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

Mitchell D. Ramuta[1‡], Nicholas R. Minor[2‡], Miranda R. Stauss[2], Olivia E. Harwood[1], Savannah F. Brakefield[1], Alexandra Alberts[1], William C. Vuyk[1], Max J. Bobholz[1], Jenna R. Rosinski[1], Sydney Wolf[1], Madelyn Lund[1], Madison Mussa[1], Lucas J. Beversdorf[3], Matthew T. Aliota[4], Shelby L. O'Connor[1,2], David H. O'Connor[1,2]*

[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

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 [1]. 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 [2–5]. Air sampling captures bioaerosols, making it possible to identify aerosolized viruses, bacteria, fungi, or other biological material[6–9]. 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 [10–14]. 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[11]. 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 [15]. 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 [16]. Sampling for longer durations or at higher flow rates can increase biomass

59 collected but might negatively impact virus particle stability and genetic material integrity [17].
60 Furthermore, bacteria and fungi have universal genetic marker regions (16S and internal
61 transcribed spacer (ITS) ribosomal RNA, respectively) that are typically used for studying these
62 microbial communities. Viruses lack a conserved viral gene marker for studying viral
63 communities, which requires unbiased amplification of genetic material [18]. Using metagenomic
64 sequencing techniques compatible with collected air samples could reveal underappreciated
65 viruses that are present in the environment. One especially promising method is sequence-
66 independent single-primer amplification (SISPA) [19]. SISPA sequencing has been demonstrated
67 to detect a wide range of viruses in clinical samples [20].

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

73 Here, we use SISPA to characterize a broad array of human pathogenic viruses from air
74 collected from congregate spaces. We focus on detecting and characterizing RNA viruses due
75 to their public health importance. Understanding human pathogens in built environments' air
76 may help elucidate illness trends in communities over time. This approach could enhance air
77 sampling as a tool for improving public health pathogen surveillance and preparedness against
78 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 [21]. 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 [22,23]. 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,

121 ICV is an understudied respiratory virus. There are a total of 2,475 ICV sequences available in
122 NCBI Virus (taxid:11552) and only 134 ICV sequences have been submitted from the United
123 States in the 21st century. In order to further characterize the ICV isolate from air sample
124 AE0000100A8B3C, we sequenced the sample's remaining SISPA prepared cDNA using the
125 Oxford Nanopore GridION to obtain a greater depth of coverage across the ICV genome.
126 Metagenomic sequencing on the GridION resulted in a greater depth of coverage across all
127 viruses in the sample, including all seven ICV gene segments (supplementary data 1). These
128 reads were used to create consensus sequences for each gene segment. Phylogenetic
129 analysis was performed to compare the ICV detected in the air sample to 45 other ICV viruses
130 that were obtained from GenBank (supplementary data 1). There are six genetic lineages for
131 the HE gene and two lineages for all other gene segments (Figure 2; supplementary figure 1).
132 Phylogenetic analysis was able to identify the viral lineage of each gene segment. HE grouped
133 with the C/Kanagawa/1/76 lineage. PB2, PB1, M, and NS grouped with the C/Yamagata/81
134 lineage. P3 and NP group with C/Mississippi/80 lineage. These data suggest that a genetic
135 reassortment event, the exchange of genome segments between two different strains,
136 occurred previously for this virus because it contains genetic segments from multiple different
137 lineages. Frequent reassortment events have been previously observed for circulating ICV [24].
138 However, this is not a novel reassortant since the sequence clusters closely with influenza C
139 virus C/Scottland/7382/2007 previously identified by Smith et al. (Figure 2; supplementary
140 figure 1) [25]. These data demonstrate that metagenomic sequencing can be used to characterize
141 understudied viruses collected in air samples, which could provide valuable information for
142 improving public health awareness.

143 # Longitudinal detection of human viruses in a preschool

144 Metagenomic analysis of air samples longitudinally collected from congregate settings can
145 provide insight into changes in the prevalence of pathogens over time. These data could
146 provide public health authorities valuable information to improve routine pathogen surveillance
147 programs and outbreak investigations. To track the prevalence of viral genetic material from
148 ICV and other human viruses in a preschool, we analyzed four air samples that were
149 longitudinally collected from January 5, 2021, to March 1, 2022. ICV was first detected in an air
150 sample collected from January 26th to February 1st, 2022. Viral reads in this sample mapped
151 to three out of the seven gene segments including HE, PB2, and NS. Two air samples collected
152 after February 1st, 2022, also contained reads that mapped to several ICV gene segments. ICV
153 genetic material was detected at the highest abundance in the air sample collected from
154 February 1st to the 8th with reads mapping to all seven gene segments (Table 1;
155 supplementary data 1). Viral reads mapping to five of the seven gene segments, including PB2,
156 PB1, P3, HE, and NP, were detected in an air sample collected from February 23rd to March
157 1st. In contrast, human astrovirus genetic material was detected in three air samples collected
158 between January 5th, 2021, to February 8th, 2022. Human astrovirus was not detected in the
159 air sample collected from February 23, 2022, to March 1st, 2022. Similar to ICV, we see a peak
160 and decline in viral reads for human astrovirus. These data suggest that metagenomic
161 sequencing can be used to track the prevalence of multiple respiratory and enteric viruses in a
162 congregate setting over time. This also shows that these are real virus detection signals and
163 not just artifacts of a single sample at a single time point.

