COLD SPRING HARBOR
Molecular Case Studies

RESEARCH REPORT

Genomic surveillance of SARS-CoV-2 during the first year of the pandemic in the Bronx enabled clinical and epidemiological inference

J. Maximilian Fels,[1,14] Saad Khan,[2,14] Ryan Forster,[2,14] Karin A. Skalina,[3,13] Surksha Sirichand,[4] Amy S. Fox,[3,5] Aviv Bergman,[2,3,6,7] William B. Mitchell,[5,8] Lucia R. Wolgast,[3] Wendy Szymczak,[3] Robert H. Bortz III,[1] M. Eugenia Dieterle,[1] Catalina Florez,[1,9] Denise Haslwanter,[1] Rohit K. Jangra,[1] Ethan Laudermilch,[1] Ariel S. Wirchnianski,[1,10] Jason Barnhill,[9] David L. Goldman,[1,5,11] Hnin Khine,[5,12] D. Yitzchak Goldstein,[3] Johanna P. Daily,[1,4] Kartik Chandran,[1] and Libusha Kelly[1,2]

[1]Department of Microbiology and Immunology, [2]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New York 10461, USA; [3]Department of Pathology, Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, New York 10461, USA; [4]Department of Medicine (Infectious Diseases), [5]Department of Pediatrics, [6]Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York 10461, USA; [7]The Santa Fe Institute, Santa Fe, New Mexico 87501, USA; [8]Division of Pediatric Hematology/Oncology, Children's Hospital at Montefiore, Bronx, New York 10461, USA; [9]Department of Chemistry and Life Science, United States Military Academy at West Point, West Point, New York 10996, USA; [10]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, USA; [11]Division of Pediatric Infectious Diseases, [12]Division of Pediatric Emergency Medicine, Children's Hospital at Montefiore, Bronx, New York 10467, USA

Corresponding authors: libusha.kelly@einsteinmed.org; kartik.chandran@einsteinmed.org; jdaily@montefiore.org; dogoldst@montefiore.org; hkhine@montefiore.org; dagoldma@montefiore.org

Ontology terms: immune dysregulation; severe viral infections

Published by Cold Spring Harbor Laboratory Press

doi:10.1101/mcs.a006211

Abstract The Bronx was an early epicenter of the COVID-19 pandemic in the United States. We conducted temporal genomic surveillance of 104 SARS-CoV-2 genomes across the Bronx from March to October 2020. Although the local structure of SARS-CoV-2 lineages mirrored those of New York City and New York State, temporal sampling revealed a dynamic and changing landscape of SARS-CoV-2 genomic diversity. Mapping the trajectories of mutations, we found that although some became "endemic" to the Bronx, other, novel mutations rose in prevalence in the late summer/early fall. Geographically resolved genomes enabled us to distinguish between cases of reinfection and persistent infection in two pediatric patients. We propose that limited, targeted, temporal genomic surveillance has clinical and epidemiological utility in managing the ongoing COVID pandemic.

[Supplemental material is available for this article.]

INTRODUCTION

COVID-19 continues to have a devastating effect on the health of communities across the globe, with more than 500 million reported cases and more than six million deaths since the start of the pandemic, as reported on April 24th, 2022 (World Health Organization

[13]Present address: Department of Radiation Oncology, Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, New York 10461, USA
[14]These authors contributed equally to this work.

2020). The Bronx, a borough of New York City (NYC), sustained the second highest rate of COVID-19 in early waves of the pandemic in New York City with 6035 cases per 100,000 people as of January 11, 2021; as of April 22 the rate has reached 28,974 cases per 100,000 people (Elflein 2022). To track the local spread of SARS-CoV-2, during the first year of the pandemic, we conducted a genomic epidemiologic study at Montefiore Health Systems (MHS), which offers health-care services to two million residents throughout the Bronx, one of the most diverse and poorest urban communities in the United States.

The number of COVID-19 cases peaked in the Bronx in March–April 2020 and subsided during the late spring into summer 2020. To characterize the genetic diversity of SARS-CoV-2 in the Bronx, we selected nasopharyngeal remnant clinical samples positive for SARS-CoV-2 by reverse transcription-polymerase chain reaction (RT-PCR) testing from the MHS clinical laboratory between March and October 2020.

RESULTS

Positive patient samples collected through routine clinical care and demonstrating cycle threshold (C_t) values of <30 via RT-PCR testing were removed from storage weekly. This material was logged, anonymized, aliquoted, and frozen at −70°C. Shipments provided for sequencing were generally a convenience sampling of stored frozen material. Genomic viral RNA was extracted from nasopharyngeal swabs, and sequencing libraries were prepared using the ARTIC Network protocol and analyzed on an Oxford Nanopore MinION (Quick et al. 2017; Quick 2020). The ARTIC Network bioinformatics protocol was used to quality check and annotate SARS-CoV-2 genomes with default parameterization (Lowman et al. 2019). We called mutations with the NextClade tool and annotated lineages using the October 13th, 2021 update of PANGOLIN version 3.1.14 (Hadfield et al. 2018; Rambaut et al. 2020). Samples were derived from patients who required hospitalization (48%), patients who had mild disease managed as outpatients (26%), and asymptomatic carriers (8.9%) (Fig. 1A).

