



OPEN Metagenomic sequencing identifies potential respiratory pathogens in PCR-negative subset of surveillance samples

Anne Caroline Mascarenhas^{1,3}, Rose S. Kantor^{1,3}, James Thissen¹, Car Reen Kok¹, Monica Borucki¹, Christina Morales², Sharon Messenger², Crystal Jaing¹✉ & Debra A. Wadford²

Respiratory pathogens are a significant source of global morbidity, mortality, and economic burden, with the COVID-19 pandemic driving increased interest in and funding for respiratory disease surveillance. Syndromic panel multiplex nucleic acid amplification tests (NAATs) such as the BioFire Respiratory Panel (RP) are designed to identify the most common etiologic agents of respiratory illness. Untargeted metagenomic sequencing is a powerful tool for pathogen-agnostic detection, enabling the recovery of complete genomes for genomic epidemiology and variant tracking. In this study, we performed untargeted metagenomic sequencing of 305 samples previously negative by BioFire RP and SARS-CoV-2 testing and 26 samples that were previously positive by either of the diagnostic tests. A subset of 78 samples underwent probe-capture enrichment sequencing targeting human viruses. Using these methods, we identified human respiratory viruses in 16 of the 305 previously negative samples (5%). The most common viruses identified were Influenza C virus, Human Bocavirus, Rhinovirus A and C, and SARS-CoV-2. Consensus genomes were recovered for 14 viruses with >90% coverage breadth, revealing closely related Bocavirus strains from neighboring counties and distinct Rhinovirus strains across samples. We also identified 21 samples with a single predominant bacterial or fungal species in the previous negative cohort. These findings underscore the challenges of identifying causal agents from multiplex NAAT-negative cases and highlight the utility of metagenomics for expanding the scope of pathogen surveillance.

Keywords Genomic surveillance, Metagenomics, Pathogen detection, Probe-capture, Respiratory virus

Respiratory pathogens cause significant morbidity, mortality, and economic burden with an estimated 12.8 billion new upper respiratory infections globally in the year 2021¹. The COVID-19 pandemic led to increased interest in and funding for the surveillance of respiratory diseases. This surveillance has primarily been conducted via qPCR and PCR-based panels for clinical specimens, with a focus on SARS-CoV-2, influenza A virus, and respiratory syncytial virus (RSV)^{2–4}. Wastewater-based surveillance has also provided a partial view of circulating pathogens, although some respiratory pathogens may not be shed into wastewater at detectable concentrations^{3,5–7}.

While clinical surveillance increased during and after the pandemic, the range of targets has generally been limited to commonly occurring pathogens whose genomes are conserved enough to be detectable by established PCR methods. Hence, rarer or more divergent pathogens may have been missed. Separately, use of pathogen-agnostic untargeted metagenomic and metatranscriptomic sequencing has increased for characterization of respiratory infections in clinical samples^{8–10}. Recently, several groups have developed validated clinical metagenomic workflows for respiratory viruses^{11,12}. Additionally, commercial probe-capture sequencing panels for broad-range viral enrichment, including respiratory viruses, have recently been shown to increase the sensitivity of sequencing^{13,14}. Clinical metagenomic sequencing may even produce complete pathogen genomes, which, as demonstrated for SARS-CoV-2, can be useful for variant tracking, for genomic epidemiology and for

¹Physical and Life Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA, USA. ²Viral and Rickettsial Disease Laboratory, Center for Laboratory Sciences, California Department of Public Health, Richmond, CA, USA. ³These authors contributed equally: Anne Caroline Mascarenhas and Rose S. Kantor. ✉email: jaing2@llnl.gov

predicting vaccine and treatment efficacy¹⁵. Yet despite its potential utility, metagenomics has not been widely applied in surveillance beyond SARS-CoV-2 due to its relatively high cost.

During the COVID-19 pandemic, the California SARS-CoV-2 and Respiratory Virus Surveillance System (CalSRVSS) was established by the California Department of Public Health (CDPH) to provide outpatient testing of symptomatic individuals in 10 counties across California¹⁶. From 2020 to 2023, this program collected 15,325 samples from symptomatic individuals, which were tested for SARS-CoV-2 and 22 additional pathogens included on the BioFire Respiratory Pathogens panel. Of these, 55% were negative for all 22 BioFire pathogen analytes (unpublished data). These BioFire- and SARS-CoV-2 negative samples provided a unique opportunity to examine whether less commonly recognized pathogens may have been associated with the respiratory symptoms observed from these cases. We randomly selected 305 previously negative samples and 26 previously positive samples from the CalSRVSS sample set for untargeted sequencing and analyzed a subset of 78 samples using semi-targeted metagenomics. Here, we report on the frequency of potential pathogens detected.

Methods

Sample collection

De-identified residual nasopharyngeal and other swab eluate samples ($n = 305$, identifiers “N#”) from adult and pediatric patients with upper respiratory tract infection symptoms were obtained from the California Department of Public Health (CDPH) as part of the California SARS-CoV-2 and Respiratory Virus Surveillance System (CalSRVSS) initiative (Table S1)¹⁶. These samples were collected during outpatient visits between July 2020 to January 2022 in 10 counties in California and were negative for SARS-CoV-2 and for 18 viruses and 4 bacteria included in the BioFire Respiratory Panel 2.0 or 2.1 (<https://www.biofire.com>), which covers pathogens commonly associated with upper respiratory tract infections¹⁷. A set of known positive swabs including SARS-CoV-2 positive samples from individuals previously infected with SARS-CoV-2 (“reinfections”, $n = 7$) and BioFire positive samples ($n = 19$) were included for comparison (identifiers “P#”; Table S1). Samples were received in PrimeStore Molecular Transport Medium (ThermoFisher Scientific, Waltham, MA). The CalSRVSS project was reviewed by the State of California Health and Human Services Agency Committee for the Protection of Human Subjects and determined to be considered Not Research or Exempt (Project # 2021 – 168).

All methods were carried out in accordance with California Department of Public Health and Lawrence Livermore National Laboratory (LLNL) institutional biosafety protocols. All experimental protocols were approved by California Protection of Human Subjects and LLNL's Institutional Research Board and deemed exempt for human subject research. Informed consent was not required because no research on human subjects was conducted, as confirmed by California Protection of Human Subjects Committee. Only residual respiratory specimens, not human tissues, were used in this study.