164 # Detection of SARS-CoV-2 in RT-PCR-positive air samples

165 To explore whether metagenomic sequencing can detect a human virus that is known to be
166 present in an air sample, we sequenced air samples with known SARS-CoV-2 status. Each
167 AerosolSense cartridge comes with two filter substrates. One filter substrate from each air
168 sample was tested by reverse transcription PCR (RT-PCR) to determine its SARS-CoV-2
169 status. The other substrate was eluted in 500ul of PBS and stored at -80°C until it was
170 processed for sequencing. Several different RT-PCR assays were used on samples included in
171 this study, depending on when and where they were collected, as previously described [21]. Cut-
172 off values used for determining if an air sample was positive, inconclusive, or negative for
173 SARS-CoV-2 are described in the methods section.
174
175 SISPA sequencing was able to detect SARS-CoV-2 reads in two out of 14 (14%) of the SARS-
176 CoV-2 positive samples (Table 1; supplementary data 1). The percent of genome coverage
177 varied between the two samples (6.5% and 46.7%). No SARS-CoV-2 reads were observed in
178 any of the samples that were negative or inconclusive for SARS-CoV-2 by RT-PCR testing or
179 with no template controls (supplementary data 1). An inconclusive result was defined as a
180 sample with only amplification in one of the PCR targets. These data suggest that SISPA
181 sequencing results are concordant with SARS-CoV-2 RT-PCR status. However, SISPA
182 sequencing is not as sensitive as RT-PCR for detecting viral genomic material captured in the
183 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 [11,14,26]. 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 [27,28]. Air samples can also detect microbe-containing bioaerosols from various secondary sources, such as plumbing systems (toilets, sinks, showers), resuspended dust, skin, or clothing [7]. 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 [24,29,30]. The epidemiology of ICV remains poorly understood despite recent studies

208 showing its ubiquity [24,31]. The ICV results from this study highlight the potential of using air
209 sample networks to identify and characterize unexpected and understudied viruses to improve
210 public health awareness. Throughout history, human interest and difficulties associated with
211 detecting viruses, such as ICV, have imposed biases on our knowledge of circulating viruses.
212 Detecting active influenza C virus infections used to be difficult with cell culture techniques
213 because ICV causes weak cytopathic effects. [29] This has likely led to an underestimation of
214 disease burden. Advances in molecular detection technologies, such as pathogen-agnostic
215 sequencing, could free us from these biases and provide a more comprehensive picture of
216 circulating viruses.
217
218 Our study shows that we can use metagenomics from collected air samples to detect
219 pathogens that were otherwise hidden from view. This is not surprising, as the entire field of
220 metagenomics is intended to detect pathogens whose presence was not apparent. However,
221 detection of hidden viruses in the air has not been extensively studied, despite the opportunity
222 to seek out viruses in the environment that could have pathogenic consequences. Once
223 detected by methods described in this paper, molecular detection assays can be created to
224 more sensitively monitor the presence of these viruses within individuals, collected from the air,
225 or located in other environmental samples. Metagenomics of collected air samples provides a
226 window into what other pathogens are present and might be important for public health
227 investigations.
228
229 Lessons learned during the COVID-19 pandemic have sparked interest in using environmental
230 surveillance strategies for improving pandemic preparedness and outbreak response. Unlike
231 wastewater surveillance, active air samplers are mobile, which makes it easy to quickly deploy
232 air sampling networks in settings of interest such as health clinics, airplanes, ports of entry,

233 public transit, farms, K-12 schools, long-term care facilities, or emergency housing facilities [32–
234 34]. These may be specific areas where public health agencies are interested in collecting
235 airborne pathogens for assessing local, regional, or global outbreaks. Highlighting this potential
236 application of air sampling, Mellon et al. recently deployed AerosolSense samplers in an
237 outpatient clinic for patients suspected of mpox infection to look for mpox virus in the air [35]. Air
238 sample results were concordant with the presence of patients diagnosed with mpox infections;
239 future studies could extend this work to sequencing mpox virus from air samples to track
240 evolution longitudinally.

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

248 Improvements in bioaerosol collection and sequencing technologies could significantly
249 improve the ability to characterize viruses. The SARS-CoV-2 RT-PCR and sequencing data
250 from this study shows that our current metagenomic sequencing approach is not optimally
251 sensitive. Air samples contain high amounts of human, animal, and microbial ribosomal RNA
252 (rRNA), likely associated with airborne microbes and host cells transported on dust particles [36].
253 Several studies have shown that rRNA depletion can improve the performance of unbiased
254 sequencing techniques for recovering human RNA viruses from different matrices [37], and
255 should be considered for use with air samples. Alternatively, using capture-based probe
256 hybridization may help enrich viral target sequences in the library preparation process. This

