Metagenomic sequencing detects human respiratory and enteric viruses in air samples collected from congregate settings.

Mitchell D. Ramuta¹*±, Nicholas R. Minor²±, Miranda R. Stauss², Olivia E. Harwood¹, Savannah F. Brakefield¹, Alexandra Alberts¹, William C. Vuyk¹, Max J. Bobholz¹, Jenna R. Rosinski¹, Sydney Wolf¹, Madelyn Lund¹, Madison Mussa¹, Lucas J. Beversdorf⁵, Matthew T. Aliota⁴, Shelby L. O’Connor¹, David H. O’Connor¹,²±

¹ Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison, Madison, WI, USA
² Wisconsin National Primate Research Center, Madison, WI USA
³ City of Milwaukee Health Department Laboratory, Milwaukee, WI, USA
⁴ Department of Veterinary and Biomedical Sciences, University of Minnesota, Twin Cities, Minneapolis, MN, USA

± These authors contributed equally.
* Correspondence can be addressed to:
David H. O’Connor
dhoconno@wisc.edu
555 Science Drive
Madison, WI USA 53711

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Innovative methods for evaluating viral risk and spread, independent of test-seeking behavior, are needed to improve routine public health surveillance, outbreak response, and pandemic preparedness. Throughout the COVID-19 pandemic, environmental surveillance strategies, including wastewater and air sampling, have been utilized alongside widespread individual-based SARS-CoV-2 testing programs to provide population-wide data. To date, environmental surveillance strategies have mainly relied on pathogen-specific detection methods to monitor viruses through space and time. However, this provides a limited picture of the virome present in a sample, leaving us blind to most circulating viruses. In this study, we explore whether virus-agnostic deep sequencing can improve the utility of air sampling to detect human viruses captured in air samples. We show that sequence-independent single-primer amplification sequencing of nucleic acids from air samples can detect common and unexpected human respiratory and enteric viruses, including influenza virus type A and C, respiratory syncytial virus, human coronaviruses, rhinovirus, SARS-CoV-2, rotavirus, mamastrovirus, and astrovirus.
Introduction

As of April 19th, 2023, over 1 billion SARS-CoV-2 diagnostic tests have been performed in the United States. Deploying individual testing programs at this scale is extraordinarily expensive and resource-intensive. It is not sustainable, prompting the need for new virus monitoring strategies to overcome the challenges associated with individual testing.

Environmental surveillance strategies, such as wastewater and air sampling, do not rely on individualized testing and offer rapid and efficient ways to assess infection risk in communities and congregate settings. Air sampling captures bioaerosols, making it possible to identify aerosolized viruses, bacteria, fungi, or other biological material. Developing pathogen-agnostic detection methods that take advantage of air surveillance networks could improve surveillance in the absence of expensive, pathogen-specific individual testing. Despite the importance of airborne viruses to human disease transmission, only a few studies have used metagenomic sequencing to characterize them in the air of built environments. These studies successfully detected viruses in various settings but found human-associated viruses in low abundance compared to other viruses. Prussin et al. characterized airborne viral communities in a daycare center’s HVAC system over a year. While bacteriophages and plant-associated viral families were abundant, commonly circulating human viruses accounted for less than 0.005% of the total relative abundance.

Technical challenges in characterizing human viruses in bioaerosols compared to other microbes has likely contributed to the limited research in this field. Air samples capture low amounts of viral genetic material compared to other microbes, but the captured material is less compromised by the dilution effect and the PCR-inhibitor compounds present in wastewater samples. Sampling for longer durations or at higher flow rates can increase biomass.
collected but might negatively impact virus particle stability and genetic material integrity ¹⁷. Furthermore, bacteria and fungi have universal genetic marker regions (16S and internal transcribed spacer (ITS) ribosomal RNA, respectively) that are typically used for studying these microbial communities. Viruses lack a conserved viral gene marker for studying viral communities, which requires unbiased amplification of genetic material ¹⁸. Using metagenomic sequencing techniques compatible with collected air samples could reveal underappreciated viruses that are present in the environment. One especially promising method is sequence-independent single-primer amplification (SISPA) ¹⁹. SISPA sequencing has been demonstrated to detect a wide range of viruses in clinical samples ²⁰.

In 2021, we reported the characterization of SARS-CoV-2 from air samples collected from congregate spaces ²¹. We additionally used a semi-quantitative PCR assay to assess 40 other pathogens collected from the air. Detection and sequencing of pathogens in air with unbiased approaches is key to understanding which other pathogens are present in the spaces around us.

