RAND

Research Report

SARA DUHACHEK MUGGY, MARY AVRIETTE, DEREK ROBERTS,
KRISTIN J. LEUSCHNER

Modern Biosurveillance Methods

A Lay Introduction to Effective Use of Nucleic Acid Methods

For more information on this publication, visit **www.rand.org/t/RRA3263-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2024 RAND Corporation
RAND® is a registered trademark.

Cover: Andrew Brookes / Getty Images

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, please visit www.rand.org/pubs/permissions.

About This Report

Biosurveillance systems are used to detect, monitor, and characterize health threats, or pathogens, in human and animal populations, food, wastewater, and the environment. To accomplish these goals, biosurveillance uses various techniques to identify pathogens in a sample, including nucleic acid testing to determine the biologic agents present in clinical or environmental samples. Biosurveillance often relies on rapidly advancing technologies to detect pathogens. This report seeks to inform the nonscientific community of the methods by which current and emerging nucleic acid biosurveillance techniques work, describe their capabilities and limitations, and discuss the promise of emerging technologies.

Community Health and Environmental Policy Program

RAND Social and Economic Well-Being is a division of RAND that seeks to actively improve the health, social, and economic well-being of populations and communities throughout the world. This research was conducted in the Community Health and Environmental Policy Program within RAND Social and Economic Well-Being. The program focuses on such topics as infrastructure, science and technology, industrial policy, community design, community health promotion, migration and population dynamics, transportation, energy, and climate and the environment, as well as other policy concerns that are influenced by the natural and built environment, technology, and community organizations and institutions that affect well-being. For more information, email chep@rand.org.

Funding

Funding for this research was provided by gifts from RAND supporters and income from operations.

Acknowledgments

We would like to thank the following colleagues for their efforts and assistance:

- Bianca Espinosa and Alex Demarsh for serving as quality assurance reviewers
- project principal investigators Henry Willis and Angela O'Mahony for their guidance and support during our research and analysis
- team members Dan Gerstein, Laura Faherty, Jalal Awan, and Emily Hoch for their feedback and comments throughout the research and analysis.

Summary

The devastating impact of the coronavirus disease 2019 (COVID-19) pandemic has led governments, health care institutions, and researchers to intensify their focus on developing improved means to detect, contain, and manage potential future outbreaks. As part of these efforts, it is useful for policymakers and other nontechnical personnel to understand the promise of available biosurveillance technologies, in particular nucleic acid testing for genetic markers and the technologies that support them. Biosurveillance systems are used to detect, monitor, and characterize health threats, or pathogens, in human and animal populations, food, wastewater, and the environment.

Nucleic Acid Testing Approaches

Various nucleic acid testing approaches can be used to identify if a pathogen's deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) is present in a sample. These approaches vary in terms of cost, speed, accuracy, and other features.

Pathogen-Specific Tests

Some testing approaches target a specific pathogen (or pathogens), allowing a scientist to ask, "Is pathogen X present in my sample?":

- **Polymerase chain reaction** (PCR) is a rapid, accurate method to detect and copy very small amounts of genetic material, which can then be used to determine if a pathogen's DNA is present. First developed in the 1980s, PCR approaches have been enhanced in recent years.
  - **Quantitative PCR** (qPCR) is a variation of PCR that enables much faster results and relative quantification of the target DNA than traditional PCR.
  - **Digital PCR** (dPCR) enables better quantification of the nucleic acids present, especially when there are small amounts of the DNA of interest available.
- **Loop-mediated isothermal amplification** (LAMP). LAMP is not a PCR-based technique, but, like PCR, it uses primers and special enzymes to amplify the nucleic acid in a sample to indicate the presence of the pathogen of interest. LAMP is a versatile, inexpensive technique.

There are several advantages to PCR and LAMP. PCR techniques are widely used in research and diagnostic facilities, so the equipment and skills required to perform these tests are common. LAMP offers some advantages with respect to establishing nucleic acid detection capabilities in facilities with minimal scientific staff and equipment.

Pathogen-Agnostic Tests

Other forms of nucleic acid testing are "pathogen-agnostic" (or "unbiased"), which means that the scientist can use these techniques to discover the genetic material present, without first having a list of suspected pathogens in mind. Currently, pathogen-agnostic nucleic acid techniques involve next generation sequencing (NGS). Nucleic acid sequencing is a process that determines the exact composition of the nucleic acids in a sample. Historically, sequencing methods used primers and were therefore only useful if the researcher had some information about the pathogens that might be present. However, the latest nucleic acid sequencing methods, NGS methods, are pathogen-agnostic and enable researchers to sequence all the nucleic acids found in the sample simultaneously. The final step of NGS methods involves advanced data analysis techniques that can search the sequences found in the sample against known genomic sequences to lead to identification of the genomes present in the original sample. NGS techniques allow scientists to ask, "What pathogens are present in my sample?":

- **Illumina sequencing** is a type of "short read" sequencing technology developed and sold by Illumina, Inc.; this means that it is only capable of producing very short fragments of a sequence and relies heavily on data analysis to generate complete sequences. Many parts of this method are starting to lose patent protections and we may begin to see alternatives using this basic chemistry hit the market under new names.
- **Nanopore sequencing**, such as the systems being developed and sold by Oxford Nanopore Technologies, enables "very long read" sequencing, which greatly simplifies the assembly process. Nanopore sequencing uses specialized flow cells that contain numerous nanopores—that is, microscopic holes that can be monitored for changes in the electric current.

One known issue with NGS on complex mixtures of nucleic acids is that species present in smaller amounts may be missed. In these cases, **target enrichment** can be used. Target-enrichment sequencing involves enriching the sample for a specific agent or type of agent and can be used with either Illumina or nanopore sequencing. Target enrichment can occur using PCR-like amplification of targets or through selective purification techniques to try to eliminate other contaminating species (e.g., filtering out larger bacterial species to study the viruses present). Target enrichment prioritizes the detection of previously identified pathogens or classes of pathogen over new or unexpected microbes. Target enrichment is useful if a large number of other, less relevant, microbes are expected in the sample; however, it does narrow the aperture on the pathogens that can be identified from the enriched sample.

The major advantage of using any NGS approach is that, because NGS enables researchers to detect and identify whichever species are present in the sample, it is capable of identifying truly novel agents, novel variants of known agents, and unexpected agents. Additionally, NGS can provide more information about the pathogen beyond its presence in the sample. For example, NGS can indicate whether a pathogenic bacterium contains known antibiotic resistance genes that may affect the clinical or public health response.

However, in samples with many microbes and therefore a large mixture of DNA sequences, PCR and LAMP technologies are more sensitive than NGS in detecting agents that are not present in high concentrations. In addition, NGS can be complicated to use. Not every laboratory has the hardware to run the sequencing, limiting its accessibility and prolonging the time needed to obtain results. Also, detecting an microbe's genome in a clinical sample does not necessarily imply causation of disease, which is important to keep in mind if a novel agent is detected. NGS can also be limited by characteristics of the sample; samples teeming with microbes, for example, require additional processing.

Metagenomic Analysis Using Next Generation Sequencing

NGS can be used to support **metagenomic analysis**, which uses advanced computing to simultaneously analyze genomic sequences from multiple species in a single sample. **Metagenomics** involves sequencing the nucleic acids from the mixed sample to identify the genetic sequences that are present and assembling the sequences into genomes to determine which species are present.

In recent years, the cost of sequencing a full genome has been driven down significantly because of the high data output from NGS technologies. For these reasons, metagenomic studies are now almost exclusively performed using NGS. However, metagenomic sequencing approaches have some important limitations, including longer time requirements for analysis, difficulty in characterizing novel agents as potential threats, and dilution of smaller genomes by larger genomes. Thus, strong experimental design—which takes into account the research questions of interest and any resource constraints (e.g., time, funding)—is critical to obtaining the best-quality data when choosing a biosurveillance approach.

Selecting a Technology for Wastewater Biosurveillance

We examined a scenario involving testing for pathogens in wastewater using biosurveillance technology. Wastewater offers a method for collecting samples without individuals presenting in clinics for care and thus can offer data prior to symptoms occurring and from those who have lower access to clinical care. During the COVID-19 pandemic, this field showed promise in identifying hotspots for COVID-19 outbreaks before other traditional epidemiological indicators.

Policymakers interested in establishing a wastewater biosurveillance (WBS) program should evaluate their specific research questions and their resource limitations to determine which technology is best suited for their needs. Selecting a single technology that can accomplish all biosurveillance goals is not necessary. Table S.1 provides some information and guidance toward selecting the best technique; however, relevant stakeholders, including laboratory personnel and wastewater utilities, should be included in these discussions.

TABLE S.1

Selecting a Technology by Biosurveillance Goal

What is your goal?	Technologies Available	Speed (time from collection to analyzed data)	Restrictions
Identifying whether known agents are present in wastewater	Illumina NGS Nanopore	Slow Slow	Target enrichment may be required prior to sequencing for reliable detection of pathogens.
	PCR LAMP	Fast Very fast	Agent mutation may reduce or eliminate efficacy of the test.
Monitoring a known agent for mutations	Illumina NGS Nanopore	Slow Slow	Best-use case involves PCR-based amplification for target enrichment prior to sequencing.
Monitoring a known agent for levels of agent in the population	Illumina NGS Nanopore	Slow Slow	Results are less quantifiable than PCR/LAMP, especially in cases with target enrichment prior to sequencing.
	PCR LAMP	Fast Very fast	Agent mutation may reduce or eliminate test efficacy.
Identifying novel agents that are present in wastewater	Illumina NGS Nanopore	Slow Slow	Target enrichment may be required prior to sequencing for reliable detection of pathogens; however, these methods do limit the scope of the resulting analysis.
Identifying unexpected agents that are present in wastewater	Illumina NGS Nanopore	Slow Slow	Target enrichment may be required prior to sequencing for reliable detection of pathogens.
	PCR	Fast	Large numbers of individual reactions will be needed to accomplish this goal. Agents not specifically tested for will be missed.
	LAMP	Very fast	Large numbers of individual reactions will be needed to accomplish this goal. Agents not included in the test will be missed. Speed may not be the same for all agents; only those with tests that work on raw sewage will be very fast.

Contents

About This Report iii
Summary v
Figures and Tables xi

Modern Biosurveillance Methods 1
Introduction 1
Biosurveillance and Nucleic Acid Detection Techniques 3
Applying Nucleic Acid Sequencing Through Wastewater Metagenomics 18
Understanding the Technical Limitations of Metagenomics 22
Conclusion 27

Appendix. Glossary 29
Abbreviations 31
References 33

Figures and Tables

Figures

1. Structure of RNA and DNA 6
2. Examples of Some Pathogens Found in Wastewater Summarized by Key Properties of Their Genomes 24
3. Viral Genomes Are Diluted by Larger Genomes from the Sample 25

Tables

S.1. Selecting a Technology by Biosurveillance Goal viii
1. Comparing Illumina and Nanopore Sequencing 14
2. Selecting a Technology by Biosurveillance Goal 28
A.1. Glossary of Terms Used in This Report 29

Modern Biosurveillance Methods

Introduction

The devastating impact of the coronavirus disease 2019 (COVID-19) pandemic has left an enduring mark on societies worldwide and ignited a desire to prevent future pandemics, or, at the very least, to mitigate their toll on human lives and global economic stability. The tragic loss of life and unprecedented disruption to economies, education, and daily routines have underscored the urgent need for proactive measures and robust systems to combat infectious diseases effectively and efficiently. Governments, health care institutions, and researchers have intensified their focus on developing improved biosurveillance systems,[1] rapid response mechanisms,[2] and global collaboration to detect, contain, and manage potential outbreaks that could become pandemics.[3]

These efforts will likely not be able to stop pandemics altogether due to the lag between the initial phases of an outbreak and the detection and response, but, ultimately, earlier detection may mitigate loss of life, lessen economic disruption, and potentially hasten society's ability to return to normal following a pandemic.[4] As part of an effort to prevent or mitigate the next pandemic, it is worthwhile to examine biosurveillance technologies and to understand their promise for keeping societies safe from potential pandemics.