257 could limit the amount of background genomic material from host and other microbes.
258 However, this is a more biased approach using probes specific for viral sequences of the target
259 pathogens [37]. Commercial kits are available containing probes that cover 3,153 different viral
260 genomes targeting ssRNA, dsRNA, dsDNA, and ssDNA viruses. These kits have been used
261 with several different sample types to detect novel and common viruses, including human and
262 animal specimens (nasal swabs and plasma), mosquitoes, and wastewater [38–41]. This approach
263 shows promise for enriching viral reads in samples that contain high amounts of host genetic
264 material. Future studies should evaluate using target enrichment approaches to improve the
265 broad detection of human viruses in air samples.
266
267 Virus-agnostic sequencing methods for air surveillance can be substantially improved, as
268 technologies are evolving rapidly. Sequencing workflows need to be optimized to be high-
269 throughput, cost effective, and have rapid result turnaround for widespread use with air
270 surveillance programs. In this study we ran two Oxford Nanopore sequencing runs on the
271 PromethION 24. The runs multiplexed 16 and 9 samples and had an output of 85 and 67
272 Gbases per flow cell, respectively. The manufacturer estimates a maximal output of 290
273 Gbases per flow cell when using newer sequencing kit chemistries [42]. Improving the
274 sequencing yield to 200 Gbases could allow for multiplexing up to 32 air samples, while
275 maintaining an average of 6 million Gbases per air sample. This could make sequencing more
276 cost effective while maintaining a similar per sample output obtained in this study. Additionally,
277 Oxford Nanopore sequencing enables real-time processing of sequencing data. This could
278 help decrease the turnaround time for results since you do not have to wait the full 72 hour run
279 time before downloading and analyzing the data. This rapid result turnaround time could be
280 beneficial during outbreak response, when real-time data is essential.
281

282 Recent advances in metagenomic sequencing technologies have increased efforts to study
283 microbial communities in built environments. This study demonstrates that metagenomic
284 sequencing approaches can be used to track the prevalence of human respiratory and enteric
285 viruses of public health importance in real-world settings. This approach could provide a more
286 rapid and efficient method for providing community-wide pathogen surveillance data without
287 relying on test-seeking behavior or pathogen-specific detection assays.

288 Methods

289 Air sample collection and processing

290 AerosolSense instruments (Thermo Fisher Scientific) were installed in a variety of indoor
291 congregate settings to collect bioaerosols for pathogen surveillance from December 2021 to
292 December 2023. AerosolSense instruments were placed on flat surfaces 1-1.5 meters off the
293 ground in high-traffic areas of an athletics training facility, preschool, emergency housing
294 facility, brewery taproom, and five K-12 schools in the Upper Midwestern States of Wisconsin
295 and Minnesota. Air samples were collected using AerosolSense cartridges (Thermo Fisher
296 Scientific) according to the manufacturer's instructions. The iOS and Android Askidd mobile
297 app was used to collect air cartridge metadata and upload it to a centralized Labkey database,
298 as previously described in Ramuta et al [21]. After the air samples were removed from the
299 instruments, they were transferred to the lab for further processing. Two air sample substrates
300 were removed from each of the AerosolSense cartridges using sterile forceps to place them in
301 two separate 1.5 mL tubes containing 500 µL of PBS. The tubes were vortexed for 20 seconds,
302 centrifuged for 30 seconds, and stored at -80°C until RNA extraction and complementary DNA
303 (cDNA) preparation.

304

305 Air sample total nucleic acid extraction and concentration

306 Total nucleic acids were extracted from air samples using the Maxwell 48 Viral Total Nucleic
307 Acid Purification Kit (Promega) according to the manufacturer's recommendations. Briefly, 300
308 µL of air sample eluate was added to a 1.5 µL tube containing 300 µL of lysis buffer and 30 µL
309 of Proteinase K. An unused air cartridge was processed with each Maxwell run to be used as a
310 no-template control. The reaction mix was vortexed for 10 seconds and incubated at 56°C for
311 10 minutes. Following the incubation, the tubes were centrifuged for 1 minute. Then 630 µL of
312 the reaction mix was added to the Maxwell 48 cartridges, which were loaded into a Maxwell 48
313 instrument and processed with the Viral Total Nucleic Acid program. Nucleic acids were eluted
314 in a final volume of 50 µL of nuclease-free water. To clean and concentrate the viral RNA, 30 µL
315 of extracted total nucleic acids were treated with TURBO DNase (Thermo Fisher Scientific) and
316 concentrated to 10 µL with the RNA Clean & Concentrator-5 kit (Zymo Research) according to
317 the manufacturer’s protocols.

318 Air sample sequencing

319 A modified sequence-independent single primer amplification (SISPA) approach previously
320 described by Kafetzopoulou et al. was used to generate cDNA from the air samples [20,43]. First,
321 1 µL of Primer A (Table 2) was added to 4 µL of concentrated viral RNA and incubated in a
322 thermocycler at 65°C for 5 minutes, followed by 4°C for 5 minutes. To perform reverse
323 transcription, 5 µL of Superscript™ IV (SSIV) First-Strand Synthesis System (Invitrogen) master
324 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
325 buffer, and 0.5 µL of SSIV RT) was added to the reaction mix and incubated in a thermocycler

326 at 42°C for 10 minutes. To perform second-strand cDNA synthesis, 5 µL of Sequenase Version
327 2.0 DNA polymerase (Thermo Fisher Scientific) master mix (3.85 µL of nuclease-free water, 1
328 µL of 5X Sequenase reaction buffer, and 0.15 µL of Sequence enzyme) was added to the
329 reaction mix and incubated at 37°C for 8 minutes. After the incubation, 0.45 µL of the
330 Sequenase dilution buffer and 0.15 µL of Sequenase were added to the reaction mix and
331 incubated at 37°C for 8 minutes. To amplify the randomly primed cDNA, 5 µL of the cDNA was
332 added to 45 µL of the Primer B reaction mix (5 µL of AccuTaq LA 10x buffer, 2.5 µL of dNTP
333 (10mM), 1 µL of DMSO, 0.5 µL of AccuTaq LA DNA polymerase, 35 µL of nuclease-free water,
334 and 1 µL of Primer B (Table 2)). The following thermocycler conditions were used to amplify the
335 cDNA: 98°C for 30 seconds, 30 cycles (94C for 15 seconds, 50°C for 20 seconds, and 68°C for
336 2 minutes), and 68°C for 10 minutes. The amplified PCR product was purified using a 1:1 ratio
337 of AMPure XP beads (Beckman Coulter) and eluted in 25 µL of nuclease-free water. The
338 purified PCR products were quantified with the Qubit dsDNA high-sensitivity kit (Invitrogen).
339

