Feasibility of neighborhood and building scale wastewater-based genomic epidemiology for pathogen surveillance

Rachel R. Spurbeck [a,*], Angela Minard-Smith [a], Lindsay Catlin [b]

[a] Health Outcomes and Biotechnology Solutions, Battelle Memorial Institute, 505 King Ave, Columbus, OH 43201, United States of America
[b] National Security Bioscience Center, Battelle Memorial Institute, 505 King Ave, Columbus, OH 43201, United States of America

HIGHLIGHTS

- Neighborhood or building level wastewater analysis accurately detects SARS-CoV-2.
- SARS-CoV-2 was detected in wastewater from one infected person out of 60 residents.
- Total RNAseq detected SARS-CoV-2 and other pathogens in wastewater.
- Targeted whole genome sequencing of wastewater samples identified Spike mutations.

GRAPHICAL ABSTRACT
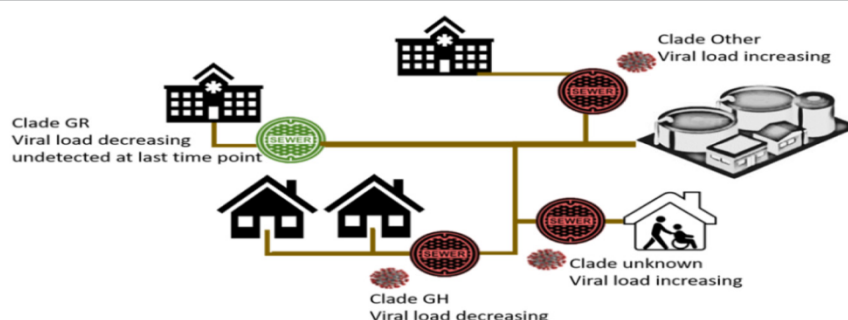
ABSTRACT

The benefits of wastewater-based epidemiology (WBE) for tracking the viral load of SARS-CoV-2, the causative agent of COVID-19, have become apparent since the start of the pandemic. However, most sampling occurs at the wastewater treatment plant influent and therefore monitors the entire catchment, encompassing multiple municipalities, and is conducted using quantitative polymerase chain reaction (qPCR), which only quantifies one target. Sequencing methods provide additional strain information and also can identify other pathogens, broadening the applicability of WBE to beyond the COVID-19 pandemic. Here we demonstrate feasibility of sampling at the neighborhood or building complex level using qPCR, targeted sequencing, and untargeted metatranscriptomics (total RNA sequencing) to provide a refined understanding of the local dynamics of SARS-CoV-2 strains and identify other pathogens circulating in the community. We demonstrate feasibility of tracking SARS-CoV-2 at the neighborhood, hospital, and nursing home level with the ability to detect one COVID-19 positive out of 60 nursing home residents. The viral load obtained was correlative with the number of COVID-19 patients being treated in the hospital. Targeted wastewater-based sequencing over time demonstrated that nonsynonymous mutations fluctuate in the viral population. Clades and shifts in mutation profiles within the community were monitored and could be used to determine if vaccine or diagnostics need to be adapted to ensure continued efficacy. Furthermore, untargeted RNA sequencing identified several other pathogens in the samples. Therefore, untargeted RNA sequencing could be used to identify new outbreaks or emerging pathogens beyond the COVID-19 pandemic.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Wastewater based epidemiology (WBE) is a rapidly growing field for the surveillance of pathogen load within human populations. Since it was discovered that SARS-CoV-2, the causative agent of COVID-19,

* Corresponding author.
 E-mail address: spurbeck@battelle.org (R.R. Spurbeck).

can be quantified from feces (Wu et al., 2020b), quantitative polymerase chain reaction (qPCR)-based wastewater testing has been implemented by several countries and states to track the COVID-19 pandemic by monitoring the viral load in wastewater (Achak et al., 2021; Ahmed et al., 2020; Bivins et al., 2020; Crits-Christoph et al., 2021; Foladori et al., 2020; Gonzalez et al., 2020; Gormley et al., 2020; Hamouda et al., 2020; Haramoto et al., 2020; Hart and Halden, 2020; Kitajima et al., 2020; Medema et al., 2020; Murakami et al., 2020; Weidhaas et al., 2020; Westhaus et al., 2021; Wu et al., 2020a). However, while qPCR provides viral load data, it is not maximizing the information that could be gathered from wastewater. Next generation sequencing techniques can be used to provide full SARS-CoV-2 genomic sequences from wastewater samples to identify the dominant viral strain in circulation, including mutations that could affect diagnostic or therapeutic strategies. With vaccines being rolled out worldwide, it is imperative to monitor newly emerging viral variants which could bear mutations affecting vaccine efficacy (England, 2020; Kemp et al., 2020; Tu et al., 2021). Furthermore, monitoring COVID-19 prevalence in smaller regions than an entire catchment, such as at the level of a neighborhood or building enables localized and swift reactions to trends in COVID-19. For instance, the University of Arizona was able to identify and isolate two infected students via monitoring wastewater at the level of student dormitory buildings (Peiser, 2020). While other universities without a dormitory wastewater monitoring program have had to shut down in person classes periodically due to large COVID-19 outbreaks, the University of Arizona was able to quarantine the two students and continue classes.

More than one pathogen can be monitored at a time when utilizing next generation sequencing techniques such as RNA sequencing which sequences all of the RNA present in a sample. By broadening the net to include all RNA in a sample, genomic WBE can monitor other communicable diseases such as influenza, norovirus, Salmonella, or nosocomial infections. This technique has been commonly utilized in environmental microbiology and virology to study the microbiome or virome of soil, water, and other niches such as the human gut. Recently, RNA sequencing was utilized to determine the impact of wastewater effluent on river water, showing that wastewater impacts the RNA virus diversity in rivers, and identified norovirus (Adriaenssens et al., 2018). By utilizing these metagenomic techniques on wastewater over time, pathogen dynamics in a community will be identified and can be used to predict or monitor for local outbreaks of foodborne illness, influenza, or identify an emerging pathogen. With governments and non-government organizations calling for building a pandemic surveillance network, metagenomic wastewater-based epidemiology would be an ideal method to identify emerging pathogens, especially as pathogens circulating in populations can be monitored using non-invasive sampling techniques (Murakami et al., 2020).

