Title: Genomic surveillance of SARS-CoV-2 in the Bronx enables clinical and epidemiological inference

Authors:

J. Maximilian Fels[1#], Saad Khan[2#], Ryan Forster[2#], Karin A. Skalina[3§], 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[1,10], Denise Haslwanter[1], Rohit K. Jangra[1], Ethan Laudermilch[1], Ariel S. Wirchnianski[1,9], Jason Barnhill[10], David L. Goldman[1,5,12*], Hnin Khine[5,11*], D. Yitzchak Goldstein[3*], Johanna P. Daily[1,4*], Kartik Chandran[1*], Libusha Kelly[1,2*]

Affiliations:

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

[#]These authors contributed equally

[§]Current affiliation: Department of Radiation Oncology, Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, NY 10461, USA

*Corresponding authors: libusha.kelly@einsteinmed.org (L.K.); kartik.chandran@einsteinmed.org (K.C.); jdaily@montefiore.org (J.P.D.); dogoldst@montefiore.org (D.Y.G.); hkhine@montefiore.org (H.K.); dagoldma@montefiore.org (D.L.G.)

Abstract:

The Bronx was an early epicenter of the COVID-19 pandemic in the USA. We conducted temporal genomic surveillance of SARS-CoV-2 genomes across the Bronx from March-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 variants, we found that while some have become 'endemic' to the Bronx, other, novel variants rose in prevalence in the late summer/early fall. Geographically resolved genomes enabled us to distinguish between a case of reinfection and a case of persistent infection. We propose that limited, targeted, temporal genomic surveillance has clinical and epidemiological utility in managing the ongoing COVID pandemic.

Significance:

The ongoing emergence of novel SARS-CoV-2 variants has highlighted the need for continual genomic surveillance in order to track their spread and limit introductions into new areas. An understanding of circulating viral strains also provides a powerful tool that can be used to make clinical inferences. Here, we employ temporally and geographically resolved sequencing of SARS-CoV-2 samples in order to describe the local landscape of viral variants in the Bronx and to differentiate between cases of re-infection and persistent infection. We propose that local and targeted sequencing of viral isolates is an underutilized approach for managing the COVID pandemic.

Main text:

COVID-19 has had a devastating effect on the health of communities across the globe, with over 79 million reported cases and greater than 1.7 million deaths since the start of the pandemic (*1*). Until vaccines become widely available, understanding and interrupting SARS-CoV-2 transmission to prevent infection is the mainstay of public health efforts. The Bronx, a borough of New York City (NYC) has sustained the second highest rate of COVID-19 in New York City with 6,035 cases per 100,000 people as of January 11, 2021 (*2*). To track the local spread of SARS-CoV-2, we conducted a genomic epidemiologic study at Montefiore Health Systems (MHS), which offers healthcare 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, we randomly selected nasopharyngeal samples that were positive for SARS-CoV-2 by RT-PCR testing at the MHS clinical laboratory between March and October 2020. 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 (*3*, *4*). The ARTIC Network bioinformatics protocol was used to quality check and annotate SARS-CoV-2 genomes with default parameterization (*5*). We called variants with the NextClade tool and annotated lineages using the constructed PANGOLIN guide tree from 05/29/2020 (*6*, *7*). Samples were derived from patients who required hospitalization (48%), 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 (**Fig. S1 and S2)**. Sequence data were derived from residents throughout the Bronx and were associated with 22 of 25 zip codes (**Fig. 1B)**. Genomic sampling

was greatest at the onset of the COVID-19 pandemic in March and April, but intermittent sampling continued as caseloads declined over the summer and fall (**Fig. 1C**).

Analysis of the resulting 104 SARS-CoV-2 genome sequences revealed that B.1 and B.1.3 lineages were the most prevalent during the early months of the pandemic in the Bronx; however, several other lineages were also present at low frequencies (**Fig. 2A**). Although B.1.3 plateaued after the first wave, B.1 continued to be sampled, and a new lineage, B.1.1, arose in late August. We observed no major differences between Bronx SARS-CoV-2 lineages and other SARS-CoV-2 lineages in NYC and New York State (NYS) (**Fig. 2B**) (*8*, *9*). We note that "A" lineage SARS-CoV-2 viruses are less prevalent in the Bronx, NYC, and NYS compared to the rest of the USA and the world. To determine how the Bronx sequences compared with 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. We found that Bronx SARS-CoV-2 sequences represented subsets of different clades of the global tree (**Fig. 2C**).

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

In the spike protein, we found amino acid variants D614G (core), N501T in 5 patients, and both N501Y and P681R in one patient. We note that P314L in Orf1b is also a core variant in our dataset, reflecting observations in other studies that this variant is in linkage disequilibrium

with D614G (*10*). We did not observe the B.1.1.7 variant strain, first identified in the United Kingdom in the fall of 2020 which also contains the N501Y variant 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 variants at this position may influence ACE2 receptor binding (*11*). In comparing Bronx variants to those found in the rest of the world, we find that some variants, such as the spike protein D614G variant, are prevalent both in our set and in the world; however, some 'core' Bronx variants such as C1059T (T265I in Orf1ab) and G25563T (Q57H in Orf3a) are not as prevalent in the rest of the world **(top bar Fig. 3A, Fig. 3C, Fig S3)**. The geographic specificity of variants creates a fingerprint that can be useful for tracing the spread of particular variants; a lineage containing the variant C2416T, linked to the Boston Biogen COVID-19 outbreak, could be traced to infections around the world (*12*). The C2416T variant was also observed in three patients in our dataset. We note that rare variants are uniformly distributed throughout the sampling period (**Fig. S4**) and further that the functional impact of these variants is not well resolved.

A phylogenetic tree of SARS-CoV-2 shows that strains collected earlier in the pandemic are distinguishable from strains collected later, suggesting that new strains are being continuously introduced into the Bronx (**Fig. 4**, inner ring, red indicates earlier samples, green newer samples). In considering the lineages of SARS-CoV-2, there was evidence of ongoing presence of some B.1 lineage-associated strains, throughout the study period, starting from the onset of pandemic until the end of the study period (**Fig. 4**, outer ring indicates lineage). We found that the B.1.1 lineage had increasing presence in the latter part of the study period and that newer B.1 strains, which cluster away from older B.1 sequences, appear at later sampling dates. Newer B.1 and B.1.1 lineages 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 are circulating locally and those that were newly introduced. We consider SARS-CoV-2 isolates that group on the downsampled global tree and group on the local Bronx

tree with our first wave pandemic sampling to be 'circulating.' We continue to observe isolates that fall into this 'first wave' clade during the summer, post-first wave, and therefore consider these to have persisted in the Bronx. We consider 'introduced' isolates those that are newer sequences in the local Bronx tree that are also spread out in different clades across the global tree; it remains to be seen if these introduced isolates will form the basis of a second set of circulating SARS-CoV-2 strains during a new wave of COVID-19 in the Bronx (**Fig. 2C and 4).**

This local phylogenetic framework of SARS-CoV-2 strains 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 10-15-y.o. female who was initially seen in April 2020 in the emergency department with 3 days of fever, sore throat, anosmia, and ageusia. SARS-CoV-2 infection in this patient was confirmed by RT-PCR. She had a total of 6 days of symptoms and was in general good health until the second presentation. In August 2020, she presented again to the emergency department with two days 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. A respiratory pathogen panel was negative but her SARS-CoV-2 RT-PCR was positive, as was her SARS-COV-2 IgM Immune Status Ratio (ISR) (2.1, with less than 1 considered negative). Her IgG ISR was negative, 8.7 (normal range ISR < 9). The patient had a total of three days of fever with complete resolution of all other symptoms by day four of illness.