Here, we use SISPA to characterize a broad array of human pathogenic viruses from air collected from congregate spaces. We focus on detecting and characterizing RNA viruses due to their public health importance. Understanding human pathogens in built environments' air may help elucidate illness trends in communities over time. This approach could enhance air sampling as a tool for improving public health pathogen surveillance and preparedness against emerging and re-emerging viruses.
Results

Study design

From July 2021 to December 2022, we deployed active air samplers in several community settings in the Upper Midwestern states of Wisconsin and Minnesota for routine pathogen monitoring. Thermo Fisher AerosolSense Samplers were used to collect air samples from high-traffic areas in several different congregate settings, including a preschool, campus athletic facility, emergency housing facility, brewery taproom, household, and five K-12 schools. Air samples were collected at weekly and twice-weekly intervals as previously described. To demonstrate the feasibility of using pathogen-agnostic sequencing to detect human viruses captured in air samples in real-world settings, we analyzed a total of 22 air samples across the 10 congregate settings (Table 1). We also processed three air sample filter substrates from unused AerosolSense cartridges, as no-template controls. Viral RNA was extracted from air samples, and complementary DNA (cDNA) was prepared using sequence-independent single primer amplification (SISPA) for Oxford Nanopore deep sequencing and metagenomic analysis. Sequencing reads were filtered for host and reagent contaminants and mapped to a viral RefSeq file downloaded from Genbank containing 835 human-associated viral genome sequences to look for common circulating RNA viruses (available on GitHub at https://github.com/dholab/pathogen-agnostic-sequence-analysis/blob/main/resources/ncbi_human_virus_refseq_20221011.masked.fasta).
Detection of human respiratory and enteric viruses

Deep sequencing detected human viruses in 19 out of 22 (86%) air samples. No human viruses were detected in any of the no-template controls. A viral hit was defined as multiple reads mapped to the viral reference sequence in at least two unique locations of the genome. A total of 13 human RNA viruses were detected in air samples (Table 1). We were able to detect several respiratory pathogens associated with frequent and seasonal illnesses that cause a burden on the healthcare system, including influenza virus type A and C, respiratory syncytial virus subtypes A and B, human coronaviruses (NL63, HKU1, and 229E), rhinovirus, and SARS-CoV-2 (Figure 1). Additionally, metagenomic analysis was able to identify human viruses associated with enteric disease, including rotavirus, human astrovirus, and mamastrovirus.

Enteric viruses were detected in ten out of 22 (45%) air samples in this study. We observed the greatest abundance of human respiratory and enteric viruses at the preschool testing site. This is consistent with previous studies that have observed higher frequencies of infectious diseases among children in daycare and preschool facilities, especially respiratory diseases. This highlights the ability of metagenomic sequencing to detect viruses that frequently circulate, as well as unexpected and understudied viruses.

Characterizing influenza C virus lineage in a preschool air sample

Using metagenomic analysis of air sample AE0000100A8B3C collected from the preschool, we detected sequences of all seven gene segments of influenza C virus (ICV), including hemagglutinin-esterase (HE), each of the genes encoding proteins for the polymerase complex (PB2, PB1, and P3), nucleoprotein (NP), matrix (M), and nonstructural protein (NS) (supplementary data 1). We wanted to further study this sample because of the unexpected number of ICV reads, which gave near-full genome sequences for each segment. Additionally,
ICV is an understudied respiratory virus. There are a total of 2,475 ICV sequences available in NCBI Virus (taxid:11552) and only 134 ICV sequences have been submitted from the United States in the 21st century. In order to further characterize the ICV isolate from air sample AE0000100A8B3C, we sequenced the sample's remaining SISPA prepared cDNA using the Oxford Nanopore GridION to obtain a greater depth of coverage across the ICV genome. Metagenomic sequencing on the GridION resulted in a greater depth of coverage across all viruses in the sample, including all seven ICV gene segments (supplementary data 1). These reads were used to create consensus sequences for each gene segment. Phylogenetic analysis was performed to compare the ICV detected in the air sample to 45 other ICV viruses that were obtained from GenBank (supplementary data 1). There are six genetic lineages for the HE gene and two lineages for all other gene segments (Figure 2; supplementary figure 1). Phylogenetic analysis was able to identify the viral lineage of each gene segment. HE grouped with the C/Kanagawa/1/76 lineage. PB2, PB1, M, and NS grouped with the C/Yamagata/81 lineage. P3 and NP group with C/Mississippi/80 lineage. These data suggest that a genetic reassortment event, the exchange of genome segments between two different strains, occurred previously for this virus because it contains genetic segments from multiple different lineages. Frequent reassortment events have been previously observed for circulating ICV\textsuperscript{24}. However, this is not a novel reassortant since the sequence clusters closely with influenza C virus C/Scottland/7382/2007 previously identified by Smith et al. (Figure 2; supplementary figure 1)\textsuperscript{25}. These data demonstrate that metagenomic sequencing can be used to characterize understudied viruses collected in air samples, which could provide valuable information for improving public health awareness.
Longitudinal detection of human viruses in a preschool

