Sequencing analysis of the spread of SARS-CoV2 in the Greater New York City Region

Matthew T. Maurano[1,2,*], Sitharam Ramaswami[3], Gael Westby[3], Paul Zappile[2,3], Dacia Dimartino[3], Guomiao Shen[2], Xiaojun Feng[2], André M. Ribeiro-dos-Santos[1,2], Nicholas A. Vulpescu[1,2], Margaret Black[2], Megan Hogan[1,2], Christian Marier[3], Peter Meyn[3], Yutong Zhang[3], John Cadley[1,2], Raquel Ordoñez[1,2], Raven Luther[1,2], Emily Huang[1,2], Emily Guzman[3], Antonio Serrano[2], Brendan Belovarac[2], Tatyana Gindin[2], Andrew Lytle[2], Jared Pinnell[2], Theodore Vougiouklakis[2], Ludovic Boytard[4], John Chen[5], Lawrence H. Lin[2], Amy Rapkiewicz[2], Vanessa Raabe[6], Marie I. Samanovic-Golden[7], George Jour[2,8], Iman Osman[4,8], Maria Aguero-Rosenfeld[2], Mark J. Mulligan[6], Paolo Cotzia[2,4], Matija Snuderl[2,*], Adriana Heguy[2,3,*]

1 Institute for Systems Genetics, NYU School of Medicine, New York, USA
2 Department of Pathology, NYU School of Medicine, New York, USA
3 Genome Technology Center, Division of Advanced Research Technologies, NYU School of Medicine, New York, USA
4 Center for Biospecimen Research and Development, NYU Langone Health, New York, USA
5 Medical Center IT, NYU Langone Health, New York, USA
6 Division of Infectious Diseases and Immunology, Department of Medicine and NYU Langone Vaccine Center, NYU School of Medicine, New York, USA
7 Department of Medicine, NYU School of Medicine, New York, USA
8 Department of Dermatology, NYU School of Medicine, New York, USA

* Correspondence:
maurano@nyu.edu
matija.sunderl@nyulangone.org
adriana.heguy@nyulangone.org

Abstract

Effective public response to a pandemic relies upon accurate measurement of the extent and dynamics of an outbreak. Viral genome sequencing has emerged as a powerful means to link seemingly unrelated cases, and large-scale sequencing surveillance can inform on critical epidemiological parameters. Here, we report analysis of 156 SARS-CoV2 sequences from individuals in the New York City metropolitan area during the initial stages of the 2020 COVID-19 outbreak. The majority of samples had no recent travel history or known exposure. Comparison to global viral sequences showed that the majority of sequences were most related to samples from Europe. Our data are consistent with numerous seed transmissions and a period of unrecognized community spreading. This work highlights the complementary role of genomic surveillance to traditional epidemiological indicators.

Introduction

In December of 2019, the novel pneumonia COVID-19 emerged in the city of Wuhan, in Hubei province, China. Shotgun metagenomics rapidly identified the SARS-CoV2 as a betacoronavirus related to the etiological agent of the 2002 SARS outbreak, SARS-CoV (Zhou et al. 2020) and of possible bat origin (Andersen et al. 2020). In the ensuing months, genomic epidemiology has been applied to track the worldwide spread of SARS-CoV2 using mutations in viral genomes to link otherwise unrelated infections (Grubaugh et al. 2019; Zhang and Holmes 2020). Rapid development of targeted sequencing protocols (Quick et al. 2017; ARTIC Network), open sharing of sequences through the GISAID (Global Initiative on Sharing All Influenza Data) repository (Shu and McCauley 2017), and rapid analysis and visualization of viral phylogenies using Nextstrain (Hadfield et al. 2018) have provided unprecedented and timely insight into the spread of the pandemic. Most notably, surveillance sequencing in the Seattle area linked a community-acquired case to a travel-related case sampled 5 weeks prior, implying widespread undetected community transmission (Bedford et al. 2020).

The New York City metropolitan region has rapidly become an epicenter of the pandemic since the first community acquired case was detected on March 3rd (a resident of New Rochelle in nearby Westchester County who worked in Manhattan). By mid-April, New York State had >200,000 cases – nearly a third of known cases in the United States, and 10% of the worldwide total. The five boroughs of New York City lead in numbers of cases in NY State (>110,000 cases), followed by Nassau and Suffolk counties to the east on Long Island (>47,000 cases) (NYS DOH). The current epicenter overlaps with the catchment area of the NYU Langone Health hospital system, including hospitals on the east side of Manhattan, one in Brooklyn (NYU Lutheran), and one in Nassau County (NYU Winthrop). We therefore aimed to characterize the urban landscape of early COVID-19 transmission as well as to identify the provenance of the circulating strains.