340 Oxford Nanopore sequencing

341 SISPA-prepared cDNA were submitted to the University of Wisconsin-Madison Biotechnology
342 Center for sequencing on the Oxford Nanopore PromethION. Upon arrival, the PCR product
343 concentrations were confirmed with the Qubit dsDNA high-sensitivity kit (Invitrogen). Libraries
344 were prepared with up to 100-200 fmol of cDNA according to the Oxford Nanopore ligation-
345 based sequencing kit SQK-LSK109 and Native Barcoding kit EXP-NBD196. The quality of the
346 finished libraries was assessed using an Agilent Tapestation (Agilent) and quantified again
347 using the Qubit® dsDNA HS Assay Kit (Invitrogen). Samples were pooled and sequenced with
348 an FLO-PRO002 (R9.4.1) flow cell on the Oxford Nanopore PromethION 24 for 72 hours. Data

349 were basecalled using Oxford Nanopore's Guppy software package (6.4.6) with the high
350 accuracy basecalling model (read filtering parameters: minimum length 200 bp, minimum
351 Qscore=9). Air sample AE0000100A8B3C was also sequenced on the Oxford Nanopore
352 GridION to obtain a greater depth of coverage across all seven influenza C virus gene
353 segments. A sequencing library was prepared for AE0000100A8B3C according to the Oxford
354 Nanopore ligation-based sequencing kit SQK-LSK110 instructions. The sample was
355 sequenced with an FLO-MIN106 (R9.4) flow cell on the Oxford Nanopore GridION for 72 hours.
356 Data were basecalled using Oxford Nanopore's Guppy software package (6.4.6) with the high
357 accuracy basecalling model (read filtering parameters: minimum length 20 bp, minimum
358 Qscore=9).

359 Sequencing analysis

360 Sequencing data generated from air samples were deposited in the Sequence Read Archive
361 (SRA) under bioproject PRJNA950127. The removal of host reads was requested at the time of
362 SRA submission using the Human Read Removal Tool (HRRT). The sequencing data were
363 analyzed using a custom workflow. To ensure reproducibility and portability, we implemented
364 the workflow in NextFlow and containerized all software dependencies with Docker. All
365 workflow code and replication instructions are publicly available at
366 (https://github.com/dholab/pathogen-agnostic-sequence-analysis). Briefly, the workflow starts
367 by automatically pulling the study fastq files from SRA, though it has the option of merging
368 locally stored demultiplexed fastq files as well. Then, reads are filtered to a minimum length
369 (200bp) and quality score (Qscore=9), and adapters and barcodes are trimmed from the ends
370 of the reads, all with the reformat.sh script in bbmap (39.01-0). The filtered fastq files for each
371 air sample are then mapped to contaminant FASTA files containing common contaminants

372 with minimap2 (v2.22). Reads that do not map to the contaminant FASTA files are retained and
373 mapped to their sequencing run's negative control reads to further remove contaminants
374 present from library preparation. The cleaned fastq files for each air sample are then mapped
375 to a RefSeq file containing human viruses downloaded from NCBI Virus using minimap2
376 (v2.22). The human virus reference file contains 835 viral genome sequences and was
377 processed using the bbmask.sh command in bbmap (39.01-0) with default parameters to
378 prevent false-positive mapping to repetitive regions in viral genomes. SAM files for each
379 sample are converted to BAM format, again with reformat.sh. The workflow then completes by
380 generating a pivot table of pathogen "hits", which lists the number of reads supporting each
381 mapped pathogen for each sample. For this study, we then imported the BAM alignments into
382 Geneious Prime (2023.0.4) to inspect the mapping results visually. Genome coverage plots
383 were created for several respiratory and enteric viruses detected in air samples using ggplot2
384 (3.4.1) with a custom R script (4.2.3) in RStudio (2023.03.0+386).

385 Phylogenetic analysis

386 To compare the influenza C virus detected in the preschool air sample AE0000100A8B3C we
387 downloaded 45 influenza C virus genome sequences for each of the seven gene segments
388 from Genbank (HE, PB2, PB1, P3, NP, M, and NS). Accession numbers for each segment can
389 be found in supplementary data 1. Consensus sequences were generated from
390 AE0000100A8B3C with a minimum coverage of 20X. Sections with low coverage were masked
391 with N and trimmed to the reference sequence length. Next, each set of influenza C virus gene
392 segment sequences was aligned using MUSCLE (5.1) implemented in Geneious Prime
393 (2023.0.4) with the PPP algorithm. We then used the Geneious Tree Builder (2023.0.4) to

394 construct a phylogeny for each gene segment using the Neighbor-joining method and Tamura-
395 Nei model with 100 bootstrapped replicates.