Total nucleic acid extraction

Total nucleic acid (DNA and RNA) was extracted from 500 μ L of each eluate sample as follows. First, cells were lysed in 100 μ L ATL Buffer (Qiagen, Redwood City, CA) with bead-beating for 40 s at 6 m/s with FastPrep-24 5G equipment using lysing matrix F tubes (MP Biomedicals, Burlingame, CA). Total nucleic acid was extracted from 200 μ L of the resulting lysate using the IndiSpin Pathogen Kit (Indical Bioscience, Leipzig, Germany) following the manufacturer's protocol with final elution in 55 μ L AVE Buffer. DNA and RNA were quantified separately using the Qubit High Sensitivity dsDNA and RNA kits, respectively, on a Qubit 3.0 fluorometer (Invitrogen, Carlsbad, CA) using the recommended instructions.

cDNA synthesis and whole genome amplification

The first strand cDNA synthesis reaction contained 11 μ L of sample, 1 μ L of random hexamer primers (50mM; Integrated DNA Technologies, Coralville, IA) and 1 μ L of dNTP mix (10mM; Thermo Fisher Scientific, Waltham, MA) and was incubated at 65 °C for 5 min, and 5 min on ice-water slurry. Synthesis was completed by addition of the remaining reaction components (4 μ L of 5x First Strand Synthesis Buffer, 1 μ L of 0.1 M DTT, 1 μ L of RNaseOUT, 1 μ L of SuperScript IV reverse transcriptase and 1 μ L of DEPC water (Thermo Fisher Scientific, Waltham, MA) to the cooled samples followed by incubation in a thermocycler (protocol: 23 °C for 10 min, 55 °C for 60 min, and 80 °C for 10 min). A total of 10 μ L from each DNA/cDNA sample was amplified using the Whole Transcriptome Amplification kit (Qiagen, Redwood City, CA) following the manufacturer's guidelines. Amplified samples were purified with AMPure XP magnetic beads (Beckman Coulter, Brea, CA) using a slightly modified version of the recommended protocol: 1.8x AMPure XP magnetic beads, 3x Ethanol 70% washes, and elution in 60 μ L of Ultra-Pure Water (UPW). Purified samples were used in downstream library preparation protocols for sequencing.

Untargeted metagenomic sequencing

The untargeted metagenomics DNA libraries were prepared from 200 to 300 ng of amplified nucleic acid with the Illumina DNA Prep kit (Illumina, San Diego, CA) following the standard manufacturer's protocol but using the Illumina UD 10-bases Nextera Indexes from IDT with 5 PCR cycles and 30 μ L of RSB for final library elution. Final libraries were quantified, and quality was assessed with the High Sensitivity D5000 assay on a TapeStation 4200 instrument (Agilent Technologies, Santa Clara, CA). Libraries were pooled in equimolar quantities for multiplexing and sequenced in 300 cycles, with 2 $\times$ 151 bp inserts and 10 bp indexes in paired-end mode on an Illumina NextSeq2000 using its on-board denaturing function.

Twist CVRP library preparation and sequencing

Samples that were selected for enrichment with the Twist comprehensive viral research panel (Twist Bioscience, South San Francisco, CA) were prepared as follows. First, DNA libraries from the whole genome amplified

samples were prepared with the Twist Total Nucleic Acids Library Preparation EF Kit 2.0 for Viral Pathogen Detection and Characterization starting from Step 3 “Perform DNA fragmentation, end repair and dA-tailing” of the protocol. The whole-genome amplified samples (1–2 ng in 25 µL of UPW) were fragmented, repaired and dA-tailed under the following thermal cycler conditions: (1) hold at 4 °C, (2) incubation at 37 °C for 20 min to generate 180–220 bp inserts, (3) 65 °C for 30 min and (4) hold at 4 °C. Then, the Twist universal adapters were ligated to the dA-tailed DNA fragments, and libraries were purified with magnetic beads, as per the instructions. The adapter-ligated libraries were amplified with the Twist UDI Primers (10 bp in length) using 10 cycles for the Polymerase Reaction (PCR), purified using the manufacturer’s instructions but eluted in 25 µL of UPW. The resulting libraries were quantified with the High Sensitivity dsDNA Kit on a Qubit 2.0 fluorometer.

Libraries were enriched with the Twist Comprehensive Viral Research Panel (CVRP) following the Twist Target Enrichment Standard Hybridization v2 Protocol. Briefly, libraries were pooled in equal quantities (187.5 ng each) in groups of 8 samples for a multiplexed 16 h hybridization with the probes in the enrichment panel. The hybridized targets were captured with streptavidin beads, amplified using a panel size of 10–50 Mb using 8 PCR cycles for multiplexed samples, followed by a purification step with DNA Purification Beads and eluted in 35 µL of UPW. The viral enriched libraries were quantified with a High Sensitivity dsDNA Kit on a Qubit 2.0, and quality was measured using a High Sensitivity D5000 assay on a TapeStation 4200 instrument. The Twist library pools were multiplexed in equal quantities and sequenced with 300 cycles in paired-end mode on an Illumina NextSeq2000 using its on-board denaturing function.

Dataset pre-processing

Illumina paired-end metagenomic sequencing reads were demultiplexed and converted from bcl to fastq file format with bcl2fastq v2.20.0.422 (Illumina, San Diego, CA) using default parameters. Reads were filtered and trimmed with Fastp v0.23.4^{18,19} using the command line switches ‘-y -g --poly_g_min_len = 5 -x --poly_x_min_len = 5’. Human genomic content was removed from the metagenomic datasets by mapping the sequencing reads to the reference human genome (GRCh38.p14; RefSeq Assembly: GCF_000001405.40) using minimap2 v2.26-r1175²⁰ configured for short-read alignment with ‘-x sr’. The human-depleted sam files were converted to bam with Samtools v1.13^{21,22} using the view algorithm and the command line options ‘-bSh -f 4’. Decontaminated bam files were sorted and converted to fastq with samtools.

All of the paired-end sequencing reads for untargeted and Twist-enriched metagenomic datasets were de-identified as recommended by the National Center of Biotechnology Information (NCBI) prior to deposition in the Sequence Read Archive (SRA)²³. Briefly, Illumina paired-end reads were interleaved with SeqFu v1.22.3²⁴ using the built-in interleave function and careful mode enabled (-c). Interleaved datasets were de-identified using the NCBI’s Human Read Removal Tool (HRRT) v2.2.1 tool (SRA Human Scrubber; <https://github.com/ncbi/sra-human-scrubber>)²⁵ with the ‘-x’ command line switch. The ‘scrubbed’ reads were de-interleaved with SeqFu using its deinterleave function and careful mode enabled (-c).