Results

To assess the spread of SARS-CoV2 within the NYU Langone Health (NYULH) COVID-19 inpatient and outpatient population, we implemented and optimized a viral sequencing and analysis pipeline. Samples were randomly selected from individuals testing positive between March 12 and April 1, 2020. Illumina RNA-seq libraries were generated from ribo-depleted total RNA. We experimented with several library construction approaches, ultimately selecting the robotic library construction approach used for later batches (Methods). Hybridization capture with custom biotinylated baits was then used to enrich RNA-seq libraries for viral cDNA for Illumina sequencing (**Supplementary Fig. 1**). We extend our existing informatics processing and analysis pipeline for sequencing quality control (QC), mapping of reads to the reference, identification of variants, and generation of consensus sequences (Methods). Due to the infeasibility of repeating sub-optimal libraries from remnant samples, we used a two-tiered release criteria with minimum coverage cutoffs of 6x and 30x. Most samples yielded a successful sequence, although success rates were lower for viral loads below 1000 copies/μL (**Supplementary Fig. 2a-b**). We observed that high-quality sequences could be generated directly from shotgun libraries for viral loads above 10^5 copies/μL, thereby simplifying pooling and logistical constraints for capture. Up to 18 samples at a time were multiplexed in a given capture pool (**Supplementary Fig. 2c-d**). Samples with similar viral loads were grouped to minimize the range of target cDNA in any given capture pool (**Supplementary Fig. 2e-f**).

We analyzed a set of 156 sequences passing quality control (**Fig. 1a**). Analysis of medical records found no recorded exposures for more than half the individuals (**Fig. 1c**). Samples were collected throughout the NYULH system, comprising hospitals in the New York City boroughs of Manhattan and Brooklyn, and Nassau County, a suburb to the east of the city on Long Island (**Fig. 1d**). The majority of samples derived from Brooklyn and Manhattan residents, followed by Nassau County (**Fig. 1e**). Analysis of residential ZIP codes showed that most residences were concentrated in the hospital catchment area throughout the New York metropolitan region (**Fig.** 2). Notably, there were no individuals from Westchester County to the north of the city, outside of our catchment area, where the earliest detected regional outbreak was concentrated.

We performed a phylogenetic analysis of the sampled sequences to assess their relatedness (**Fig. 3a**). Coloring individual sequences by their county of residence within the New York region showed substantial mixing within the region from the time where our sampling began, the first week in March. We then assessed their relatedness to sequences from across the world by comparison with analysis of 7,627 sequences from the GISAID EpiCov repository. Using a global phylogeny generated via Nextstrain (**Supplementary Fig. 3**), each of the sequences we generated was colored by the location of the most similar sequence not originating from the region (NY, NJ, CT; **Fig. 3b**). The most related sequence was from Europe for 41% of the samples, while for 46% of individuals the most related sequence had been sampled in the US or Canada (**Fig. 3c**). We also assessed the collection date of the most recent common ancestor broken

down by different global regions (**Fig. 3d**). Two thirds of our samples could be linked to European samples collected as early as February 24. It is important to caution that, given the paucity of sequences relative to the spread of the pandemic, these data alone do not directly implicate specific transmission events or timelines.

We detected 189 nucleotide and 97 amino acid mutations across all samples (**Fig. 4**). Notably, detection of novel variants within our sample set continued as additional samples were sequenced (**Supplementary Fig. 4**), suggesting that further surveillance by regional, national, and international groups will be needed to monitor pandemic spread and public responses (**Supplementary Fig. 5**).

Discussion

The rapid growth of viral genome sequencing, enabled by real-time worldwide sharing of these data and scalable analysis platforms support an important role of genomic epidemiology in the continuing management of the COVID-19 pandemic.

Our conclusions must be considered in the context of the limited availability of sequences from relevant timepoints and locations and the general undersampling of infected individuals worldwide. Sampling density and incomplete availability of exposure history preclude fine-scale delineation of transmission directions and routes. In particular, few sequences from the key periods of January and February, thus limiting the direct observation of transmission events. While epidemiological metrics like Influenza-like illness surveillance and clinical parameters of COVID-19-like pneumonias put a low upper limit on possible prevalence in that time frame, it is possible that screening and sequencing of archival samples may clarify the initial spread. Nevertheless, viral sequencing reflects the broad spread of SARS-CoV2 into the New York region, either directly or indirectly.

In the case of seasonal flu, virus mutations play a key role in virulence, vaccine efficacy, and oseltamivir resistance (Miller and Palese 2014; Bloom et al. 2010; Lee et al. 2019). However, given the rapid spread of SARS-CoV2, it is expected that most mutations identified thus far will not have detectable function, but rather are the result of genetic drift. Further sequencing will increase statistical power to search for signals of positive selection. Functional analysis will be required to determine whether function can be ascribed to any of these mutations, and what role mutations might play in shaping the pandemic.