396 SARS-CoV-2 RT-PCR

397 Air samples collected between December 2021 and May 2022 were tested for SARS-CoV-2
398 viral RNA using three different SARS-CoV-2 RT-PCR assays depending on their collection
399 location as previously described [21]. Air samples collected after May 2022 were tested for
400 SARS-CoV-2 viral RNA using an RT-PCR protocol as previously described [44]. Briefly, viral RNA
401 was isolated from the air sample substrate using 300 µL of eluate and the Viral Total Nucleic
402 Acid kit for the Maxwell 48 instrument (Promega), following the manufacturer's instructions.
403 RNA was eluted in 50 µL of nuclease-free water. Reverse transcription qPCR was performed
404 using primer and probes from an assay developed by the Centers for Disease Control and
405 Prevention to detect SARS-CoV-2 (N1 and N2 targets). The 20 µL reaction mix contained 5 µL
406 of 4x TaqMan Fast Virus 1-Step Master Mix, 1.5 µL of N1 or N2 primer/probe mix (IDT), 5 µL of
407 sample RNA, and 8.5 µL of nuclease-free water. The RT-PCR amplification was run on a
408 LightCycler 96 at the following conditions: 37°C for 2 minutes, 50°C for 15 minutes, 95°C for 2
409 minutes, 50 cycles of 95C for 3 seconds and 55°C for 30 seconds, and final cool down at 37°C
410 for 30 seconds. The data were analyzed in the LightCycler 96 software 1.1 using absolute
411 quantification analysis. Air samples were called positive when N1 and N2 targets both had
412 cycle threshold (Ct) values <40, inconclusive when only one target had Ct <40, and negative if
413 both targets had Ct >40.

414

References

1. CDC. COVID data tracker. *Centers for Disease Control and Prevention* https://covid.cdc.gov/covid-data-tracker/ (2020).
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).
3. Kirby, A. E. *et al.* Using Wastewater Surveillance Data to Support the COVID-19 Response - United States, 2020-2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 1242–1244 (2021).
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. *MMWR Morb. Mortal. Wkly. Rep.* **71**, 103–105 (2022).
5. McClary-Gutierrez, J. S. *et al.* SARS-CoV-2 Wastewater Surveillance for Public Health Action. *Emerg. Infect. Dis.* **27**, 1–8 (2021).
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 aeroplanes. *J. R. Soc. Interface* **8**, 1176–1184 (2011).
7. Prussin, A. J., 2nd & Marr, L. C. Sources of airborne microorganisms in the built environment. *Microbiome* **3**, 78 (2015).
8. Wang, C. C. *et al.* Airborne transmission of respiratory viruses. *Science* **373**, (2021).
9. Myatt, T. A. *et al.* Detection of airborne rhinovirus and its relation to outdoor air supply in office environments. *Am. J. Respir. Crit. Care Med.* **169**, 1187–1190 (2004).
10. Rosario, K., Fierer, N., Miller, S., Luongo, J. & Breitbart, M. Diversity of DNA and RNA Viruses in Indoor Air As Assessed via Metagenomic Sequencing. *Environ. Sci. Technol.* **52**, 1014–1027 (2018).

439 11. Prussin, A. J., 2nd *et al.* Seasonal dynamics of DNA and RNA viral bioaerosol communities
440 in a daycare center. *Microbiome* **7**, 53 (2019).
441 12. Be, N. A. *et al.* Metagenomic analysis of the airborne environment in urban spaces.
442 *Microb. Ecol.* **69**, 346–355 (2015).
443 13. Hall, R. J. *et al.* Metagenomic detection of viruses in aerosol samples from workers in
444 animal slaughterhouses. *PLoS One* **8**, e72226 (2013).
445 14. Brisebois, E. *et al.* Human viral pathogens are pervasive in wastewater treatment center
446 aerosols. *J. Environ. Sci.* **67**, 45–53 (2018).
447 15. Prussin, A. J., 2nd, Marr, L. C. & Bibby, K. J. Challenges of studying viral aerosol
448 metagenomics and communities in comparison with bacterial and fungal aerosols. *FEMS*
449 *Microbiol. Lett.* **357**, 1–9 (2014).
450 16. Baaijens, J. A. *et al.* Lineage abundance estimation for SARS-CoV-2 in wastewater using
451 transcriptome quantification techniques. *Genome Biol.* **23**, 236 (2022).
452 17. Pan, M., Lednicky, J. A. & Wu, C.-Y. Collection, particle sizing and detection of airborne
453 viruses. *J. Appl. Microbiol.* **127**, 1596–1611 (2019).
454 18. Prussin, A. J., 2nd *et al.* Viruses in the Built Environment (VIBE) meeting report.
455 *Microbiome* **8**, 1 (2020).
456 19. Reyes, G. R. & Kim, J. P. Sequence-independent, single-primer amplification (SISPA) of
457 complex DNA populations. *Mol. Cell. Probes* **5**, 473–481 (1991).
458 20. Kafetzopoulou, L. E. *et al.* Assessment of metagenomic Nanopore and Illumina
459 sequencing for recovering whole genome sequences of chikungunya and dengue viruses
460 directly from clinical samples. *Euro Surveill.* **23**, (2018).
461 21. Ramuta, M. D. *et al.* SARS-CoV-2 and other respiratory pathogens are detected in
462 continuous air samples from congregate settings. *Nat. Commun.* **13**, 4717 (2022).