Here we demonstrate that WBE can be utilized at the local level to monitor SARS-CoV-2 viral load, SARS-CoV-2 viral variants circulating in the population, and other pathogens of interest to public health. The sampling locations in this study were chosen specifically to enable comparison of viral load in wastewater to known case load or resident population numbers mapped to the sewer junction points. Two manholes used for sampling were located directly outside of major hospitals known to be treating COVID-19 patients, one was located outside of a nursing home, and a fourth manhole was located downstream of a residential neighborhood within Toledo, OH, USA. Our results show that our methods were able to detect SARS-CoV-2 when 1 out of 60 residents were infected. Multiple mutations were identified, with strains belonging to clades GH and GR. Untargeted RNAseq identified several other communicable diseases fluctuating within the samples over time, including uropathogenic *E. coli*, tetanus, and bacteria known to cause nosocomial pneumonia. Furthermore, two different RNA extraction methods were evaluated in conjunction with qPCR to determine the most robust method for COVID-19, demonstrating that the Qiagen Viral RNA Kit produced more reproducible results than TRIzol™ extraction by qPCR.

2. Materials and methods

2.1. Sampling site selection

Two hospitals known to be treating COVID-19 patients (Hospitals V and P), a Nursing Home (N) that had previously had COVID-19 positive residents, and a Residential Neighborhood Community (C) within Toledo, OH, USA within a zip code with the highest COVID-19 positive case load in June were identified by the Toledo-Lucas County Public Health Department. Manhole sewer access points were identified City of Toledo, Division of Environmental Services to ensure that the effluent sampled was only from the buildings or neighborhood of interest to the study. Positive case numbers for the nursing home were obtained for each week from the Toledo-Lucas County Public Health Department and the management of the nursing home. Positive Case counts were obtained directly from the management at Hospital V, and Hospital P declined to provide this information. No specific case count data was available for the residential neighborhood site.

2.2. Wastewater sample collection

One-liter sample collections were conducted at four specific locations in Toledo, OH, USA from manhole access once a week for three weeks. 24-hour composite samples were collected using an autosampler (ISCO 3700C), with some exceptions due to low flow. In low flow situations, a bucket was used to collect wastewater prior to use of the autosampler to ensure enough volume was present for the composite collection. Table 1 shows which samples were grab or composite samples, and other metadata such as sample collection time, temperature and pH. All samples were collected between 10 AM and 12 PM EST and placed on ice in coolers during transportation to the analytical laboratory. As back up during analysis, 250 mL of each wastewater sample was stored. The following sample metadata was gathered each week during each sample collection: time of collection, method of collection, pH of samples, ambient temperature, and water temperature at time of collection.

2.3. Viral particle concentration

To encourage detachment of virions from organic material in the wastewater, 107 mL of glycine buffer (0.05 M glycine, 3% beef extract, pH 9.6) was added to 750 mL of wastewater and mixed for 1 min (Vlok et al., 2019). The samples were then centrifuged at 8000 $\times g$ for 30 min. The supernatant was filtered through 0.22 micron filters to remove bacteria, fungi, or eukaryotic cells present in the samples. The virions were precipitated from the filtrate using a PEG-8000/NaCl solution ((PEG/NaCl) 320 g/L PEG-8000, 70 g/L (w/v) NaCl) at a ratio of one-part PEG/NaCl to three parts filtrate (Wu et al., 2020a). Mixtures were gently stirred for 12 h at 4 °C. Precipitated virions were collected by centrifugation for 90 min at 13,000 $\times g$ at 4 °C. Pellets were resuspended in 600 μL 1× PBS. The volume changes throughout processing were captured to enable back calculation of total viral load.

2.4. Viral RNA extraction

Two methods were used for RNA extraction. First, RNA was extracted from the collected viral preparations using TRIzol Reagent (ThermoFisher, Cat# 15596026) following the manufacturer's instructions. RNA quantity and quality were measured by NanoDrop. Second, viral RNA was extracted by QIAamp Viral RNA Mini Kit (Qiagen, Cat# 52904) following the manufacturer's protocol.

2.5. qPCR

Each RNA sample was analyzed by reverse transcription quantitative polymerase chain reaction (RT-qPCR) targeting the nucleocapsid gene segment (N1) to determine the SARS-CoV-2 viral load. Briefly, 5 μL of

Table 1
Sample site metadata.

Date	7/14/2020				7/22/2020				7/29/2020			
Site	P	V	N	C	P	V	N	C	P	V	N	C
Sampling method	Composite	Grab	Grab	Grab	Composite	Composite	Composite	Grab	Grab	Composite	Composite	Composite/grab[a]
Sampling start	10:50 AM	N/A	N/A	N/A	10:56 AM	10:08 AM	11:15 AM	N/A	N/A	10:00 AM	10:49 AM	N/A
Sampling end	9:50 AM	10:20 AM	11:10 AM	11:35 AM	9:56 AM	9:08 AM	10:15 AM	11:20 AM	10:50 AM	9:00 AM	9:05 AM	11:46 AM
Ambient temp range	17.8 °C–32.2 °C				20 °C–28.9 °C				21.1 °C–31.1 °C			
Manhole water temp	ND	ND	ND	ND	27.7 °C	24.3 °C	27.3 °C	19.2 °C	22.3 °C	25.1 °C	27.3 °C	20.1 °C
Chilled temp	ND	N/A	N/A	N/A	7.7 °C	8.0 °C	14.2 °C	N/A	N/A	6.2 °C	14 °C	7.1 °C[a]
pH	ND	ND	ND	ND	7.44	8.01	7.73	7.6	8.02	7.83	10.19	7.75

[a] Partial composite obtained. Grab sampling was used to increase the volume to the 1 L required.