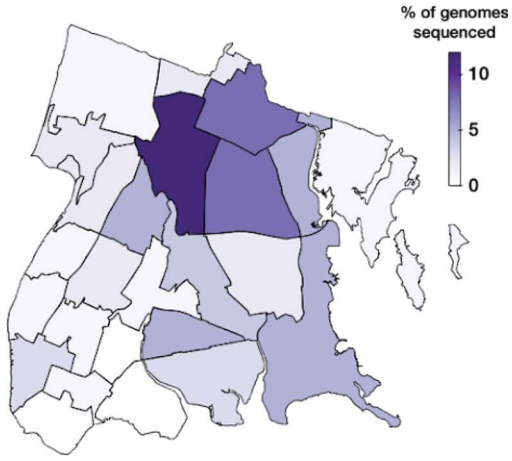
We collected 137 samples, and from these generated 104 high-quality genomes from 101 patients with >95% coverage to ensure optimal base calling and lineage assignment; this is in accordance with the high coverage threshold used by GISAID instead of the standard 90% coverage (Supplemental Figs. S1 and S2). Sequence data were derived from residents across the Bronx and were associated with 22 of 25 zip codes (Fig. 1B), we note that our sampled sequences are not identically distributed with caseloads during the sampling period (Supplemental Fig. S3). Genomic sampling was greatest at the onset of the COVID-19 pandemic in March and April, and intermittent sampling continued as caseloads declined over the summer and fall. The sampling period of the study gives insight into SARS-CoV-2 viral diversity in MHS patients during the month of April and limited insight into the later months. Although our samples track temporally with caseload, we do not expect our sequenced genomes to cover the total viral diversity of SARS-CoV-2 in the Bronx during the period of sampling (Fig. 1C).

Analysis of the resulting 104 SARS-CoV-2 genome sequences revealed that the B.1 lineage was the most prevalent during the early months of the pandemic in the Bronx; however, several other lineages were also present at low frequencies throughout the sampling period (Fig. 2A). Many low-frequency B.1 sublineages were sampled at different time points. Some of these lineages were first observed elsewhere before being observed in our cohort. Two lineages, B.1.604 and B.1.448, were first observed in the MHS cohort and subsequently spread to other areas of the United States. We sampled the first five of 48 B.1.604 assigned genomes, which demonstrated a dissemination within the Northeastern United States. B.1.448 represented a larger sublineage of B.1 that appeared to arise from the Bronx, having

A

Characteristic	N=101
Age, yr – median (IQR)	59 (41-71)
Female – no. (%)	47 (46.5)
Race – no. (%)	
White	6 (5.9)
Black	19 (18.8)
Other	19 (18.9)
Unknown	57 (56.4)
Ethnicity – no. (%)	
Hispanic	39 (38.6)
Non-Hispanic	44 (43.5)
Unknown	18 (17.8)
Disposition – no. (%)	
Outpatient	23 (22.8)
Inpatient, COVID-19 related	49 (48.5)
Inpatient, COVID-19 unrelated	10 (9.9)
Disposition unknown	19 (18.9)
Clinical presentation – no. (%)	
Asymptomatic	9 (8.9)
Mild COVID-19, outpatient	23 (22.8)
COVID-19, hospitalized	49 (48.5)
Not documented	20 (19.8)
Comorbidities – no. (%)	n=91
Hypertension	47 (51.6)
Coronary Artery Disease/Congestive Heart Failure	14 (15.3)
Diabetes Mellitus	21 (23.1)
Asthma/Chronic Obstructive Pulmonary Disease	13 (14.2)
Cancer	17 (18.7)
Immunocompromised	15 (16.4)
Chronic Kidney Disease ≥ 3	8 (8.8)
Body Mass Index ≥30 kg/m^2	22 (24.2)
Length of stay, days – median (IQR)[1]	6 (4-13)
Maximum Oxygen Requirement – no. (%)[1]	n=49
Room Air	14 (28.6)
Nasal Cannula	11 (22.4)
Non-Rebreather/High-flow nasal oxygen	13 (26.5)
Mechanical Ventilations	11 (22.4)
30-day outcome - no. (%)	
Alive	62 (61.4)
Deceased	17 (16.8)
Not Available	22 (21.8)

B



C

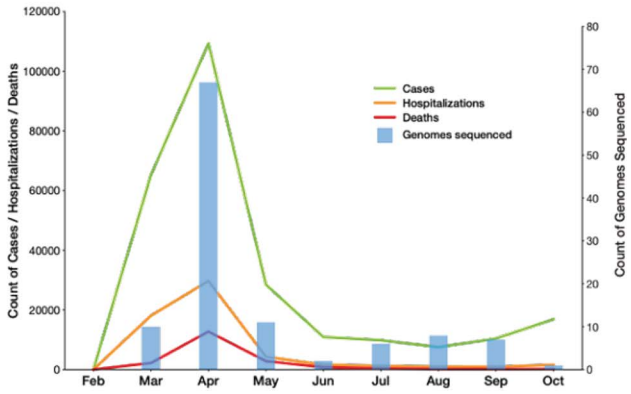


Figure 1. Surveilling SARS-CoV-2 genomes in the Bronx. (*A*) Table of clinical characteristics of sampled patients. (*B*) SARS-CoV-2 genomes sequenced per zip code in NYC; darker colors indicate heavier sampling. (*C*) SARS-CoV-2 genomes sequenced over time during the COVID-19 pandemic. The date is indicated on the *x*-axis. Blue bars and the associated right-hand *y*-axis indicate the number of genomes sequenced. The left-hand *y*-axis represents different features of COVID-19 in the Bronx; green lines indicate COVID-19 cases, the red line deaths associated with COVID-19, and the orange line hospitalizations associated with COVID-19 in the Bronx.

332 other genomes sharing lineage-defining single-nucleotide polymorphisms (SNPs). Both of these lineages were detected until early 2021, and B.1.604 was only sampled at the MHS during the time frame of March to October. The majority of Bronx genomes were classified as sublineages of B.1 or B.1.1; the B.1 lineage continued to be sampled until late August. From March to October the Bronx SARS-CoV-2 lineages represent a subset of SARS-CoV-2 lineage diversity present in NYC, New York State (NYS), the United States (US), and the world (Fig. 2B; Gonzalez-Reiche et al. 2020; Maurano et al. 2020). B.1.448 was found in other areas of the US during the sampling period, prevalent in the western US 2 months and even later. Although the first reported sequence was in the Bronx, the lineage B.1.448's prevalence may indicate its origins remain elsewhere, whereas B.1.604 was only observed in the Bronx. We noted that most B.1 and B.1.1 sublineages were low in number compared to the B.1 and B.1.1 lineages, whereas the B.1.1 relative representation at MHS, in NY, and in the US was higher than the global representation. To determine how the Bronx