Biosurveillance aims to provide early warning of disease outbreak or biological attack. Biosurveillance systems are used to detect, monitor, and characterize health threats, or **pathogens**, in human and animal populations, food, water, and the environment. Biosurveillance has many uses: it is performed as a part of food safety programs, disease monitoring by public health

---

[1] Centers for Disease Control and Prevention (CDC), "National Wastewater Surveillance System (NWSS)," webpage, March 10, 2023a.

[2] National Science and Technology Council, White House Steering Committee for Pandemic Innovation, *First Annual Report on Progress Towards the Implementation of the American Pandemic Preparedness Plan*, September 2022.

[3] International Monetary Fund, "Policy Responses to COVID-19: Policy Tracker by Country," webpage, July 2, 2021.

[4] Sneha Kannoth, Sasikiran Kandula, and Jeffrey Shaman, "The Association Between Early Country-Level COVID-19 Testing Capacity and Later COVID-19 Mortality Outcomes," *Influenza and Other Respiratory Viruses*, Vol. 16, No. 1, January 2022.

departments, environmental testing for vector-borne diseases, and even Department of Homeland Security monitoring for potential biological attacks.

Biosurveillance can use various techniques to identify pathogens in a sample, including **nucleic acid testing**, which assesses nucleic acids to determine the biologic agents present in clinical samples (i.e., a sample collected by a medical laboratory from a symptomatic human) or environmental samples (e.g., wastewater).[5] Biosurveillance often relies on rapidly advancing technologies to detect pathogens. For example, in cases where scientists are looking for evidence that a specific known pathogen is present in a sample, **polymerase chain reaction** (PCR) can be used. PCR can detect very small amounts of genetic material in a sample, allowing researchers to determine if a pathogen's genetic material is present. This approach was used in early tests for COVID-19, which required a doctor's office visit and a laboratory analysis of the sample.

When a patient's clinical symptoms are unusual or a suspected pathogen is unknown, a newer type of technology known as **next generation sequencing** (NGS) can be used to "read" the genetic material present in the sample and match the genomic sequence to other known sequences to determine all microbes (e.g., bacteria or viruses) that are present in a sample simultaneously. This form of testing is **pathogen-agnostic** and therefore allows the scientist to discover the genetic material present without having a list of suspected pathogens in mind first.[6] NGS can be used to support **metagenomic analysis**, which uses advanced computing to provide the simultaneous genomic sequencing of multiple microbes in a single sample.

Nucleic acid testing approaches can vary in terms of cost, speed, accuracy, and other features. Policymakers and other nontechnical personnel can thus benefit from a basic understanding of biosurveillance approaches and the technologies that support them. Biosurveillance programs must distribute limited resources efficiently, and policymakers must also be able to communicate effectively to the public about biosurveillance techniques.

This report seeks to inform the nonscientific community of the methods by which nucleic acid biosurveillance methods work, describe their capabilities and limitations, and discuss the promise of emerging technologies. We also present a scenario involving testing for pathogens in wastewater using detection technology within a comprehensive and pathogen-agnostic biosurveillance system. Wastewater testing offers a method for collecting samples without individuals presenting in clinics for care. It can therefore offer data prior to symptoms occurring, and data from those who have less access to clinical care. We outline the current state of these nucleic acid detection methods and the wastewater use case to demonstrate the

[5] Other approaches include **bacterial cultures**, in which samples are tested by attempting to grow bacteria on a food source that is specific to a particular bacterial species; and **protein tests**, which require the user to select a pathogen to assess and perform the test for that single agent or group of similar agents.

[6] This statement is qualified later: NGS samples are often prepared in a way that minimizes host DNA to better determine what other genetic material is present. As a result, there is some selection process early in sample processing, and additional selection is used during algorithmic analysis of the results. A researcher might not have a list of pathogens in mind (*E. coli*, *Salmonella*, etc.), but likely has a class (bacteria) in mind. In this way, NGS is not completely unbiased but is still termed pathogen-agnostic.

advancements needed to develop a truly pathogen-agnostic, cost-effective wastewater biosurveillance (WBS) system.

Organization of This Report

The remainder of this report is organized as follows. The next section presents a primer on biosurveillance testing and on some of the key nucleic acid methods that are available today. The third section discusses metagenomics and considerations for collecting and preparing wastewater samples. The fourth section delves further into the pathogen-agnostic technologies (or "unbiased") and discusses known limitations of those techniques and other important considerations. A concluding section summarizes the report.

Biosurveillance and Nucleic Acid Detection Techniques

In this section we provide a detailed overview of the technologies in use for modern biosurveillance efforts and address the advantages and disadvantages of each technology. We begin with a brief discussion of biosurveillance approaches and then describe nucleic acids and nucleic acid detection techniques. The section ends with a brief discussion of metagenomics, an analysis technique that can be used with more advanced nucleic acid detection approaches and which is the focus of the wastewater scenario discussed later in the report.

Biosurveillance Testing

Because most of the uses for biosurveillance involve preventing widespread human disease, there are optimal locations for performing biosurveillance testing. These locations tend to be sites associated with significant risk, the threat of outbreak, or those with a large number of individuals passing through. Such sites include regions in close proximity to a large population center, high-traffic areas such as transportation hubs, locations with a strong animal-human interface such as animal husbandry facilities or wet markets,[7] and wastewater. Many human pathogens are shed in the feces, urine, or saliva that all end up in household wastewater, and therefore wastewater provides an excellent source of information for diseases that may be percolating through the population.

There are various forms of biosurveillance that can be used to detect outbreaks early. For example, **syndromic surveillance** seeks to identify clusters of illness without necessarily determining the causal pathogen, such as a sudden onset of diarrheal disease among many individuals within a geographic area. In addition to monitoring for clusters of patients presenting at medical clinics with similar diseases, many other methods are used for syndromic

[7] A wet market is a market where fresh meat is butchered and sold. Wet markets in mainland China specifically have been of particular interest to emerging disease researchers, and a Wuhan wet market dominated the news for several months after COVID-19 began spreading in the Wuhan area.

surveillance, such as monitoring over-the-counter medication sales, internet searches, and tracking worker absenteeism.[8] The value of syndromic surveillance is its potential to detect illness patterns rapidly, potentially before aggregated laboratory results detect a growing outbreak.[9] However, syndromic surveillance, by definition, does not diagnose the causative agent involved. In other words, it cannot be used to identify genetic markers or to identify the pathogen involved.

However, syndromic surveillance can be paired with other forms of biosurveillance that focus on identifying the pathogens involved in an outbreak. For example, syndromic surveillance can be used in conjunction with **environmental biosurveillance** methods such as WBS. In this case, syndromic surveillance would detect a cluster of similar illness symptoms in a geographic area, and public health authorities would then sample the wastewater from this area to determine the causative agent of the illness.

Syndromic surveillance can also be paired with **wildlife biosurveillance** to detect so-called spillover events, in which diseases circulating among wildlife develop the ability to infect and spread among humans.[10] Genomic analysis of wildlife biosurveillance samples can reveal viruses present in animals,[11] many of which do not become ill when harboring the pathogen. This method of anticipating spillover events can be resource-intensive without specifying any clear path forward. However, partnering wildlife biosurveillance techniques with **human biosurveillance** data has potential to close this information gap. If a microbe is isolated from hospital wastewater and not from community wastewater, there may be a connection between the microbe and illness.[12] Further understanding of where humans may

[8] For a more complete discussion of methods, see Rajeev Ramchand, Sangeeta C. Ahluwalia, Mary Avriette, Gary Cecchine, Monika Cooper, Christy Foran, Daniel Hicks, Natasha Lander, and Sarita D. Lee, *Syndromic Surveillance 2.0: Emerging Global Surveillance Strategies for Infectious Disease Epidemics,* RAND Corporation, RR-A1812-1, 2023; and Kenneth D. Mandl, J. Marc Overhage, Michael M. Wagner, William B. Lober, Paola Sebastiani, Farzad Mostashari, Julie A. Pavlin, Per H. Gesteland, Tracee Treadwell, Eileen Koski, Lori Hutwagner, David L. Buckeridge, Raymond D. Aller, and Shaun Grannis, "Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience," *Journal of the American Medical Informatics Association*, Vol. 11, No. 2, March–April 2004.

[9] C. C. van den Wijngaard, W. van Pelt, N. J. Nagelkerke, M. Kretzschmar, and M. P. Koopmans, "Evaluation of Syndromic Surveillance in the Netherlands: Its Added Value and Recommendations for Implementation," *Eurosurveillance*, Vol. 16, No. 9, March 3, 2011.

[10] Neil M. Vora, Lee Hannah, Chris Walzer, Mariana M. Vale, Susan Lieberman, Ashley Emerson, Jonathan Jennings, Robyn Alders, Matthew H. Bonds, Jo Evans, Bhavana Chilukuri, Sonila Cook, Nigel C. Sizer, and Jonathan H. Epstein, "Interventions to Reduce Risk for Pathogen Spillover and Early Disease Spread to Prevent Outbreaks, Epidemics, and Pandemics," *Emerging Infectious Diseases*, Vol. 29, No. 3, March 2023.

[11] There are myriad ways to detect viruses or other pathogens in wildlife. For more, see Kim M. Pepin, Ryan S. Miller, and Mark Q. Wilber, "A Framework for Surveillance of Emerging Pathogens at the Human-Animal Interface: Pigs and Coronaviruses as a Case Study," *Preventive Veterinary Medicine*, Vol. 188, March 2021.

[12] The relationship between pathogens and illness is complex, and infected individuals may not present for medical care for a host of reasons (mild illness, expense of treatment, or delay between exposure and illness). Thus, environmental surveillance nearly always requires clarifying work to establish the true relationship between pathogen and illness.

have encountered the microbe, such as from existing wildlife biosurveillance data, may illuminate epidemiological links earlier than other methods.

Selection of the appropriate biosurveillance strategy should be based on the question to be answered through biosurveillance, *is pathogen X present in wastewater and in what concentrations?, how have concentrations of pathogen X changed in response to public health interventions,* or *what pathogens are causing an outbreak in the human population of city Y?* As part of a biosurveillance strategy, it is also necessary to choose the most appropriate nucleic acid detection method. A variety of nucleic acid detection methods are available, but each is associated with various capabilities, required inputs, and purposes. In the following subsections, we will discuss what nucleic acids are and why we test for them, and then outline the most important technologies currently being used for biosurveillance.

What Are Nucleic Acids?

Nucleic acids are large molecules that store and transmit the development plan of all living microbes and viruses, and can therefore be used to identify species present in a mixed sample.

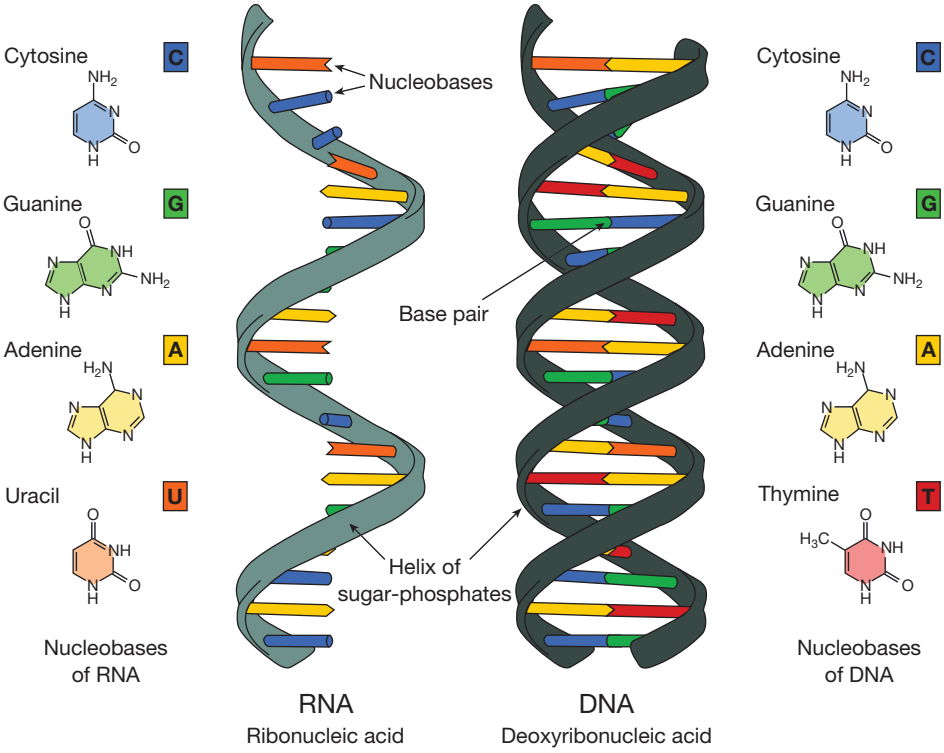
The predominant types of nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is the main genetic material of life, and functions as a sort of operations manual for each cell in a microbe. Nucleic acids are comprised of chains of nucleotides, each of which contains a nitrogenous base, a sugar molecule, and a phosphate group. The primary difference between DNA and RNA is the type of sugar molecule present in the nucleotide. Additionally, in living cells, DNA generally exists in a double-stranded form, while RNA is single-stranded. Figure 1 shows the structure of DNA and RNA.