Taxonomic classification and analysis

The filtered metagenomic reads were taxonomically classified using a Lowest Common Ancestor (LCA) approach by Centrifuge v1.0.4²⁶ through read alignment against the NCBI nucleotide (NT) database²⁷ curated in January 2023²⁸. Centrifuge was configured with a minimum hit length of 15 and best-hit reporting using ‘--min-hitlen 15’ and ‘-k 1’. Results were refined with Recentrifuge v1.12.0²⁹ using a minimum hit length of 40. Recentrifuge read count data was imported into R v4.2.0 and Python v3.13.0 (pandas v2.2.3) for downstream microbiome (bacterial, fungal, and viral) analyses. Within R, taxonomic compositional analysis was carried out using the phyloseq package v1.48.0³⁰. Species-level hits were cross-referenced with a curated list of human infectious pathogens compiled from the Swiss Institute of Bioinformatics ViralZone database (De Castro et al., 2024), the NCBI Pathogen Detection database (Information, 2016) and two previously published compilations of bacterial (Bartlett et al., 2022) and fungal (Bartoszewicz et al., 2022) pathogens. Excluded from consideration were non-respiratory viruses including polyomaviruses, papillomaviruses, herpesviruses, erythroviruses, Molluscum contagiosum virus, GB virus C, rotaviruses, Torque teno virus, and picobirnaviruses (Supplementary Figure S2). We also removed pervasive low read-count hits to human immunodeficiency virus (HIV) caused by human-contaminated viral genomes in the database. Viral detection in previously negative and positive samples used a threshold of ≥ 10 reads per virus based on Recentrifuge. Two no-template control libraries were also sequenced and found to contain no reads hitting to any human pathogenic virus. Heatmaps were generated showing species-level read counts (with the exception of SARS-CoV-2, which was “no rank” within the taxonomy database used), using plotnine v0.13.6.

Viral genome analysis

For samples found by Recentrifuge to contain ≥ 10 reads hitting to Bocavirus, Rhinovirus A, Rhinovirus C, Influenza C virus, or SARS-CoV-2, additional analysis was conducted. First, all complete genomes classified as Bocaparvovirus (taxID 1507401), Rhinovirus A (taxID 147711), and Rhinovirus C (taxID 463676) were downloaded from Genbank using the NCBI Virus portal (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>; accessed April 7, 2025). For each of these groups, Genbank genomes were clustered at 95% identity and 75% coverage using mmseqs2³¹ (version 18d8ddc2f468b8a00cec9addb3d2328eb790ac9e). Reads from each sample containing each virus were mapped to cluster representatives, using minimap2 (v2.28-r1209)²⁰ without secondary hits. Mapping results were filtered to remove supplementary hits (samtools view -F 0 × 800), sorted, indexed, and used as input for coverage calculations by samtools (v1.21). The genome with highest coverage breadth for each virus in each sample was used for a final mapping step, which provided coverage information for untargeted and targeted sequencing (Table 1). Masked consensus genomes were extracted using samtools (v1.21) with default settings (requiring a frequency of 0.75 to make a base call) from the final mapping results using targeted

Sample information			Best BLAST hit				Untargeted cov.				Targeted cov.			
Sample ID	County	Date Collected	Best-hit strain in NCBI	Accession	Collection information	% ID*	Breadth	Depth	Rel. abund.	Breadth	Depth	Rel. abund.		
N145 ^a	Contra Costa	7/7/21	Human bocavirus 1, strain unidentified due to low coverage	-	-	-	19	38	0	33	29,141	86.4		
N149 ^a	Contra Costa	7/19/21	Human bocavirus 1 isolate x005-boca-1	OL519570.1	China, 2018	99.58	99	34	0.1	100	99,295	94.6		
N248 ^a	Marin	9/14/22	Human bocavirus 1 isolate x005-boca-1	OL519570.1	China, 2018	99.54	100	126	2.8	100	124,814	99.3		
N278 ^a	Marin	12/5/22	Human bocavirus 1 isolate x005-boca-1	OL519570.1	China, 2018	99.54	98	61	1.2	100	278,654	98.8		
N212	Marin	7/11/22	Rhinovirus C26 strain RvC26(USA/2021/SVTSZ)	MZ363457.1	USA, 2021	98.39	95	97	0.3	100	79,152	82.5		
N261	Marin	10/4/22	Rhinovirus A46 isolate RV-A46(USA/269,075/2022	PP194011.1	USA, 2022	98.93	80	3	0	77	218	32.3		
N325	Marin	5/16/22	Rhinovirus A58 strain RvA58(USA/2021/B6GY49	OK017922.1	USA, 2021	99.07	100	43	0.7	100	47,948	86.3		
N325	Marin	5/16/22	Rhinovirus C59 isolate RV-C59(USA/070666/2023	PP187409.1	USA, 2023	99.05	95	6	0.1	79	1293	2.3		
P333	Imperial	10/11/21	Rhinovirus A49 strain RvA49(USA/2021/Z392UL	OK017969.1	USA, 2021	99.49	100	18	0.7	100	14,758	96.6		
P1008	Alameda	11/13/23	Rhinovirus C15/human/USA/MA-Broad_BWH-19,937/2023	PV178289.1	USA, 2023	98.97	100	40	0.5	-	-	-		
P1009	Contra Costa	11/6/23	Rhinovirus C32 strain USA/CA/RGDS-2016-1008	MK520815.1	USA, 2016	95.51	92	7	0	93	153	33.5		
P1012	Placer	11/6/23	Rhinovirus A isolate S0937XLQ	MW587089.1	China, 2016/17	97.56	96	8	0.1	91	383	21.7		
N212 ^b	Marin	7/11/22	Influenza C virus, strain unidentified due to low coverage	-	-	98.97	43	6	0	79	6838	7.8		
N293 ^b	Imperial	2/10/22	Influenza C virus (C/Minnesota/28/2015)**	CY239471.1	USA, 2015	98.38	31	2	0	99	1391	68.1		
N297 ^c	Contra Costa	2/16/22	Influenza C virus (C/Minnesota/28/2015)**	CY239471.1	USA, 2015	98.28	29	2	0	100	817	49.5		
N311	Imperial	4/5/22	Influenza C virus (C/Minnesota/28/2015)**	CY239471.1	USA, 2015	97.84	99	54	2.2	100	48,700	81.1		
N314	Marin	4/13/22	Influenza C virus, strain unidentified due to low coverage	-	-	-	3	2	0	31	207	13		
N324 ^c	Marin	5/10/22	Influenza C virus (C/Minnesota/28/2015)**	CY239471.1	USA, 2015	98.38	97	36	0.2	100	20,691	89.9		
N055	-	12/8/20	SARS-CoV-2 Epsilon; B.1.429; clade 21 C	MZ636205.1	USA, 2021	99.97	100	178	15.3	100	82,207	99.7		
N198	Contra Costa	5/23/22	SARS-CoV-2 Omicron; BA.2.12.1; 22 C	ON851644.1	USA, 2022	99.93	100	163	7.5	100	144,981	99.2		
N323	Imperial	5/3/22	SARS-CoV-2, lineage unidentified due to low coverage	-	-	-	14	18	0.1	34	11,237	51.3		
P348	Contra Costa	1/24/22	SARS-CoV-2 Omicron; BA.1; clade 21 K	OX690065.1	Multiple, 2022	99.9	98	30	1.4	-	-	-		
P349	Contra Costa	1/31/22	SARS-CoV-2 Omicron; BA.1.1; clade 21 K	ON619211.1	USA, 2022	99.86	98	24	2.3	-	-	-		