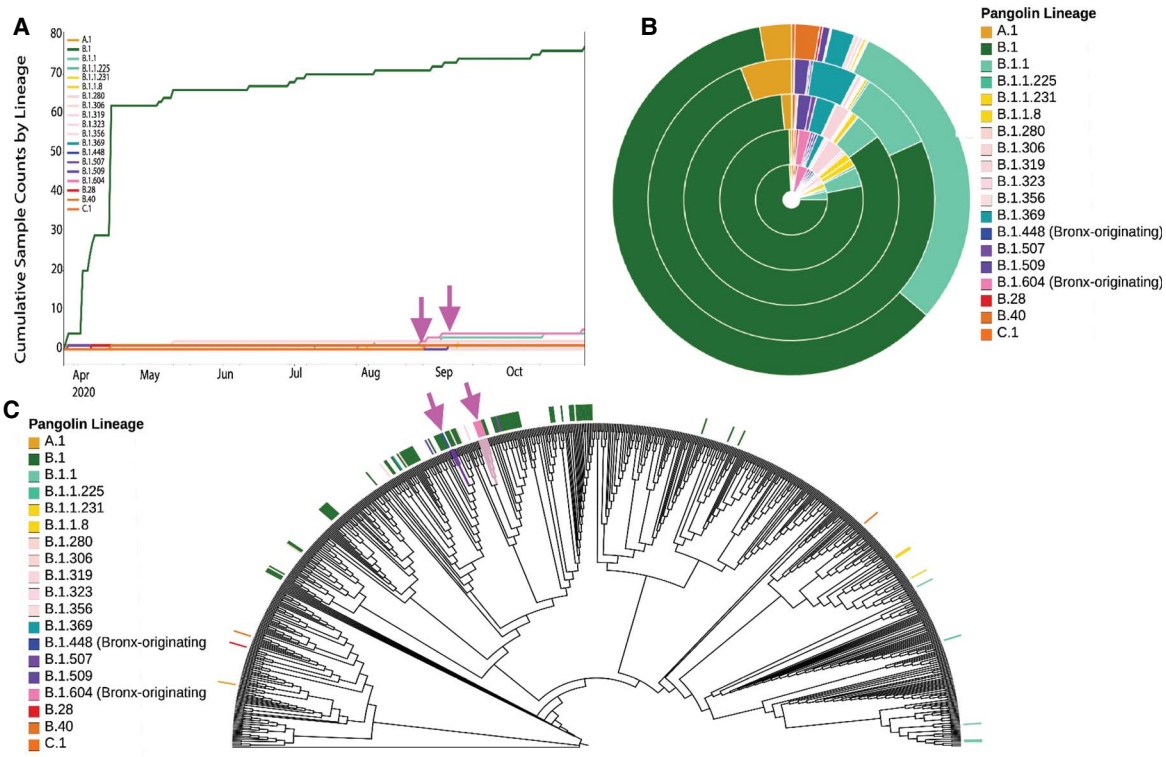
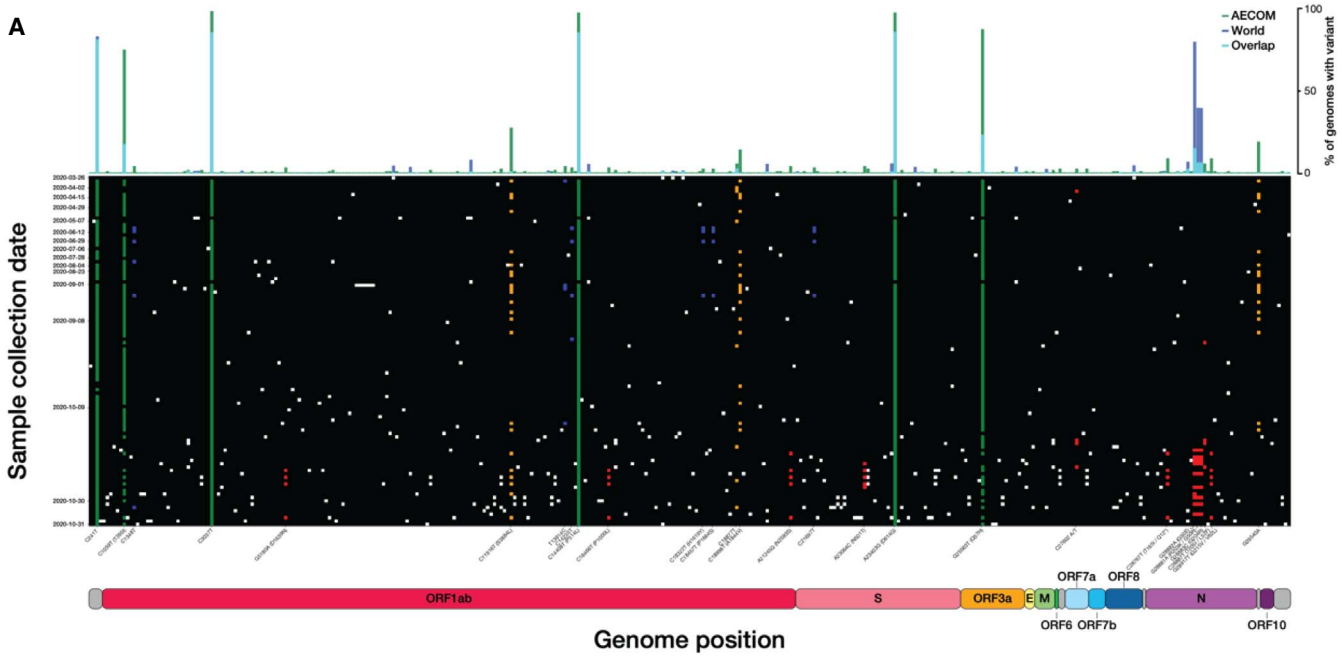