All living microbes with active metabolisms have DNA and are generating RNA so both types of nucleic acids can be isolated from these samples. Viruses, which are neither alive nor metabolically active, generally contain a single type of nucleic acid. Human pathogenic viruses like smallpox, adenoviruses, hepatitis B virus, and chicken pox virus are all DNA viruses, meaning that they contain DNA but not RNA. RNA viruses include influenza, coronaviruses, polio, and Ebola virus and are also of concern for public health.

Nucleic Acid Detection Methods

DNA and RNA can both be detected using standard nucleic acid techniques; however, most methods use DNA as a starting material, not RNA. RNA can be converted to a type of DNA called complementary DNA (cDNA), which can be used for DNA-specific methods, but this represents an additional step in the workflow. For all methods discussed here, sample preparation or purification is required before analyzing the nucleic acid. In this simplest form, this means isolating, or extracting, the nucleic acids from the sample. For example, in a nasal swab sample there is likely some human cells and mucus present in addition to the potential virus or bacterial pathogen. Extracting the nucleic acids (either DNA or RNA) means removing the other cellular material and the mucus to ensure that purified nucleic acid is used for testing.

FIGURE 1
Structure of RNA and DNA



SOURCE: Reproduced from Sponk, "Difference of DNA and RNA," 2010 (CC BY-SA 3.0). This is an English translation of an original image.

Nucleic acid detection methods are diverse in their abilities and limitations. In this subsection, we present information on a range of nucleic acid detection methods. The first set of methods—**pathogen-specific techniques**—depend upon knowing the specific pathogen that is being tested. The second set of methods—**pathogen-agnostic techniques**—are unbiased and can be used to identify any agent in a sample.

Pathogen-Specific Techniques

There are two main categories of pathogen-specific techniques that will be discussed here: **PCR** and **loop-mediated isothermal amplification** (LAMP). Each of these technologies can be used to answer a question about a specific pathogen—"Is pathogen X present in my sample?"—rather than asking the more general question "What pathogens are present in my sample?"

Polymerase Chain Reaction

PCR is a rapid, accurate method to **amplify** (i.e., cause genetic material to undergo massive replication) very small amounts of genetic material in almost any sample; if the genetic material was amplified, researchers can then conclude that a pathogen's DNA or RNA is present in the sample. **Simplex PCR** determines if a single pathogen is present in a sample. **Multiplex PCR** can detect several known and specified pathogens in a single sample simultaneously. The latter form of testing is most useful when a physician has a short rule-out list of pathogens to test for based on a patient's symptoms. There are three main types of PCR tests: traditional PCR, quantitative PCR,[13] and digital PCR. Traditional PCR is rarely used today but we discuss it because it formed the basis for newer types of PCR.

Traditional PCR is a foundational technique of modern biological detection. Originally developed in the 1980s, traditional PCR was used to enable amplification of a particular region of DNA. However, traditional PCR is rarely used anymore as modern improvements to the basic technique offer greater sensitivity and flexibility.

PCR creates copies of a specific region of DNA using a short reaction that is repeated, or cycled, up to 40 times. Since the technique is specific to a region of DNA, it requires the user to know which agent they are testing for, and to have some information about the genetic sequence. A positive result is obtained if the reaction is able to amplify the selected region, meaning it was able to make additional copies of the selected nucleic acid region indicating that the region had been present in the sample, even if it was present at very low quantities.

Because PCR techniques are so heavily utilized in research and industry, there are many software resources available to assist in identifying an appropriate region to amplify and to create **primers**. Primers are reagents that are designed to initiate the amplification and are extremely specific for the region of interest. Primers determine the beginning and end of the region that will be amplified. However, all primers must be tested and validated to ensure that they are working properly prior to relying on them for any type of testing.

PCR tests can be developed rather quickly once a genetic sequence is available. PCR requires a special instrument to complete the cycled reactions and a few other special reagents to be used up during the test. Both the instrument and the reagents are relatively inexpensive and widely available. Cycled reactions are typically run for 45–60 minutes. However, traditional PCR does not have a direct output to determine whether amplification occurred; thus, another detection test must be performed after cycling, which adds additional time, equipment, and consumable reagents.

Quantitative PCR (qPCR) is a more modern form of PCR. COVID-19 PCR tests used in the first year of the pandemic were qPCR tests. Like traditional PCR, a qPCR test for a new pathogen can be designed quickly once the sequence is known. qPCR differs from the original PCR technique in two ways: it includes a fluorescent marker and utilizes a specialized instrument that can read this marker throughout the entire reaction process.

[13] This technique is sometimes called "real-time PCR."

qPCR includes a special reagent called a **probe** that makes this technique much faster and more flexible than traditional PCR. As noted above, under traditional PCR, another test must be run after the cycles to see if a product or a signal was created. In contrast, qPCR probes are designed to recognize part of the region that is to be amplified, and the probe becomes fluorescent if the region is amplified. The use of a probe and primers offers an additional point of sequence recognition, which reduces the likelihood of false positives because now one more reagent must bind for a positive signal to be produced. The cycling time for qPCR is similar to that for traditional PCR, but qPCR eliminates the need to run another test upon completion. The **specialized instrument** for qPCR is more expensive than traditional PCR instruments but still cheaper than other nucleic acid techniques. Probes are relatively inexpensive but are more expensive than the primers alone.

Digital PCR (dPCR) was recently developed through modification of qPCR and enables better quantification of the nucleic acids present, especially in scenarios with low abundance of the DNA or RNA of interest. The major advantage of using digital PCR is the improved quantitation compared with other PCR techniques.

Building on qPCR, dPCR uses almost identical chemistry: the primers, probe, and cycled reactions. The major modification for dPCR is that, before the reactions begin, the solution containing all the elements for amplification is partitioned into approximately 20,000 microscopic reactions. These reactions are typically less than 1 billionth of a liter in volume.[14]

In one partitioning method, a specialized microfluidic plate is used to separate the reaction into tiny chambers. Once partitioned, the samples are cycled normally but the fluorescence signal is not monitored during the cycles. Once the cycled reactions are complete, the mini reactions are analyzed for fluorescence. In this chambered-plate dPCR method, the same instrument is used to analyze fluorescence in each of the mini reactions. While in qPCR the whole sample is read together, with dPCR each mini reaction is read and added together. This enables better detection of very rare regions since those weaker signals might not be detected reliably in the whole solution. Additionally, dPCR allows for a better ability to quantify the number of copies of the selected DNA region (and therefore, the pathogen it comes from) since it counts the mini reactions with a signal. Most chambered dPCR solutions involve a single unique dPCR instrument and special consumable chambered plates.

Another method for partitioning the sample involves **emulsification**, whereby the sample is mixed with an oil to create tiny droplets that are invisible to the naked eye. Each tiny droplet is now a unique, mini qPCR reaction, and this technique is called digital droplet PCR (ddPCR). ddPCR reactions are cycled without reading the fluorescence just as in chambered-plate dPCR. However, in ddPCR, the solution is passed through another machine to read the fluorescence of each droplet individually. ddPCR uses different instruments to

[14] Michael Bussmann, Miriam Hesse, Oezlem Karalay, Rose Nash, and Andreas Missel, "Impact of Template Addition Volume and Analyzed Volume on Digital PCR Sensitivity," Application Note PROM-20715-001 1127766, June 2022.

emulsify the sample, perform the cycled reactions, and read the fluorescence of the droplets. dPCR reactions are often complete within 90 minutes of starting the reactions.

Loop-Mediated Isothermal Amplification

LAMP is another nucleic acid detection technique. Although not a PCR-based technique, it also uses primers and special enzymes to amplify the nucleic acid in the sample and uses amplification to indicate the presence of the pathogen of interest. Whereas PCR uses two primers for each amplification, LAMP uses 4–6 primers. Since there are so many, the primers bind in several locations on the target DNA and create a nonuniform set of reactions that occur simultaneously, rapidly creating many copies of the nucleic acid. The products of LAMP are also nonuniform as a result of the multiple sites of primer binding.

The reaction is significantly more complex than PCR; however, it offers some major advantages. LAMP occurs at a single temperature, which means that, in its simplest form, LAMP can be done at room temperature and without the use of any instruments. Some modifications of LAMP require instruments but, even in these cases, the instruments can be designed to be small, extremely inexpensive, and quite mobile.[15] Once the reaction is complete, usually in 20–30 minutes, the sample must be interpreted to determine whether amplification has occurred. LAMP can be adapted to create several different outputs, some of which are readable by the naked eye or even a cell phone camera–based application, while others require more advanced instruments.

LAMP is a versatile, inexpensive technique to detect the presence of a pathogen. Designing LAMP tests is a little more challenging than PCR ones due to the number of primers needed and the optimization required to produce strong signals with minimal background noise. As with PCR, there is plenty of available software for aiding in primer design, but ultimately the primers must be tested and validated in real experiments prior to use for pathogen detection.

Advantages and Limitations of PCR and LAMP

There are several advantages to PCR and LAMP. PCR techniques are widely used in research and diagnostic facilities so the equipment and skills required to perform these tests are common. LAMP offers some advantages with respect to establishing nucleic acid detection capabilities in facilities with minimal scientific staff and equipment. Additionally, LAMP is less affected by inhibitors that are commonly found in clinical samples like blood, feces, and saliva.[16]

---

[15] "WEF Recognizes Two Leaders in Rapid Wastewater Testing," *Water Environment & Technology*, May 2022.

[16] May Khat Nwe, Nisachon Jangpromma, and Lapatrada Taemaitree, "Evaluation of Molecular Inhibitors of Loop-Mediated Isothermal Amplification (LAMP)," *Scientific Reports*, Vol. 14, 2024.

Pathogen-Agnostic Detection Using Nucleic Acid Sequencing

When a patient's clinical symptoms are unusual or a physician's list of suspected pathogens has been ruled out, other techniques can be used to look for pathogens that are unknown or unexpected. Pathogen-agnostic tests allow the scientist to discover the genetic material present without having a list of suspected pathogens in mind first.[17] The nucleic acid pathogen-agnostic methods are a newer type of technology called NGS, which can be used to "read" the genetic material present in the sample and match the genomic sequence to other known sequences to determine all microbes present in a sample simultaneously.

DNA sequencing is the process of determining the exact sequence of nitrogenous bases in the DNA sample. Early techniques for DNA sequencing emerged in the mid-1970s and early 1980s.[18] These same techniques were used widely for many years; in fact, the chemistry from the Sanger sequencing methodology published in 1977 is still practiced in specific applications, albeit with improved analysis technology. While sequencing techniques have been available for many years, in the early 2000s new sequencing technologies became available. NGS collectively refers to all such novel sequencing technologies and does not differentiate based upon the specific method or reaction chemistry being used.

NGS methods are not dependent on primers or other sequence information; additionally, they are massively parallelized sequencing reactions. NGS often runs millions of simultaneous reactions that can be individually distinguished and analyzed. This dramatically increases the throughput of sequencing reactions. Therefore, using NGS, a mixed sample containing DNA from multiple microbes can be sequenced without the need to know any sequence information, and a single sequencing experiment can provide sequences for most or all the species present in the sample.

There are several NGS technologies available and currently being developed; however, the most widely used NGS technology at this time is Illumina. Another technology known as Nanopore sequencing is a more recent method that has some potential advantages over Illumina. We discuss both below.