4× TaqMan Fast Viral One-Step Master Mix (ThermoFisher, Cat# 4444432) 0.33 µL 60× custom TaqMan Gene Expression Assay primer/probe mix (Forward primer 5′-GAC CCC AAA ATC AGC GAA AT-3′; Reverse primer 5′-TCT GGT TAC TGC CAG TTG AAT CTG-3′; Probe 5′-FAM-ACC CCG CAT TAC GTT TGG TGG ACC-NFQ-MGB-3′) (ThermoFisher, Cat# 4331348), 9.67 µL molecular biology grade water, and 5 µL sample were combined for a 20 µL reaction volume. Each plate contained an 8 standard (SARS-CoV-2 synthetic RNA from Biosynthesis, Inc.) serial dilution (1E7 to 1E0) run in triplicate, no template control in triplicate, and samples were in duplicate. Plates were run according to the following thermal cycling parameters: 50 °C for 5 min, 95 °C for 20 s, and 40 cycles of 95 °C for 3 s followed by 60 °C for 30 s in an ABI 7500 fast quantitative thermocycler. This method is sensitive down to one genome copy per µL.

2.6. RNA virus metagenomics

Total RNA was sequenced from each sample. RNA libraries were prepared using the Rapid RNAseq kit (Swift, Cat# R2096) from Swift Biosciences and sequenced on an Illumina NextSeq using NextSeq 500/550 High Output Kit v2.5 (150 Cycles, 75 × 75 bp). No template control libraries were prepared alongside RNAseq samples, but not sequenced. Data were analyzed following the pipeline outlined in the flowchart depicted in Fig. 1.

2.7. COVID-19 amplicon sequencing

Samples were tested specifically for SARS-CoV-2 by amplicon sequencing using the SNAP SARS-CoV-2 Kit (Swift, Cat# SN-5X296 and COVG1-96) from Swift Biosciences. Briefly, first strand cDNA was synthesized from 11 µL of total RNA from each sample using the Superscript IV First Strand Synthesis System with random hexamers (Invitrogen, Cat# 18091050) following the manufacturer's instructions. The resulting cDNA (10 µL) was used as template for the SARS-CoV-2 Kit following the manufacturer's instructions. Due to adapter dimer issues encountered with SNAP SARS-CoV-2 Kit, the manufacturer's protocol had to be modified to include a second bead clean up prior to the final library amplification step to enable the SARS-CoV-2 sample libraries to be sequencable from wastewater. This extra step reduced the adapter dimers prior to amplification. Prior to sequencing, the samples were quantified by Qubit HS DNA kit (Invitrogen, Cat# Q32851), pooled, and sequenced on an Illumina MiSeq 300 cycle V3 flow cell (Illumina, Cat# MS-102-2002). Samples were batched by week to enable high depth of coverage for single nucleotide polymorphism analysis. The amplicon data were aligned to the NCBI Reference Sequence NC_045512.2 (Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome) and analyzed using the pipeline provided by Swift Biosciences.
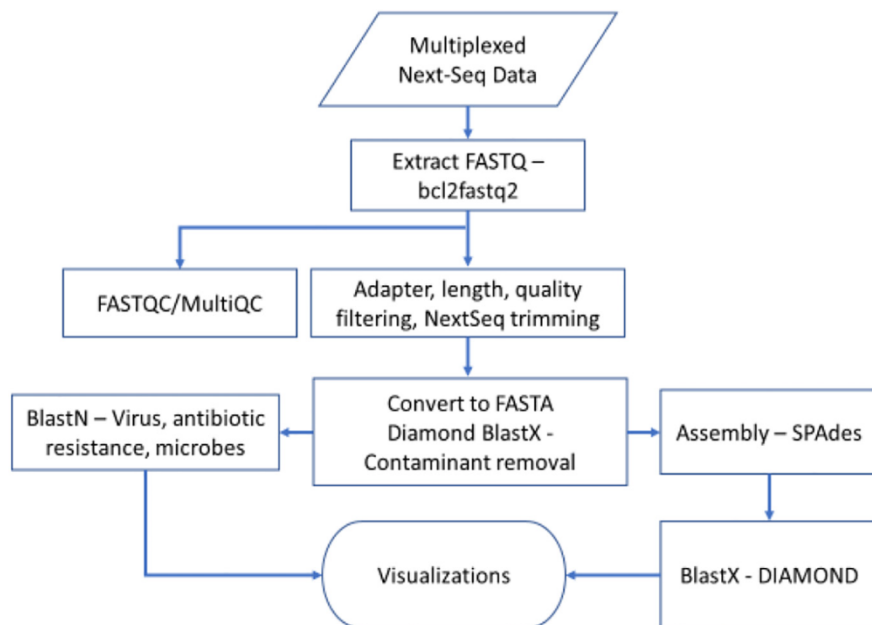

Fig. 1. Processing flowchart for Total RNA-Seq data. RNA-Seq data was obtained from runs on an Illumina NextSeq instrument. Files were demultiplexed, quality checked, and trimmed. Sequences were assembled using SPAdes into contigs and then compared to known reference sequences in the refseq database using BlastN or DIAMOND BlastX to identify pathogens in the wastewater samples.