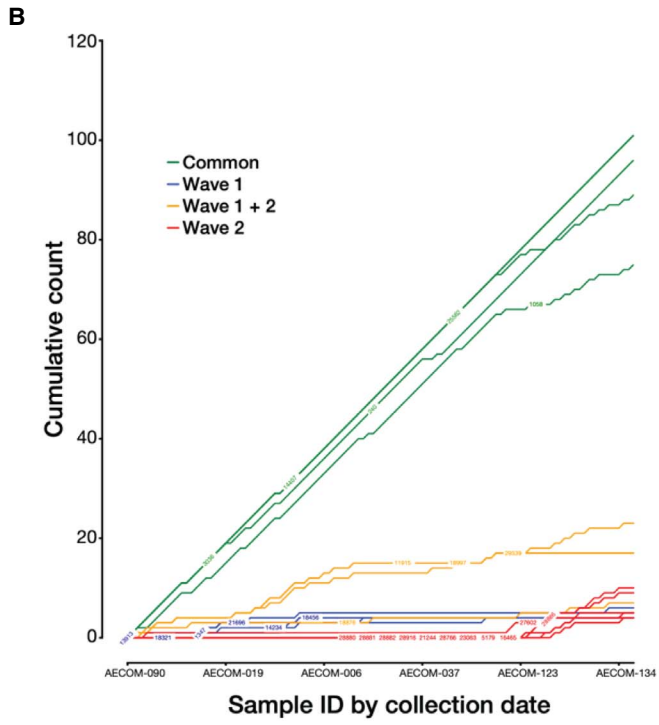
Figure 2. Bronx SARS-CoV-2 genome lineages in the context of local and global sampling. (*A*) Cumulative counts of PANGOLIN guide tree–based lineage assignments plotted against time (first detection of Bronx-originating lineages indicated by purple arrows). (*B*) Prevalence of lineages seen in the Bronx compared to their prevalence in other regions. The *inner* to *outer* rings represent the Bronx, New York City, New York State, the United States, and the world, respectively. Lineage coloring is the same as in *A*. (*C*) Phylogeny of the Bronx isolates in the context of SARS-CoV-2 isolates from around the world. Bronx isolates and their associated lineages are indicated with colored lines; Bronx-originating lineages are indicated by purple arrows.

sequences of these lineages compared to those sampled across the world, we created a downsampled SARS-CoV-2 tree from 613 high-quality SARS-CoV-2 genomes deposited in GISAID with available location and collection dates from March to October 2020. We found that Bronx SARS-CoV-2 sequences represented subsets of different clades of the global tree (Fig. 2C).

We next examined the mutation patterns in nucleotide positions observed in our data. We found that variation in mutations are distributed across the SARS-CoV-2 genome and that some mutations are present in almost all Bronx genomes sequenced—these can be described as "core" to the Bronx at present (Fig. 3A). Core mutations include the spike protein mutation A23403G (D614G), as well as mutations C241T, C1059T (T265I), C3037T, C14408T (P314L) in *Orf1ab* and G25563T (Q57H) in *Orf3a*. We next examined the dynamics of individual SARS-CoV-2 mutations. Although the core mutations continued to increase in prevalence as we sequenced new genomes, we also observed mutations novel to the Bronx whose prevalence increased, whereas others plateaued (Fig. 3B).

In the spike protein sequence, we found amino acid mutations D614G (core), N501T in five patients, and both N501Y and P681R in one patient. We noted that P314L in *Orf1b* is also a core mutation in our data set, reflecting observations in other studies that this mutation is in linkage disequilibrium with D614G (Ogawa et al. 2020). We did not observe the B.1.1.7 lineage first identified in the United Kingdom in the fall of 2020, or any other WHO-classified





C

	Nucleotide	Amino acid	Gene	Type	Count
3037	C3037T		ORF1a	Common	101
14408	C14408T	P314L	ORF1b	Common	101
23403	A23403G	D614G	S	Common	101
241	C241T			Common	96
25563	G25563T	Q57H	ORF3a	Common	89
1059	C1059T	T265I	ORF1a	Common	75
1348	C1348T		ORF1a	Wave 1	6
13914	T13914C		ORF1a	Wave 1	4
14235	C14235T		ORF1a	Wave 1	4
18322	C18322T	H1619Y	ORF1b	Wave 1	4
18457	C18457T	P1664S	ORF1b	Wave 1	4
21697	C21697T		S	Wave 1	4
28881	G28881A	R203K / G50R	N / ORF14	Wave 2	10
28882	G28882A	G50E	N / ORF14	Wave 2	9
28883	G28883C	G204R	N / ORF14	Wave 2	9
21245	A21245G	N2593S	ORF1b	Wave 2	5
23064	A23064C	N501T	S	Wave 2	5
28767	C28767T	T165I / Q12*	N / ORF14	Wave 2	5
28917	G28917T	G215V / V62L	N / ORF14	Wave 2	5
5180	G5180A	D1639N	ORF1a	Wave 2	4
16466	C16466T	P1000L	ORF1b	Wave 2	4
27603	C27603A		ORF7a	Wave 2	4
28887	C28887T	T205I / L52F	N / ORF14	Wave 2	4
11916	C11916T	S3884L	ORF1a	Wave 1+2	23
18998	C18998T	A1844V	ORF1b	Wave 1+2	17
29540	G29540A			Wave 1+2	17
18877	C18877T		ORF1b	Wave 1+2	7

Figure 3. SARS-CoV-2 mutations and their trajectories in the Bronx. (*A*) Individual SARS-CoV-2 mutations plotted across the viral genome (*x*-axis), with genomes sorted by sampling date (*y*-axis). Positions that are variable with respect to the reference SARS-CoV-2 isolate are shown with a white (low-frequency), green (common), yellow (wave 1 + 2), blue (wave 1), or red (wave 2) squares. The histogram across the *top* plots the prevalence of a given mutation across all Bronx SARS-CoV-2 genomes in this study relative to the world. (*B*) Rarefaction curve of cumulative mutation counts over time for mutations observed at least four times in the Bronx SARS-CoV-2 genomes set. (*C*) Table showing details for mutations in 3B.