Illumina Sequencing

Illumina sequencing is a method developed by Illumina Inc., and is a "short read" sequencing technology. This means that it is only capable of producing very short fragments of a sequence,[19] and relies heavily on data analysis to generate complete sequences. While this technique is only available from Illumina Inc., at this time, the patents are expiring. Therefore,

[17] As mentioned previously, NGS samples are often prepared in a way which minimizes host DNA to better determine what other genetic material is present. As a result, there is some selection process early in sample processing, and additional selection is used during algorithmic analysis of the results. Therefore, NGS may not be completely unbiased in real applications but is still termed pathogen-agnostic.

[18] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel, "Overview of Next Generation Sequencing Technologies," *Current Protocols in Molecular Biology*, Vol. 122, No. 1, April 2018.

[19] Illumina sequencing generates fragments that are 50–600 bases in length.

similar approaches may be emerging soon from alternate suppliers. The Illumina sequencing workflow occurs in four primary steps:

1. **Sample preparation** occurs using a commercially available reagent kit and some instruments that are routinely found in molecular biology laboratories. Sample preparation involves fragmenting the DNA into short sequences and adding additional synthetic sequences to the ends known as tags. These tags are short pieces of DNA with a known sequence. Tags enable sequencing to initiate without knowledge about the sample sequence since primers can be used to recognize the tag to initiate sequencing reactions. Initiating sequencing using primers that recognize the tag sequence enables subsequent pathogen-agnostic analysis, since no sequence information about the sample is necessary to begin the analysis.
2. **Cluster amplification** leads to the creation of a cluster of identical fragments that can be sequenced. These clusters amplify the signal that is created during sequencing and enable more sensitive detection. The cluster amplification and sequencing steps both occur inside the specialized sequencing instrument and on a consumable device known as a **flow cell**. Each flow cell allows several million simultaneous reactions to occur and be analyzed by the instrument. During cluster amplification the DNA sample is placed onto the flow cell, where it is immobilized and copied many times.
3. **Sequencing** is the process of amplifying the sample using fluorescent markers to indicate which nitrogenous base is being added. The fluorescent markers are monitored, and the sequence of the DNA can be established by observing the fluorescence over time. Each base has a unique color of fluorescence and therefore can be used to re-create the sequence of bases that were incorporated into the growing DNA fragment. Illumina sequencing runs can take from 24 to 48 hours to complete but generate a very large amount of data. Each cluster on the flow cell, which contains between 170,000 and 1.4 million clusters per square millimeter,[20] results in a short DNA sequence.
4. **Data analysis** using computational approaches makes sense of the massive amount of data generated. The short sequences are compared and assembled to create longer regions of sequence. These longer sequences can be searched in a database to identify the microbe from which they came, and thus identify the microbe present in the sample; for unknown agents, this analysis will produce a sequence with no known matching microbe. In some of these cases, it will be related to another known microbe and enable researchers to make some predictions about the new microbe.[21]

[20] Illumina Inc., "Cluster Density Guidelines for Illumina Sequencing Platforms Using Non Patterned Flow Cells," webpage, undated.

[21] While data analysis is always a key step of the Illumina workflow, the type of data analysis discussed here is relevant in a biosurveillance system but there are other ways to analyze and use Illumina data that are not discussed in this primer.

Illumina sequencing can detect unknown or unexpected agents since no primers are needed against a certain pathogen. Illumina sequencing is considerably slower and more expensive than LAMP or PCR to run and requires more advanced expertise to perform the sample preparation, sequencing, and data analysis steps than is required for PCR or LAMP techniques. However, Illumina sequencing is quite commonly used in some scientific fields, and the number of scientists trained to perform these steps is increasing. Advancements in data analysis techniques and computing may shorten the time needed to complete the analysis and may result in more automated workflows that can be used to both standardize sequencing data analysis and decrease the time required to generate results.

One major limitation of Illumina sequencing results from the need to generate short DNA pieces and therefore short "reads," which are the raw sequences derived from each cluster. Each read is usually 50–150 bases in length. For comparison, the SARS-CoV-2 viral genome, which is quite small compared with most microbes, is approximately 30,000 bases in length. These short reads increase the data analysis time and the complexity of the task, but they also lead to other challenges and limitations of the technique, which is discussed later in the report.

Nanopore Sequencing

Another sequencing technology representing a newer wave of innovation is nanopore sequencing, specifically of the type pioneered by Oxford Nanopore Technologies. Nanopore sequencing has emerged over the last decade and uses pioneering technology with some complementary qualities to those of Illumina NGS. Using an entirely different workflow, nanopore sequencing enables "very long read" sequencing,[22] which greatly simplifies the assembly process, and allows for direct sequencing of RNA without first converting it to cDNA. The nanopore sequencing workflow occurs in three primary steps:

1. **Prepare:** Throughout sample extraction and purification, nanopore sequencing workflows are optimized to produce the longest possible nucleic acid fragments.[23] A variety of kits to accomplish this are available catering to different sample types. Finally, adapter sequences are added to both ends of all fragments to enable strand capture at the nanopore.[24]
2. **Sequence:** Following sample preparation, samples are loaded onto flow cells for sequencing. Nanopore sequencing uses specialized flow cells that contain numerous nanopores, which are microscopic holes that can be monitored for changes in the electric current around that nanopore. The flow cell is placed inside another specialized

[22] Nanopore sequencing can sequence contiguous fragments 10,000–1,000,000 bases in length.

[23] Oxford Nanopore Techologies, "Assembling Microbial Genomes from Complex Metagenomic Samples Using Long Nanopore Reads," Workflow 1137, October 3, 2022.

[24] Miten Jain, Hugh E. Olsen, Benedict Paten, and Mark Akeson, "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community," *Genome Biology*, Vol. 17, 2016.

sequencing instrument where the sequencing occurs. Individual fragments are pulled into position above a nanopore using the adapters (added during the Prepare step) and are then pulled into the nanopore. As they pass through, an ionic current change occurs that is distinct for each nucleotide base. These changes are detected in order and used to determine the DNA or RNA sequence. Currently, nanopore sequencing can be performed using instruments ranging from small mobile devices that can run a single sample at a time to larger benchtop machines that can run 48 samples simultaneously. All devices function through the combination of two key technologies: the ability to pass long single-stranded nucleic acid chains through a nanopore, and the ability to identify the specific nucleotide bases that pass through by detecting changes in the ionic current.

3. **Analyze:** Genome assembly of sequence data produced by nanopore sequencing is simpler than that produced by Illumina sequencing because reads are usually much longer. This makes it easier, but usually not trivial, to identify overlaps and put the pieces together in the proper order. Once assembled, the genomes are run through databases to enable identification of the pathogen.[25]

Nanopore methods require a similar amount of time to run to Illumina NGS, but also require less sample preparation and less data analysis. Additionally, nanopore sequencing is the only technique discussed here that can be performed on RNA directly; therefore, analyzing a sample for potential RNA viruses does not require an additional step to convert the RNA to cDNA first.

Like the data analysis in Illumina sequencing, the sequences that result from nanopore sequencing are searched for in genomic databases to identify the microbes present in the sample. However, because of the longer sequences, the complex assembly processes are greatly reduced or eliminated with nanopore sequencing.

The major drawback of nanopore sequencing is that it is more error-prone than Illumina sequencing. Additionally, it is less sensitive, meaning that rare sequences are less likely to be identified using nanopore sequencing than Illumina sequencing. These issues may be improved in future iterations of this technology, but currently Illumina is the more accurate and sensitive NGS technology.

Concerted Utilization of Illumina and Nanopore Sequencing

Both Illumina sequencing and nanopore sequencing have unique advantages and best-use cases. In many ways, they cover the weaknesses of each other. Therefore, a strategy involving use of both is likely optimal, but may be cost-prohibitive. Table 1 highlights the advantages and disadvantages of both technologies.

[25] Again, this type of data analysis is most relevant in a biosurveillance context; however, nanopore sequencing can be used in a variety of ways that may necessitate different analysis steps and outcomes.

TABLE 1
Comparing Illumina and Nanopore Sequencing

Area of Comparison	Illumina	Nanopore
Laboratory time (sample prep)	3 hours	3 hours
Laboratory time (generation of sequences)	20–56 hours	48 hours
Amount of DNA required for input	1–50 ng	0.1–1 µg
Sensitivity (derivation of sequences from low viral load samples)	Medium	Low
Genome coverage	100% (with predictable areas of lower coverage)	100%
Depth of sequencing	Reliable detection of mutations, but may require enrichment steps, which can produce bias	Currently limited by need for enrichment, which can introduce bias
Likely error rate	0.2–0.7% per read	–12% per read, 0.3% (after correction)
Long read vs. short read	Short (estimated 150–600 base pairs per read)	Long (only limited by size of library loaded onto flow cell)
Key attributes	Representation of complete diversity present within a sample	Generation of full-length viral genomes; portable test offers potential for point-of-care diagnosis
Key concerns	Potential errors in reconstruction of complete viral genomes. Expensive setup; requires lab infrastructure	May under-represent diversity; poor sensitivity in the absence of an amplification step

SOURCE: Data are reproduced in part and adapted from Anna L. McNaughton, Hannah E. Roberts, David Bonsall, Mariateresa de Cesare, Jolynne Mokaya, Sheila F. Lumley, Tanya Golubchik, Paolo Piazza, Jacqueline B. Martin, Catherine de Lara, Anthony Brown, M. Azim Ansari, Rory Bowden, Eleanor Barnes, and Philippa C. Matthews, "Illumina and Nanopore Methods for Whole Genome Sequencing of Hepatitis B Virus (HBV)," *Scientific Reports*, Vol. 9, No. 1, May 8, 2019.

Given these advantages and weaknesses, Illumina is the optimal technology in some scenarios whereas nanopore is best in others. Illumina can run with less sample DNA than is required for nanopore; therefore, if sample quantity is an issue, Illumina is probably the better technology to use. If DNA quality is poor or the DNA is already damaged and fragmented into smaller pieces, then Illumina is also the better choice.

However, given the complexities of the data analysis for Illumina results, this technology is also better suited for situations in which the DNA is from known species that have been well characterized and have **reference genomes** available. A reference genome is a genome that has previously been sequenced to a high level of quality; reference genomes are often available in online databases. Therefore, the assembly of read sequences into a genome can be done via reference-guided genome assembly, which is the assembly of previously sequenced

genomes that is greatly aided by reference genomes. However, nanopore produces much longer sequences and is better suited for what is called *de novo* sequence assembly or assembly of novel or never-before-sequenced genomes. These longer reads are also better if the sequence is extremely repetitive or very similar to sequences found in many other species. Geneticists call these sequences (which are found in many species) **conserved sequences**. Repetitive and conserved regions create challenges when assembling the shorter reads obtained from Illumina sequencing because it is difficult to assemble them correctly. In the box, we summarize the best-use cases for Illumina versus Nanopore sequencing.

Best-Use Cases for Illumina versus Nanopore Sequencing

Illumina is optimal for

- low nucleic acid mass samples
- highly fragmented samples
- reference-guided sequencing to identify mutations or polymorphisms.

Nanopore is optimal for

- *de novo* sequencing
- RNA sequencing
- sequencing of highly conserved or repetitive elements.

Ideally, both systems could be used on the same sample: nanopore to reconstruct the full genomic sequence and Illumina to reveal relative levels of genetic diversity. This is, however, generally cost-ineffective, and the data produced can be redundant.

Target-Enrichment Sequencing

Another sequencing strategy that could be used in tandem with the (relatively) unbiased sequencing performed by Illumina and nanopore is **target-enrichment sequencing**. Target-enrichment sequencing makes some assumptions about the pathogen before the unbiased sequencing approach is used. One method functions through the enrichment of specific targeted nucleic acid sequences that constitute the genomes of pathogens (generally viruses due to their smaller genome size). One such platform, ViroFind, does this by creating thousands of viral probes that hybridize with their complementary sequences in samples and selectively amplify them.[26] This deep sequencing of targets greatly improves detection and allows thorough identification of genetic variation in target populations. Data thus produced could inform vaccine development for pathogen variants. This type of approach is

[26] Spyros Chalkias, Joshua M. Gorham, Erica Mazaika, Michael Parfenov, Xin Dang, Steve DePalma, David McKean, Christine E. Seidman, Jonathan G. Seidman, and Igor J. Koralnik, "ViroFind: A Novel Target-Enrichment Deep-Sequencing Platform Reveals a Complex JC Virus Population in the Brain of PML Patients," *PLoS One*, Vol. 13, No. 1, January 23, 2018.

also being used to monitor for novel variants of the SARS-CoV-2 virus.[27] In this case, the viral genome is amplified using qPCR and then the amplified nucleic acids are sequenced. Using a PCR amplification of a particular pathogen first sacrifices the ability to sequence all other microbes present but does enable mutation surveillance for a particular known agent, as demonstrated in the SARS-CoV-2 example.