A positive control and a no template control were processed and quantified. The positive control (SARS-CoV-2 Standard from Exact Diagnostics) was sequenced demonstrating full genome coverage of reference sequence NC_045512.2. Due to high levels of adapter dimer present in the no template controls even after the addition of another clean up step, sequencing the no template control was impossible. The high level of adapter dimers would cause the flow cell to over cluster and abort sequencing mid run. Sequencing was achieved for 9 out of 12 samples by only sequencing the positive control and the samples known to have COVID-19 based on quantitative PCR results. The manufacturer claims ability to obtain full genome coverage from 10 genome copies from clinical samples, however there was no information on ability to sequence from wastewater in July–August of 2020 when these samples were assessed.

3. Results

3.1. Sampling site selection

Unlike most wastewater epidemiology studies which sample at the influent going into the wastewater treatment plants servicing a particular region of a city, our project studied the biological components in raw sewage derived from manholes in four specific locations in the city of Toledo, Ohio. The sites were carefully chosen through discussion with the city wastewater management team and the Toledo-Lucas County Public Health Department to ensure the sites were in locations with a high probability of SARS-CoV-2 circulation. Two sites were the direct effluent from hospitals known to be actively treating COVID-19 patients at the time of the study, Hospital P(P, Fig. 2A), and Hospital V (V, Fig. 2B). A third site was outside a Nursing Home N (N, Fig. 2C) and a fourth was outside a residential neighborhood, Community C (C, Fig. 2D). The type of sampling method, sampling time, the pH, ambient, and water temperatures were collected (Table 1). During the first collection on 7/14/2020, the ambient temperature ranged from a high of 32.2 °C to a low of 17.8 °C.

Table 2
Percent of positive cases compared to the viral load detected in wastewater.

Site	7/14/20		7/22/20		7/28/20	
	Percent positive (%)	Viral load (gc/mL)	Percent positive (%)	Viral load (gc/mL)	Percent positive (%)	Viral load (gc/mL)
C	Unknown	39.1	Unknown	4.5	Unknown	2.9
N	0	ND	0	ND	2	0.6
P	DNP	2.3	DNP	1.0	DNP	ND
V	8	17.4	12	9.0	15	110.6

DNP = Data not provided, ND = not detected.

3.2. Known case load at sampling sites by week and qPCR detection of SARS-CoV-2

The RNA extracted by the TRIzol reagent did not show evidence of any SARS-CoV-2 in the samples; however, the QIAamp Viral/Pathogen RNA preparations produced detectable levels of SARS-CoV-2 by qPCR in 9 of the 12 samples (Table 2). For comparison, the percentage of known COVID-19 cases in Hospital V and Nursing Home N were determined by dividing the number of known COVID-19 positive patients by the total number of inpatients, and a weak positive correlation is observed between the number of infected and viral load in the wastewater. Hospital P did not provide the data to enable the calculation of percent positive cases in that hospital. The percent positive cases in Community C were also unknown, as this residential neighborhood is smaller than the smallest reportable case counts for the region. In Lucas County, Ohio, USA, where these samples are taken, data was reported for community positive cases at the zip code level.

3.3. Non-targeted RNAseq detection of SARS-CoV-2

In theory, RNAseq should afford direct SARS-CoV-2 signature detection in wastewater without requiring enrichment. However, despite following viral concentration protocols, a large amount of the sequence

Fig. 2. Sampling sites within Toledo, Ohio, USA. Orange circles indicate manholes sampled. Brown lines indicate sewer lines. A. Hospital P. B. Hospital V. C. Community C. D. Nursing home N. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
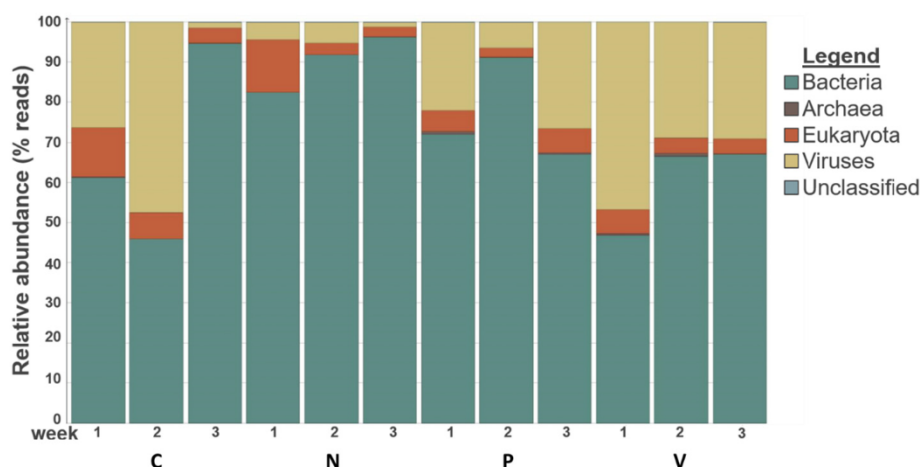
Fig. 3. Relative abundance of viruses, bacteria, archaea, and eukaryotes identified by total RNA sequencing of wastewater. Collection dates are indicated as 1 (7/14/20), 2 (7/21/20), or 3 (7/28/20).

data was found to map to other organisms such as bacteria, bacteriophage, and eukaryotic organisms (Fig. 3). Reads mapping to SARS-CoV-2 were detected in only five out of the twelve samples, with at most 3 reads mapping to the viral genome, despite there being an average of 11 M raw reads per sample and 3 M filtered reads per sample. SARS-CoV-2 was detected in Community C samples collected week 1 (3 reads) and week 2 (3 reads), and Hospital V samples collected in week 1 (1 read), week 2 (2 reads), and week 3 (2 reads). This low coverage is not sufficient for confidence in reporting presence of the pathogen at the local level. Due to the relatively small genome and other larger genomes present in the wastewater, the SARS-CoV-2 reads were likely obscured.