Table 1. Identification, coverage, and relative abundance of viral consensus genomes recovered. *%Identity of alignment between the best BLAST hit in NCBI core-nt (accessed April 14, 2025) and the unmasked portion of the consensus, where coverage breadth for the consensus was ≥80% for untargeted or targeted sequencing. **For influenza C virus, best BLAST hit was reported for segment 4 (hemagglutinin esterase fusion gene). ^a, ^b, ^cGroups of sequences with >99% nucleotide identity.

sequencing data. Best BLAST hits were identified using BLASTn against the NCBI core-nt database (accessed April 14, 2025). For influenza C virus and SARS-CoV-2, the RefSeq entries (strains C/Ann Arbor/1/50 and Wuhan Hu-1, respectively) were used as the reference genomes in a single mapping step followed by coverage calculations and extraction of the consensus. Mapping achieved near 100% coverage breadth for most samples, suggesting these reference genomes were acceptable starting points. For SARS-CoV-2, variant lineages were identified from consensus sequences using the Nextclade web interface (<https://clades.nextstrain.org/> accessed April 28, 2025).

Results

We randomly selected 305 BioFire-negative samples for sequencing to detect less common respiratory pathogens not included as part of the BioFire Respiratory Pathogens panel. Samples were collected from individuals with respiratory symptoms as part of CalSRVSS across five counties in California between July 2020 and December 2022¹⁶. In parallel, a set of known viral positive samples and SARS-CoV-2 reinfection samples was selected based on positive results for one or more respiratory pathogens (“previous positives”, $n=26$). Untargeted Illumina sequencing produced an average of 10.8 million reads per sample, and probe-capture enrichment (Twist CVRP) on a subset of 78 samples (56 previous negative, 22 previous positives) produced an average of 4.9 million reads per sample (Supplementary Table S1). In nearly all samples, a majority of untargeted sequencing reads were derived from the human host, with a few exceptions in which bacteria dominated (Supplementary Figure S1). After host read removal, untargeted sequence data were dominated by bacterial sequences and by sequences that did not match to the database with a stringent minimum hit length of 40 nucleotides (at a minimum hit length of 15 nucleotides, just 6% of untargeted sequences were unclassified). Meanwhile, targeted sequencing via probe-capture enrichment using the Twist CVRP yielded a relatively high fraction of viral reads for a subset of previous negative and most previous positive samples (Fig. 1A).

Viral pathogen detection

We first employed Centrifuge and Recentrifuge as a highly sensitive method for detecting viral reads, cross-referencing hits against a list of human pathogenic viruses. In previous positive samples, expected viruses were largely confirmed by both untargeted and probe-capture sequencing, with some exceptions (Fig. 2A). In particular, of 7 SARS-CoV-2 previous positive samples, just two were successfully sequenced with untargeted sequencing. The remaining 5 samples underwent targeted sequencing to look for additional infections. One of these had a known co-infection with human coronavirus 229E, which was confirmed by sequencing while SARS-CoV-2 was not detected. Rhinovirus was detected in one other SARS-CoV-2 non-detect sample with probe-capture enrichment. Expected viruses were not detected in three other previous positives even with targeted sequencing (Fig. 2A).

Human respiratory viruses were detected in 16 of 305 (5%) of previous negative samples via untargeted and targeted sequencing. Probe-capture enrichment improved viral sequence recovery relative to untargeted sequencing (Fig. 2B). The most commonly detected viruses in previous negative samples were Influenza C virus (6 samples), Human Bocavirus (4 samples), Rhinovirus A and C (3 samples), and SARS-CoV-2 (3 samples). One additional sample contained Human parvovirus B19 (Fig. 2B).

To assess genomic novelty and potential strain-level relationships between the most common viruses in previous negative samples, we mapped untargeted and targeted sequencing reads to corresponding reference sequences (see Methods). Consensus genomes were extracted for 14 viruses with >90% coverage breadth. We recovered 3 complete genomes and one partial genome for Human Bocavirus 1 from samples collected from the nearby counties of Contra Costa and Marin between 2021 and 2022 (Table 1). All genomes were closely related, differing from one another at a total of 25 nucleotide positions (99.6% identity). Sequences were also closely related to the nearest GenBank reference (99.5% identity).

We detected Rhinovirus A (RhV-A) from two previous negative samples (N261 and N325; Fig. 2B), one known positive sample (P1012), and unexpectedly from two samples previously positive for other viruses (P1016 and P333; Fig. 2A). Rhinovirus C (RhV-C) was identified from two previous negative samples (N212 and N325; Fig. 2B), two previous positive samples (P1008 and P1009), and one SARS-CoV-2 positive sample using Centrifuge. From these samples, we recovered 7 RhV genomes with >90% coverage breadth and one partial RhV genome. Notably, each sample contained distinct RhV strain(s), and the percent identity to the nearest reference genome ranged from 95.5% to 99.5%. One sample (N325) contained a co-infection of Rhinovirus A58 and C59, of which both sequences were recovered (Fig. 2B; Table 1). This sample also contained Enterovirus A, according to Centrifuge-based analysis, although this was not analyzed further as no other samples contained Enterovirus A. We also detected a co-infection in a previous negative sample (N212), which had both RhV-C and Influenza C virus (Fig. 2B).

From the six samples in which Influenza C virus was identified, we recovered 4 near-complete (>99% coverage breadth) and 2 partial viral genomes. BLAST results against NCBI core-nt for the consensus sequences of the hemagglutinin esterase fusion gene (HEF), suggested that the four near-complete genomes are likely most closely related to Influenza C virus strain C/Minnesota/28/2015. To assess the relationship between Influenza C virus genomes, we aligned the consensus sequences of the hemagglutinin esterase fusion (HEF) and M segments. This alignment showed two sequence groups, each with >99% within-group identity (Table 1). All samples positive for Influenza C were collected between February and July 2022 and represented 3 different counties. However, the near-identical sequences in each group were not from the same county (Table 1).