variants that were circulating during this time frame, known to contain the N501Y mutation and similarly the P681H in our samples. The N501 residue of the spike protein is part of the receptor binding domain and the receptor binding motif, and mutations at this position may influence *ACE2* receptor binding (Wan et al. 2020). In comparing Bronx mutations to those found in the rest of the world, we found that some mutations, such as the spike protein D614G mutation, are prevalent both in our set and in the world; however, some "core" Bronx mutations such as C1059T (T265I in *Orf1ab*) and G25563T (Q57H in *Orf3a*) are not as prevalent in the rest of the world at the study time period (top bar Fig. 3A,C; Supplemental Fig. S4). The geographic specificity of mutations creates a fingerprint that can be useful for tracing the spread of particular mutations; a lineage containing the mutation C2416T, linked to the Boston Biogen COVID-19 outbreak, could be traced to infections around the world (Lemieux et al. 2021). The C2416T mutation was also observed in three patients in our data set. We note that rare mutations are uniformly distributed throughout the sampling period (Supplemental Fig. S5) and further that the functional impact of these mutations is not well-resolved.

A phylogenetic tree of SARS-CoV-2 shows that samples collected earlier in the pandemic are distinguishable from isolates collected later, suggesting that new isolates were being continuously introduced into the Bronx (Fig. 4, inner ring, red indicates earlier samples, green newer samples). There was evidence of ongoing presence of B.1 lineage throughout the study period, starting from the onset of the pandemic until the end of the study period (Fig. 4, outer ring indicates lineage). We found that the B.1 sublineages, such as the parent

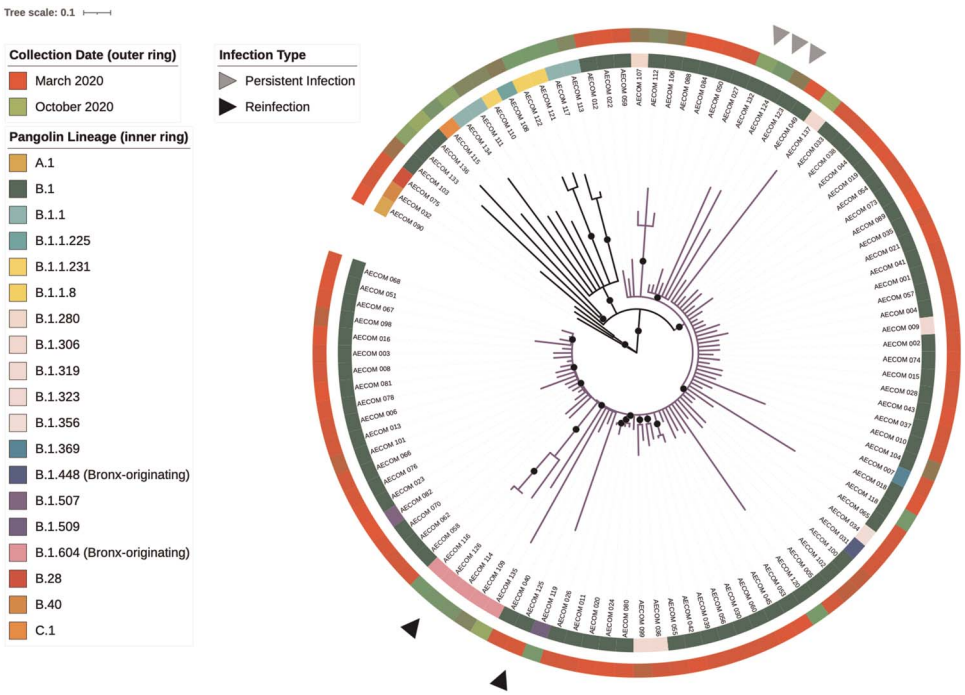


Figure 4. Clinical relevance of the changing genomic landscape of SARS-CoV-2 in the Bronx. Phylogenetic tree based on whole-genome alignments of Bronx isolates. Colored rings around the tree indicate SARS-Cov-2 lineage (*inner* ring) and the date of sampling (*outer* ring, red = earlier, green = later). "Introduced" isolates are black branches; "circulating" isolates are purple branches. Samples from the same patient are indicated with symbols; a reinfection case is indicated with black arrows and a putative persistent infection case is indicated with gray arrows. Black circles on the branches indicate bootstrap values of 85 or greater. The tree was generated with TimeTree and visualized with iTOL (Sagulenko et al. 2018; Letunic and Bork 2019).

lineage of variants of concern Alpha and Omicron B.1.1 and Bronx originating lineages B.1.448 and B.1.604, had increasing presence in the latter part of the study period and that newer collections of B.1 isolates, which cluster away from older B.1 sequences, appear at later sampling dates. There are newer B.1 and B.1 sublineages that form a distinct clade from older B.1 lineages in the Bronx SARS-CoV-2 tree. We posit that these two clades reflect two different types of SARS-CoV-2 isolates: those that were circulating locally and those that were newly introduced. We considered SARS-CoV-2 isolates grouping on the downsampled global tree and the local Bronx tree with our first wave pandemic sampling to be "circulating." We observed that the Bronx-originating lineages pair with what we consider circulating isolates of B.1 on the global tree. However, we continue to observe isolates that fall into this "first wave" clade of B.1 during the summer, post–first wave, and therefore consider these to have persisted in the Bronx. There is no indication that any genomes represent an outbreak, as they were observed throughout the sampling period. We consider "introduced" isolates —that is, those that are newer sequences in the local Bronx tree that are also spread out in different clades across the global tree or arising from lineages that originate outside the Bronx (Figs. 2C and 4).