Despite removal of microorganisms and enrichment for viral particles, a large portion of the sequencing data (51–90%) mapped to bacteria, and a fraction of the data (1–19%) mapped to macroorganisms, which demonstrates our viral preparation also included RNA from lysed cells (Fig. 3). Although not the target of this study, the sequenced microbial RNA identified several commensal species that are nonpathogenic to their hosts, as well as infectious and opportunistic pathogens (Fig. 4). While not pertaining to the COVID-19 pandemic, this finding demonstrates that other pathogenic organisms can be identified and tracked using wastewater sequencing, and therefore non-targeted

wastewater sequencing could be continued after the pandemic wanes to improve our understanding of pathogen distribution within local communities and healthcare facilities. Several pathogenic categories of interest were detected in the hospitals, nursing home, and community, including microbes associated with nosocomial infections, gastroenteritis, cystitis, pneumonia, septicemia, and tetanus (Fig. 5) from assembled contigs. The etiological agents identified for nosocomial infections included the *Acinetobacter* species: *A. baumannii A. junii SH205*, and *A. Iwoffii SH145. Campylobacter jejuni* subsp. *jejuni BH-01-0142*, a known cause of gastroenteritis, was identified in the Community C neighborhood, Nursing Home N, and Hospital V. The causative agent of cystitis identified in the first week of testing at Nursing Home N was *Escherichia coli UTI89*. This data suggests that WBE can be utilized to track contagions outside of a pandemic and could be a beneficial tool for future epidemiological and biosurveillance studies.

3.4. Amplicon sequencing

The whole genome of SARS-CoV-2 was enriched from the wastewater samples by targeted amplicon sequencing, enabling single nucleotide polymorphisms (SNP) and insertion/deletion mutations (indels) to be identified when aligned to the reference strain SARS-CoV-2
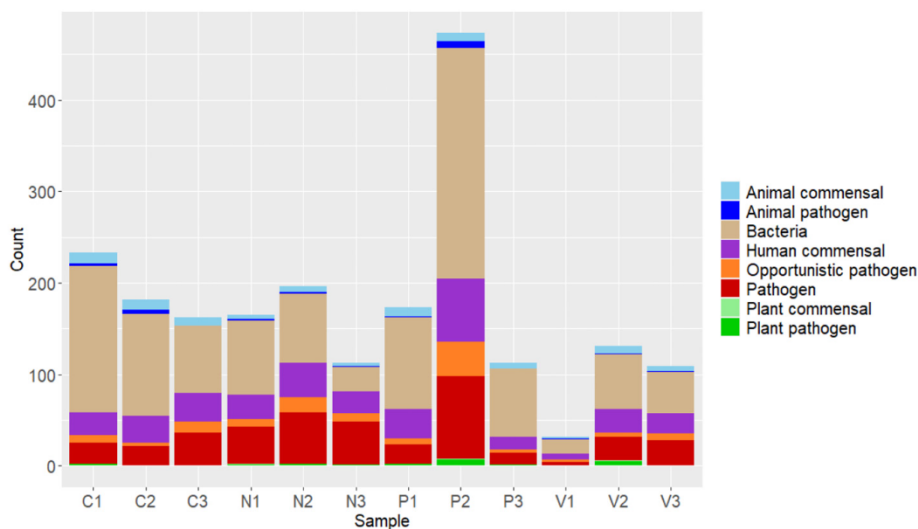

Fig. 4. Ecological niches associated with the microorganisms identified by total RNA sequencing identified several human, animal and pl ant pathogens, as well as potential and opportunistic pathogens. Sample sites are indicated as Nursing home N (N), Hospital V (V), Hospital P (P), and the Community C (C). Collection dates are indicated as 1 (7/14/20), 2 (7/21/20), or 3 (7/28/20).
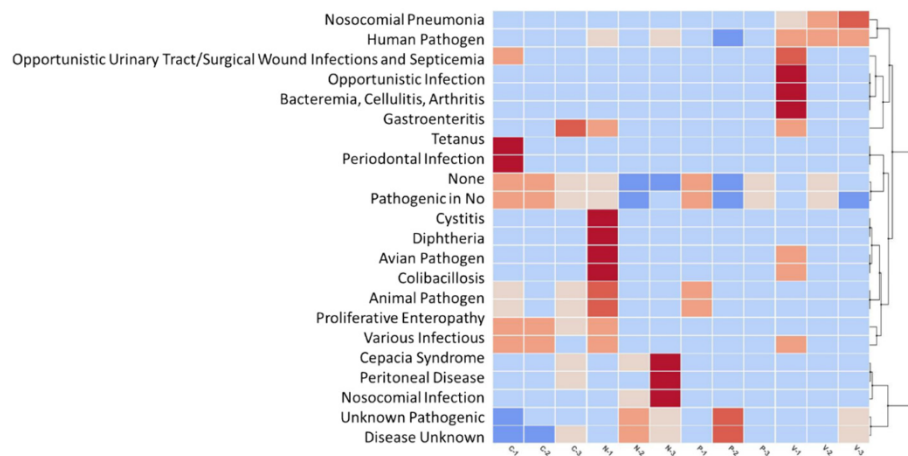
Fig. 5. The abundance of pathogens identified by total RNA sequencing varied by sampling site and over time. Notably, pathogens associated with nosocomial infections were found in the effluent from the Nursing Home (N) and Hospital V (V), but not in Hospital P (P) or the Community C (C). Weeks of collection are indicated as 1 (7/14/20), 2 (7/21/20), or 3 (7/28/20).

Genome Wuhan HU-1. Table 3 presents non-synonymous mutations identified in the Toledo samples that are unique (previously unobserved) or existing mutations (previously observed). Of the twelve samples analyzed in this study, nine were COVID-19 positive based on qPCR. Sequencing the positive samples, only five produced full genomes with enough coverage for confident calls: weeks 1 and 2 from the Community C and weeks 1, 2, and 3 from Hospital V (Table 3). The mutations identified enabled Clade classification. Community C samples fell into Clade GH, which carries the Spike D614G mutation. Hospital V data showed several mutations at variable penetrance indicating multiple strains are present in these samples, therefore the Clade was described

as Other. Despite having multiple strains present, all three weeks had SARS-CoV-2 sequences carrying the Spike D614G mutation.