The two SARS-CoV-2 genomes sequenced from this patient were 142 days apart and differed in nucleotide sequence at 17 different positions. The first and second samples from this patient fall in different local phylogenetic clades in the Bronx phylogenetic tree, supporting the hypothesis that we are observing a new infection and not prolonged shedding from the original SARS-CoV-2 infection (**Fig. 4, purple arrows**). To our knowledge, this is the first case of symptomatic reinfection in a child who had prior symptomatic SARS-CoV-2 infection. 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 case involved a 15-20-y.o. female with an incompletely characterized immunodeficiency who presented in July 2020 with an oral lesion. She had no fever, or respiratory or gastrointestinal symptoms, and had neutropenia (absolute neutrophil count 700 cells/ul). After admission for further evaluation, she was found to be SARS-CoV-2 positive. During the admission, she was intermittently febrile and neutropenic and was treated with broad spectrum antibiotics. She developed a buttock lesion that was biopsied, revealing a thrombotic vasculopathy with infarction. Due to concern that the lesion could represent COVID-19–associated vasculopathy, and in the setting of persistent fever and intermittent neutropenia, she was treated with a 10-day course of remdesivir. The patient continued to have positive nasopharyngeal swabs for SARS-CoV-2 from early July to the end of September (**Table S1 and Fig. S5**). Her 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 fall in the same clade (**Fig. 4, orange 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 for (SH-aLRT replicates) (*13*, *14*). We therefore posit that the three strains 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 strain. Other examples of persistent infection with SARS-CoV-2 have been reported, but not, to our knowledge, in children (*15*–*17*). A woman diagnosed with chronic lymphocytic leukemia who was sampled 5 times, had SARS-CoV-2 sequences displaying intrahost variation despite the SARS-CoV-2 being polytomic, similar to

what we observe here (*18*). The polytomy that encompasses this persistent case also contains independent local strains of SARS-CoV-2 that do not separate on the global tree, suggesting that some variants seen in this patient are also shared locally in the Bronx (**Fig. 2C and 4**).

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, ~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 vs. persistent infection and community-level observations of older vs. newly introduced strains. The targeted utilization of small numbers of stored swabs for temporally resolved viral genomic surveillance could thus resolve clinical questions related to persistent vs. reinfection. With the introduction and spread of recently identified United Kingdom and South African SARS-Cov-2 variants with potentially different epidemiological features from existing strains, there has been speculation that we may observe a selective sweep of existing viral genotypes in the months to follow (*19*, *20*). Temporally and geographically resolved sequencing of SARS-CoV-2 genotypes provides a background against which introduction of these or other new genotypes into our local community can be observed in real time. Given the lack of a national sequencing effort, we suggest that decentralized, small-scale sequencing coupled with rapid data sharing to public databases provides an alternative and more practical tool to monitor and curtail the introduction and spread of SARS-CoV-2.

Figures:

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)

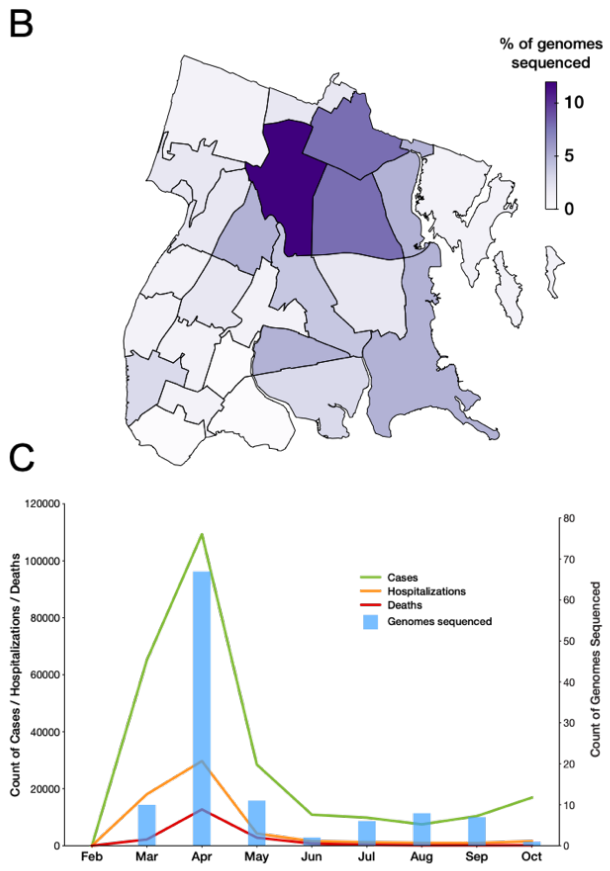


Figure 1. Surveilling SARS-CoV-2 genomes in the Bronx. A) Table of clinical characteristics of sampled patients. **B)** SARS-CoV-2 genomes sequenced per Zip code in NYC. Darker colors indicate heavier sampling; **C)** SARS-CoV-2 genomes sequenced over time during the COVID-19 pandemic. Date is indicated 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.