This local phylogenetic framework of SARS-CoV-2 isolates in the Bronx enabled us to distinguish between a case of reinfection and a case of persistent infection in two pediatric patients. The first case is a 12- to 18-yr-old patient who was initially seen in April 2020 in the emergency department with 3 d of fever, sore throat, anosmia, and ageusia in the setting of the death of the patient's father at home from suspected COVID-19. SARS-CoV-2 infection in this patient was confirmed by RT-PCR. The patient had a total of 6 d of symptoms and was in general good health until the second presentation. In August 2020, the patient presented again to the emergency department with 2 d of fever, severe postprandial abdominal cramps, watery diarrhea, and generalized body aches. All other reviews of symptoms were negative. The patient had no known COVID-19 exposures and limited outside exposure with visits only to supermarkets and parks near home. A respiratory pathogen panel was negative but the patient's SARS-CoV-2 RT-PCR was positive, as was the SARS-COV-2 IgM Immune Status Ratio (ISR) (2.1, with <1 considered negative). The patient's IgG ISR was negative, 8.7 (normal range ISR < 9). The patient had a total of 3 d of fever with complete resolution of all other symptoms by day 4 of illness. Long-term antibody levels in children have not been well-characterized and durability of antibody responses may depend on several factors including the assay being used, patient age, and severity of disease (Toh et al. 2022), potentially explaining why the patient's IgG ISR was negative.

The two SARS-CoV-2 genomes sequenced from this patient were 142 d apart and differed in nucleotide sequence at 17 different positions. The first and second samples from this patient have different lineage designations and fall in different local phylogenetic clades in the Bronx phylogenetic tree, supporting the hypothesis that this represents a new infection and not prolonged shedding from the original SARS-CoV-2 infection. In fact, the patient was reinfected with one of the first few isolates of the Bronx-originating lineage B.1.604, suggesting they belong to a separate transmission chain (Fig. 4, black arrows). Given the history of limited exposures to high-risk activities for this patient between the two episodes and the overall low incidence of SARS-CoV-2 infection in New York at the time of the second presentation in August, genomic and phylogenetic analysis provided key confirmatory evidence in support of the clinical inference of a reinfection.

The second patient was a 12- to 18-yr-old patient who presented with an oral lesion in July 2020 and found to be SARS-CoV-2 positive. The patient's past medical history includes *DIAPH1* deletions associated with seizure disorder, cerebral palsy, and cortical blindness. The patient had an incompletely characterized immunodeficiency, thought to be autoimmune in nature and characterized by cytopenia, hypogammaglobulinemia, and thrombocytopenia. Medications at the time of admission included IVIG infusions, dapsone,

fluconazole, azithromycin, sirolimus, and multiple antiseizure medications. The patient was not febrile at admission and had no respiratory or gastrointestinal symptoms, but did have neutropenia (absolute neutrophil count: 700 neutrophils/μL). After admission for further evaluation, the patient was found to be SARS-CoV-2-positive. During admission, the patient was intermittently febrile and neutropenic and was treated with broad spectrum antibiotics. The patient developed a buttock lesion that was biopsied, revealing a thrombotic vasculopathy with infarction. Because of concern that the lesion could represent COVID-19-associated vasculopathy, and in the setting of persistent fever and intermittent neutropenia, the patient was treated with a 10-d course of remdesivir. The patient continued to have positive nasopharyngeal swabs for SARS-CoV-2 from early July to the end of September (Supplemental Table S1; Supplemental Fig. S5). The patient's SARS-CoV-2 IgG (Abbott) was negative in mid-August.

For this patient, the three sequenced SARS-CoV-2 genomes sampled in July, August, and October are members of the B.1 lineage and fall in the same clade (Fig. 4, gray arrows). This clade is polytomic by TimeTree, meaning that it is not possible to resolve the relationships between sequences within this clade, but the clade itself is supported by a bootstrap value of 870/1000 (SH-aLRT replicates) (Sagulenko et al. 2018; Nguyen et al. 2015). We therefore posit that the three isolates sequenced from this patient, despite having some variation, are more likely to represent a single SARS-CoV-2 infection rather than multiple infections. Together, these genomic, phylogenetic, and clinical observations strongly suggest that this patient has been unable to clear a single infection of SARS-CoV-2, as opposed to being reinfected with a distinct isolate. Other examples of persistent infection with SARS-CoV-2 have been reported, but not, to our knowledge, in children (Abu-Raddad et al. 2020; To et al. 2020; Sevillano et al. 2021; Tillett et al. 2021). A woman diagnosed with chronic lymphocytic leukemia who was sampled five times had SARS-CoV-2 sequences displaying intrahost variation despite the SARS-CoV-2 being polytomic, similar to what we observe here (Avanzato et al., 2020). The polytomy that encompasses this persistent case also contains independent local isolates of SARS-CoV-2 that do not separate on the global tree, suggesting that some mutations seen in this patient are also shared locally in the Bronx (Figs. 2C, 4).

DISCUSSION

Our work supports guiding principles for practical and clinical applications of SARS-CoV-2 sequencing in the COVID-19 pandemic. How many genomes do you need to sequence for a local community to resolve clinical questions? In our case, approximately 100 genomes were sufficient to place new patients into the context of the variability of SARS-CoV-2 during the pandemic and to be able to answer coarse-grained questions to determine reinfection versus persistent infection and community-level observations of older versus newly introduced isolates. The targeted utilization of small numbers of stored swabs for temporally resolved viral genomic surveillance could thus resolve clinical questions related to persistence versus reinfection. This localized molecular and temporal description of SARS-CoV-2 genomes during the first wave of the COVID-19 pandemic in the Bronx, New York demonstrates the value of local sequencing efforts to guide clinical inference and serves as a valuable baseline for ongoing studies of the pandemic in this underserved urban community.

METHODS

RNA Isolation

Viral RNA was isolated from nasopharyngeal swabs using the MagMAX Viral RNA isolation kit (Applied Biosystems AM1939) according to the manufacturer's specification. An amount of

400 µL of viral transport medium was used as input for each sample. Isolated RNA was then stored at −80°C prior to sequencing library generation.

