that are the central theme of this supplemental issue of *Emerging Infectious Diseases*.

The Starry Night was one of more than 150 paintings that van Gogh produced in 1889 and shared with a doctor friend during a year of recovery and intense personal turmoil in Saint-Rémy-de-Provence. The work captures the swirling clouds and interconnected stars of the night sky with a

Author affiliations: Centers for Disease Control and Prevention, Atlanta, Georgia, USA (D. MacCannell, S. Santibanez); The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA (B. MacInnis)

DOI: https://doi.org/10.3201/eid3113.250641

staccato of short and purposeful brush strokes. Up close, these individual strokes seem chaotic and disjointed, but they resolve into coherent dimensions of time and space when viewed from afar. In pathogen genomics, each genetic sequence, much like a brushstroke, provides only a glimpse of a pathogen's genomic structure, characteristics, or origins. When viewed through a bioinformatic lens, those sequence fragments can be aligned and assembled into a more complete picture, with further resolution of the genomic landscape, and a new understanding of its features that emerges when each fragment is appreciated in the context of hundreds or even thousands of others.

The painting also confers a sense of movement and unity: the whirling clouds, punctuated by celestial bodies that capture the viewer's attention and evoke feelings of connection and interaction. Similarly, genomic epidemiology integrates molecular and epidemiologic data to illuminate complex patterns of disease transmission and the interactions among pathogens, hosts, and entire populations. Just as van Gogh used the clouds to connect the stars in his sky, genomic epidemiologists draw links between cases, mapping transmission chains and identifying sources of infection.

The shapes and colors of this painting echo the data visualization methods that have emerged over the past decade to help scientists communicate complex phylogenetic and phylogeographic data to other public health professionals, policymakers, and the public. Visualization tools such as phylogenetic trees, flowcharts, heatmaps, and transmission networks help to translate nuanced genomic data into understandable narratives through color and form.

For van Gogh, this painting also reflects an important period of introspection and discovery, as well as an attempt to capture and interpret his environment. Applied public health and infectious disease research share a similar intention, focused on better understanding and responding to infectious disease threats. In a sense, both represent an effort to find order in complex systems, and to reveal the hidden patterns that connect them.

Bibliography

1. du Plessis A. 'Starry Night' van Gogh—in-depth analysis and facts. Art in Context. October 21, 2021 [cited 2025 Apr 10]. https://artincontext.org/starry-night-van-gogh
2. Kumekina V. Masterpiece story: The Starry Night by Vincent van Gogh. Daily Art Magazine. 2025 May 4 [cited 2025 May 4]. https://www.dailyartmagazine.com/masterpiece-story-the-starry-night-by-vincent-van-gogh
3. MoMA. Vincent van Gogh. The Starry Night. Saint Rémy, June 1889 [cited 2025 Apr 10]. https://www.moma.org/collection/works/79802
4. Saatchi C. The Starry Night [cited 2025 Apr 10]. https://www.charlessaatchi.com/the-starry-night
5. The Metropolitan Museum of Art. The Starry Night [cited 2025 Apr 10]. https://www.metmuseum.org/art/collection/search/828514
6. The Van Gogh Gallery. Vincent Van Gogh: Starry Night [cited 2025 Apr 10]. https://www.vangoghgallery.com/painting/starry-night.html
7. The Vincent Van Gogh Gallery. Starry Night [cited 2025 Apr 10]. https://www.vggallery.com/painting/p_0612.htm

Address for correspondence: Duncan MacCannell, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop H24-11, Atlanta, GA 30029-4018, USA; email: fms2@cdc.gov