463 22. Martin, E. T., Fairchok, M. P., Stednick, Z. J., Kuypers, J. & Englund, J. A. Epidemiology of
464 multiple respiratory viruses in childcare attendees. *J. Infect. Dis.* **207**, 982–989 (2013).
465 23. Bradley, R. H. & National Institute of Child Health and Human Development (NICHD) Early
466 Child Care Research Network. Child care and common communicable illnesses in children
467 aged 37 to 54 months. *Arch. Pediatr. Adolesc. Med.* **157**, 196–200 (2003).
468 24. Matsuzaki, Y. *et al.* Genetic Lineage and Reassortment of Influenza C Viruses Circulating
469 between 1947 and 2014. *J. Virol.* **90**, 8251–8265 (2016).
470 25. Smith, D. B., Gaunt, E. R., Digard, P., Templeton, K. & Simmonds, P. Detection of
471 influenza C virus but not influenza D virus in Scottish respiratory samples. *J. Clin. Virol.* **74**,
472 50–53 (2016).
473 26. Dennehy, P. H., Nelson, S. M., Crowley, B. A. & Saracen, C. L. Detection of Rotavirus RNA
474 in Hospital Air Samples by Polymerase Chain Reaction (PCR) • 828. *Pediatr. Res.* **43**, 143–
475 143 (1998).
476 27. Ghosh, S. *et al.* Enteric viruses replicate in salivary glands and infect through saliva.
477 *Nature* **607**, 345–350 (2022).
478 28. Zhuo, R. *et al.* Identification of Enteric Viruses in Oral Swabs from Children with Acute
479 Gastroenteritis. *J. Mol. Diagn.* **20**, 56–62 (2018).
480 29. Sederdahl, B. K. & Williams, J. V. Epidemiology and Clinical Characteristics of Influenza C
481 Virus. *Viruses* **12**, (2020).
482 30. Atkinson, K. V. *et al.* Influenza C in Lancaster, UK, in the winter of 2014-2015. *Sci. Rep.* **7**,
483 46578 (2017).
484 31. Daniels, R. S. *et al.* Molecular Characterization of Influenza C Viruses from Outbreaks in
485 Hong Kong SAR, China. *J. Virol.* **94**, (2020).
486 32. Anderson, B. D. *et al.* Bioaerosol Sampling in Modern Agriculture: A Novel Approach for

487 Emerging Pathogen Surveillance? *J. Infect. Dis.* **214**, 537–545 (2016).

488 33. Bailey, E. S., Choi, J. Y., Zemke, J., Yondon, M. & Gray, G. C. Molecular surveillance of
489 respiratory viruses with bioaerosol sampling in an airport. *Trop Dis Travel Med Vaccines* **4**,
490 11 (2018).

491 34. Coleman, K. K. *et al.* Bioaerosol Sampling for Respiratory Viruses in Singapore's Mass
492 Rapid Transit Network. *Sci. Rep.* **8**, 17476 (2018).

493 35. Mellon, G. *et al.* Air detection of monkeypox virus in a dedicated outpatient clinic room for
494 monkeypox infection diagnosis. *J. Infect.* **86**, 256–308 (2023).

495 36. Rai, S., Singh, D. K. & Kumar, A. Microbial, environmental and anthropogenic factors
496 influencing the indoor microbiome of the built environment. *J. Basic Microbiol.* **61**, 267–
497 292 (2021).

498 37. Fitzpatrick, A. H. *et al.* High Throughput Sequencing for the Detection and
499 Characterization of RNA Viruses. *Front. Microbiol.* **12**, 621719 (2021).

500 38. Terwilliger, A. *et al.* HIV Detection in Wastewater as a Potential Epidemiological
501 Bellwether. (2022) doi:10.2139/ssrn.4257909.

502 39. Kim, K. W. *et al.* Respiratory viral co-infections among SARS-CoV-2 cases confirmed by
503 virome capture sequencing. *Sci. Rep.* **11**, 3934 (2021).

504 40. Wang, Y. *et al.* SARS-CoV-2 Exposure in Norway Rats (Rattus norvegicus) from New York
505 City. *MBio* e0362122 (2023).

506 41. Yamaguchi, J. *et al.* 548. Target Enriched NGS Reveals Wide Breadth of Viruses Causing
507 Acute Undifferentiated Fever in Thailand. *Open Forum Infect Dis* **9**, ofac492.601 (2022).

508 42. *PromethION 2 Solo Technical Specification*.
509 https://community.nanoporetech.com/requirements_documents/promethion-2s-spec.pdf
510 (2022).

511 43. Moreno, G. K. *et al.* Revealing fine-scale spatiotemporal differences in SARS-CoV-2
512 introduction and spread. *Nat. Commun.* **11**, 5558 (2020).
513 44. Newman, C. M. *et al.* Initial Evaluation of a Mobile SARS-CoV-2 RT-LAMP Testing
514 Strategy. *J. Biomol. Tech.* **32**, 137–147 (2021).

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.

535 # Data Availability

536 The air sample sequencing data generated in this study have been deposited in the Sequence
537 Read Archive (SRA) under bioproject PRJNA950127. The accession numbers for influenza C
538 virus samples used in the phylogenetic analysis are provided in Supplementary Data 1.