The presence of SARS-CoV-2 in three previous negative samples was unexpected. We recovered complete genomes from two of these samples and were able to classify these as belonging to the Epsilon (sample N055) and Omicron BA.2 (sample N198) lineages, respectively (Table 1). We searched the two complete and one partial SARS-CoV-2 genomes for the primer and probe sequences of the CDC RT-PCR diagnostic assay for SARS-

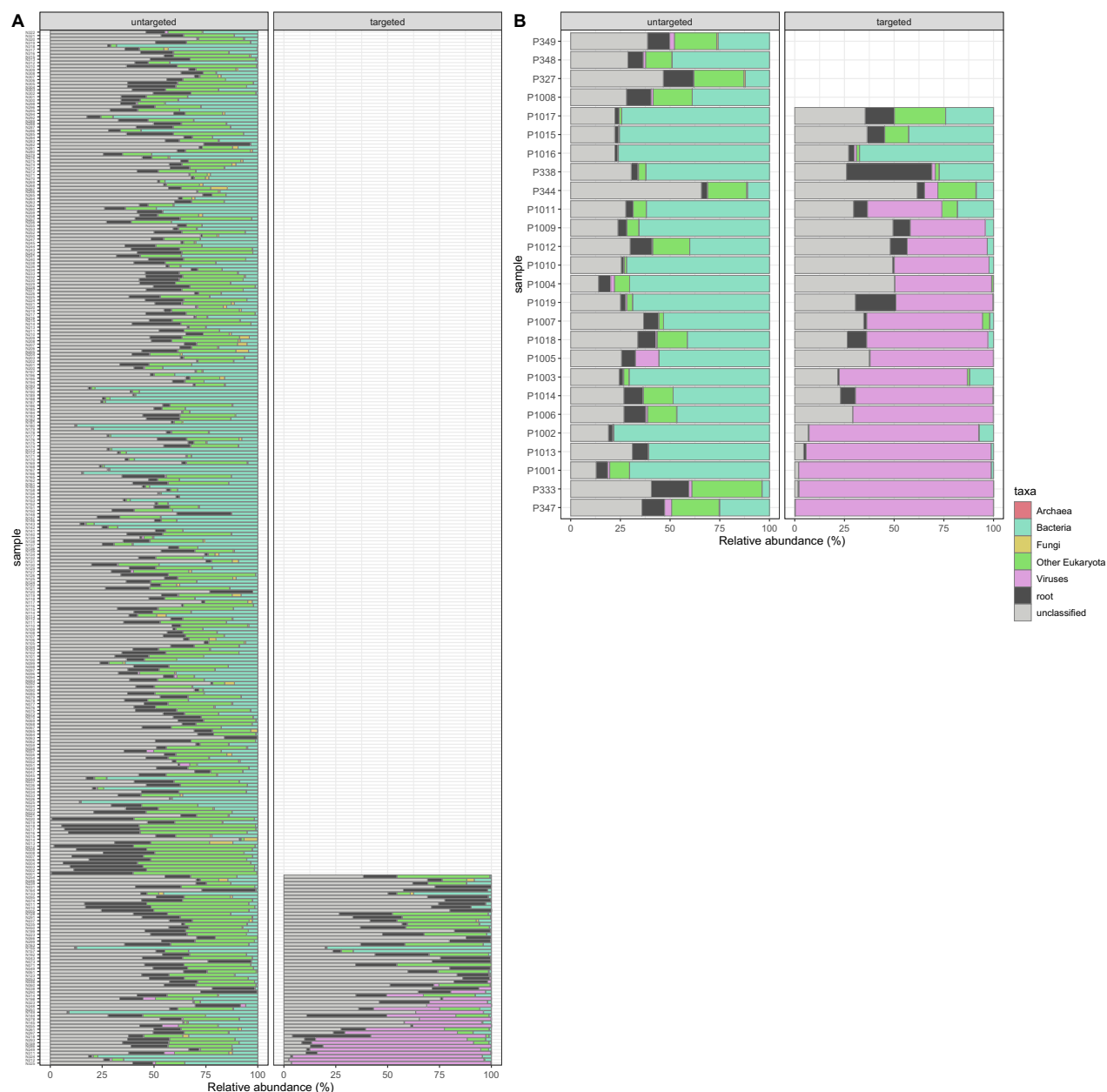


Fig. 1. Relative abundance of viral, archaeal, bacterial, fungal, non-human and non-fungal eukaryotic reads across the sample set. Previous **A** positive and **B** negative samples were sequenced with untargeted metagenomics (top panel), while a subset also underwent viral enrichment (bottom panel). Data are presented after human host removal (see Supplementary Figure S1 for inclusion of host data) and taxonomy classification with Centrifuge and Recentrifuge. Bar segments are colored by taxonomy, while black represents reads that matched to taxa from multiple domains and gray represents reads that did not match to the database at the high-stringency minimum hit length ≥ 40 nucleotides.

CoV-2³². For one genome (N055), primer and probe sequences for targets N1, N2, and N3 were perfect matches, and for the other two genomes (samples N198 and N323), the primer and probe sequences for N3 were exact matches while the N1 target had a single mismatch near the 5' end of the probe. Genomes were also recovered from two previous positive samples via untargeted sequencing, allowing classification of these as Omicron BA.1 and BA1.1 (Table 1). We also tested five samples previously positive for SARS-CoV-2 by NAAT. These samples had failed to amplify with the ARTIC tiled-amplicon sequencing protocol (data not shown) and probe-capture enrichment likewise did not recover SARS-CoV-2 (Fig. 2).

Use of probe-capture enrichment improved genome coverage by as much as 71%, producing the biggest gains for influenza C virus (Table 1). For three Rhinovirus genomes, coverage breadth decreased after probe-capture enrichment although the fraction of total reads mapped to this genome (relative abundance) increased.