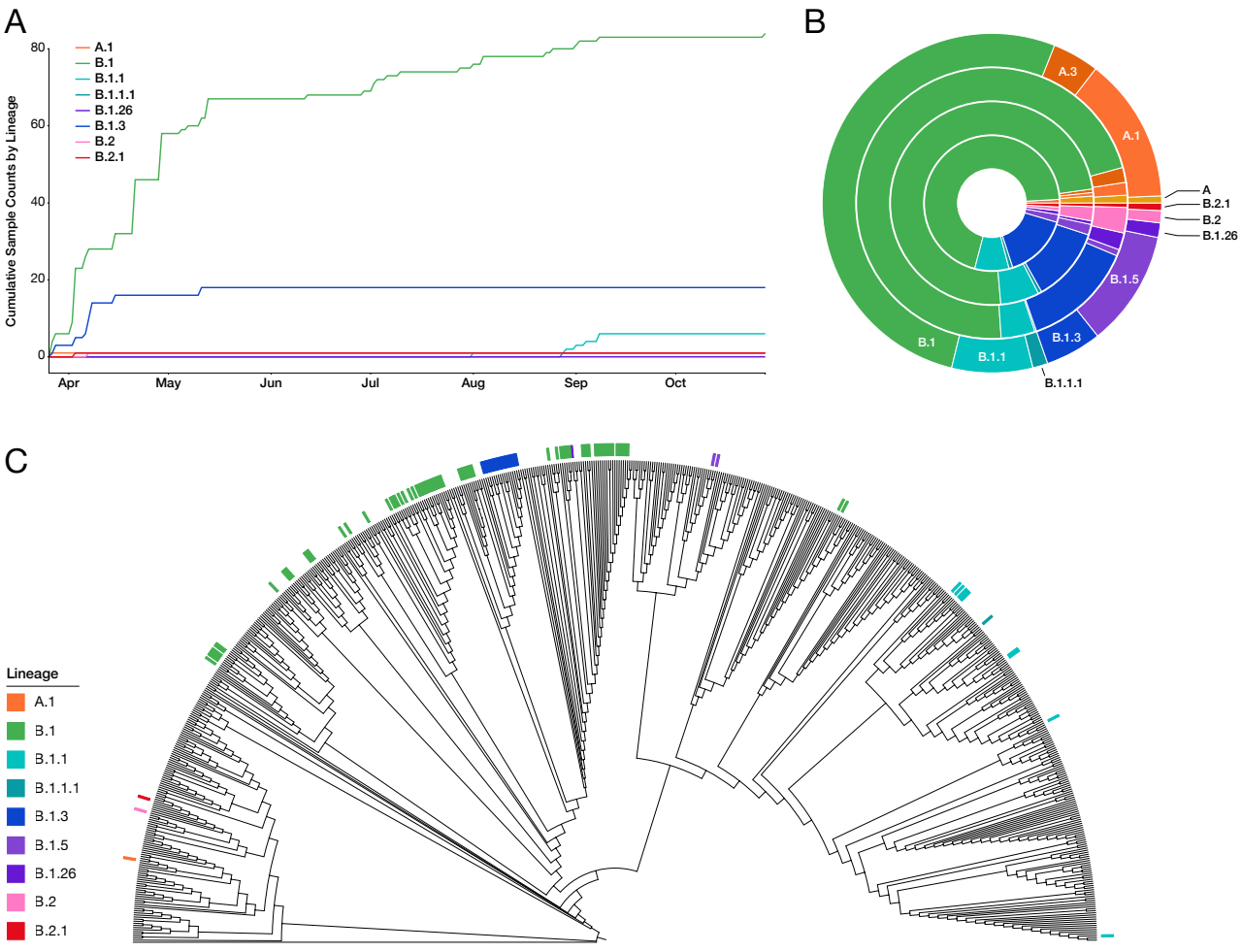
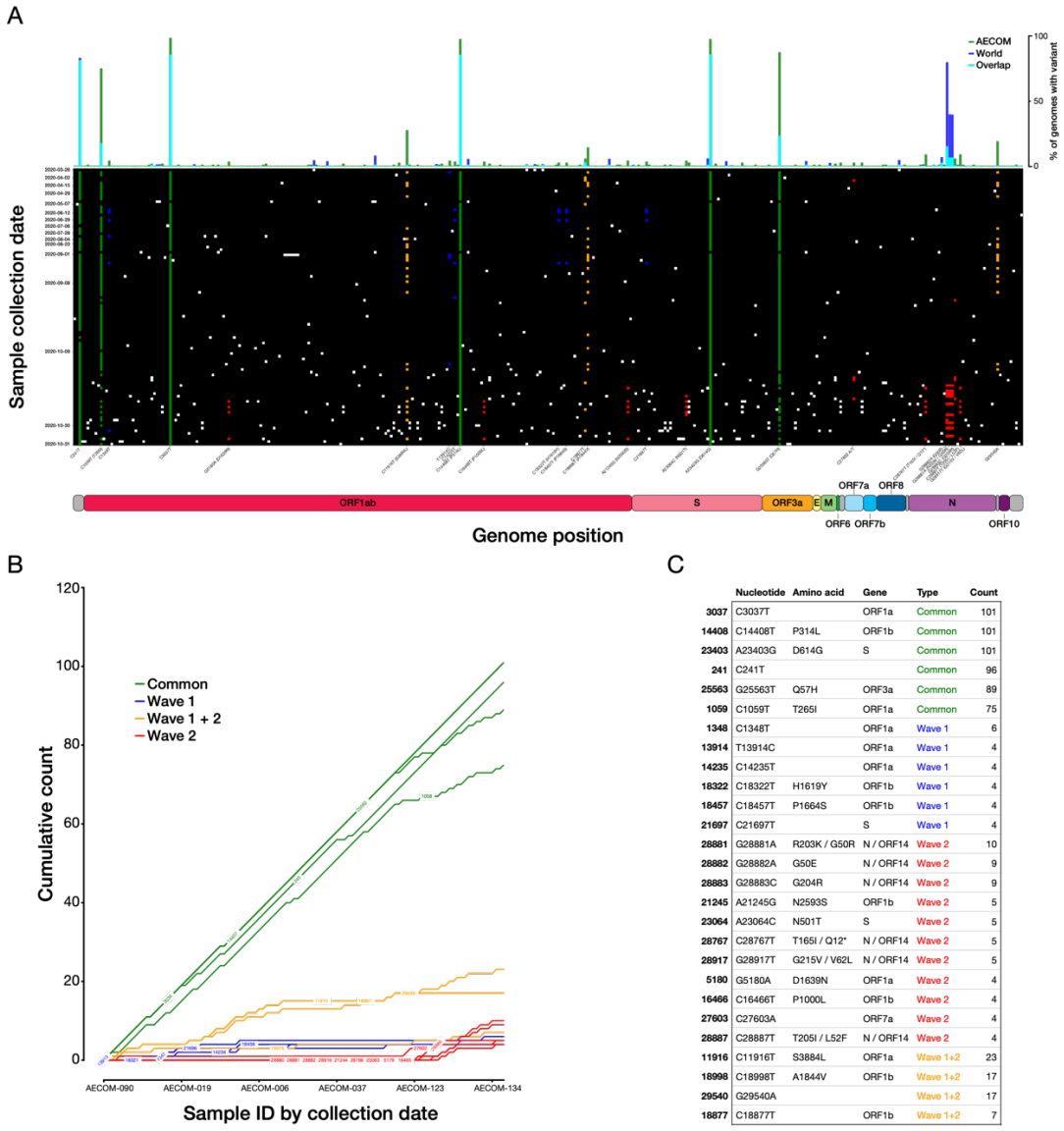


Figure 2. Bronx SARS-CoV-2 genome lineages in the context of local and global sampling. A) Cumulative counts of PANGOLIN guide tree-based lineage assignments plotted against time; **B)** Prevalence of lineages seen in the Bronx compared to their prevalence in other regions. Inner to outer rings represent the Bronx, New York State, USA, the world, respectively. Lineage coloring is the same as A); **C)** Phylogeny of the Bronx strains in the context of SARS-CoV-2 strains from around the world. Bronx strains and their associated lineages are indicated with colored lines.