539 # Code Availability

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

Figures and Tables

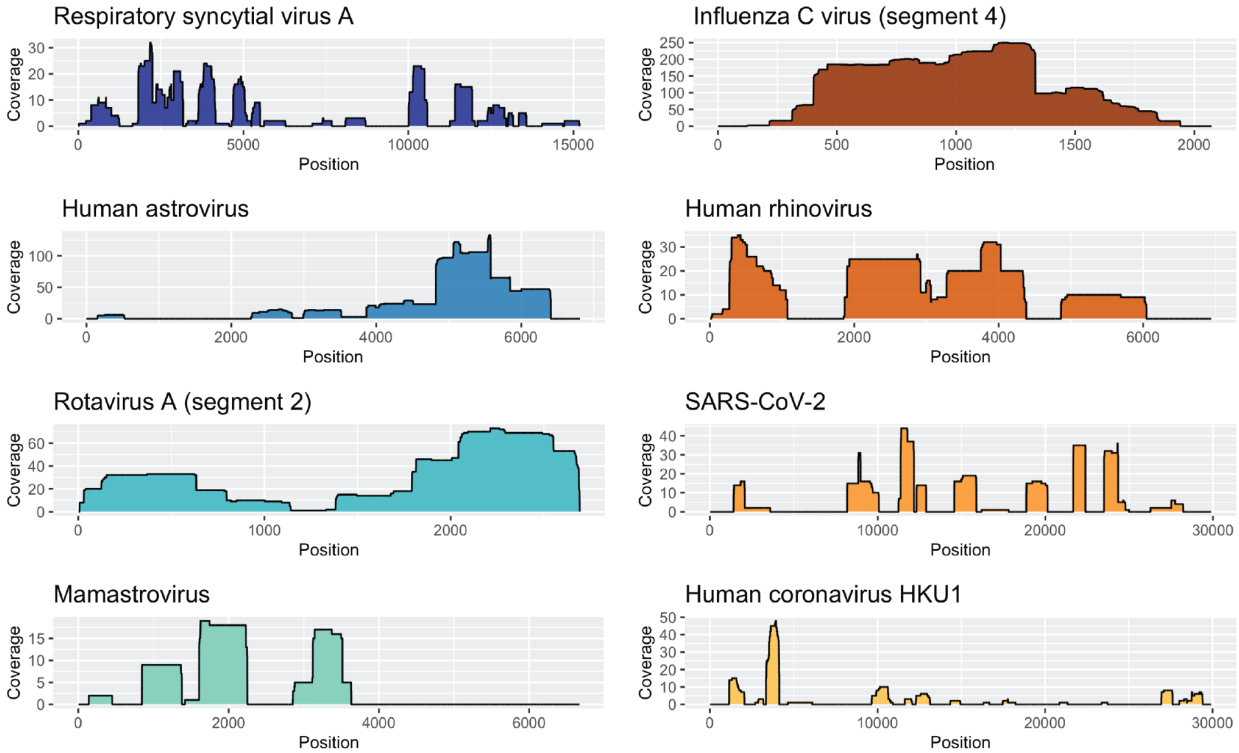

Figure 1. Human respiratory and enteric viruses detected in air samples by SISPA amplification and Oxford Nanopore sequencing. Genome coverage plots showing read depth across eight human RNA viruses detected in air samples. The depth of coverage is shown on the y-axis, and the genome position is shown 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).