Metagenomic analysis of air samples longitudinally collected from congregate settings can provide insight into changes in the prevalence of pathogens over time. These data could provide public health authorities valuable information to improve routine pathogen surveillance programs and outbreak investigations. To track the prevalence of viral genetic material from ICV and other human viruses in a preschool, we analyzed four air samples that were longitudinally collected from January 5, 2021, to March 1, 2022. ICV was first detected in an air sample collected from January 26th to February 1st, 2022. Viral reads in this sample mapped to three out of the seven gene segments including HE, PB2, and NS. Two air samples collected after February 1st, 2022, also contained reads that mapped to several ICV gene segments. ICV genetic material was detected at the highest abundance in the air sample collected from February 1st to the 8th with reads mapping to all seven gene segments (Table 1; supplementary data 1). Viral reads mapping to five of the seven gene segments, including PB2, PB1, P3, HE, and NP, were detected in an air sample collected from February 23rd to March 1st. In contrast, human astrovirus genetic material was detected in three air samples collected between January 5th, 2021, to February 8th, 2022. Human astrovirus was not detected in the air sample collected from February 23, 2022, to March 1st, 2022. Similar to ICV, we see a peak and decline in viral reads for human astrovirus. These data suggest that metagenomic sequencing can be used to track the prevalence of multiple respiratory and enteric viruses in a congregate setting over time. This also shows that these are real virus detection signals and not just artifacts of a single sample at a single time point.
Detection of SARS-CoV-2 in RT-PCR-positive air samples

To explore whether metagenomic sequencing can detect a human virus that is known to be present in an air sample, we sequenced air samples with known SARS-CoV-2 status. Each AerosolSense cartridge comes with two filter substrates. One filter substrate from each air sample was tested by reverse transcription PCR (RT-PCR) to determine its SARS-CoV-2 status. The other substrate was eluted in 500ul of PBS and stored at -80°C until it was processed for sequencing. Several different RT-PCR assays were used on samples included in this study, depending on when and where they were collected, as previously described. Cut-off values used for determining if an air sample was positive, inconclusive, or negative for SARS-CoV-2 are described in the methods section.

SISPA sequencing was able to detect SARS-CoV-2 reads in two out of 14 (14%) of the SARS-CoV-2 positive samples (Table 1; supplementary data 1). The percent of genome coverage varied between the two samples (6.5% and 46.7%). No SARS-CoV-2 reads were observed in any of the samples that were negative or inconclusive for SARS-CoV-2 by RT-PCR testing or with no template controls (supplementary data 1). An inconclusive result was defined as a sample with only amplification in one of the PCR targets. These data suggest that SISPA sequencing results are concordant with SARS-CoV-2 RT-PCR status. However, SISPA sequencing is not as sensitive as RT-PCR for detecting viral genomic material captured in the air samples.
Discussion

In this study, we used metagenomic sequencing to detect human RNA viruses captured in air samples collected from various congregate settings. Our results show that pathogen-agnostic sequencing can detect common and understudied respiratory and enteric viruses that circulate in the human population. Several studies have also detected enteric viruses in air samples from a daycare, wastewater treatment facility, and hospital. Air samples can detect virus-containing bioaerosols that are produced by infected individuals when they breathe, speak, sneeze, cough, or talk. Enteric viruses are mainly transmitted through the fecal-oral route, but viral genetic material has been detected in oral swabs and saliva of infected individuals and in animal models. Air samples can also detect microbe-containing bioaerosols from various secondary sources, such as plumbing systems (toilets, sinks, showers), resuspended dust, skin, or clothing. This makes it difficult to determine the source of pathogens detected in the air samples from this study. Regardless of the source of the nucleic acids, the detection of enteric viruses in air samples demonstrates the potential for expanding air sampling as an environmental surveillance approach beyond respiratory viruses.

Additionally, we demonstrate SISPA amplification and Oxford Nanopore sequencing can characterize viral genetic material in air samples to track circulating virus variants. Here, we successfully identified the lineages of all seven gene segments of an influenza C virus detected in a preschool. Influenza C virus is a lesser-studied influenza virus that is often excluded from routine respiratory pathogen surveillance programs, and highlights the limitations of pathogen-specific surveillance. Previous studies have shown a high seroprevalence of ICV in children increasing in age, suggesting that this is a common yet underestimated cause of respiratory illness. The epidemiology of ICV remains poorly understood despite recent studies.
showing its ubiquity. The ICV results from this study highlight the potential of using air sample networks to identify and characterize unexpected and understudied viruses to improve public health awareness. Throughout history, human interest and difficulties associated with detecting viruses, such as ICV, have imposed biases on our knowledge of circulating viruses. Detecting active influenza C virus infections used to be difficult with cell culture techniques because ICV causes weak cytopathic effects. This has likely led to an underestimation of disease burden. Advances in molecular detection technologies, such as pathogen-agnostic sequencing, could free us from these biases and provide a more comprehensive picture of circulating viruses.

Our study shows that we can use metagenomics from collected air samples to detect pathogens that were otherwise hidden from view. This is not surprising, as the entire field of metagenomics is intended to detect pathogens whose presence was not apparent. However, detection of hidden viruses in the air has not been extensively studied, despite the opportunity to seek out viruses in the environment that could have pathogenic consequences. Once detected by methods described in this paper, molecular detection assays can be created to more sensitively monitor the presence of these viruses within individuals, collected from the air, or located in other environmental samples. Metagenomics of collected air samples provides a window into what other pathogens are present and might be important for public health investigations.