	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 variants and their trajectories in the Bronx. A) Individual SARS-CoV-2 variants 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 strain 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 variant across all Bronx SARS-CoV-2 genomes in this study relative to the world; **B)** Rarefaction curve of cumulative variant counts over time for variants observed at least four times in the Bronx SARS-CoV-2 genomes set; **C)** Table showing details for variants in 3B.

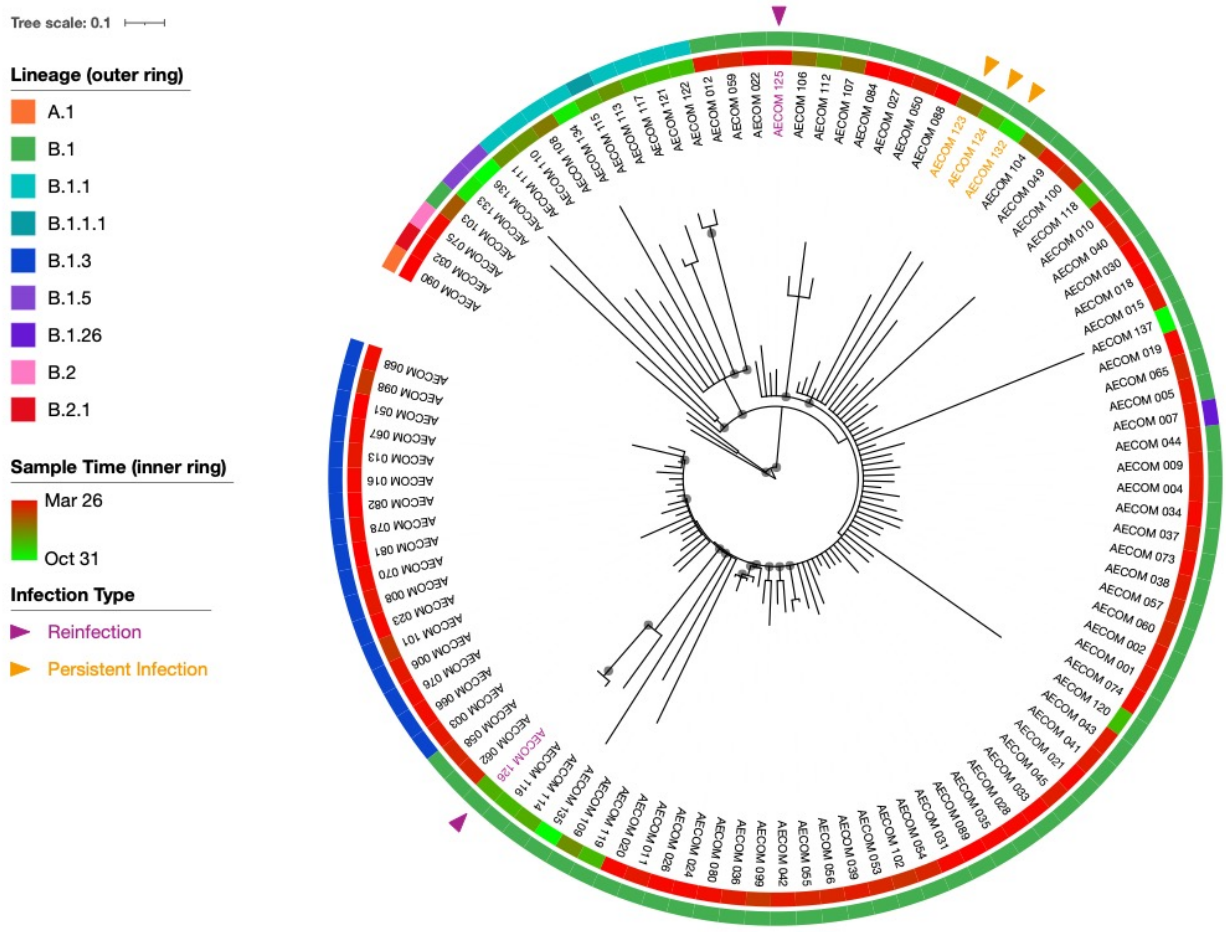


Figure 4. Clinical relevance of the changing genomic landscape of SARS-CoV-2 in the Bronx. Phylogenetic tree based on whole genome alignments of Bronx isolates. Colored rings around the tree indicate SARS-Cov-2 lineage (outer ring) and the date of sampling (inner ring, red=earlier, green=later). Samples from the same patient are indicated with symbols; a reinfection case is indicated with purple arrows and a putative persistent infection case is indicated with orange arrows. Grey circles on the branches indicate bootstrap values of 85 or greater. Tree was generated with TimeTree and visualized with iTOL (*13*, *21*)**.**

References and notes

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

2. Statista, Rates of COVID-19 cases in New York City as of January 11, 2021, by borough (2021), (available at https://www.statista.com/statistics/1109817/coronavirus-cases-rates-by-borough-new-york-city/).

3. J. Quick *et al.*, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).

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

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

6. J. Hadfield *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. **34**, 4121–4123 (2018).

7. A. Rambaut *et al.*, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).

8. A. S. Gonzalez-Reiche *et al.*, Introductions and early spread of SARS-CoV-2 in the New York City area. *Science*. **369**, 297–301 (2020).

9. M. T. Maurano *et al.*, Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region. *Genome Res.* **30**, 1781–1788 (2020).

10. J. Ogawa *et al.*, The D614G mutation in the SARS-CoV2 Spike protein increases

infectivity in an ACE2 receptor dependent manner. *BioRxiv* (2020), doi:10.1101/2020.07.21.214932.

11. Y. Wan, J. Shang, R. Graham, R. S. Baric, F. Li, Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J. Virol.* **94** (2020), doi:10.1128/JVI.00127-20.

12. J. E. Lemieux *et al.*, Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* (2020), doi:10.1126/science.abe3261.

13. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

14. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

15. K. K.-W. To *et al.*, COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. *Clin. Infect. Dis.* (2020), doi:10.1093/cid/ciaa1275.

16. L. J. Abu-Raddad *et al.*, Assessment of the risk of SARS-CoV-2 reinfection in an intense re-exposure setting. *Clin. Infect. Dis.* (2020), doi:10.1093/cid/ciaa1846.

17. R. L. Tillett *et al.*, Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect. Dis.* **21**, 52–58.

18. V. A. Avanzato *et al.*, Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell*. **183**, 1901-1912.e9 (2020).

19. H. Tegally *et al.*, Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in

South Africa. *medRxiv* (2020), doi:10.1101/2020.12.21.20248640.