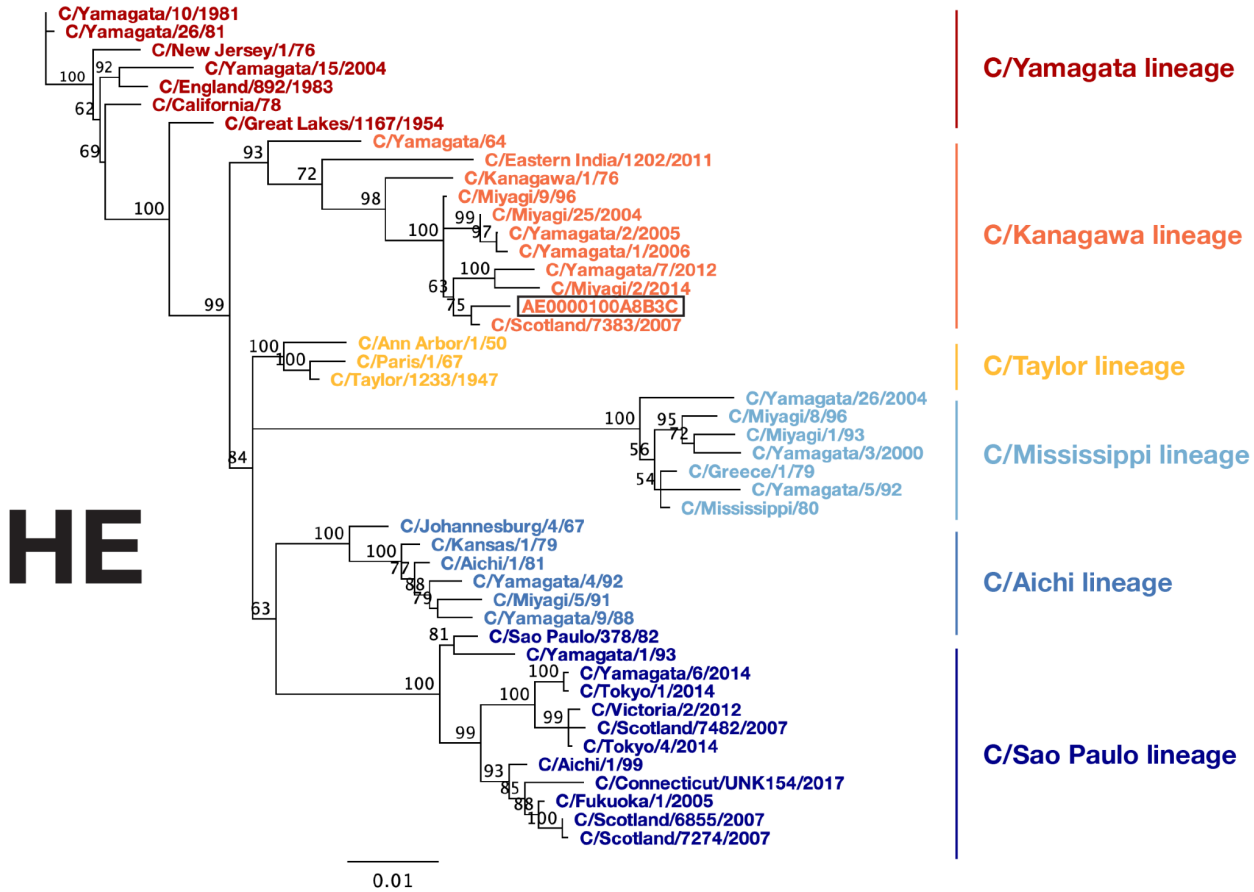

Figure 2. Phylogenetic analysis of influenza C virus hemagglutinin-esterase gene. Phylogenetic trees for influenza C virus hemagglutinin-esterase (HE) gene segment. Nucleotide sequences were aligned using MUSCLE (5.1). The phylogenetic tree was constructed with the Geneious Tree Builder (2023.0.4) using 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/Kanagwa lineage in orange, C/Yamagata lineage in red.

560 **Table 1. Human RNA viruses detected in air samples through deep sequencing with the**
561 **Oxford Nanopore PromethION.**
562

Air sample ID	Location	Start	End	Total number of reads	Pathogen	Number of mapped reads	Percent genome coverage
HC211129	Brewery Taproom	11.22.21	11.29.21	4,113,547	Rotavirus A	276	57.2%
AE000010795B42	Emergency housing shelter	12.21.21	1.7.22	4,898,090	SARS-CoV-2	266	6.5%
AE00001004492C	Campus athletic facility	12.22.21	12.23.21	4,566,986	Respiratory syncytial virus B	182	16.7%
AE0000100A9F46	Preschool	1.5.22	1.19.22	7,037,819	Human astrovirus	70	10.6%
AE0000100A8839	Preschool	1.26.22	2.1.22	4,770,197	Human astrovirus	207	66.3%
					Respiratory syncytial virus A	220	68.9%
					Mamastrovirus	41	26.2%
					Influenza C virus	15	9.2%
AE0000100A8B3C	Preschool	2.1.22	2.8.22	4,022,365	Influenza C virus	1,826	95.0%
					SARS-CoV-2	220	46.7%
					Human rhinovirus	102	68.6%
					Human astrovirus	48	40.2%
					Mamastrovirus	47	36.8%
AE0000100A9532	Preschool	2.23.22	3.1.22	8,252,402	Human coronavirus NL63	921	19.3%
					Influenza C virus	884	38.83%
AE0000100C3430	Elementary school 1	4.11.22	4.19.22	3,205,370	Human coronavirus 229E	8	13.2%
AE0000100CA130	Elementary school 2	3.7.22	3.14.22	4,606,042	Human astrovirus	12	46.2%
					Human Coronavirus NL63	7	11%
					Human coronavirus 229E	4	6%
AE0000100CD638	Elementary school 2	4.4.22	4.18.22	3,945,143	Rotavirus A	50	14.93%
AE0000100CA232	Elementary school 2	4.18.22	5.2.22	3,212,852	Human coronavirus 229E	11	5.8%
AE0000100C7A40	Elementary school 2	5.2.22	5.9.22	4,958,920	NA	NA	NA
AE0000100C702C	Elementary school 2	5.9.22	5.16.22	3,788,598	Rotavirus A	31	37.8%
AE0000100CC536	Elementary school 2	5.16.22	5.23.22	3,920,258	Rotavirus A	105	40.7%
AE0000100CC63D	Elementary school 2	5.23.22	5.31.22	4,836,345	Rotavirus A	40	7.2%
					Simian Agent 10	7	8.9%
AE000010014930	Elementary school 2	11.10.22	11.14.22	8,999,110	Human coronavirus HKU1	5	2.2%
AE000010011724	Elementary school 2	11.22.22	12.1.22	4,946,530	Rotavirus A	3	10.2%
AE00001073172E	Elementary school 3	12.5.22	12.8.22	7560971	Human coronavirus HKU1	113	36.1%
AE00000FEC714A	Middle school	10.31.22	11.3.22	7,421,046	NA	NA	NA
AE00000FEC2245	Middle school	12.5.22	12.8.22	9,380,984	NA	NA	NA
AE0000106DC33E	High school	11.28.22	12.1.22	6,669,497	Influenza A virus	30	10.7%
AE0000100DE93E	Household	4.22.22	4.26.22	6,810,106	Human rhinovirus	17	17.8%
NTC_27269_1	No template control	NA	NA	26,833	NA	NA	NA
NTC2_27269_2	No template control	NA	NA	5,158	NA	NA	NA
NTC_28210	No template control	NA	NA	253,439	NA	NA	NA

563
564 Percent genome coverage was calculated by summing the total number of basepairs with any amount of read coverage
565 divided by the total number of basepairs for the given virus target in the RefSeq file. Dates are listed as MM.DD.YY.
566 Abbreviations: NA, not applicable.

567 **Table 2. Sequence independent single primer amplification primers.**

Primer Name	Sequence	Concentration
Primer A	5′-GTT TCC CAC TGG AGG ATA-(N9)-3′	40 pmol/μL
Primer B	5′-GTT TCC CAC TGG AGG ATA -3’	100 pmol/μL

568