Lessons learned during the COVID-19 pandemic have sparked interest in using environmental surveillance strategies for improving pandemic preparedness and outbreak response. Unlike wastewater surveillance, active air samplers are mobile, which makes it easy to quickly deploy air sampling networks in settings of interest such as health clinics, airplanes, ports of entry,
public transit, farms, K-12 schools, long-term care facilities, or emergency housing facilities. These may be specific areas where public health agencies are interested in collecting airborne pathogens for assessing local, regional, or global outbreaks. Highlighting this potential application of air sampling, Mellon et al. recently deployed AerosolSense samplers in an outpatient clinic for patients suspected of mpox infection to look for mpox virus in the air. Air sample results were concordant with the presence of patients diagnosed with mpox infections; future studies could extend this work to sequencing mpox virus from air samples to track evolution longitudinally.

Could a nationwide air surveillance network have been deployed in health clinics and other congregate settings to improve public health awareness early on in the COVID-19 pandemic response? It is tempting to speculate that air samplers at critical locations such as airports, hospitals, and aged care facilities could have provided sequence-agnostic detection of SARS-CoV-2 in a critical window when PCR testing of clinical samples was extremely limited. These data could have helped public health agencies make data-driven decisions early in the response.

Improvements in bioaerosol collection and sequencing technologies could significantly improve the ability to characterize viruses. The SARS-CoV-2 RT-PCR and sequencing data from this study shows that our current metagenomic sequencing approach is not optimally sensitive. Air samples contain high amounts of human, animal, and microbial ribosomal RNA (rRNA), likely associated with airborne microbes and host cells transported on dust particles.

Several studies have shown that rRNA depletion can improve the performance of unbiased sequencing techniques for recovering human RNA viruses from different matrices, and should be considered for use with air samples. Alternatively, using capture-based probe hybridization may help enrich viral target sequences in the library preparation process. This
could limit the amount of background genomic material from host and other microbes.

However, this is a more biased approach using probes specific for viral sequences of the target pathogens\textsuperscript{37}. Commercial kits are available containing probes that cover 3,153 different viral genomes targeting ssRNA, dsRNA, dsDNA, and ssDNA viruses. These kits have been used with several different sample types to detect novel and common viruses, including human and animal specimens (nasal swabs and plasma), mosquitoes, and wastewater\textsuperscript{38–41}. This approach shows promise for enriching viral reads in samples that contain high amounts of host genetic material. Future studies should evaluate using target enrichment approaches to improve the broad detection of human viruses in air samples.

Virus-agnostic sequencing methods for air surveillance can be substantially improved, as technologies are evolving rapidly. Sequencing workflows need to be optimized to be high-throughput, cost effective, and have rapid result turnaround for widespread use with air surveillance programs. In this study we ran two Oxford Nanopore sequencing runs on the PromethION 24. The runs multiplexed 16 and 9 samples and had an output of 85 and 67 Gbases per flow cell, respectively. The manufacturer estimates a maximal output of 290 Gbases per flow cell when using newer sequencing kit chemistries\textsuperscript{42}. Improving the sequencing yield to 200 Gbases could allow for multiplexing up to 32 air samples, while maintaining an average of 6 million Gbases per air sample. This could make sequencing more cost effective while maintaining a similar per sample output obtained in this study. Additionally, Oxford Nanopore sequencing enables real-time processing of sequencing data. This could help decrease the turnaround time for results since you do not have to wait the full 72 hour run time before downloading and analyzing the data. This rapid result turnaround time could be beneficial during outbreak response, when real-time data is essential.
Recent advances in metagenomic sequencing technologies have increased efforts to study microbial communities in built environments. This study demonstrates that metagenomic sequencing approaches can be used to track the prevalence of human respiratory and enteric viruses of public health importance in real-world settings. This approach could provide a more rapid and efficient method for providing community-wide pathogen surveillance data without relying on test-seeking behavior or pathogen-specific detection assays.

Methods

Air sample collection and processing

AerosolSense instruments (Thermo Fisher Scientific) were installed in a variety of indoor congregate settings to collect bioaerosols for pathogen surveillance from December 2021 to December 2023. AerosolSense instruments were placed on flat surfaces 1-1.5 meters off the ground in high-traffic areas of an athletics training facility, preschool, emergency housing facility, brewery taproom, and five K-12 schools in the Upper Midwestern States of Wisconsin and Minnesota. Air samples were collected using AerosolSense cartridges (Thermo Fisher Scientific) according to the manufacturer’s instructions. The iOS and Android Askidd mobile app was used to collect air cartridge metadata and upload it to a centralized Labkey database, as previously described in Ramuta et al. After the air samples were removed from the instruments, they were transferred to the lab for further processing. Two air sample substrates were removed from each of the AerosolSense cartridges using sterile forceps to place them in two separate 1.5 mL tubes containing 500 µL of PBS. The tubes were vortexed for 20 seconds, centrifuged for 30 seconds, and stored at -80°C until RNA extraction and complementary DNA (cDNA) preparation.
Air sample total nucleic acid extraction and concentration