Preparation of Sequencing Libraries

Sequencing libraries were prepared according to the protocol established by the ARTIC network (ARTIC Network 2019, 2020). Briefly, cDNA was generated from viral RNA using SuperScript IV reverse transcriptase (Thermo Scientific 18090010). Four hundred nucleotides of tiled amplicons were generated using the V3 primer pool, divided into four subpools for increased efficiency. Amplification was performed using Q5 High-Fidelity polymerase (New England Biolabs M0491S) with cycle numbers optimized for each subpool. Following amplicon cleanup using AMPure XP beads (Beckman Coulter A63880), 5 ng of input DNA, quantified using Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen P7589), was natively barcoded using the Native Barcoding Expansion (Nanopore EXP-NBD104). After another round of amplicon cleanup using AMPure XP beads, sequencing adapters were ligated to pooled barcoded amplicons using NEBNext Quick Ligation Module (New England Biolabs E6056). Following an additional step of cleanup and quantification, the final libraries were sequenced.

Nanopore MinION Sequencing

Sequencing libraries were diluted in elution buffer (QIAGEN 19086) to a concentration corresponding to ∼20 ng of library per sequencing run. MinION flow cells (Oxford Nanopore FLO-MIN106D) were prepared using the Ligation Sequencing Kit (Oxford Nanopore SQK-LSK109). Libraries were then loaded onto the flow cell and sequencing allowed to proceed for 10–20 h depending on library size.

Sequencing Analysis

ONT MinION output files in fast5 format were processed using an implementation of the ARTIC sequencing pipeline on Google Cloud Platform. Briefly, this pipeline consists of the following steps: (1) Base call reads using Oxford Nanopore's Guppy tool; (2) detect barcodes to sort out reads from different samples using Guppy; (3) remove chimeric reads and small contaminations by filtering out all reads not within 400–700 nt in length; and (4) align reads to the Wuhan reference genome (NCBI identifier MN908947.3) using minimap2, generate a consensus genome, and call variants using the nanopolish tool. The pipeline was run using the workflow tool Argo running on a Kubernetes cluster in the cloud. Data was stored on a cloud storage bucket between steps (see Supplementary Information). Low-coverage sequences were improved by combining passed reads from multiple sequencing runs before generating consensus sequences.

Quality Control

We included in our analysis only sequences that had 95% or higher coverage, a criterion 104 out of 132 sequences satisfied (Supplemental Fig. S2). We also looked for signs of biases in the base-calling pipeline that would result in higher or lower likelihood of gaps in certain regions. We found that the probability of a gap being present in the consensus is strongly correlated with the coverage level in the BAM file generated by the pipeline. In particular, we found that a coverage of 20× was almost always sufficient to result in a base call being made at a given position but that the majority of positions had coverage >400×. Thus, any biases in the pipeline are more likely to arise from biases in the nanopore sequencer itself or its base caller rather than the consensus generation software.

Variant Annotation and Global Analysis of Variants

We used the NextClade command line tool to assign variant calls to each of the samples. This tool performs a pairwise alignment between an assembled genome and the Wuhan reference genome and reports the differences as variant calls. NextClade was also used to determine the amino acid changes implied by each variant. This method of variant calling was chosen over the one provided in the ARTIC pipeline in order to maintain consistency with our comparative analysis of global variant distributions.

We downloaded all of the 139,676 genomes available from GISAID as of November 14, 2020 and used the NextClade command line tool to annotate each of their variants. This tool automatically rejects sequences that it deems of low quality, and this yielded variant calls from 139,590 genomes from around the world. We used this output to compute the frequency of a variant as the percentage of samples in the world/AECOM data set containing a given variant.

Creating the Local Phylogenetic Tree

Individual FASTA files of ≥95% coverage were collected after output by the ARTIC pipeline. The multi-FASTA was aligned using MAFFT on the Nextstrain command line interface version 1.16.7 (Katoh and Standley 2013; Hadfield et al. 2018). The resulting alignment FASTA generated was constructed into a maximum likelihood tree with 1000 SH-aLRT bootstraps using a TIM + F + I substitution model via iqtree-2.1.1-Windows (Nguyen et al. 2015). The tree was rooted on AECOM 90, the oldest outgroup sequence, and the entire tree was branch length–corrected based on a fixed mutation rate of 0.0008 nucleotides/site/year with a standard deviation of 0.0004 using treetime 0.7.6 (Sagulenko et al. 2018). The tree was visualized on iTOL and annotated with the iTOL annotation editor (Letunic and Bork 2019).

Creating the Global Phylogenetic Tree

The GISAID database GISAID—Initiative limited to 95% coverage and higher was used as an input for this analysis. The multi-FASTA of 11/14/2020 was filtered using the Nextstrain command line interface version 1.16.7 filter command. The specifications entailed and inclusion criteria used to construct a globally and temporally representative multi-FASTA was adapted from the criteria used to construct the Nextstrain global tree. An inclusion and exclusion text file was used to remove and keep strains that Nextstrain deemed important and is located here: https://github.com/nextstrain/ncov. The entire GISAID database was purged of any sequence with inconsistent metadata and grouped based on the country sequenced, the year, and the month collected, making 612 distinct groups from which one sequence was randomly chosen out of each group. The resultant multi-FASTA was aligned using MAFFT on the Nextstrain command line interface version 1.16.7 (Katoh and Standley 2013; Hadfield et al. 2018). A maximum likelihood tree was constructed with 1000 SH-aLRT bootstraps using a GTR substitution model via iqtree-2.1.1-Windows (Nguyen et al. 2015). The tree was visualized on iTOL and annotated with the iTOL annotation editor (Letunic and Bork 2019).

Identifying Lineages

To identify PANGOLIN lineages, the PANGOLIN command line tool 2.0.8 was used in legacy mode, relying upon the 05/29/2020 update of the guide tree to assign lineages to local sequences via bootstrapping. The browse function of the GISAID database was used to count the lineages present in New York State. United States and global data were retrieved from SARS-CoV-2 lineages (cov-lineages.org) (Rambaut et al. 2020).

ADDITIONAL INFORMATION

Data Deposition and Access

All sequences generated in this study have been made publicly available through the GISAID hCoV-19 sequence database. The source code used for sequencing, analysis, and figure generation is hosted on Github at https://github.com/kellylab/genomic-surveillance-of-the-bronx.

Ethics Statement

Remnant nasopharyngeal swabs were collected and deidentified at Montefiore Medical Center. This work was approved by the Institutional Review Board of Albert Einstein College of Medicine under IRB number 2016-6137, and informed consent was waived because the study was retrospective and involved no more than minimal risk to subjects.