Target enrichment is also useful if a large number of other, less relevant, microbes are expected in the sample. For example, environmental samples are known to contain a wide variety of microbes while the pathogenic species of interest may be present in small concentrations. A highly sensitive approach would be needed to detect such a pathogen, and some form of target enrichment could improve sensitivity. This has been shown to be the case for evaluating wastewater for human pathogenic viruses. The overwhelming number of bacteria present obscures the ability to reliably detect human pathogenic viruses, and some viral enrichment is needed.[28] Removal of these nontarget microbes can be done during sample preparation due to the known properties of bacterial cells and viral particles. An example of this is discussed in the next section on using pathogen-agnostic approaches for WBS. One caveat to target enrichment is that it does require some assumptions to be made and can therefore introduce bias or inadvertently eliminate a pathogen or sequence of interest.

Advantages and Limitations of NGS Approaches

The major advantage of using any NGS approach as compared with PCR or LAMP is that NGS enables researchers to ask different questions about their sample. Instead of having a suspected pathogen or a pathogen of particular concern to be monitored, NGS enables researchers to detect whichever species are present in the sample. Thus, NGS is capable of detecting truly novel agents, novel variants of known agents, and unexpected agents. Additionally, NGS can provide more information about the pathogen beyond its presence in the sample. The sequence can indicate whether a pathogenic bacterium contains known antibiotic resistance genes that may affect the clinical or public health response or can be used to track mutations and perform forensic analyses on the origin of a pathogen.

Since NGS generates a sequence, any sequence-based follow-on analyses can proceed immediately after the data analysis is complete, whereas with PCR or LAMP, the sample would need to be sent for DNA sequencing first. However, PCR and LAMP technologies are more sensitive than NGS for detecting agents that are not present in high concentrations;

[27] Matthew F. Smith, Steven C. Holland, Mihyun B. Lee, James C. Hu, Nghia C. Pham, Regan A. Sullins, LaRinda A. Holland, Tianchen Mu, Alexis W. Thomas, Remington Fitch, Erin M. Driver, Rolf U. Halden, Michelle Villegas-Gold, Sheri Sanders, Jennifer L. Krauss, Lora Nordstrom, Mary Mulrow, Michael White, Vel Murugan, and Efrem S. Lim, "Baseline Sequencing Surveillance of Public Clinical Testing, Hospitals, and Community Wastewater Reveals Rapid Emergence of SARS-CoV-2 Omicron Variant of Concern in Arizona, USA," *mBio*, Vol. 14, No. 1, January 9, 2023.

[28] Harry T. Child, George Airey, Daniel M. Maloney, Abby Parker, Jonathan Wild, Suzie McGinley, Nicholas Evens, Jonathan Porter, Kate Templeton, Steve Paterson, Ronny van Aerle, Matthew J. Wade, Aaron R. Jeffries, and Irene Bassano, "Comparison of Metagenomic and Targeted Methods for Sequencing Human Pathogenic Viruses from Wastewater," *mBio*, Vol. 14, No. 6, November–December 2023.

this is especially true in samples with many microbes and therefore a large mixture of DNA sequences. This may result in a workflow where NGS is used for biosurveillance to identify novel agents and a PCR or LAMP test is produced to monitor ongoing levels of the agent once it has been identified as an agent of interest.

At first impression, NGS seems to be the ideal diagnostic technology, but it rapidly becomes complicated in use. Not every laboratory has the hardware to run the sequencing, limiting its accessibility and prolonging time to obtain results due to shipping requirements. Also, detecting a microbe's genome in a clinical sample does not necessarily imply causation of disease—humans are home to trillions of microbes.[29] Therefore, isolating an microbe from a sick individual does not indicate the isolated microbe caused the illness. As a result, much of the utility of NGS comes from comparing results with genetic sequence databases that catalog microbes and their known association with diseases. NGS can also be limited by characteristics of the sample; samples teeming with microbes require additional processing. Finally, the advanced algorithmic processing of the sequence information can take several days to complete.

Metagenomic Analysis Using Next-Generation Sequencing

The advent of NGS technologies, more-comprehensive genomic databases, and improved bioinformatic algorithms for performing the data analysis has increased the potential of the field of study known as **metagenomics**. Metagenomics is the study of the nucleic acids present in a sample where multiple species are present, generally microbial species. Metagenomics involves sequencing the nucleic acids from the mixed sample to identify the genetic sequences that are present and assembling the sequences into genomes to determine which species are present in the sample. Therefore, any proposed NGS studies for biosurveillance would be classified as metagenomic studies since the expectation is that there will be multiple microbes present in the test sample.

Metagenomics is often used in samples that are expected to contain a lot of species, like handrail swabs from public transit stations, pond water, fecal samples, and many others. However, metagenomics also applies to samples in which only a few genomes are expected, such as clinical samples that should be free from microbes and only contain the patient DNA and a possible pathogen, such as a blood sample from a patient with a suspected blood infection.

In recent years, the cost of generating a full genome has been driven down significantly because of the high data output from NGS technologies. For these reasons, metagenomic studies now are almost exclusively performed using NGS. However, strong experimental design—which takes into account the research questions of interest and any resource constraints (e.g., time, funding)—is critical to obtaining the best-quality data.

[29] "The Idea of 'Holobiont' Represents a Paradigm Shift in Biology," *The Economist*, June 14, 2023.

Applying Nucleic Acid Sequencing Through Wastewater Metagenomics

The ability to apply the nucleic acid detection methods we have just discussed to a real-world biosurveillance example can illustrate some of the ongoing challenges with the technology and the use of data in decisionmaking. In this section we describe a scenario that applies metagenomics to WBS.

What Is Wastewater Biosurveillance?

It has long been known that human pathogens can be detected in domestic wastewater (or sewage). In the early 1900s, researchers discovered that poliovirus could be detected in human feces.[30] Beginning in the 1950s, Russian authorities began assessing wastewater to gain a population-level view of enteroviruses, like polio, in municipalities.[31] This is one of the earliest examples of a large program for WBS. WBS is a collection of techniques that involve using samples of wastewater to test for a variety of biological agents or biologically active agents. Since the 1950s, the field of WBS has been slowly developing, growing in prevalence and scope as technological methods have advanced and become cheaper. During the COVID-19 pandemic, the field showed promise in identifying hotspots for COVID-19 outbreaks even before other traditional epidemiological indicators.[32]

The ultimate goals for WBS efforts are to complement other biosurveillance efforts by assisting in targeting public health messaging, vaccination, or testing campaigns, and potentially forecasting outbreaks within a region, which enables proactive resource allocation.

What Can Be Detected Through Wastewater Biosurveillance?

Historically, most WBS efforts have focused on assessing illicit drug use at a population level and prevalence of antibiotic resistance genes in urban waste.[33] More recently, there has been renewed interest in WBS in relation to monitoring biologic agents, specifically human pathogens. WBS is extremely valuable in this context because it offers a population-level view of infection rates. The approach negates the need for individuals to undergo diagnostic testing to identify localized outbreaks. This is especially important in regions where health care access is low and in situations where diagnostic testing is not always performed. Such was

[30] Mildred G. Jimenez-Rodriguez, Fernando Silva-Lance, Lizeth Parra-Arroyo, D. Alejandra Medina-Salazar, Manuel Martínez-Ruiz, Elda M. Melchor-Martínez, María Adriana Martínez-Prado, Hafiz M. N. Iqbal, Roberto Parra-Saldívar, Damià Barceló, and Juan Eduardo Sosa-Hernándeza, "Biosensors for the Detection of Disease Outbreaks Through Wastewater-Based Epidemiology," *Trends in Analytical Chemistry*, Vol. 155, October 2022.

[31] Jimenez-Rodriguez et al., 2022.

[32] CDC, 2023a.

[33] Maria Lorenzo and Yolanda Pico, "Wastewater-Based Epidemiology: Current Status and Future Prospects," *Current Opinion in Environmental Science and Health*, Vol. 9, June 2019.

the case in the later stages of the COVID-19 pandemic when at-home testing became widespread. WBS is also valuable for detecting other viruses that are routinely circulating but do not lead to severe disease such as adenoviruses or the less severe human coronaviruses, so clinical testing is less likely to be performed. Additionally, WBS for biologic agents can provide information about a population with presymptomatic or asymptomatic disease,[34] giving public health officials an edge as they move to contain the infection or respond to those who are experiencing symptomatic disease.

Significant progress has been made in WBS during the last few years, specifically related to the COVID-19 pandemic and the other emerging infectious diseases that cropped up during that time, such as the polio reemergence in upstate New York and the international outbreak of human-transmissible Mpox. The CDC launched a nationwide program to fund, support, and provide data sharing for state and local government who are interested in leveraging WBS to monitor the amount of pathogens, particularly SARS-CoV-2, circulating in their watershed. The National Wastewater Surveillance System (NWSS) offers information on uses for WBS data in public health, guidance on sample collection, testing, and data interpretation.[35] Additionally, NWSS established three communities of practice including end users, testing facilities, and wastewater utilities. NWSS also offers a data portal for data sharing with the public.[36] NWSS utilizes dPCR as a technique and therefore must select the agent or gene for which they are surveilling prior to performing their analysis. The primary use for the NWSS so far has been SARS-CoV-2 surveillance.[37] Some municipalities have added other pathogens of interest locally, but this data has not been incorporated into the NWSS data sharing center. NWSS is planning to expand to an additional two dozen new targets that include gastrointestinal and respiratory pathogens as well as antibiotic resistance genes.[38]

WBS does have a few key limitations, the primary one being that it requires homes and offices to be linked to municipal sewer systems. Homes using septic tanks are not linked to municipal sewer systems and therefore cannot be evaluated using WBS. Similarly, when considering a global perspective, many countries and communities with poor sanitation or wastewater infrastructure may also suffer from poor data quality using WBS since a number of individuals or communities are not integrated into the wastewater network. Additionally,

[34] Scott W. Olesen, Maxim Imakaev, and Claire Duvallet, "Making Waves: Defining the Lead Time of Wastewater-Based Epidemiology for COVID-19," *Water Research*, Vol. 202, September 1, 2021.

[35] CDC, "National Wastewater Surveillance System (NWSS) Wastewater Surveillance," webpage, March 14, 2023d.

[36] CDC, "Wastewater Surveillance," webpage, undated; CDC, "Building COVID-19 Wastewater Surveillance with the National Wastewater Surveillance System," fact sheet, March 7, 2022.

[37] Amy E. Kirby, Maroya Spalding Walters, Wiley C. Jennings, Rebecca Fugitt, Nathan LaCross, Mia Mattioli, Zachary A. Marsh, Virginia A. Roberts, Jeffrey W. Mercante, Jonathan Yoder, and Vincent R. Hill, "Using Wastewater Surveillance Data to Support the COVID-19 Response—United States, 2020–2021," *Morbidity and Mortality Weekly Report*, Vol. 70, No. 36, September 10, 2021.

[38] Madeleine Johnson, "CDC Greatly Expands Number of Infectious Disease Targets for National Wastewater Surveillance," *GenomeWeb*, June 14, 2023.

regions with large transient populations such as tourism centers or commuter traffic may generate unreliable data for the local response since the sick individuals may simply be passing through. Finally, WBS also involves samples with many species present, and the pathogens are typically quite rare in this milieu. Therefore, the limits of detection are not well understood in this context even for SARS-CoV-2, which has been extensively studied using WBS.[39]

Sample Collection

Sampling involves identifying the sample type and sample location for your desired metagenomic analysis. In this scenario, wastewater samples will be collected. However, clarity on the optimal types of wastewater samples and from what sized catchment area for public health decisionmaking are not clear currently. Many states and municipalities perform collection at the influent to a water treatment plant.[40] Sampling itself is done by either collecting a sample at a single defined time (a grab sample) or by using a specialized sampler that collects a small amount of influent at regular intervals over the test period (a composite sample). For example, if the biosurveillance plan is to test weekly, the sample can be set to collect a fraction of the sample every few hours for the entire week, which will provide a better understanding of the pathogens passing through the wastewater treatment plant during the entire week.