Total nucleic acids were extracted from air samples using the Maxwell 48 Viral Total Nucleic Acid Purification Kit (Promega) according to the manufacturer’s recommendations. Briefly, 300 μL of air sample eluate was added to a 1.5 μL tube containing 300 μL of lysis buffer and 30 μL of Proteinase K. An unused air cartridge was processed with each Maxwell run to be used as a no-template control. The reaction mix was vortexed for 10 seconds and incubated at 56°C for 10 minutes. Following the incubation, the tubes were centrifuged for 1 minute. Then 630 μL of the reaction mix was added to the Maxwell 48 cartridges, which were loaded into a Maxwell 48 instrument and processed with the Viral Total Nucleic Acid program. Nucleic acids were eluted in a final volume of 50 μL of nuclease-free water. To clean and concentrate the viral RNA, 30 μL of extracted total nucleic acids were treated with TURBO DNase (Thermo Fisher Scientific) and concentrated to 10 μL with the RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer’s protocols.

Air sample sequencing

A modified sequence-independent single primer amplification (SISPA) approach previously described by Kafetzopoulou et al. was used to generate cDNA from the air samples. First, 1 μL of Primer A (Table 2) was added to 4 μL of concentrated viral RNA and incubated in a thermocycler at 65°C for 5 minutes, followed by 4°C for 5 minutes. To perform reverse transcription, 5 μL of Superscript™ IV (SSIV) First-Strand Synthesis System (Invitrogen) master mix (1 μL of dNTP (10mM), 1 μL of nuclease-free water, 0.5 μL of DTT (0.1 M), 2 μL of 5X RT buffer, and 0.5 μL of SSIV RT) was added to the reaction mix and incubated in a thermocycler.
at 42°C for 10 minutes. To perform second-strand cDNA synthesis, 5 µL of Sequenase Version 2.0 DNA polymerase (Thermo Fisher Scientific) master mix (3.85 µL of nuclease-free water, 1 µL of 5X Sequenase reaction buffer, and 0.15 µL of Sequence enzyme) was added to the reaction mix and incubated at 37°C for 8 minutes. After the incubation, 0.45 µL of the Sequenase dilution buffer and 0.15 µL of Sequenase were added to the reaction mix and incubated at 37°C for 8 minutes. To amplify the randomly primed cDNA, 5 µL of the cDNA was added to 45 µL of the Primer B reaction mix (5 µL of AccuTaq LA 10x buffer, 2.5 µL of dNTP (10mM), 1 µL of DMSO, 0.5 µL of AccuTaq LA DNA polymerase, 35 µL of nuclease-free water, and 1 µL of Primer B (Table 2)). The following thermocycler conditions were used to amplify the cDNA: 98°C for 30 seconds, 30 cycles (94°C for 15 seconds, 50°C for 20 seconds, and 68°C for 2 minutes), and 68°C for 10 minutes. The amplified PCR product was purified using a 1:1 ratio of AMPure XP beads (Beckman Coulter) and eluted in 25 µL of nuclease-free water. The purified PCR products were quantified with the Qubit dsDNA high-sensitivity kit (Invitrogen).

Oxford Nanopore sequencing

SISPA-prepared cDNA were submitted to the University of Wisconsin-Madison Biotechnology Center for sequencing on the Oxford Nanopore PromethION. Upon arrival, the PCR product concentrations were confirmed with the Qubit dsDNA high-sensitivity kit (Invitrogen). Libraries were prepared with up to 100-200 fmol of cDNA according to the Oxford Nanopore ligation-based sequencing kit SQK-LSK109 and Native Barcoding kit EXP-NBD196. The quality of the finished libraries was assessed using an Agilent Tapestation (Agilent) and quantified again using the Qubit® dsDNA HS Assay Kit (Invitrogen). Samples were pooled and sequenced with an FLO-PRO002 (R9.4.1) flow cell on the Oxford Nanopore PromethION 24 for 72 hours. Data
were basecalled using Oxford Nanopore’s Guppy software package (6.4.6) with the high
accuracy basecalling model (read filtering parameters: minimum length 200 bp, minimum
Qscore=9). Air sample AE0000100A8B3C was also sequenced on the Oxford Nanopore
GridION to obtain a greater depth of coverage across all seven influenza C virus gene
segments. A sequencing library was prepared for AE0000100A8B3C according to the Oxford
Nanopore ligation-based sequencing kit SQK-LSK110 instructions. The sample was
sequenced with an FLO-MIN106 (R9.4) flow cell on the Oxford Nanopore GridION for 72 hours.
Data were basecalled using Oxford Nanopore’s Guppy software package (6.4.6) with the high
accuracy basecalling model (read filtering parameters: minimum length 20 bp, minimum
Qscore=9).