A week-to-week comparison of Community C samples yielded highly similar existing nonsynonymous mutation loads with ten mutations observed in both weeks. In addition, different unique nonsynonymous mutations were observed, and one existing nonsynonymous mutation was lost from week 1 to week 2. All three weeks from Hospital V had five existing nonsynonymous mutations in common (NSP2_T85I, NSP3_M1788I, NSP12_P323L, NSP16_R216C, and Spike_D614G). However, Hospital V weeks 1 and 2 had an additional five existing nonsynonymous mutations in common (NSP5_L89F, NSP14_N129D,

Table 3
Consensus sequences identified nonsynonymous mutations and categorized the strains by clade.

Sample	Genome coverage (%)	Length (AA)	#Mutations (%)	Unique mutations	Existing mutations	Clade
C-1	97.43%	9688	13 (0.13%)	NSP14_V115I	NSP2_T85I, **NSP3_M1788I,** NSP5_L89F, NSP12_P323L, NSP14_N129D, NSP16_R216C, Spike_D614G, NS3_Q57H, NS3_G172V, NS8_S24L, N_P67S, N_P199L,	GH
C-2	86.09%	8262	13 (0.16%)	NSP13_G170R NSP14_S418C Spike_N370K	NSP2_T85I, NSP5_L89F, NSP14_N129D, NSP16_R216C, Spike_D614G, NS3_Q57H, NS3_G172V, NS8_S24L, N_P67S, N_P199L	GH
C-3	1.88%	61	–	–	–	–
N-3	8.87%	752	1 (0.13%)			–
P-1	30.66%	2459	4 (0.16%)			GR
P-2	26.31%	2164	7 (0.32%)			GH
V-1	92.78%	9642	14 (0.15%)	NSP3_A416V NSP6_K151R	NSP2_T85I, NSP3_M1788I, NSP5_L89F, NSP12_P323L, NSP14_N129D, NSP16_R216C, Spike_D614G, NS3_Q57H, **NS3_G172V,** NS8_S24L, N_P67S, **N_P199L**	Other
V-2	91.66%	8826	15 (0.17%)	NSP3_A1262S NSP3_N1263T NSP14_S418C M_V10I	**NSP1_H110Y,** NSP2_T85I, NSP3_M1788I, NSP5_L89F, NSP12_P323L, NSP14_N129D, NSP16_R216C, Spike_D614G, NS3_Q57H, NS8_S24L, N_P67S	Other
V-3	99.57%	9710	9 (0.09%)	NSP3_F1107L NSP14_N129B	NSP2_T85I, NSP3_M1788I, NSP12_P323L, NSP16_R216C, Spike_D614G, **NS3_G172V,** **M_A2V**	Other

Gray rows indicate samples with partial genome coverage, which may not reliably call the clade. Bold mutations are ones that were not present in all weeks at the same site.

Table 4
SNP frequency at genes targeted for diagnostics or therapeutics and number of indels.

	C		V		
	Week 1	Week 2	Week 1	Week 2	Week 3
orf1ab	0.056%	0.038%	0.042%	0.103%	0.211%
S	0.078%	0.078%	0.026%	0.052%	0.235%
N	0.159%	0.159%	0.317%	0.635%	0.238%
All other	0.170%	0.085%	0.142%	0.227%	0.340%
#indels	0	2	3	9	17

NS3_Q57H, NS8_S24L, and N_P67S), demonstrating that nonsynonymous mutation load has changed over time.

Similarly, the number of SNPs and indels in the whole genome sequencing data changed over time at each site (Table 4). The three genes most targeted by diagnostics were prone to both SNPs and indels, demonstrating the importance of continuous monitoring by sequencing as the pandemic continues. Over the three weeks of this study, the viral population present in the effluent from Hospital V underwent the most striking changes, with the total number of SNPs and indels increasing each week. Importantly, two deletions were found in the S protein in the third week of sampling (7/28/20) at Hospital V. One was a deletion of a full codon, ΔACC23493-23496, which was found in 9% of the data. The other S protein deletion, ΔG24696, was found in 25% of the data.

4. Discussion

The objective of this work was to first demonstrate the utility of wastewater surveillance at the neighborhood or hospital level and then to identify the best methods for tracking SARS-CoV-2 and other pathogens to enable a population-based approach to pandemic biosurveillance. Our data have demonstrated that while total RNAseq, targeted whole genome sequencing, and qPCR methodologies can be used to detect SARS-CoV-2 in wastewater, a combination of qPCR to identify viral load and targeted whole genome sequencing to detect mutations provides deeper insights to support public health decisions during pandemic surveillance than qPCR alone. RNAseq, while identifying other pathogens present in the samples such as microbes that cause nosocomial infections, periodontitis, gastroenteritis, pneumonia, or cystitis, did not recover enough reads mapping to SARS-CoV-2 to detect the viral pathogen with any confidence. Therefore, in a targeted effort to understand the viral variants in circulation during the pandemic, RNAseq would be limited in utility. However, the non-targeted surveillance technique would effectively alert hospitals and long-term care facilities to potential issues with nosocomial pathogens or residents that may have infections that need to be treated and can otherwise go undiagnosed such as cystitis due to asymptomatic or recurrent infections. Use in community settings would alert neighborhoods to potential norovirus or *Salmonella* outbreaks and help public health officials pinpoint restaurants or grocery stores selling contaminated products. Further method development will be necessary to increase the utility of untargeted RNAseq for surveillance of emerging pathogens.