Another important consideration will be to determine an appropriate time interval between sample collection events. Too-frequent collection dramatically increases both cost and the burden on the wastewater technicians, and will create a large amount of data that require processing and storage. Too-infrequent collection runs the risk of missing early information that will enable identification of a novel pathogen as it begins circulating in the population. Determining the appropriate interval to balance all these considerations is a critical task. The New York State Wastewater Surveillance Network recommends twice a week and no less than weekly for their SARS-CoV-2 biosurveillance efforts.[41]

Sample Preparation

After collecting the sample, the next step for nearly all nucleic acid detection methods discussed here is to prepare the sample for analysis. The NGS techniques used in metagenomic studies require a highly purified sample for analysis. However, the wastewater samples that are collected will be quite dirty samples, with significant contribution from human excreta and any other household waste, including cooking/food waste, shower water, and potentially pet waste. In areas with commercial or industrial activities, this waste may be combined with household waste.

[39] Kirby et al., 2021.

[40] CDC, "National Wastewater Surveillance System (NWSS) Sampling Strategy," webpage, March 9, 2023b.

[41] New York State Wastewater Surveillance Network, "Sample Collection," webpage, undated-b.

Sample preparation includes purifying the nucleic acid from this raw sewage. This can be performed in a multistep process where the first step removes or significantly decreases the number of larger microbes such as fungi, parasites, human cells, and even bacteria; however, if the goal is to observe genetic contribution from nonviral microbes, this step may be skipped. The next step is to perform a DNA (or RNA) extraction from the cleaned sample, or perhaps both if viral agents are being investigated.

As an example, the widely used methods for purifying sewage for SARS-CoV-2 analysis by dPCR include steps that specifically work for the small viral particles of SARS-CoV-2 and result in loss of many of the other species present in the raw sewage.[42] Careful planning of the purification method is necessary to ensure any agents of interest are being retained in the purified product. One interesting item to note is that LAMP tests are being developed to proceed directly to testing using raw sewage without any sample preparation.[43] If these tests continue to show promise, this represents a major step toward decreasing the time necessary to generate pathogen-specific results.

DNA purification and RNA purification are common techniques for any laboratory that performs nucleic acid detection. There are a multitude of commercially available purification kits to isolate nucleic acids. These kits are generally inexpensive, especially if purchased in bulk, and produce very high-quality purified DNA or RNA. Many research labs that perform extensive nucleic acid purification have their own protocols and reagents to perform the same steps, often using the same steps as the commercial kits. DNA or RNA purification can be completed in as little as 45 minutes, but some protocols take longer depending on the specific method used and how large of a sample is being purified.

Nucleic Acid Detection and Analysis

Once a sample has been collected and prepared for sequencing, the metagenomic analysis can begin. In the context of wastewater analysis, all NGS will be metagenomics since the sample will contain a mixture of genomes that will need to be correctly assembled and identified during the data analysis steps. The data analysis steps for NGS require access to genomic databases against which to compare the assembled genomes. The ability to identify microbes is dependent on access to high-quality and large databases. Access to such databases may represent a challenge to some WBS programs.

The technique used to conduct the sequencing may be altered depending on the goals of the biosurveillance program. Additionally, if an agent is being monitored over a longer course of time and the main research question is related to how much of the agent is currently

[42] CDC, "National Wastewater Surveillance System (NWSS) Testing Methods," March 10, 2023c; New York State Wastewater Surveillance Network, "Analysis Methods," webpage, undated-a.

[43] Jesmin Akter, Wendy J. M. Smith, Metasebia Gebrewold, Ilho Kim, Stuart L. Simpson, Aaron Bivins, and Warish Ahmed, "Evaluation of Colorimetric RT-LAMP for Screening of SARS-CoV-2 in Untreated Wastewater," *Science of the Total Environment*, Vol. 907, January 10, 2024.

present, a pathogen-specific technique can be substituted here instead of a sequencing technology. While converting to PCR or LAMP during ongoing monitoring of an outbreak may be best with respect to cost and time from sampling to data, these techniques do not produce metagenomic data. Metagenomic data may or may not be necessary at all stages of a biosurveillance program, and careful evaluation of the needs and research questions should be considered.

Understanding the Technical Limitations of Metagenomics

Overall, metagenomic sequencing of wastewater for biosurveillance offers great promise for identifying a novel pathogen that is circulating in the population. NGS allows researchers and public health officials the ability to monitor for novel agents as well as observe unexpected agents that would not have been included in a PCR experiment, or novel mutations of known agents. This type of analysis is not possible using methods that are pathogen-specific. However, the NGS techniques do have some important limitations.

Sequencing Data Analysis Is Slow When Compared with PCR and LAMP

Metagenomic sequencing and data analysis can be very slow. Using NGS technology will require a significant delay from sample collection until a report is generated listing the agents and strains identified in a particular sample. Both Illumina and nanopore sequencing require longer run and analysis times than other nucleic acid detection methods. For Illumina, the data analysis can be very time-consuming since the genome assembly process is so complex. Once sequences are obtained, they must be checked to determine from which known microbes they originate; any unknown microbes can be moved forward as potential novel agents. For known agents, a more rapid screen can be performed using PCR techniques. While PCR cannot be used to identify novel agents and cannot reliably identify novel strains of any of the identified microbes, PCR results can be obtained in a matter of hours, which is substantially faster than the several days required for metagenomic sequencing analyses. Additionally, if sequencing identifies a novel agent to monitor, new primers and probes can be designed and rolled out into a PCR testing panel very quickly.

Novel Agents Will Be Difficult to Characterize as Potential Threats

One of the challenges in characterizing novel agents is determining whether they represent a threat. This is especially true in the proposed scenario where samples from the environment are being tested. There is no evidence that these agents will infect humans or cause serious disease. In the case of SARS-CoV-2, the causative agent was identified in human patient samples so the ability to infect humans and cause severe disease was established clinically. Many

metagenomic sequencing studies have identified "unknown" sequences,[44] yet we have not experienced a major pandemic resulting from any of these unknown sequences. The presence of an microbe alone is insufficient to predict whether it will be particularly dangerous or transmissible. Modeling and predictive technologies attempting to categorize an unknown agent based on nucleic acid sequence alone are constantly improving. In the future, predictive abilities may be more advanced and more reliable, but for now there is still a degree of uncertainty that makes raising an alarm off these predications unwise. There is ongoing research about using temporal changes in the detection of novel microbes to help identify potential novel pathogens.[45] For example, novel microbes that appear to be increasing in concentration over time may be interesting, but without some clinical indication that there is a pathogen circulating in the population, a larger public health response seems unwarranted.

Additionally, establishing causation is challenging. The simple presence of an unknown agent, even in a clinical sample, does not necessarily mean that the agent is responsible for the disease state. The gold standard for determining causative agents in biology involves isolating the agent and exposing a healthy host to the agent and replicating the disease state. Obviously, this type of experimentation is not ethical for human diseases and therefore more evidence is required to establish a causal link using observational epidemiology and laboratory testing on cells and in animals. Ideally, novel agents can be identified by NGS and monitored for increased presence in wastewater either by continued NGS analysis to monitor abundance and changes due to mutation, or by using a faster technique such as a PCR array to monitor abundance only. The data obtained can then be combined with traditional epidemiological methods for assessing clusters of unknown infectious disease. The combination of techniques will give the highest certainty that the novel agent is present, increasing in incidence or prevalence, and potentially causing human disease.

Small Genomes Are Often Diluted by Larger Genomes

Another technical challenge is related to the potential for dilution of viral sequences by other microbe genomes. This problem arises from the disparity in the sizes of the various genomes and by the mechanism for creating reads in Illumina sequencing. Briefly, viral genomes are different than the genomes of other potential infectious agents such as bacteria and protozoa. Figure 2 shows some agents that are commonly found or known to be found in wastewater. Viruses contain either DNA or RNA but not both, unlike all living microbes. Additionally, in bacteria and protozoa (as well as mammals) the DNA is double-stranded (dsDNA) and the RNA is single-stranded (ssRNA). However, viral genomes are less rigid, and double-stranded RNA (dsRNA) and single-stranded DNA (ssDNA) genomes are known as well as ssRNA and

[44] Andrew Maltez Thomas and Nicola Segata, "Multiple Levels of the Unknown in Microbiome Research," *BMC Biology*, Vol. 17, 2019.

[45] Will Bradshaw and Simon Grimm, "Comparing Sampling Strategies for Early Detection of Stealth Biothreats," Nucleic Acid Observatory, February 19, 2024.

FIGURE 2

Examples of Some Pathogens Found in Wastewater Summarized by Key Properties of their Genomes

	Bacteria	Protozoa	Viruses
Names	Salmonella, Shigella, Campylobacter, Escherichia	Plasmodium, Giardia, Cryptosporidium	Coronaviruses, Norovirus, Polio virus
Nucleic Acids	Both dsDNA and ssRNA	Both dsDNA and ssRNA	Either DNA or RNA, both can be ss or ds
Genome size	2–5 million bp	10–15 million bp	1000–35,000 bases 300–2.5 million bp

dsDNA. The most striking difference in viral genomes is their size. Viral genomes are typically several orders of magnitude smaller than the bacterial, protozoan, fungal, and animal genomes that also exist in these wastewater samples.[46]

The challenge with identifying a small genome is that is can be significantly diluted once the sample is fragmented. Figure 3 illustrates how the viral genes can be poorly represented in the fragmented sample simply because the larger genomes are broken into more pieces during fragmentation than the smaller viral genome. Therefore, each single bacterial or protozoan cell is contributing many times more fragments to the sample than each single viral particle, leading to dilution of the viral genome by the larger genomes.[47]

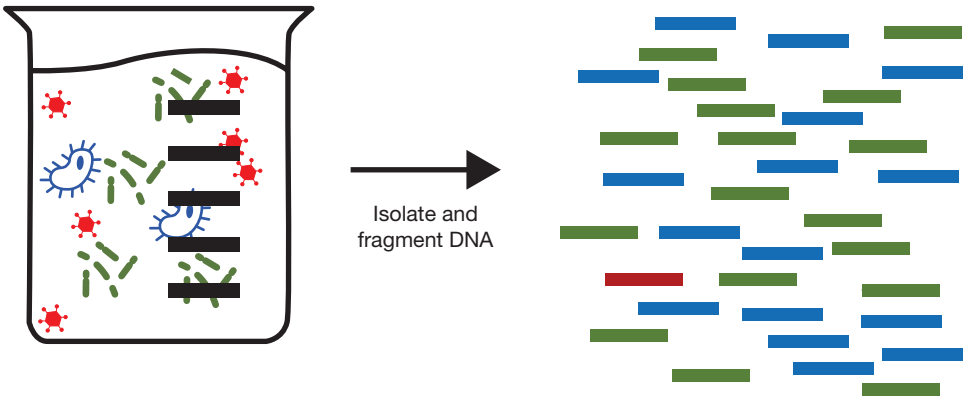
A particular complication with this dilution effect is that viruses must not be excluded from a global observatory for several reasons. The *Characteristics of Pandemic Pathogens* report identifies RNA viruses as being particularly concerning,[48] and recent history has had several global pathogenic concerns, all of which are viral (mpox, Ebola, SARS-CoV-2,

[46] For reference, the human genome is approximately 3.2 billion base pairs.

[47] Karol Perlejewski, Iwona Bukowska-Ośko, Małgorzata Rydzanicz, Agnieszka Pawełczyk, Kamila Caraballo Cortès, Sylwia Osuch, Marcin Paciorek, Tomasz Dzieciątkowski, Marek Radkowski, and Tomasz Laskus, "Next-Generation Sequencing in the Diagnosis of Viral Encephalitis: Sensitivity and Clinical Limitations," *Scientific Reports*, Vol. 10, No. 1, September 30, 2020.

[48] Amesh A. Adalja, Matthew Watson, Eric S. Toner, Anita Cicero, and Thomas V. Inglesby, *The Characteristics of Pandemic Pathogens*, Johns Hopkins Center for Health Security, May 2018.