20. N. G. Davies *et al.*, Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. *medRxiv* (2020), doi:10.1101/2020.12.24.20248822.

21. I. Letunic, P. Bork, Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

22. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780 (2013).

Acknowledgements:

We thank Isabel Gutierrez, Estefania Valencia, and Laura Polanco for laboratory management and technical assistance and the Chandran and Kelly labs for helpful comments on the manuscript. We thank NextStrain, GISAID, and all labs who contributed SARS-CoV-2 sequences for public access. We thank the healthcare workers and patients of the Montefiore Healthcare System. **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). S.K. is supported by the Einstein Medical Scientist Training Program (2T32GM007288-45) and by a National Institutes of Health T32 Fellowship in Geographic Medicine and Emerging Infectious Diseases (2T32AI070117-13). K.S. is supported by a National Institutes of Health F30 Fellowship (F30CA200411) and T32 Fellowship (T32GM007288). **Author contributions:** Study design: LK, KC, JPD, DG, HK, DG. Data collection and analysis: SK, RF, JMF, LK, KC, AB. Clinical and experimental support: KAS, SS, ASF, WBM, LRW, RHBIII, MED, CF, DH, RKJ, EL, ASW, JB. Wrote paper: LK, KC, JMF, SK, RF. Edited paper: all authors**. Competing interests:** KC is a member of the scientific advisory boards of Integrum Scientific, LLC and the Pandemic Security Initiative of Celdara Medical, LLC.. **Data availability:** 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.

Supplementary materials:

Materials and Methods

Figs. S1 to S5

Table S1

Supplementary Materials

Genomic surveillance of SARS-CoV-2 in the Bronx enables clinical and epidemiological inference

J. Maximilian Fels[1#], Saad Khan[2#], Ryan Forster[2#], Karin A. Skalina[3§], 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[1,10], Denise Haslwanter[1], Rohit K. Jangra[1], Ethan Laudermilch[1], Ariel S. Wirchnianski[1,9], Jason Barnhill[10], David L. Goldman[1,5,12]*, Hnin Khine[5,11]*, D. Yitzchak Goldstein[3]*, Johanna P. Daily[1,4]*, Kartik Chandran[1]*, Libusha Kelly[1,2]*

Correspondence to: libusha.kelly@einsteinmed.org (L.K.); kartik.chandran@einsteinmed.org (K.C.); jdaily@montefiore.org (J.P.D.); dogoldst@montefiore.org (D.Y.G.); hkhine@montefiore.org (H.K.); dagoldma@montefiore.org (D.L.G.)

This PDF file includes:

Material and methods

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.

Data availability

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.

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

Preparation of sequencing libraries

Sequencing libraries were prepared according to the protocol established by the ARTIC network (*3*, *4*). Briefly, cDNA was generated from viral RNA using SuperScript IV reverse transcriptase (Thermo Scientific, #18090010). 400 nt tiled amplicons were generated using the V3 primer

pool, divided into 4 sub-pools for increased efficiency. Amplification was performed using Q5 High-Fidelity polymerase (New England Biolabs, #M0491S) with cycle numbers optimized for each sub-pool. 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 approximately 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 to 20 hours 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) Basecall 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 400-700nt in length; 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 supplemental code). 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 (**Fig S2)**. We also looked for signs of biases in the base calling pipeline which 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 20x was almost always sufficient to result in a basecall being made at a given position but that the majority of positions had coverage above 400x. Thus, any biases in the pipeline are more likely to arise from biases in the nanopore sequencer itself or its basecaller 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 139676 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 139590 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 dataset containing a given variant.

Creating the local phylogenetic tree

Individual FASTA files ≥ 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 (*6*, *22*). 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 (*14*). 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 (*13*). The tree was visualized on iTOL and annotated with the iTOL annotation editor (*21*).

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 month

collected making 612 distinct groups from which 1 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 (*6*, *22*). A maximum likelihood tree was constructed with 1000 SH-aLRT bootstraps using a GTR substitution model via iqtree-2.1.1-Windows (*14*). The tree was visualized on iTOL and annotated with the iTOL annotation editor (*21*).

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) (*7*).

Commands

Cat function and set encoding:

Local: cat *.fasta > MSA_finale.txt

cat *txt > MSA_finale.fasta

Get-Content MSA_finale.fasta | Set-Content -Encoding utf8 MSA_finale_last.fasta

Global: cat *.fasta > Global_utf16.txt

cat *txt > Global_utf16.fasta

Get-Content Global_utf16.fasta | Set-Content -Encoding utf8 Global.fasta

Nextstrain Augur Align:

Local: augur align \

--sequences MSA_finale_last.fasta\

--reference-sequence config/ MN908947.3.gb \

--output aligned_MSA_finale_last.fasta \

--fill-gaps

--remove-reference

Global: augur align \

--sequences Global.fasta\

--reference-sequence config/MN908947.3.gb\

--output aligned_Global.fasta \

--fill-gaps

Output reads/ mafft specifications: using mafft to align via:

mafft --reorder --anysymbol --nomemsave --adjustdirection --thread 1 aligned_MSA_finale_last.fasta.to_align.fasta 1> aligned_MSA_finale_last.fasta 2> aligned_MSA_finale_last.fasta.log

Katoh et al, Nucleic Acid Research, vol 30, issue 14

https://doi.org/10.1093%2Fnar%2Fgkf436

Iqtree:

Local: bin\iqtree2 -s aligned_MSA_finale_last.fasta -m MFP -bb 10000 -alrt 1000

Global: bin\iqtree2 -s aligned_Global.fasta -m MFP -bb 10000 -alrt 1000

treetime:

treetime --tree MSA_finale_last.nwk --dates Swab_samples_metadata.tsv --aln aligned_MSA_finale_last.fasta --outdir Finale! --reroot AECOM_090 --clock-rate .0008 --clock-std-dev .0004 --keep-polytomies

Nextstrain Augur Filter:

```
augur filter \
  --sequences sequences.fasta \
  --metadata metadata.tsv \
  --exclude exclude.txt \
  --include include.txt\
  --output filtered.fasta \
  --group-by country year month \
  --min-length 27000 \
  --subsample-max-sequences 1000 \
  --exclude-where "division='USA' date='2020' date='2020-01-XX' date='2020-02-XX' date='2020-03-XX' date='2020-04-XX' date='2020-05-XX' date='2020-06-XX' date='2020-07-XX' date='2020-08-XX' date='2020-09-XX' date='2020-10-XX' date='2020-11-XX' date='2020-12-XX' date='2020-01' date='2020-02' date='2020-03' date='2020-04' date='2020-05' date='2020-06' date='2020-07' date='2020-08' date='2020-09' date='2020-10' date='2020-11' date='2020-12'"\
  --min-date 2019.74
```

Pangolin Command Line Tool:

pangolin MSA_finale_last.fasta --legacy --outfile Lineages.csv

Figures and guide to scripts

All figures were conjoined and post-processed to adjust colors and layout in Adobe Illustrator.