Utilizing amplicon sequencing to target and sequence the full SARS-CoV-2 genome enabled mutation analysis of the viral population identified in the wastewater samples. As demonstrated recently, new strains of SARS-CoV-2 are emerging which are hyper transmissible, and potentially could carry mutations in the Spike protein that could be vaccine or convalescent plasma therapy escape mutants (England, 2020; Kemp et al., 2020; Tu et al., 2021). Also, the Spike, N, or Orf1ab genes are diagnostic targets so any SNP or indel identified in these genes that could affect primer or probe binding of a diagnostic test are of high interest and need to be reported to enable improved diagnostics. Indeed, several mutations have been identified that affect different COVID-19 diagnostic tests (Wang et al., 2020). Furthermore, nonsynonymous mutations in the Spike protein could indicate the potential for vaccine escape

and will need to be characterized to understand impact on vaccine efficacy. Sequencing from wastewater samples within a city or at hospital effluent can aid in surveillance of these mutations and reduces the burden on supplies, as instead of sampling, quantifying and sequencing SARS-CoV-2 in individuals, wastewater enables population sampling, reducing the number of assays and reagents used during surveillance.

With both convalescent plasma and the vaccines targeting the spike protein, it is imperative that nonsynonymous mutations be monitored in this gene. Indeed, it has been demonstrated that spike protein mutations increase in SARS-CoV-2 as a patient is treated with convalescent plasma, indicating the antibodies are placing evolutionary pressure on the virus that selects for mutations in this therapeutic target (Kemp et al., 2020). All Toledo wastewater samples sequenced carried the Spike G614D mutation, which may confer higher transmissibility to the virus, and is characteristic of clade G. Indeed, three of the sequenced samples were categorized as GH and one as GR. In a study in Asia, clades GH and GR were found to contain high levels of SNPs, indicating a lot of diversity in these strains (Sengupta et al., 2020). Mutation analysis of wastewater samples collected in Toledo OH during July 2020 did not find the signatures for the two SARS-CoV-2 clade 20C/G variants that were later identified in Columbus, OH (Tu et al., 2021). Likewise, the strain UK-B.1.1.7 (clade 20I/501Y.V1) (England, 2020; Galloway et al., 2021), also thought to be hyper transmissible and bearing mutations in the S protein, was not present in July 2020 Toledo, OH wastewater samples. However, another spike mutation, Spike_N370K, was observed in week 1 of the Community C, and two deletions (ΔACC23493-23496 and ΔG24696) were observed in week 3 at Hospital V. Therefore, wastewater-based sequencing of SARS-CoV-2 can identify spike protein mutations which may impact transmission and response to the vaccine or convalescent plasma therapy.

Most wastewater-based epidemiology takes place at the influent to a wastewater treatment facility, providing a large catchment that encompasses multiple municipalities. In contrast, our work focused on sampling effluent from a neighborhood, two hospitals, and a nursing home. One limitation of local sampling was flow rate issues (some of our sample locations could not provide 24-hour composites due to low flow); however, sampling at regions within a city provided more refined information on where strains are circulating in the community. For example, strains circulating in Community C were all in the clade GH, while a strain identified in Hospital P belonged to clade GR, and strains at Hospital V were categorized as Other. While all strains bore some mutations in common, each week's sample had mutations unique to that sampling, demonstrating that mutations are common in SARS-CoV-2, and several strains are present in one city.

While population size is a concern, qPCR from effluent of our smallest population, that of Nursing Home N, was able to detect the single case of COVID-19 in week 3, from a population of 60 residents. Since Ohio has implemented weekly testing of nursing home residents, this detection was confirmed by the public health department's data. The case load data from Hospital V demonstrated an increase in hospitalized COVID-19 cases over time, and our qPCR data correlates with this trend, indicating that with more sampling, a predictive model could be developed and validated to identify an approximate number of infected people present in the catchment over time. Using data from the ongoing study, we are currently working to develop and validate candidate models.

Another highly discussed aspect of wastewater-based epidemiology is the degradation of target viral RNA during its transit through the sewer system. While SARS-CoV-2 RNA is intact enough at the inlet to the wastewater treatment plant for qPCR identification, sequencing samples taken closer to the source, such as the effluent from a building or a neighborhood may provide less degraded samples for whole genome sequencing, and therefore provide a better representation of the viral mutations in circulation. Future work will explore the impact of the sewer system on RNA sample quality from different locations within a catchment to optimize viral sequencing.

4.1. Conclusions

This study has demonstrated that wastewater epidemiology at the local level can be used to determine the viral load in a neighborhood or building and identify strain variants in circulation. Total RNAseq from wastewater samples, while effective in identifying bacterial pathogens in circulation, would require more sequencing depth to consistently detect SARS-CoV-2 for biosurveillance purposes, due to the low abundance of the viral RNA in comparison with all other RNA in the sample. Therefore, it is recommended that a combination of qPCR and targeted sequencing of SARS-CoV-2 be utilized to track the pandemic and provide early warning of new strains circulating in the population which could otherwise evade detection by diagnostics or indicate vaccine escape mutants. The use of wastewater in to track mutations is limited by the methods to bioinformatically deconvolute the viral sequences, and therefore in this work we presented the mutations identified without separating the sequences into individual genomes. Future directions will incorporate long read sequencing or linked read sequencing to enable deconvolution into individual viral strains. Furthermore, a comparison of genomic sequences from samples at wastewater treatment plants versus the local manholes utilized in this study would determine if the time between viral shedding in waste and collection impacts viral RNA integrity limiting the use of genomic sequencing to building or neighborhood wastewater samples for high quality full genome sequences.