Sequencing analysis

Sequencing data generated from air samples were deposited in the Sequence Read Archive
(SRA) under bioproject PRJNA950127. The removal of host reads was requested at the time of
SRA submission using the Human Read Removal Tool (HRRT). The sequencing data were
analyzed using a custom workflow. To ensure reproducibility and portability, we implemented
the workflow in NextFlow and containerized all software dependencies with Docker. All
workflow code and replication instructions are publicly available at

(https://github.com/dholab/pathogen-agnostic-sequence-analysis). Briefly, the workflow starts
by automatically pulling the study fastq files from SRA, though it has the option of merging
locally stored demultiplexed fastq files as well. Then, reads are filtered to a minimum length
(200bp) and quality score (Qscore=9), and adapters and barcodes are trimmed from the ends
of the reads, all with the reformat.sh script in bbmap (39.01-0). The filtered fastq files for each
air sample are then mapped to contaminant FASTA files containing common contaminants
with minimap2 (v2.22). Reads that do not map to the contaminant FASTA files are retained and
mapped to their sequencing run’s negative control reads to further remove contaminants
present from library preparation. The cleaned fastq files for each air sample are then mapped
to a RefSeq file containing human viruses downloaded from NCBI Virus using minimap2
(v2.22). The human virus reference file contains 835 viral genome sequences and was
processed using the bbmask.sh command in bbmap (39.01-0) with default parameters to
prevent false-positive mapping to repetitive regions in viral genomes. SAM files for each
sample are converted to BAM format, again with reformat.sh. The workflow then completes by
generating a pivot table of pathogen “hits”, which lists the number of reads supporting each
mapped pathogen for each sample. For this study, we then imported the BAM alignments into
Geneious Prime (2023.0.4) to inspect the mapping results visually. Genome coverage plots
were created for several respiratory and enteric viruses detected in air samples using ggplot2
(3.4.1) with a custom R script (4.2.3) in RStudio (2023.03.0+386).

Phylogenetic analysis

To compare the influenza C virus detected in the preschool air sample AE0000100A8B3C we
downloaded 45 influenza C virus genome sequences for each of the seven gene segments
from Genbank (HE, PB2, PB1, P3, NP, M, and NS). Accession numbers for each segment can
be found in supplementary data 1. Consensus sequences were generated from
AE0000100A8B3C with a minimum coverage of 20X. Sections with low coverage were masked
with N and trimmed to the reference sequence length. Next, each set of influenza C virus gene
segment sequences was aligned using MUSCLE (5.1) implemented in Geneious Prime
(2023.0.4) with the PPP algorithm. We then used the Geneious Tree Builder (2023.0.4) to
construct a phylogeny for each gene segment using the Neighbor-joining method and Tamura-Nei model with 100 bootstrapped replicates.

**SARS-CoV-2 RT-PCR**

Air samples collected between December 2021 and May 2022 were tested for SARS-CoV-2 viral RNA using three different SARS-CoV-2 RT-PCR assays depending on their collection location as previously described. Air samples collected after May 2022 were tested for SARS-CoV-2 viral RNA using an RT-PCR protocol as previously described. Briefly, viral RNA was isolated from the air sample substrate using 300 μL of eluate and the Viral Total Nucleic Acid kit for the Maxwell 48 instrument (Promega), following the manufacturer’s instructions. RNA was eluted in 50 μL of nuclease-free water. Reverse transcription qPCR was performed using primer and probes from an assay developed by the Centers for Disease Control and Prevention to detect SARS-CoV-2 (N1 and N2 targets). The 20 μL reaction mix contained 5 μL of 4x TaqMan Fast Virus 1-Step Master Mix, 1.5 μL of N1 or N2 primer/probe mix (IDT), 5 μL of sample RNA, and 8.5 μL of nuclease-free water. The RT-PCR amplification was run on a LightCycler 96 at the following conditions: 37°C for 2 minutes, 50°C for 15 minutes, 95°C for 2 minutes, 50 cycles of 95°C for 3 seconds and 55°C for 30 seconds, and final cool down at 37°C for 30 seconds. The data were analyzed in the LightCycler 96 software 1.1 using absolute quantification analysis. Air samples were called positive when N1 and N2 targets both had cycle threshold (Ct) values <40, inconclusive when only one target had Ct <40, and negative if both targets had Ct >40.
References

1. CDC. COVID data tracker. Centers for Disease Control and Prevention

2. Polo, D. et al. Making waves: Wastewater-based epidemiology for COVID-19 -
   approaches and challenges for surveillance and prediction. Water Res. 186, 116404
   (2020).


4. Kirby, A. E. et al. Notes from the Field: Early Evidence of the SARS-CoV-2 B.1.1.529
   (Omicron) Variant in Community Wastewater - United States, November-December 2021.