FIGURE 3
Viral Genomes Are Diluted by Larger Genomes from the Sample



MERS, and H1N1 influenza). Therefore, additional care should be taken to ensure that viral sequences are enriched or, minimally, are not lost during purification (see discussions about target-enrichment sequencing in the second section and about sample preparation in the third section).

Nucleic Acid Techniques Do Not Assess Viability

One caveat for using any nucleic acid detection method is that nucleic acids often survive decontamination efforts, especially DNA. Therefore, PCR or sequencing offers information about the nucleic acids present in the sample, but not necessarily the nucleic acids present in living microbes or intact viruses that are still capable of causing disease. For this reason, some organizations that perform nucleic acid–based biosurveillance also perform tests such as bacterial culturing or attempt to grow the virus in cell cultures to confirm the viability of the microbe that was detected via nucleic acid methods. This limitation is of little concern in the context of using metagenomics on wastewater in a public health context since presence of these nucleic acids represents presence of these microbes somewhere in the catchment area.

Other Limitations and Caveats

Privacy Considerations

One important consideration with using metagenomics to characterize the composition of wastewater samples that is not a concern with LAMP or PCR is that human genetic material will also be present in these samples. There are sample preparation techniques to reduce the human DNA in a sample, but some will likely remain. Therefore, the human nucleic acids will also be sequenced, and the raw data will contain human genomic data from individuals who did not consent to genetic sequencing for research purposes. It is currently unclear how

much human genetic information can be reconstructed from environmental sample or the impact of this information, but existing studies have been able to partially identify sample donors.[49] Future technologies are likely to be able discern human genomic material more fully from such samples. Interestingly, even the human microbiome—the microbes that live inside our bodies—has been shown to be unique enough to be a promising marker for forensic identification.[50] Luckily, the pooling of human excreta observed in wastewater will prevent this particular type of identification at this time, but future studies may uncover truly individual genetic signatures from gut microbes that can be traced from a pooled sample.

Using purification steps for viral samples or target-enrichment sequencing will also minimize, but not eliminate, the sequencing of human nucleic acids. DNA sequences are unable to be truly de-identified and are therefore personally identifiable information. As such, protecting this information is extremely important. There should be clear protocols related to handling human genomic sequences that will be obtained during metagenomic analyses. Chan Zuckerberg ID, a global metagenomic service that specializes in metagenomic analysis of clinical samples, addresses these concerns by throwing out all reads that map to the human genome to protect the privacy of individuals. This methodology is dependent upon having a clear data retention policy in place.

Additionally, a case study using wastewater to identify SARS-CoV-2 mutants followed the mutants upstream by collecting samples from sites with smaller and smaller catchment areas to identify a single business where the infected person was likely employed.[51] While an interesting and powerful application of wastewater nucleic acid biosurveillance, this study may be seen as invasive or intimidating to the general public.

Implications of Technological Advancement Rate

Since the Human Genome Project, advances in sequencing technology have emerged with rapidity. Sanger sequencing has been mostly discarded in favor of NGS technologies such as Illumina sequencing, which in turn is facing stiff competition from so called third- and fourth-generation sequencing technologies. This instability necessitates flexibility when choosing sequencing strategies for long-term projects.

---

[49] Yoshihiko Tomofuji, Kyuto Sonehara, Toshihiro Kishikawa, Yuichi Maeda, Kotaro Ogawa, Shuhei Kawabata, Takuro Nii, Tatsusada Okuno, Eri Oguro-Igashira, Makoto Kinoshita, Masatoshi Takagaki, Kenichi Yamamoto, Takashi Kurakawa, Mayu Yagita-Sakamaki, Akiko Hosokawa, Daisuke Motooka, Yuki Matsumoto, Hidetoshi Matsuoka, Maiko Yoshimura, Shiro Ohshima, Shota Nakamura, Hidenori Inohara, Haruhiko Kishima, Hideki Mochizuki, Kiyoshi Takeda, Atsushi Kumanogoh, and Yukinori Okada, "Reconstruction of the Personal Information from Human Genome Reads in Gut Metagenome Sequencing Data," *Nature Microbiology*, Vol. 8, May 15, 2023.

[50] Jarrad T. Hampton-Marcell, Jose V. Lopez, and Jack A. Gilbert, "The Human Microbiome: An Emerging Tool in Forensics," *Microbial Biotechnology*, Vol. 10, No. 2, 2017.

[51] Ewen Callaway, "These Scientists Traced a New Coronavirus Lineage to One Office—Through Sewage," *Nature*, Vol. 609, No. 7929, September 2022.

Recently, the rate of progress on the computational side in sequence assembly and bioinformatics has dwarfed that of sequencing. Improvements in processing power, algorithmic modeling, machine learning, and database accessibility have been harnessed to enable increasingly deep insights. The limiting factor is no longer acquiring data but knowing what to do with it and how to read it. The amount of information that can be extracted in theory from genomic data and gene expression data is truly enormous. It is therefore important to consider that the value of sequence data collected in the present will grow as analytic techniques improve.

Conclusion

There are a variety of nucleic acid detection techniques that are currently available to researchers and policymakers. Biosurveillance efforts often employ nucleic acid detection techniques, including the current wastewater detection efforts that have been operating since the COVID-19 pandemic. Policymakers interested in establishing a WBS program should evaluate their specific research questions and their resource limitations to determine which technology is best suited for their needs. For example, the current WBS effort for tracking SARS-CoV-2 in wastewater utilizes dPCR for monitoring the levels of virus over time, it also uses a target-enrichment sequencing effort to monitor the mutations that are most prevalent. Therefore, selecting a single technology that can accomplish all biosurveillance goals is not necessary. Table 2 provides some information and guidance toward selecting the best technique, but relevant stakeholders, including laboratory personnel and wastewater utilities, should be included in these discussions.

TABLE 2

Selecting a Technology by Biosurveillance Goal

What Is Your Goal?	Technologies Available	Speed (time from collection to analyzed data)	Restrictions
Identifying whether known agents are present in wastewater	Illumina NGS Nanopore	Slow Slow	Target enrichment may be required prior to sequencing for reliable detection of pathogens.
	PCR LAMP	Fast Very fast	Agent mutation may reduce or eliminate efficacy of the test.
Monitoring a known agent for mutations	Illumina NGS Nanopore	Slow Slow	Best-use case involves PCR-based amplification for target enrichment prior to sequencing.
Monitoring a known agent for levels of agent in the population	Illumina NGS Nanopore	Slow Slow	Results are less quantifiable than PCR/LAMP, especially in cases with target enrichment prior to sequencing.
	PCR LAMP	Fast Very fast	Agent mutation may reduce or eliminate test efficacy.
Identifying novel agents that are present in wastewater	Illumina NGS Nanopore	Slow Slow	Target enrichment may be required prior to sequencing for reliable detection of pathogens; however, these methods do limit the scope of the resulting analysis.
Identifying unexpected agents that are present in wastewater	Illumina NGS Nanopore	Slow Slow	Target enrichment may be required prior to sequencing for reliable detection of pathogens.
	PCR	Fast	Large numbers of individual reactions will be needed to accomplish this goal. Agents not specifically tested for will be missed.
	LAMP	Very fast	Large numbers of individual reactions will be needed to accomplish this goal. Agents not included in the test will be missed. Speed may not be the same for all agents, only those with tests that work on raw sewage will be very fast.

Appendix. Glossary

TABLE A.1
Glossary of Terms Used in This Report

Term	Definition
complementary DNA (cDNA)	DNA that is generated in a laboratory from an RNA sample
conserved sequences	sequences that are found in many species with minimal variation
de novo genome assembly	assembly of novel or never-before-sequenced genomes
digital PCR (dPCR)	a form of PCR that produces highly quantitative results and is capable of detecting nonabundant sequences more quantitatively than even qPCR
DNA	nucleic acids that store and transfer genetic information; typically exist in a double-stranded form
enteroviruses	a group of RNA viruses that infect and target the intestinal lining, the most well-known enterovirus is the polio virus
flow cell	a small slide into which nucleic acids are loaded; often used in sequencing machines
fragmentation	the process (often mechanical or enzymatic) by which nucleic acids are sheared into smaller sequences or fragments
genome assembly	the process of properly recombining nucleic acid fragments into their original genomic form
Illumina sequencing	short read NGS technology developed by the private company, Illumina Inc.
known agent	previously identified microbes
long read sequencing	a sequencing approach that seeks to minimize fragmentation and allows direct sequencing of long sequences
loop-mediated isothermal amplification (LAMP)	a rapid, inexpensive technique to amplify nucleic acids, LAMP can often be done at room temperature or in simple, inexpensive instruments
nanopore sequencing	a recent sequencing technology developed by the private company Oxford Nanopore Technologies; known for very long read lengths
metagenomics	the study of nucleic acids present in a bulk, often messy, sample
next generation sequencing (NGS)	a catch-all term for sequencing technologies developed post-2000
nitrogenous bases	biological compounds present in nucleotides; primarily being adenine, thymine, guanine, cytosine, and uracil (replacing thymine in RNA)
novel agent	unknown or previously unidentified agents
nucleic acid	a complex molecule consisting of chains of nucleotides
nucleotides	building blocks of DNA and RNA; consisting of a sugar molecule, a phosphoric acid molecule, and one of five nitrogenous bases

(*continued*)

TABLE A.1—Continued

Term	Definition
oligonucleotides (oligos)	short nucleic acid fragments; often synthetically generated
polymerase chain reaction (PCR)	an optimized process by which nucleic acid sequences are amplified in vitro; usually occurring in a thermocycler machine
primer	An oligonucleotide that is specific to a particular region of DNA and is used to initiate DNA amplification reactions in PCR and LAMP
probe	An oligonucleotide that is specific to a region in between the primers and contained a fluorescent marker than is activated when this region is amplified during qPCR and dPCR
quantitative PCR (qPCR)	a form of PCR that provides real-time and quantitative results, also sometimes called “real-time PCR”
reference genome	a genome that has previously been sequenced to a high level of quality; often available in online databases
reference-guided genome assembly	assembly of previously sequenced genomes; greatly aided by reference genomes
RNA	nucleic acids that can store genetic information but often function as messengers; typically exist in a single-stranded form
sequence coverage depth	the number of times each nucleotide in a given sequence is sequenced on average; more depth means higher accuracy
sequencing	the process of determining the identity and order of nucleic acid sequences
short read sequencing	a sequencing approach in which nucleic acids are fragmented into small sequences before assembly; also called “shotgun sequencing”
tagging	the process of attaching oligos to nucleic acid sequences; usually for identification purposes
wastewater biosurveillance (WBS)	a method of using wastewater to monitor for potential public health concerns related to biological agents

Abbreviations

CDC	U.S. Centers for Disease Control and Prevention
cDNA	complementary deoxyribonucleic acid
COVID-19	coronavirus disease 2019
ddPCR	digital droplet polymerase chain reaction
DNA	deoxyribonucleic acid
dPCR	digital polymerase chain reaction
LAMP	loop-mediated isothermal amplification
NGS	next generation sequencing
NWSS	National Wastewater Surveillance System (CDC)
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
WBS	wastewater biosurveillance

References

Adalja, Amesh A., Matthew Watson, Eric S. Toner, Anita Cicero, and Thomas V. Inglesby, *The Characteristics of Pandemic Pathogens*, Johns Hopkins Center for Health Security, May 2018.

Akter, Jesmin, Wendy J. M. Smith, Metasebia Gebrewold, Ilho Kim, Stuart L. Simpson, Aaron Bivins, and Warish Ahmed, "Evaluation of Colorimetric RT-LAMP for Screening of SARS-CoV-2 in Untreated Wastewater," *Science of the Total Environment*, Vol. 907, January 10, 2024.

Bradshaw, Will, and Simon Grimm, "Comparing Sampling Strategies for Early Detection of Stealth Biothreats," Nucleic Acid Observatory, webpage, February 19, 2024. As of March 31, 2024:
https://naobservatory.org/reports/comparing-sampling-strategies-for-early-detection-of-stealth-biothreats/

Bussmann, Michael, Miriam Hesse, Oezlem Karalay, Rose Nash, and Andreas Missel, "Impact of Template Addition Volume and Analyzed Volume on Digital PCR Sensitivity," Application Note PROM-20715-001 1127766, June 2022. As of July 13, 2023:
https://www.qiagen.com/us/resources/resourcedetail?id=625d0d8b-278b-4796-bff4-2cb04533d712&lang=en

Callaway, Ewen, "These Scientists Traced a New Coronavirus Lineage to One Office—Through Sewage," *Nature*, Vol. 609, No. 7929, September 2022.