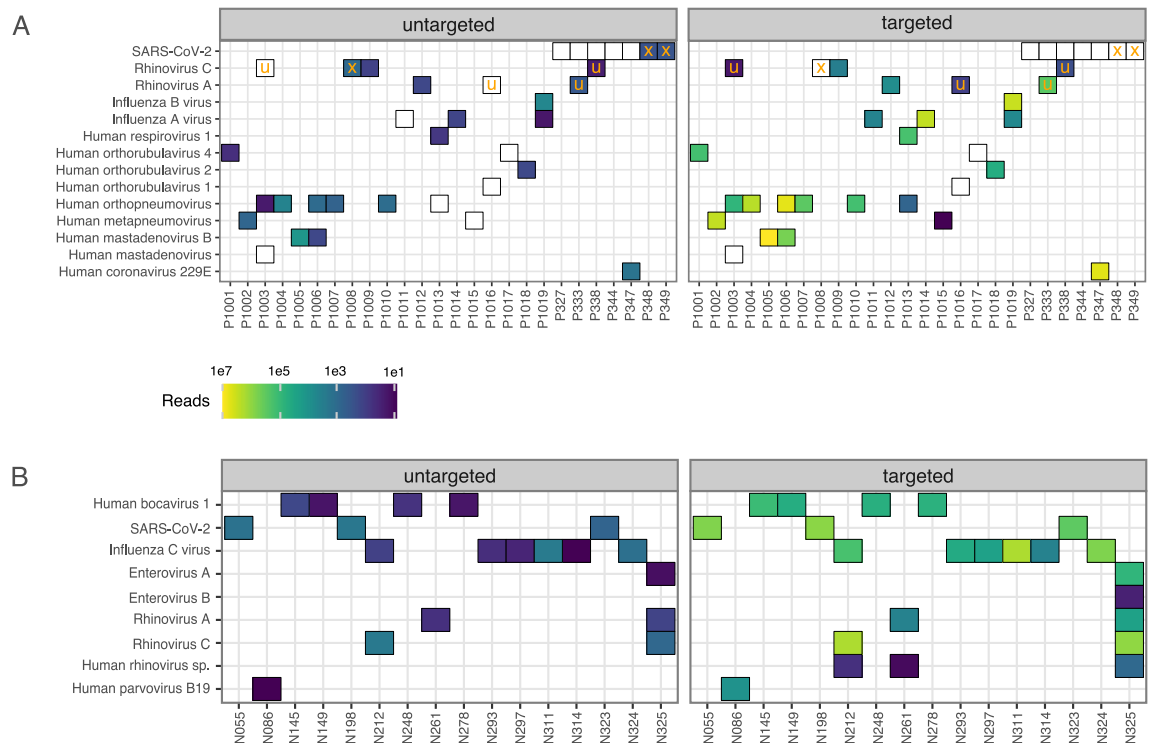


Fig. 2. Human respiratory virus detection. Detection in **A** previous positive samples and **B** previous negative samples shows substantial improvement with probe-capture enrichment (right). Fill color indicates read counts based on Recentrifuge, shown where ≥ 10 reads per virus, per sample; note log10 scale. Samples that did not undergo probe-capture enrichment are indicated with “x”, while unexpected viral detections are indicated with “u”. White tiles for previous positive samples indicate viruses expected based on previous results that were not detected by sequencing. Note that some samples contained multiple viruses.

Bacterial and fungal pathogens

Following viral assessment, we screened the 305 previous negative samples to identify those in which a single bacterial or fungal species was detected at $\geq 50\%$ relative abundance after host-read removal (Fig. 3). Twenty samples had high prevalence of a single bacterial species, such as *Dolosigranulum pigrum*, *Moraxella* spp., *Staphylococcus epidermidis* and *Acinetobacter junii*. Surprisingly, 10 samples with high bacterial content were dominated by *Pseudomonas azotoformans*. Additionally, four bacteria-dominated samples were also positive for viral human pathogens, suggesting possible viral and bacterial co-infections. Lastly, a single sample (N014) was found to be dominated by fungi, with a fungal relative abundance of 72%, most of which were classified as *Penicillium* and *Aspergillus* spp. (Fig. 3). We also analyzed two sequenced no-template controls (NTCs) to assess potential bacterial and fungal contamination. One NTC yielded no classifiable reads, while the other contained 480 reads, of which only 9 were classified as bacterial and none were classified as fungal (Supplementary Figure S3). Neither control contained any of the highly abundant bacterial species identified in our study.

Lastly, we investigated how the bacterial component of the data changed as a result of probe-capture viral sequence enrichment in 78 samples sequenced with and without enrichment (Supplementary Figure S4). Here, we compared species that occurred in at least 20% of samples within each dataset (enriched and unenriched) and had a minimum of 10 reads per sample. Of the bacterial species-level taxa observed in the unenriched samples, the majority had < 10 reads in the corresponding enriched samples. Meanwhile, across the paired samples, four taxa were enriched by an average of > 2 -fold after probe-capture, potentially suggesting off-target capture by the panel. Further work across a larger sample set and/or comparison to proprietary probe sequences would be required to understand whether these data represent true off-target enrichments.

Discussion

Pathogen detection and limitations

Our metagenomic analysis identified Influenza C virus and Human Bocavirus 1 from previous negative samples. Recent epidemiological studies have shown that these viruses are relatively common, especially in children³³, yet are not included in routine test respiratory panels. Our results suggest that both Influenza C virus and Human Bocavirus 1 circulated broadly and locally within California, based on closely related genomes of each virus and counties where the respective samples were collected. The Human Bocavirus 1 detections should be interpreted with caution as this virus is often shed well after acute infection³⁴ and has been detected in healthy controls³⁵, and as a common coinfection. Notably, we did not identify any other viruses in these samples, and recent work has shown that Bocavirus may cause respiratory symptoms as a single infection.

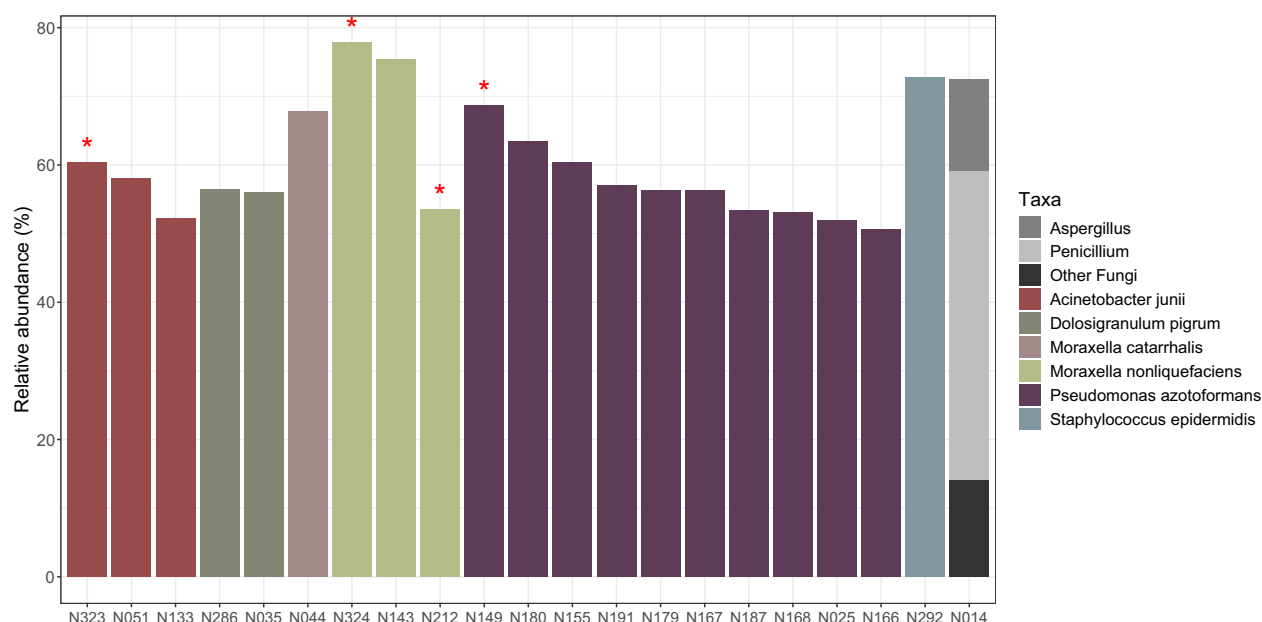


Fig. 3. Previous negative samples dominated by a single bacterial or fungal species. Only samples with bacterial or fungal relative abundance $\geq 50\%$ are shown. Red asterisks indicate samples in which respiratory viruses were also detected.