Figure 1

a) The table was generated by manually pulling EMR records for the patients whose samples came from the hospital. Not all of the fields were available for every patient.
b) The choropleth was generated by tabulating the zip codes for each of the patient samples which passed the 95% coverage threshold. We used the Python library Geopandas to then generate the choropleth using a map of the Bronx which we downloaded from NYC Open Data (https://data.beta.nyc/en/dataset/nyc-zip-code-tabulation-areas/resource/894e9162-871c-4552-a09c-c6915d8783fb).
c) This chart contains a histogram showing the distribution of samples in our dataset by month alongside three line plots showing the number of cases, hospitalizations, and deaths compiled by NYC Health (https://github.com/nychealth/coronavirus-data).

Figure 2

a) Chart of the cumulative density of lineages in our samples which passed the quality threshold as assigned using our Pangolin based method.
b) Data for this donut plot not derived from this study was acquired from SARS-CoV-2 lineages (cov-lineages.org).
c) Phylogenetic tree annotated by lineage using the online tool iTOL.

Figure 3

a) The top histogram shows the percentage of samples in the world / our dataset carrying a variant at a given position. For this histogram, we did not differentiate between different

variants showing up at a position. The heatmap was generated by ordering our samples by date of sample collection and coloring in on the x-axis wherever a variant was found on a given position. Note that the x-axis is nonlinear in the sense that only positions where a variant was found in at least one sample in our cohort are included (this also holds true for the top histogram). We labeled on the x-axis all variants which showed up at least four times in our dataset and in parentheses indicated what the amino acid changes implied by those variants were. Some of the variants were in regions where multiple open reading frames overlap, so we indicated amino acid changes for both reading frames.

For all three subfigures, we separated our variants into three categories:

i) Rare: Positions where a variant is found less than four times.

ii) Uncommon: Positions where a variant is found at least four times but less than 26 times (25% of our samples that were used in the analysis). This was further broken down into wave 1 and wave 2 categories, where a variant was considered to be ‘wave-1-associated’ if it showed up at least four times in the first half of the samples and at most once in the latter half (and vice-versa for wave 2). Samples that showed up at least once in both halves were labelled as wave 1 + 2.

iii) Common: Positions where a variant is found at least 26 times.

b) For the variants that showed up at least four times in our sampling, we constructed rarefaction curves showing the cumulative count of those variants when samples were ordered by collection date. Here, we did not have to consider if two samples might have different variants at a given position, as that scenario did not occur after we only considered uncommon and common variants. We color coded the rarefaction curves

based on if the variant was common, wave-1, wave-2, or wave-1+2. We labeled each line by variant position using a Matplotlib extension called matplotlib-label-lines.

c) This table simply shows the data in the previous figures in text format for the more frequent variants.

Figure 4

This tree was visualized using iTOL. See above for how the local phylogenetic tree was constructed.

Supplemental figures

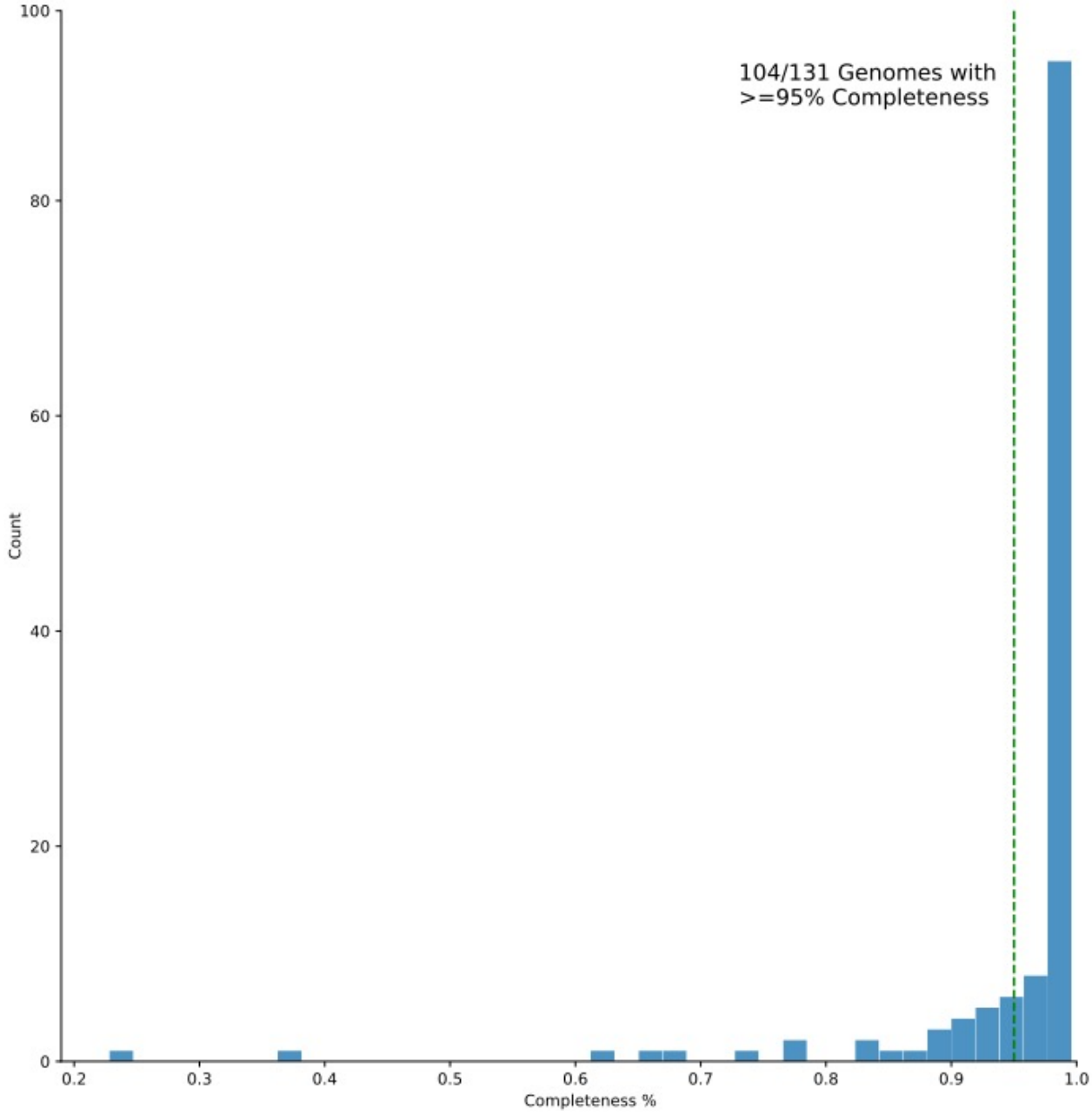


Figure S1. Histogram showing the distribution of completeness for all 131 genomes sequenced for this study.