Acknowledgments

We thank Isabel Gutierrez, Estefania Valencia, and Laura Polanco for laboratory management and technical assistance and the Chandran and Kelly laboratories for helpful comments on the manuscript. We thank Nextstrain, GISAID, and all laboratories who contributed SARS-CoV-2 sequences for public access. We thank the health-care workers and patients of the Montefiore Healthcare System.

Author Contributions

L.K., K.C., J.P.D., D.Y.G., H.K., and D.L.G. designed the study. S.K., R.F., J.M.F., L.K., K.C., and A.B. collected and analyzed the data. K.A.S., S.S., A.S.F., W.B.M., L.R.W., R.H.B. III, M.E.D., C.F., D.H., R.K.J., E.L., A.S.W., and J.B. provided clinical and experimental support. L.K., K.C., J.M.F., S.K., and R.F. wrote the paper. All authors edited the paper.

Competing Interest Statement

K.C. is a member of the scientific advisory boards of Integrum Scientific, LLC and the Pandemic Security Initiative of Celdara Medical, LLC.

Received March 18, 2022; accepted in revised form June 24, 2022.

Funding

L.K. is supported in part by a Peer Reviewed Cancer Research Program Career Development Award from the United States Department of Defense (CA171019) and a grant from the Ullmann Family Foundation. Computational resources were supported by an award from the Google Cloud Research Credits program (GCP19980904) to L.K. S.K. is supported by the Einstein Medical Scientist Training Program (2T32GM007288-45) and by a National Institutes of Health (NIH) T32 Fellowship in Geographic Medicine and Emerging Infectious Diseases (2T32AI070117-13). K.A.S. is supported by an NIH F30 Fellowship (F30CA200411) and T32 Fellowship (T32GM007288).

REFERENCES

Abu-Raddad LJ, Chemaitelly H, Malek JA, Ahmed AA, Mohamoud YA, Younuskunju S, Ayoub HH, Al Kanaani Z, Al Khal A, Al Kuwari E, et al. 2020. Assessment of the risk of SARS-CoV-2 reinfection in an intense re-exposure setting. *Clin Infect Dis* **73:** e1830–e1840. doi:10.1093/cid/ciaa1846

ARTIC Network. 2019. ARTIC Network SARS-CoV-2 sequencing (available at https://artic.network/ncov-2019).

ARTIC Network. 2020. nCoV-2019 novel coronavirus bioinformatics protocol (available at https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html).

Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, Barbian K, Judson SD, Fischer ER, Martens C, et al. 2020. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell* **183:** 1901–1912.e9. doi:10.1016/j.cell.2020.10.049

Elfien J. 2022. *New York: COVID-19 case rate by borough*. Statista. https://www.statista.com/statistics/1109817/coronavirus-cases-by-borough-new-york-city

Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, et al. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369:** 297–301. doi:10.1126/science.abc1917

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34:** 4121–4123. doi:10.1093/bioinformatics/bty407

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30:** 772–780. doi:10.1093/molbev/mst010

Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, Adams G, Fink T, Tomkins-Tinch CH, Krasilnikova LA, et al. 2021. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371:** eabe3261. doi:10.1126/science.abe3261

Letunic I, Bork P. 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucl Acids Res* **47:** W256–W259. doi:10.1093/nar/gkz239

Lowman N, Rowe W, Rambaut A. 2019. *Arctic Network*. Arctic.network. https://arctic.network/ncov.2019/ncov2019-bioinformatics-sop.html

Maurano MT, Ramaswami S, Zappile P, Dimartino D, Boytard L, Ribeiro-Dos-Santos AM, Vulpescu NA, Westby G, Shen G, Feng X, et al. 2020. Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region. *Genome Res* **30:** 1781–1788. doi:10.1101/gr.266676.120

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32:** 268–274. doi:10.1093/molbev/msu300

Ogawa J, Zhu W, Tonnu N, Singer O, Hunter T, Ryan AL, Pao GM. 2020. The *D614G* mutation in the SARS-CoV2 Spike protein increases infectivity in an *ACE2* receptor dependent manner. bioRxiv doi:10.1101/2020.07.21.214932

Quick J. 2020. *Arctic Network*. Arctic.network. https://artic.network/ncov-2019

Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, et al. 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc* **12:** 1261–1276. doi:10.1038/nprot.2017.066

Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5:** 1403–1407. doi:10.1038/s41564-020-0770-5

Sagulenko P, Puller V, Neher RA. 2018. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol* **4:** vex042. doi:10.1093/ve/vex042

Sevillano G, Ortega-Paredes D, Loaiza K, Zurita-Salinas Z, Zurita J. 2021. Evidence of SARS-CoV-2 reinfection within the same clade in Ecuador: a case study. *Intl J Infect Dis* **108:** 53–56. doi:10.1016/j.ijid.2021.04.073

Toh ZQ, Anderson J, Mazarakis N, Neeland M, Higgins RA, Rautenbacher K, Dohle K, Nguyen J, Overmars I, Donato C, et al. 2022. Comparison of seroconversion in children and adults with mild COVID-19. *JAMA Netw Open* **5:** e221313. doi:10.1001/jamanetworkopen.2022.1313

To KK, Hung IF, Ip JD, Chu AW, Chan WM, Tam AR, Fong CH, Yuan S, Tsoi HW, Ng AC, et al. 2020. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 isolate confirmed by whole genome sequencing. *Clin Infect Dis* doi:10.1093/cid/ciaa1275

Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, Laverdure C, Verma SC, Rossetto CC, Jackson D, et al. 2021. Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect Dis* **21:** 52–58. doi:10.1016/S1473-3099(20)30764-7

Wan Y, Shang J, Graham R, Baric RS, Li F. 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol* **94:** e00127-20. doi:10.1128/JVI.00127-20

World Health Organization. 2020. Weekly epidemiological update—29 December 2020 (available at https://www.who.int/publications/m/item/weekly-epidemiological-update—29-december-2020).