5. McClary-Gutierrez, J. S. et al. SARS-CoV-2 Wastewater Surveillance for Public Health

6. Yang, W., Elankumaran, S. & Marr, L. C. Concentrations and size distributions of airborne
   influenza A viruses measured indoors at a health centre, a day-care centre and on

7. Prussin, A. J., 2nd & Marr, L. C. Sources of airborne microorganisms in the built


9. Myatt, T. A. et al. Detection of airborne rhinovirus and its relation to outdoor air supply in


32. Anderson, B. D. *et al.* Bioaerosol Sampling in Modern Agriculture: A Novel Approach for


42. *PromethION 2 Solo Technical Specification.*

https://community.nanoporetech.com/requirements_documents/promethion-2s-spec.pdf

Acknowledgments

This work was made possible by financial support through the National Institutes of Health grant (AAL4371). M.D.R. is supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number T32AI55397. The author(s) thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility (Research Resource Identifier – RRID:SCR_017759) for providing PromethION sequencing services. We would like to acknowledge Eli O’Connor’s work in developing the iOS and Android Askidd mobile app to help streamline air sample metadata collection. We would like to thank all of the participating congregate settings for their partnership during this study.

Author contributions

M.D.R contributed to the conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing—original draft preparation, writing—review and editing. N.R.M contributed to the formal analysis, investigation, methodology, writing—original draft preparation, and writing—review and editing. D.H.O and S.L.O. contributed to the conceptualization, project administration, writing—original draft preparation, and writing—review and editing. M.R.S., O.E.H., A.A., W.C.V., M.J.B., and J.R.R. contributed to data curation, logistics, organization, and writing—review and editing. L.J.B. and M.T.A. contributed to data curation, resources, project management, and writing—review and editing. S.F.B, S.W., M.L., and M.M. contributed to logistics, organization, and writing—review and editing.
The air sample sequencing data generated in this study have been deposited in the Sequence Read Archive (SRA) under bioproject PRJNA950127. The accession numbers for influenza C virus samples used in the phylogenetic analysis are provided in Supplementary Data 1.

Code Availability

Code to replicate air sample sequencing analysis is available at https://github.com/dholab/pathogen-agnostic-sequence-analysis.
Figures and Tables

**Figure 1.** Human respiratory and enteric viruses detected in air samples by SISPA amplification and Oxford Nanopore sequencing. Genome coverage plots showing read depth across eight human RNA viruses detected in air samples. The depth of coverage is shown on the y-axis, and the genome position is shown on the x-axis. The scale of the y-axis varies between plots. Coverage plots were created using ggplot2 (3.4.1) using a custom R script (4.2.3) in RStudio (2023.03.0+386).
Figure 2. Phylogenetic analysis of influenza C virus hemagglutinin-esterase gene.