CDC—*See* Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention, "Building COVID-19 Wastewater Surveillance with the National Wastewater Surveillance System," fact sheet, March 7, 2022. As of May 7, 2024:
https://www.cdc.gov/nwss/pdf/328288_National_Wastewater_Surveillance_System_508

Centers for Disease Control and Prevention, "National Wastewater Surveillance System (NWSS)," webpage, March 10, 2023a. As of June 15, 2023:
https://www.cdc.gov/nwss/index.html

Centers for Disease Control and Prevention, "National Wastewater Surveillance System (NWSS) Sampling Strategy," webpage, March 9, 2023b. As of November 15, 2023:
https://www.cdc.gov/nwss/sampling.html

Centers for Disease Control and Prevention, "National Wastewater Surveillance System (NWSS) Testing Methods," March 10, 2023c. As of November 15, 2023:
https://www.cdc.gov/nwss/testing.html

Centers for Disease Control and Prevention, "National Wastewater Surveillance System (NWSS) Wastewater Surveillance," webpage, March 14, 2023d. As of November 15, 2023:
https://www.cdc.gov/nwss/wastewater-surveillance.html

Centers for Disease Control and Prevention, "Wastewater Surveillance," webpage, undated. As of May 7, 2024:
https://covid.cdc.gov/covid-data-tracker/#wastewater-surveillance

Chalkias, Spyros, Joshua M. Gorham, Erica Mazaika, Michael Parfenov, Xin Dang, Steve DePalma, David McKean, Christine E. Seidman, Jonathan G. Seidman, and Igor J. Koralnik, "ViroFind: A Novel Target-Enrichment Deep-Sequencing Platform Reveals a Complex JC Virus Population in the Brain of PML Patients," *PLoS One*, Vol. 13, No. 1, January 23, 2018.

Child, Harry T., George Airey, Daniel M. Maloney, Abby Parker, Jonathan Wild, Suzie McGinley, Nicholas Evens, Jonathan Porter, Kate Templeton, Steve Paterson, Ronny van Aerle, Matthew J. Wade, Aaron R. Jeffries, and Irene Bassano, "Comparison of Metagenomic and Targeted Methods for Sequencing Human Pathogenic Viruses from Wastewater," *mBio*, Vol. 14, No. 6, November–December 2023.

Hampton-Marcell, Jarrad T., Jose V. Lopez, and Jack A. Gilbert, "The Human Microbiome: An Emerging Tool in Forensics," *Microbial Biotechnology*, Vol. 10, No. 2, 2017.

"The Idea of 'Holobiont' Represents a Paradigm Shift in Biology," *The Economist*, June 14, 2023.

Illumina Inc., "Cluster Density Guidelines for Illumina Sequencing Platforms Using Non Patterned Flow Cells," webpage, undated. As of November 1, 2023: https://knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000001511

International Monetary Fund, "Policy Responses to COVID-19: Policy Tracker by Country," webpage, July 2, 2021. As of June 15, 2023: https://www.imf.org/en/Topics/imf-and-covid19/Policy-Responses-to-COVID-19

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson, "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community," *Genome Biology*, Vol. 17, 2016.

Jimenez-Rodriguez, Mildred G., Fernando Silva-Lance, Lizeth Parra-Arroyo, D. Alejandra Medina-Salazar, Manuel Martínez-Ruiz, Elda M. Melchor-Martínez, María Adriana Martínez-Prado, Hafiz M. N. Iqbal, Roberto Parra-Saldívar, Damià Barceló, and Juan Eduardo Sosa-Hernándeza, "Biosensors for the Detection of Disease Outbreaks Through Wastewater-Based Epidemiology," *Trends in Analytical Chemistry*, Vol. 155, October 2022.

Johnson, Madeleine, "CDC Greatly Expands Number of Infectious Disease Targets for National Wastewater Surveillance," *GenomeWeb*, June 14, 2023.

Kannoth, Sneha, Sasikiran Kandula, and Jeffrey Shaman, "The Association Between Early Country-Level COVID-19 Testing Capacity and Later COVID-19 Mortality Outcomes," *Influenza and Other Respiratory Viruses*, Vol. 16, No. 1, January 2022.

Kirby, Amy E., Maroya Spalding Walters, Wiley C. Jennings, Rebecca Fugitt, Nathan LaCross, Mia Mattioli, Zachary A. Marsh, Virginia A. Roberts, Jeffrey W. Mercante, Jonathan Yoder, and Vincent R. Hill, "Using Wastewater Surveillance Data to Support the COVID-19 Response—United States, 2020–2021," *Morbidity and Mortality Weekly Report*, Vol. 70, No. 36, September 10, 2021.

Lorenzo, Maria, and Yolanda Pico, "Wastewater-Based Epidemiology: Current Status and Future Prospects," *Current Opinion in Environmental Science and Health*, Vol. 9, June 2019.

Mandl, Kenneth D., J. Marc Overhage, Michael M. Wagner, William B. Lober, Paola Sebastiani, Farzad Mostashari, Julie A. Pavlin, Per H. Gesteland, Tracee Treadwell, Eileen Koski, Lori Hutwagner, David L. Buckeridge, Raymond D. Aller, and Shaun Grannis, "Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience," *Journal of the American Medical Informatics Association*, Vol. 11, No. 2, March–April 2004.

McNaughton, Anna L., Hannah E. Roberts, David Bonsall, Mariateresa de Cesare, Jolynne Mokaya, Sheila F. Lumley, Tanya Golubchik, Paolo Piazza, Jacqueline B. Martin, Catherine de Lara, Anthony Brown, M. Azim Ansari, Rory Bowden, Eleanor Barnes, and Philippa C. Matthews, "Illumina and Nanopore Methods for Whole Genome Sequencing of Hepatitis B Virus (HBV)," *Scientific Reports*, Vol. 9, No. 1, May 8, 2019.

National Science and Technology Council, White House Steering Committee for Pandemic Innovation, *First Annual Report on Progress Towards the Implementation of the American Pandemic Preparedness Plan*, September 2022.

New York State Wastewater Surveillance Network, "Analysis Methods," webpage, undated-a. As of November 15, 2023:
https://nywastewatcher.io/analysis-methods

New York State Wastewater Surveillance Network, "Sample Collection," webpage, undated-b. As of November 15, 2023:
https://nywastewatcher.io/sample-collection

Nwe, May Khat, Nisachon Jangpromma, and Lapatrada Taemaitree, "Evaluation of Molecular Inhibitors of Loop-Mediated Isothermal Amplification (LAMP)," *Scientific Reports*, Vol. 14, 2024.

Olesen, Scott W., Maxim Imakaev, and Claire Duvallet, "Making Waves: Defining the Lead Time of Wastewater-Based Epidemiology for COVID-19," *Water Research*, Vol. 202, September 1, 2021.

Oxford Nanopore Techologies, "Assembling Microbial Genomes from Complex Metagenomic Samples Using Long Nanopore Reads," Workflow 1137, October 3, 2022.

Pepin, Kim M., Ryan S. Miller, and Mark Q. Wilber, "A Framework for Surveillance of Emerging Pathogens at the Human-Animal Interface: Pigs and Coronaviruses as a Case Study," *Preventive Veterinary Medicine*, Vol. 188, March 2021.

Perlejewski, Karol, Iwona Bukowska-Ośko, Małgorzata Rydzanicz, Agnieszka Pawełczyk, Kamila Caraballo Cortès, Sylwia Osuch, Marcin Paciorek, Tomasz Dzieciątkowski, Marek Radkowski, and Tomasz Laskus, "Next-Generation Sequencing in the Diagnosis of Viral Encephalitis: Sensitivity and Clinical Limitations," *Scientific Reports*, Vol. 10, No. 1, September 30, 2020.

Ramchand, Rajeev, Sangeeta C. Ahluwalia, Mary Avriette, Gary Cecchine, Monika Cooper, Christy Foran, Daniel Hicks, Natasha Lander, and Sarita D. Lee, *Syndromic Surveillance 2.0: Emerging Global Surveillance Strategies for Infectious Disease Epidemics,* RAND Corporation, RR-A1812-1, 2023. As of May 23, 2024:
https://www.rand.org/pubs/research_reports/RRA1812-1.html

Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel, "Overview of Next Generation Sequencing Technologies," *Current Protocols in Molecular Biology*, Vol. 122, No. 1, April 2018.

Smith, Matthew F., Steven C. Holland, Mihyun B. Lee, James C. Hu, Nghia C. Pham, Regan A. Sullins, LaRinda A. Holland, Tianchen Mu, Alexis W. Thomas, Remington Fitch, Erin M. Driver, Rolf U. Halden, Michelle Villegas-Gold, Sheri Sanders, Jennifer L. Krauss, Lora Nordstrom, Mary Mulrow, Michael White, Vel Murugan, and Efrem S. Lim, "Baseline Sequencing Surveillance of Public Clinical Testing, Hospitals, and Community Wastewater Reveals Rapid Emergence of SARS-CoV-2 Omicron Variant of Concern in Arizona, USA," *mBio*, Vol. 14, No. 1, January 9, 2023.

Sponk, "Difference of DNA and RNA," 2010. As of October 25, 2022:
https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg

Thomas, Andrew Maltez, and Nicola Segata, "Multiple Levels of the Unknown in Microbiome Research," *BMC Biology*, Vol. 17, 2019.

Tomofuji, Yoshihiko, Kyuto Sonehara, Toshihiro Kishikawa, Yuichi Maeda, Kotaro Ogawa, Shuhei Kawabata, Takuro Nii, Tatsusada Okuno, Eri Oguro-Igashira, Makoto Kinoshita, Masatoshi Takagaki, Kenichi Yamamoto, Takashi Kurakawa, Mayu Yagita-Sakamaki, Akiko Hosokawa, Daisuke Motooka, Yuki Matsumoto, Hidetoshi Matsuoka, Maiko Yoshimura, Shiro Ohshima, Shota Nakamura, Hidenori Inohara, Haruhiko Kishima, Hideki Mochizuki, Kiyoshi Takeda, Atsushi Kumanogoh, and Yukinori Okada, "Reconstruction of the Personal Information from Human Genome Reads in Gut Metagenome Sequencing Data," *Nature Microbiology*, Vol. 8, May 15, 2023.

van den Wijngaard, C. C., W. van Pelt, N. J. Nagelkerke, M. Kretzschmar, and M. P. Koopmans, "Evaluation of Syndromic Surveillance in the Netherlands: Its Added Value and Recommendations for Implementation," *Eurosurveillance*, Vol. 16, No. 9, March 3, 2011.

Vora, Neil M., Lee Hannah, Chris Walzer, Mariana M. Vale, Susan Lieberman, Ashley Emerson, Jonathan Jennings, Robyn Alders, Matthew H. Bonds, Jo Evans, Bhavana Chilukuri, Sonila Cook, Nigel C. Sizer, and Jonathan H. Epstein, "Interventions to Reduce Risk for Pathogen Spillover and Early Disease Spread to Prevent Outbreaks, Epidemics, and Pandemics," *Emerging Infectious Diseases*, Vol. 29, No. 3, March 2023.

"WEF Recognizes Two Leaders in Rapid Wastewater Testing," *Water Environment & Technology*, May 2022.

About the Authors

Sara Duhachek Muggy is a physical scientist at RAND. Her research interests include emerging technology, biotechnology, biosecurity, national security, and countering weapons of mass destruction. She has a Ph.D. in biochemistry.

Mary Avriette is a Ph.D. student at Pardee RAND Graduate School. Her research interests include zoonotic and emerging infectious diseases, national security, and public health policy. She has a D.V.M. and an M.P.H.

Derek Roberts is a technical analyst at RAND. He is focused on advances in biotechnology, pathogen surveillance, pandemic response, gene therapy, climate change mitigation, and community resilience. He has an M.S. in molecular biology.

Kristen Leuschner is a senior research communications analyst at RAND. She supports communication efforts in a wide range of fields, including emergency preparedness, disaster resilience, national security, health care, and logistics. She has a Ph.D. in English.