In addition to identifying less common viral pathogens by metagenomics, we also detected bacterial and fungal colonizers. Twenty samples were dominated by a single bacterial species, which included both opportunistic pathogens and environmental microorganisms (Fig. 3). Specifically, four samples contained high relative abundances of *Moraxella catarrhalis* or *Moraxella nonliquefaciens*^{36,37}. These are opportunistic pathogens but are also common members of the respiratory tract microbiome, making it difficult to interpret the significance of this finding without additional data on absolute abundance and clinical symptoms. Species such as *Acinetobacter junii* and *Pseudomonas azotoformans* were also predominant in some samples. As these species are not commonly linked to human respiratory infections or to the healthy respiratory microbiome, we speculate that their high abundance in some samples may reflect transient colonization. Likewise, the high relative abundance of fungi in one sample was surprising, but without additional data it remains unclear whether this represented an infection. Lastly, we observed instances of potential viral and bacterial co-infection. Previous studies have described the association of secondary bacterial infections during respiratory viral illness with increased disease severity and mortality^{38,39}. This is potentially important, as microbial community composition is thought to play a role in disease progression and outcome of respiratory viral infections⁴⁰.

Overall, by metagenomic sequencing of 305 previous negative samples, we were able to identify 16 (5%) viral and 35 (12%) bacterial/fungal pathogens. While it is unclear whether these detections were from outright infections or part of the normal microbiome, for bacterial detections in particular, the overwhelming majority (254/305 or 83%) of the previous negative samples from individuals who were symptomatic for respiratory illness yielded no detections of known pathogens. Possible limitations for detection in this study include: (1) non-infectious etiology of respiratory symptoms (e.g., asthma or allergy), (2) poor/low quality of sample/nucleic acid due to multiple freeze-thaw events, (3) low concentration of pathogen nucleic acid, (4) improper timing of sample collection, (5) improper sample collection technique, and/or (6) high concentration of host nucleic acid material that may have interfered with pathogen sequencing reactions.

Methods evaluation

We observed several discrepancies between sequencing results and the prior surveillance test results from BioFire and SARS-CoV-2 diagnostic assays. Of 31 expected viruses from previous positive samples, 8 were not detected by either untargeted or targeted sequencing. In these instances, the sensitivity of sequencing may have been limited due to long sample storage times (up to 43 months), freeze-thaw cycles between previous testing and sequencing, or low initial concentration of viral nucleic acids in the samples. It is also possible that the previous results were false-positive. Additionally, we identified unexpected positive samples containing SARS-CoV-2. Our sequence-based analysis indicated that these viral sequences could have been detected by the CDC RT-qPCR assay, and we are uncertain as to why SARS-CoV-2 was not detected by PCR. Separately, from previously reported negative specimens, we unexpectedly detected RhV genomes (RhV-A and RhV-C) via sequencing (Fig. 2). The RhVs we detected from previous negative samples were likely too divergent to be captured by the BioFire panel. In particular, one sample contained a sequence closely related to the novel Rhinovirus C59 genotype recently reported from western Washington state, USA (Table 1) (30).

As expected, probe-capture enrichment generally increased the coverage breadth and depth of target genomes (Table 1). However, enrichment did not always recover complete genomes even when relative abundance of target reads was >50% (e.g. samples N145 and N323), suggesting selective degradation of the target genome in the sample or incomplete capture by probes due to sequence divergence. Notably, coverage breadth for RhV genomes was lower in targeted than untargeted sequencing, suggesting the probe panel, like the BioFire panel, had limited representation of this diverse viral group.

Conclusion

While routine NAATs provide rapid (on the order of hours), cost-effective diagnoses in clinical and public health laboratories, they cannot detect novel or rare pathogens (e.g. influenza C virus) and typically do not identify specific types/strains of diverse pathogens such as rhinoviruses (Fig. 2). In contrast, untargeted metagenomic sequencing is pathogen-agnostic, can detect known or emerging pathogens in a sample, and allows for genetic analysis to identify the specific strain or characterize other aspects of the pathogen such as drug resistance markers. However, metagenomic sequencing typically requires expensive equipment, specialized data analysis pipelines, high volume data storage, and has longer turnaround times due to complex, multistep workflows (on the order of days). Additionally, it is generally much more costly to perform than NAATs, particularly if only a few samples are tested.

Here we have shown that the use of a broad probe-capture sequencing panel prior to metagenomic sequencing can increase sensitivity over untargeted sequencing, improving recovery of near-complete viral genomes for downstream analysis (Table 1). Our results suggest that for public health purposes, such as pathogen surveillance and outbreak investigations, metagenomic sequencing can provide a complementary approach to common clinical diagnostic tools, such as NAATs. Due to the complexities and higher costs of metagenomic sequencing, we envision the use of retrospective metagenomic surveys for special cases such as NAAT-negative samples, particularly from severe illnesses and outbreaks as well as for emerging public health threats where the availability of a pathogen genome is beneficial for understanding potential virulence and drug resistance markers. We expect that as metagenomic sequencing becomes more routine, assay costs will decrease, and sample workflows and data analytics will be standardized. Ultimately, this technology could become a routine biosurveillance method for public health purposes, enhancing early detection of known, rare, emerging, and novel pathogens.

Data availability

The de-identified untargeted metagenomic and TWIST-enriched Illumina sequencing datasets are publicly available in the NCBI Sequencing Read Archive under BioProject PRJNA1296491.