Phylogenetic trees for influenza C virus hemagglutinin-esterase (HE) gene segment. Nucleotide sequences were aligned using MUSCLE (5.1). The phylogenetic tree was constructed with the Geneious Tree Builder (2023.0.4) using the Neighbor-joining method and Tamura-Nei model with 100 bootstrapped replicates. Numbers above the branches indicate the bootstrap values with 100 replicates. ICV strain names are listed at the end of branches. ICV strains belonging to the C/Sao Paulo lineage are represented in dark blue, C/Aichi lineage in blue, C/Mississippi lineage in light blue, C/Taylor lineage in yellow, C/Kanagawa lineage in orange, C/Yamagata lineage in red.
<table>
<thead>
<tr>
<th>Air sample ID</th>
<th>Location</th>
<th>Start Date</th>
<th>End Date</th>
<th>Total number of reads</th>
<th>Pathogen</th>
<th>Number of mapped reads</th>
<th>Percent genome coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC211129</td>
<td>Brewery Taproom</td>
<td>11.22.21</td>
<td>11.29.21</td>
<td>4,113,547</td>
<td>Rotavirus A</td>
<td>276</td>
<td>57.2%</td>
</tr>
<tr>
<td>AE000010795B42</td>
<td>Emergency housing shelter</td>
<td>12.21.21</td>
<td>1.7.22</td>
<td>4,898,090</td>
<td>SARS-CoV-2</td>
<td>266</td>
<td>6.5%</td>
</tr>
<tr>
<td>AE00001004492C</td>
<td>Campus athletic facility</td>
<td>12.22.21</td>
<td>12.23.21</td>
<td>4,566,986</td>
<td>Respiratory syncytial virus B</td>
<td>182</td>
<td>16.7%</td>
</tr>
<tr>
<td>AE0000100A9F46</td>
<td>Preschool</td>
<td>1.5.22</td>
<td>1.19.22</td>
<td>7,037,819</td>
<td>Human astrovirus</td>
<td>70</td>
<td>10.6%</td>
</tr>
<tr>
<td>AE0000100A8B39</td>
<td>Preschool</td>
<td>1.22.21</td>
<td>2.1.22</td>
<td>4,770,197</td>
<td>Human astrovirus</td>
<td>207</td>
<td>66.3%</td>
</tr>
<tr>
<td>AE0000100A8B3C</td>
<td>Preschool</td>
<td>2.1.22</td>
<td>2.8.22</td>
<td>4,022,365</td>
<td>Influenza C virus</td>
<td>1,826</td>
<td>95.0%</td>
</tr>
<tr>
<td>AE0000100A9532</td>
<td>Preschool</td>
<td>2.23.21</td>
<td>3.1.22</td>
<td>8,252,402</td>
<td>Human Coronavirus NL63</td>
<td>921</td>
<td>19.3%</td>
</tr>
<tr>
<td>AE0000100A9S32</td>
<td>Preschool</td>
<td>4.1.22</td>
<td>4.19.22</td>
<td>3,205,370</td>
<td>Human coronavirus 229E</td>
<td>8</td>
<td>13.2%</td>
</tr>
<tr>
<td>AE0000100A130</td>
<td>Elementary school 2</td>
<td>3.7.22</td>
<td>3.14.22</td>
<td>4,606,042</td>
<td>Human astrovirus</td>
<td>12</td>
<td>46.2%</td>
</tr>
<tr>
<td>AE0000100A638</td>
<td>Elementary school 2</td>
<td>4.4.22</td>
<td>4.18.22</td>
<td>3,945,143</td>
<td>Rotavirus A</td>
<td>50</td>
<td>14.93%</td>
</tr>
<tr>
<td>AE0000100A232</td>
<td>Elementary school 2</td>
<td>4.18.22</td>
<td>5.2.22</td>
<td>3,212,852</td>
<td>Human coronavirus 229E</td>
<td>11</td>
<td>5.8%</td>
</tr>
<tr>
<td>AE0000100A7A40</td>
<td>Elementary school 2</td>
<td>5.2.22</td>
<td>5.9.22</td>
<td>4,958,920</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AE0000100C702C</td>
<td>Elementary school 2</td>
<td>5.9.22</td>
<td>5.16.22</td>
<td>3,788,598</td>
<td>Rotavirus A</td>
<td>31</td>
<td>37.8%</td>
</tr>
<tr>
<td>AE0000100C536</td>
<td>Elementary school 2</td>
<td>5.16.22</td>
<td>5.23.22</td>
<td>3,920,258</td>
<td>Rotavirus A</td>
<td>105</td>
<td>40.7%</td>
</tr>
<tr>
<td>AE0000100C63D</td>
<td>Elementary school 2</td>
<td>5.23.22</td>
<td>5.31.22</td>
<td>4,836,345</td>
<td>Rotavirus A</td>
<td>40</td>
<td>7.2%</td>
</tr>
<tr>
<td>AE000010014930</td>
<td>Elementary school 2</td>
<td>11.10.22</td>
<td>11.14.22</td>
<td>8,999,110</td>
<td>Human coronavirus HKU1</td>
<td>5</td>
<td>2.2%</td>
</tr>
<tr>
<td>AE00001001724</td>
<td>Elementary school 2</td>
<td>11.22.22</td>
<td>12.1.22</td>
<td>4,946,530</td>
<td>Rotavirus A</td>
<td>3</td>
<td>10.2%</td>
</tr>
<tr>
<td>AE000010073172E</td>
<td>Elementary school 3</td>
<td>12.5.22</td>
<td>12.8.22</td>
<td>7560971</td>
<td>Human coronavirus HKU1</td>
<td>113</td>
<td>36.1%</td>
</tr>
<tr>
<td>AE0000006C714A</td>
<td>Middle school</td>
<td>10.31.22</td>
<td>11.3.22</td>
<td>7,421,046</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AE0000006E2245</td>
<td>Middle school</td>
<td>12.5.22</td>
<td>12.8.22</td>
<td>9,380,984</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AE00001006DC33E</td>
<td>High school</td>
<td>11.28.22</td>
<td>12.1.22</td>
<td>6,669,497</td>
<td>Influenza A virus</td>
<td>30</td>
<td>10.7%</td>
</tr>
<tr>
<td>AE0000100D92E</td>
<td>Household</td>
<td>4.22.22</td>
<td>4.26.22</td>
<td>6,810,106</td>
<td>Human rhinovirus</td>
<td>17</td>
<td>17.8%</td>
</tr>
<tr>
<td>NTC_27269_1</td>
<td>No template control</td>
<td>NA</td>
<td>NA</td>
<td>26,833</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>NTC_27269_2</td>
<td>No template control</td>
<td>NA</td>
<td>NA</td>
<td>5,158</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>NTC_28210</td>
<td>No template control</td>
<td>NA</td>
<td>NA</td>
<td>253,439</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Percent genome coverage was calculated by summing the total number of basepairs with any amount of read coverage divided by the total number of basepairs for the given virus target in the RefSeq file. Dates are listed as MM.DD.YY.

Abbreviations: NA, not applicable.
**Table 2.** Sequence independent single primer amplification primers.

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer A</td>
<td>5′-GTT TCC CAC TGG AGG ATA-(N9)-3′</td>
<td>40 pmol/µL</td>
</tr>
<tr>
<td>Primer B</td>
<td>5′-GTT TCC CAC TGG AGG ATA -3′</td>
<td>100 pmol/µL</td>
</tr>
</tbody>
</table>