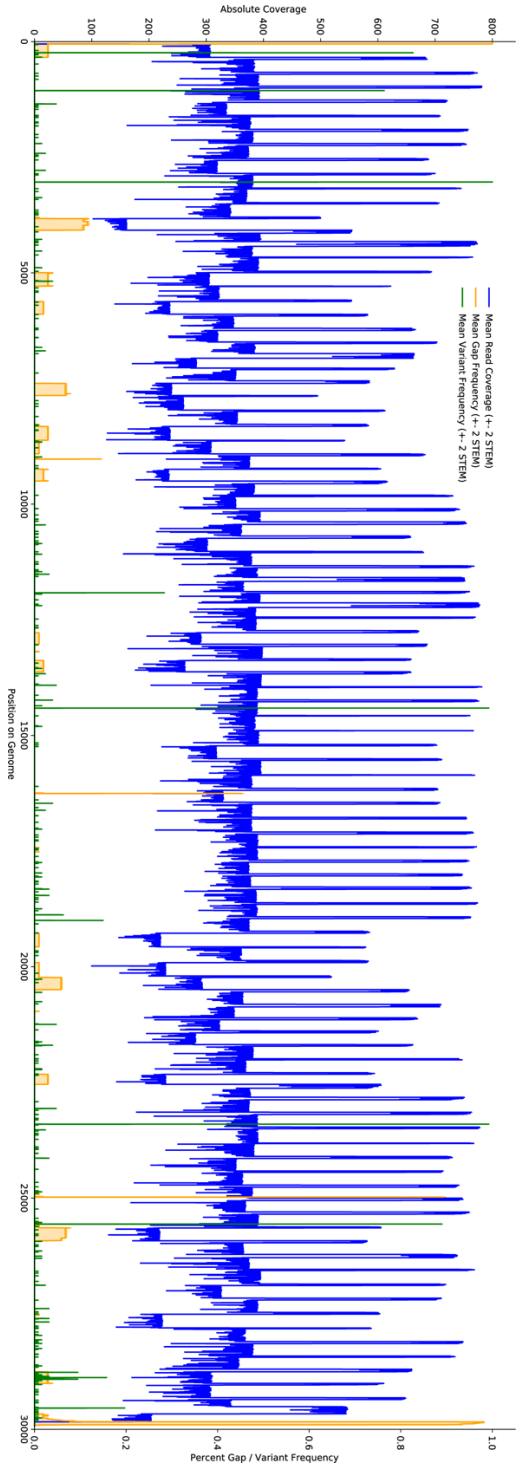


Figure S2. Visualization of coverage and gaps across all 104 included samples. In blue is the average coverage level at a given position. In yellow is the average frequency of gaps present at a given position. In green is the average frequency of variants at a given position.

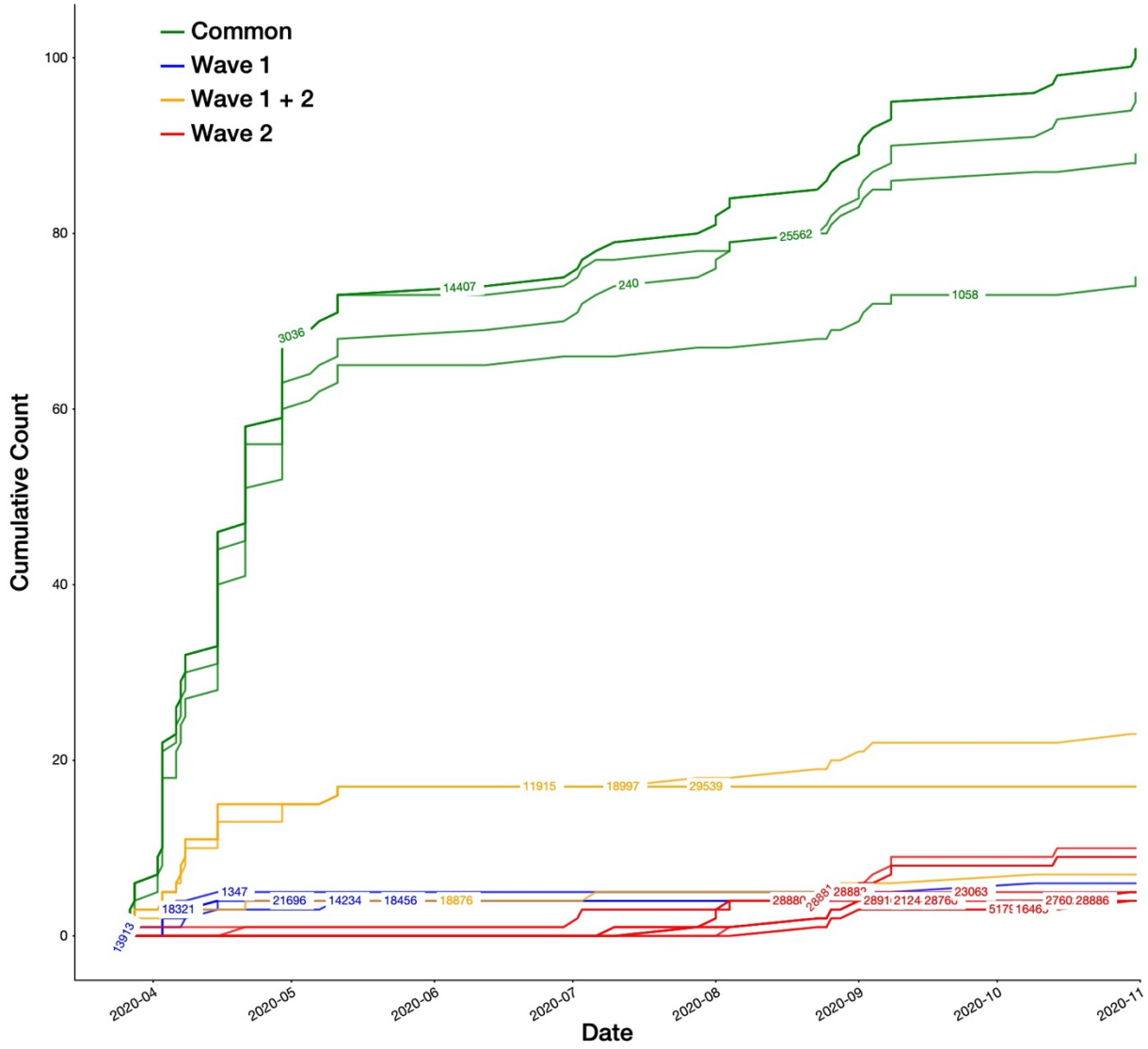


Figure S3. Distribution of cumulative counts of common and uncommon variants in the dataset. In contrast to Fig. 3B, the x-axis is ordered by date in order to demonstrate the temporal dynamics of variants in the Bronx.