CRediT authorship contribution statement

Rachel Spurbeck: Conceptualization, Methodology, Wastewater processing, Amplicon sequencing, RNA sequencing, Writing-Original Draft Preparation **Lindsay Catlin**: RNA extraction, Quantitative PCR, RNA sequencing, Amplicon sequencing, Writing- Reviewing and Editing **Angela Minard-Smith:** Bioinformatics, Data Visualization, Writing-Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

Achak, M., Alaoui Bakri, S., Chhiti, Y., M'Hamdi Alaoui, F.E., Barka, N., Boumya, W., 2021. SARS-CoV-2 in hospital wastewater during outbreak of COVID-19: a review on detection, survival and disinfection technologies. Sci. Total Environ. 761, 143192. https://doi.org/10.1016/j.scitotenv.2020.143192.

Adriaenssens, E.M., Farkas, K., Harrison, C., Jones, D.L., Allison, H.E., McCarthy, A.J., 2018. Viromic analysis of wastewater input to a river catchment reveals a diverse assemblage of RNA viruses. mSystems 3.

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J.W., et al., 2020. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community. Sci. Total Environ. 728, 138764.

Bivins, A., North, D., Ahmad, A., Ahmed, W., Alm, E., Been, F., et al., 2020. Wastewater-based epidemiology: global collaborative to maximize contributions in the fight against COVID-19. Environ. Sci. Technol. 54, 7754–7757.

Crits-Christoph, A., Kantor, R.S., Olm, M.R., Whitney, O.N., Al-Shayeb, B., Lou, Y.C., et al., 2021. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. mBio 12, e02703-20.

England PH, 2020. Investigation of Novel SARS-CoV-2 Variant: Variant of Concern 202012/01, Technical Briefing 3. Public Health England, London, United Kingdom.

Foladori, P., Cutrupi, F., Segata, N., Manara, S., Pinto, F., Malpei, F., et al., 2020. SARS-CoV-2 from faeces to wastewater treatment: what do we know? A review. Sci. Total Environ. 743, 140444.

Galloway, S.E., Paul, P., MacCannell, D., et al., 2021. Emergence of SARS-CoV-2 B.1.1.7 lineage—United States, December 29, 2020–January 12, 2021. MMWR Morb. Mortal. Wkly Rep. 70 (3), 95–99.

Gonzalez, R., Curtis, K., Bivins, A., Bibby, K., Weir, M.H., Yetka, K., et al., 2020. COVID-19 surveillance in Southeastern Virginia using wastewater-based epidemiology. Water Res. 186, 116296.

Gormley, M., Aspray, T.J., Kelly, D.A., 2020. COVID-19: mitigating transmission via wastewater plumbing systems. Lancet Glob. Health 8, e643.

Hamouda, M., Mustafa, F., Maraqa, M., Rizvi, T., Aly Hassan, A., 2020. Wastewater surveillance for SARS-CoV-2: lessons learnt from recent studies to define future applications. Sci. Total Environ. 759, 143493.

Haramoto, E., Malla, B., Thakali, O., Kitajima, M., 2020. First environmental surveillance for the presence of SARS-CoV-2 RNA in wastewater and river water in Japan. Sci. Total Environ. 737, 140405.

Hart, O.E., Halden, R.U., 2020. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: feasibility, economy, opportunities and challenges. Sci. Total Environ. 730, 138875.

Kemp, S.A., Collier, D.A., Datir, R., Ferreira, I., Gayed, S., Jahun, A., et al., 2020. Neutralising antibodies in Spike mediated SARS-CoV-2 adaptation. medRxiv.

Kitajima, M., Ahmed, W., Bibby, K., Carducci, A., Gerba, C.P., Hamilton, K.A., et al., 2020. SARS-CoV-2 in wastewater: state of the knowledge and research needs. Sci. Total Environ. 739, 139076.

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., Brouwer, A., 2020. Presence of SARS-Coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands. Environ. Sci. Technol. Lett. 7, 511–516.

Murakami, M., Hata, A., Honda, R., Watanabe, T., 2020. Letter to the editor: wastewater-based epidemiology can overcome representativeness and stigma issues related to COVID-19. Environ. Sci. Technol. 54, 5311.

Peiser, J., 2020. The University of Arizona Says It Caught a Dorm's Covid-19 Outbreak Before It Started. Its Secret Weapon: Poop. Washington Post.

Sengupta, A., Hassan, S.S., Choudhury, P.P., 2020. Clade GR and Clade GH Isolates in Asia Show Highest Amount of SNPs. bioRxiv (2020.11.30.402487).

Tu, H., Avenarius, M.R., Kubatko, L., Hunt, M., Pan, X., Ru, P., et al., 2021. Distinct Patterns of Emergence of SARS-CoV-2 Spike Variants Including N501Y in Clinical Samples in Columbus Ohio. bioRxiv (2021.01.12.426407).

Vlok, M., Lang, A.S., Suttle, C.A., 2019. Marine RNA virus quasispecies are distributed throughout the oceans. mSphere 4, e00157-19.

Wang, R., Hozumi, Y., Yin, C., Wei, G.W., 2020. Mutations on COVID-19 diagnostic targets. Genomics 112, 5204–5213.

Weidhaas, J., Aanderud, Z., Roper, D., VanDerslice, J., Gaddis, E., Ostermiller, J., et al., 2020. Correlation of SARS-CoV-2 RNA in Wastewater With COVID-19 Disease Burden in Sewersheds. Research Square.

Westhaus, S., Weber, F.A., Schiwy, S., Linnemann, V., Brinkmann, M., Widera, M., et al., 2021. Detection of SARS-CoV-2 in raw and treated wastewater in Germany - suitability for COVID-19 surveillance and potential transmission risks. Sci. Total Environ. 751, 141750.

Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W.L., Armas, F., et al., 2020a. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. mSystems 5, e00614-20.

Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., et al., 2020b. Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. Lancet Gastroenterol. Hepatol. 5, 434–435.