Received: 24 September 2025; Accepted: 23 December 2025

Published online: 16 February 2026

References

1. Sirota, S. B. et al. Global, regional, and national burden of upper respiratory infections and otitis media, 1990–2021: a systematic analysis from the Global Burden of Disease Study 2021. *Lancet Infect. Dis.* **25**, 36–51 (2025).
2. Eales, O. et al. Key challenges for respiratory virus surveillance while transitioning out of acute phase of COVID-19 pandemic. *Emerg. Infect. Disease J.* **30**, 1 (2024).
3. Clark, E. C. et al. Changes to public health surveillance methods due to the COVID-19 pandemic: scoping review. *JMIR Public. Health Surveill.* **10**, e49185 (2024).
4. WHO. *WHO Policy Brief: COVID-19 Surveillance* (World Health Organization, 2024).
5. Tisza, M. et al. Wastewater sequencing reveals community and variant dynamics of the collective human Virome. *Nat. Commun.* **14**, 6878 (2023).
6. Wyler, E. et al. Pathogen dynamics and discovery of novel viruses and enzymes by deep nucleic acid sequencing of wastewater. *Environ. Int.* **190**, 108875 (2024).
7. Smith Matthew, F. et al. Seasonality of respiratory, enteric, and urinary viruses revealed by wastewater genomic surveillance. *mSphere* **9**, e00105-00124 (2024).
8. Rajagopala, S. V. et al. Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. *Cell. Rep. Methods.* **1**, 100091 (2021).
9. Cheemarla, N. R. et al. Nasal host response-based screening for undiagnosed respiratory viruses: a pathogen surveillance and detection study. *Lancet Microbe.* **4**, e38–e46 (2023).
10. Madi, N., Al-Nakib, W., Mustafa, A. S. & Habibi, N. Metagenomic analysis of viral diversity in respiratory samples from patients with respiratory tract infections in Kuwait. *J. Med. Virol.* **90**, 412–420 (2018).
11. Gauthier, N. P. G. et al. Validation of an Automated, End-to-End metagenomic sequencing assay for agnostic detection of respiratory viruses. *J. Infect. Dis.* **230**, e1245–e1253 (2024).
12. Tan, J. K. et al. Laboratory validation of a clinical metagenomic next-generation sequencing assay for respiratory virus detection and discovery. *Nat. Commun.* **15**, 9016 (2024).
13. Ceballos-Garzon, A., Comtet-Marre, S. & Peyret, P. Applying targeted gene hybridization capture to viruses with a focus to SARS-CoV-2. *Virus Res.* **340**, 199293 (2024).
14. Mourik, K. et al. Comparison of the performance of two targeted metagenomic virus capture probe-based methods using reference control materials and clinical samples. *J. Clin. Microbiol.* **62**, e00345–e00324 (2024).
15. Wani, A. K. et al. Metagenomics in the fight against zoonotic viral infections: A focus on SARS-CoV-2 analogues. *J. Virol. Methods.* **323**, 114837 (2024).
16. Cooksey, G. L. S. et al. Severe acute respiratory syndrome coronavirus 2 and respiratory virus Sentinel Surveillance, California, USA, May 10, 2020–June 12, 2021. *Emerg. Infect. Dis.* **28**, 9–19 (2022).
17. Creager, H. M. et al. Clinical evaluation of the BioFire® respiratory panel 2.1 and detection of SARS-CoV-2. *J. Clin. Virol.* **129**, 104538 (2020).
18. Chen, S. F. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* <https://doi.org/10.1002/imt2.107> (2023).
19. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
20. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

21. Danecsek, P. et al. Twelve years of samtools and BCFtools. *Gigascience*. **10**, giab008 (2021).
22. Li, H. et al. The sequence Alignment/Map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).
23. Katz, K. et al. The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).
24. Stop neglecting fungi. *Nat. Microbiol.* **2**, 17120 (2017).
25. NCBI. *Human Read Removal Tool*, (2023). <https://ncbiinsights.ncbi.nlm.nih.gov/2023/02/02/scrubbing-human-sequences-sra-submissions/>
26. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
27. Sayers, E. W. et al. Database resources of the National center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
28. Marti Jose, M. et al. Addressing the dynamic nature of reference data: a new nucleotide database for robust metagenomic classification. *mSystems* **10**, e01239-01224 (2025).
29. Marti, J. M. & Recentrifuge Robust comparative analysis and contamination removal for metagenomics. *PLoS Comput. Biol.* **15**, e1006967 (2019).
30. McMurdie, P. J. & Holmes, S. Phyloseq: an R package for reproducible interactive analysis and graphics of Microbiome census data. *PLoS One*. **8**, e61217 (2013).
31. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
32. Lu, X. et al. US CDC Real-Time reverse transcription PCR panel for detection of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**, 1654–1665 (2020).
33. Sederdahl, B. K. & Williams, J. V. Epidemiology and clinical characteristics of influenza C virus. *Viruses*. **12**, 89–89 (2020).
34. Rayamajhi Thapa, R., Nascimento-Carvalho, C., Allander, T., Jartti, T. & Söderlund-Venermo, M. A Diagnostic approach to separate acute human bocavirus 1 respiratory tract infection from long-lasting virus shedding. *J. Infect. Dis.* jiaf130 (2025).
35. Uršič, T. et al. Acute human bocavirus 1 infection in children hospitalized for acute bronchiolitis: a 2-Year prospective study. *Pediatr. Infect. Dis. J.* **44**, 195–199 (2025).
36. Goldstein, E. J. C., Murphy, T. F. & Parameswaran, G. I. Moraxella catarrhalis, a human respiratory tract pathogen. *Clin. Infect. Dis.* **49**, 124–131 (2009).
37. Yi, H., Yong, D., Lee, K., Cho, Y. J. & Chun, J. Profiling bacterial community in upper respiratory tracts. *BMC Infect. Dis.* **14**, 583 (2014).
38. Liu, Y. et al. Outcomes of respiratory viral-bacterial co-infection in adult hospitalized patients. *eClinicalMedicine* **37** (2021).
39. Meskill, S. D. & O'Bryant, S. C. Respiratory virus Co-infection in acute respiratory infections in children. *Curr. Infect. Dis. Rep.* **22**, 3 (2020).
40. Hanada, S., Pirzadeh, M., Carver, K. Y. & Deng, J. C. Respiratory viral Infection-Induced Microbiome alterations and secondary bacterial pneumonia. *Front. Immunol.* **9** (2018).

Acknowledgements

This work was supported by the Lawrence Livermore National Laboratory's Laboratory Derived Research & Development program. We extend gratitude to Dr. Seema Jain, Dr. Jake Pry, Gail Sondermeyer Cooksey, Samuel Schildhauer, Lauren Linde, and the rest of the CDPH and local health department CalSRVSS Team. We are grateful to the CDPH VRDL Team for their work processing, managing, and testing CalSRVSS samples at CDPH including Cynthia Bernas, Brandon Brown, Alice Chen, Nick D'Angelo, Ambar Espinoza, Bianca Gonzaga, Ydelita Gonzales, Hugo Guevara, April Hatada, Blanca Molinar, Lisa Moua, Nichole Osugi, Tasha Padilla, Chao-Yang Pan, Clarence Reyes, Estela Sagar, Maria Salas, Ioana Seritan, Hilary Tamnanchit, Maria Uribe Fuentes, Chelsea Wright. The residual samples used in this study were obtained from the original community surveillance project CalSRVSS which was funded in part and by the Centers for Disease Control and Prevention Epidemiology and Laboratory Capacity for Community Surveillance (grant 6 Nu50CK000539). Disclaimer The findings and conclusions in this article are those of the authors and do not necessarily represent the views or opinions of the California Department of Public Health, the California Health and Human Services Agency or Lawrence Livermore National Laboratory. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Author contributions

RSK, ACM, CRK and CJ contributed to the writing of the of the draft manuscript. RSK, CRK, ACM and JBT analyzed and interpreted the data. CJ, DAW, CM and SM acquired the samples. JBT, ACM, MB, CJ, DAW, CM and SM designed the study and experiments. ACM, JBT, DAW, SM and CM performed experiments. CJ and DAW contributed to project management. All authors contributed to reviewing and editing the final manuscript.

Funding

The residual samples used in this study were obtained from the original community surveillance project CalSRVSS which was funded in part and by the Centers for Disease Control and Prevention Epidemiology and Laboratory Capacity for Community Surveillance (grant 6 Nu50CK000539).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-33917-4>.

Correspondence and requests for materials should be addressed to C.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026