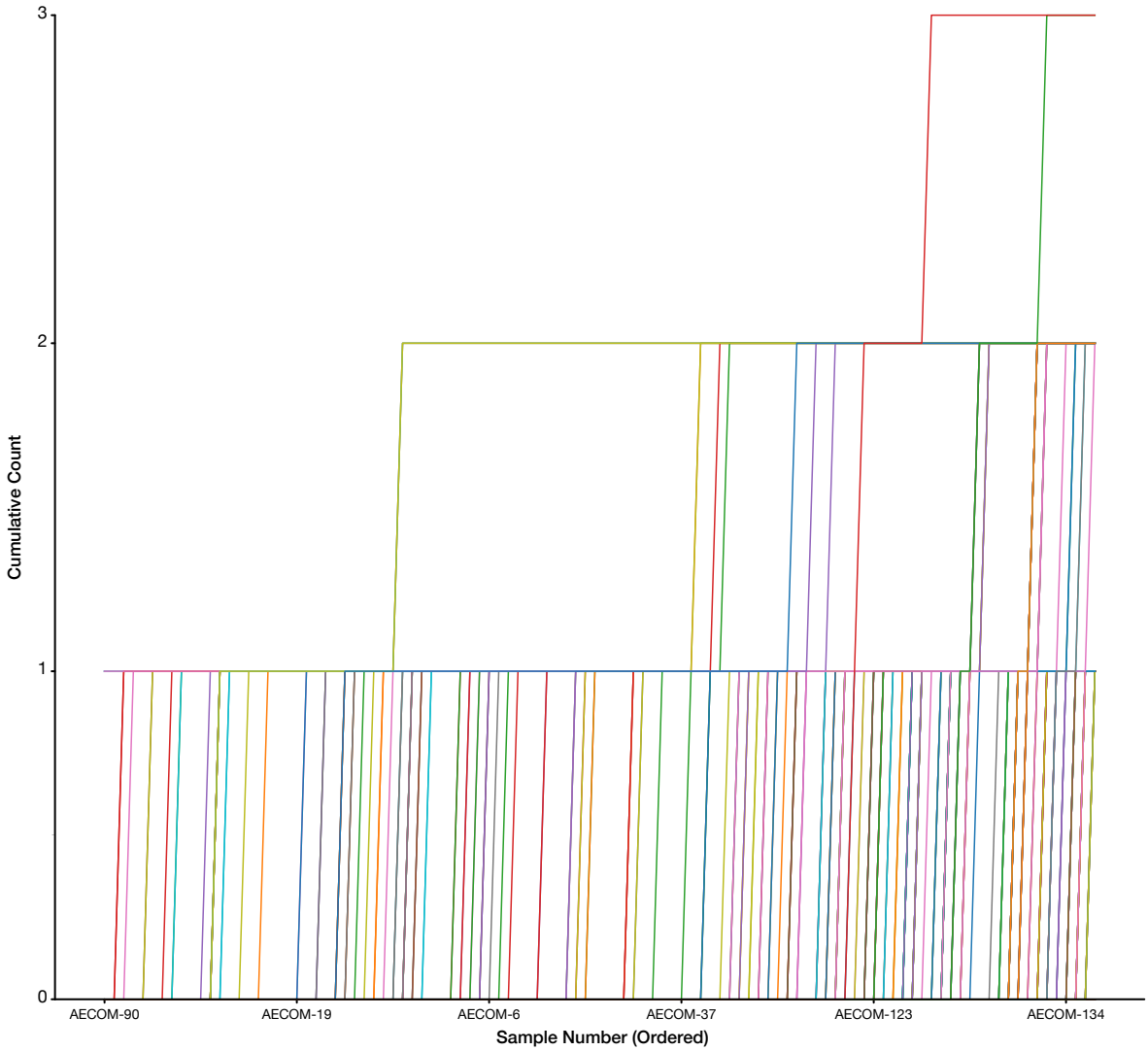


Figure S4. Rarefaction curve showing the cumulative counts of variants present less than four times total in the dataset. Samples are ordered by sample date. These rare variants were identified throughout the sampling period and do not seem to have a bias towards any particular time.

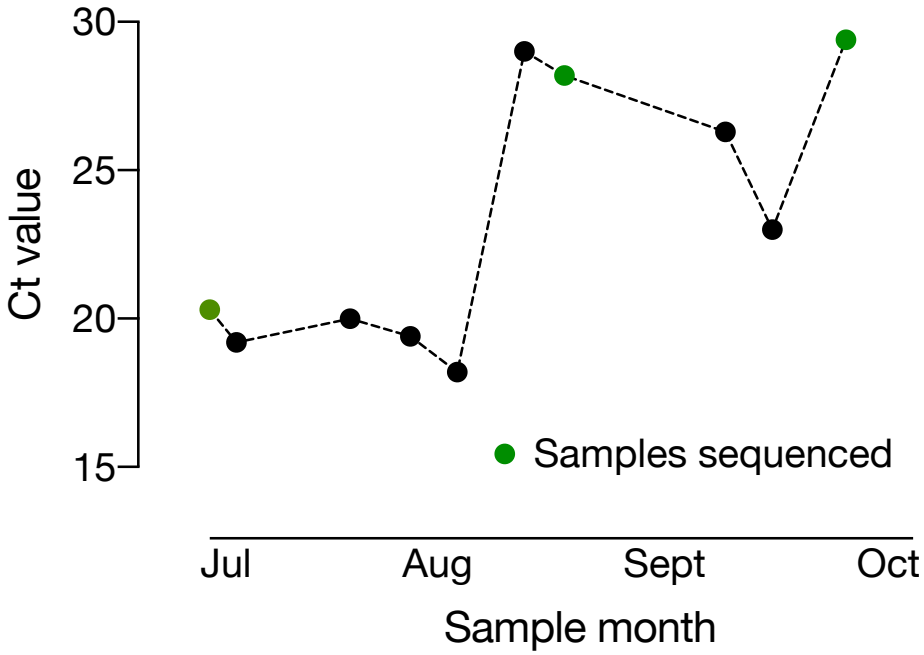


Figure S5. RT-PCR cycle thresholds of samples from Case 2 (see Table S1). Sequenced samples are highlighted in green.

	Case 1	Case 2
Age	10-15 yrs	15-20 yrs
Sex	Female	Female
Comorbidities	No	Yes, see text
Symptoms at first presentation	Fever, sore throat, anosmia, ageusia	Lip ulcer, neutropenia
Duration of first illness	6 days	Undefined, see text
Number of positive PCR tests	2	10
Symptoms at second presentation	Fever, abdominal pain, diarrhea, myalgia	See text
Admission (Y/N)	N	Y

Table S1. Patient characteristics of the re-infection (Case 1) and persistent infection (Case 2) cases.