On a regional level, genomic epidemiology can provide an independent data source to track transmission. For example, sampling from the NY region demonstrates a much broader diversity of transmission chains than initially uncovered in Seattle (Bedford et al. 2020). Retrospective analysis could illuminate the degree of undetected community spreading on a per-region basis, to assess the efficacy of policy and behavioral changes, and prospectively to inform proactive

management of an ongoing outbreak (Kissler et al. 2020). Two other groups have reported sequencing surveillance of nearby or overlapping regions (Fauver et al. 2020; Gonzalez-Reiche et al. 2020), and monitoring of a hospital system could be complemented by broader population studies (Kim et al. 2020). Given the logistical and regulatory hurdles to establishing such surveillance, it is critical to have this infrastructure already in place for future waves of COVID-19.

Acknowledgements

We recognize the NYULH clinicians involved in the care and testing of the patients in this study. We thank Lea Starita and the Seattle Flu Study for technical assistance and sharing their bait design. We would like to thank all the laboratories who have contributed sequences to GISAID, in particular Emilia Mia Sordillo, Viviana Simon, and Harm van Bakel (Mount Sinai School of Medicine) for contributing samples from the New York City area. The NYU Langone Genome Technology Center is partially supported by NIH Grant P30CA016087.

Data Availability

Sequences have been deposited into the GISAID repository immediately upon QC and can be visualized at http://nextstrain.org/ncov .

References

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nat Med* **26**: 450–452.

ARTIC Network, ed. *SARS-CoV-2 Protocol*. https://artic.network/ncov-2019 (Accessed April 15, 2020).

Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L, Nalla A, Pepper G, Reinhardt A, Xie H, et al. 2020. Cryptic transmission of SARS-CoV-2 in Washington State.

Bloom JD, Gong LI, Baltimore D. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328**: 1272–1275.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**: 2503–2505.

Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, Vogels CBF, Brito AF, Alpert T, Muyombwe A, et al. 2020. Coast-to-coast spread of SARS-CoV-2 in the United States revealed by genomic epidemiology.

Gonzalez-Reiche AS, Hernandez MM, Sullivan M, 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.

Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, Andersen KG. 2019. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4**: 10–19.

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.

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

Kim AE, Brandstetter E, Graham C, Heimonen J, Osterbind A, McCulloch DJ, Han PD, Starita LM, Nickerson DA, Van de Loo MM, et al. 2020. Seattle Flu Study - Swab and Send: Study Protocol for At-Home Surveillance Methods to Estimate the Burden of Respiratory Pathogens on a City-Wide Scale.

Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* eabb5793.

Lee JM, Eguia R, Zost SJ, Choudhary S, Wilson PC, Bedford T, Stevens-Ayers T, Boeckh M, Hurt AC, Lakdawala SS, et al. 2019. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife* **8**.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Miller MS, Palese P. 2014. Peering into the crystal ball: influenza pandemics and vaccine efficacy. *Cell* **157**: 294–299.

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

NYS DOH, ed. *COVID-19 Tracker*. https://covid19tracker.health.ny.gov/views/NYS-COVID19-Tracker/NYSDOHCOVID-19Tracker-Map?%3Aembed=yes&%3Atoolbar=no&%3Atabs=n (Accessed April 15, 2020).

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

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.

Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**: 741.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**.

United States Census Bureau, ed. *2018 Cartographic Boundary Files*. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html (Accessed April 15, 2020).

Zhang Y-Z, Holmes EC. 2020. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell*.

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**: 270–273.

Materials and Methods

Bioethics statement

The collection of COVID-19 human biospecimens for research has been approved by NYULH Institutional Review Board under the S16-00122 Universal Mechanism of human bio-specimen collection and storage for research at NYU Langone Health (NYULH), approved by the NYULH Institutional Review Board (IRB).

Sample collection

Nasopharyngeal (NP) swabs were collected and placed in 3 mL of Viral Transport medium (VTM, Copan Universal Transport Medium). NP samples in VTM were transported to the clinical laboratory at room temperature and tested the same day. For prolonged storage samples were placed at -70 °C.

Clinical tests

Testing was done in the Roche Cobas 6800 platform and the Cepheid Xpert Xpress as per manufacturer instructions. Both platforms employ real time RT-PCR technology. While Roche Cobas targets Orf 1/a and E sequences, Xpert Xpress amplifies N2 and E viral sequences. The limits of detection are 100-200 copies/mL for Roche and 250 copies/mL for Xpert Xpress.

Following RNA extraction (see below), a second confirmatory test was carried out on an ABI7500 Dx system, using the US CDC primer design, targeting three regions of the virus nucleocapsid (N) gene. An additional primer/probe targeting the human RNAse P gene (RP) was included serving as the internal control (https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html)

RNA extraction

We used the Maxwell RSC instrument (Promega, cat. AS4500) with the buccal swab DNA kit (Promega, cat. AS1640) to extract total RNA from 300 μL of viral transport medium. The following modifications were introduced to extract total RNA as opposed to total nucleic acids: samples were incubated at 65 °C for 30 min for proteinase K digestion and virus deactivation, and DNase I (Promega) was added to the reagents cartridge to remove genomic DNA during nucleic acids extraction. Total RNA was eluted in 50 μL of nuclease-free water.

Library preparation and sequencing

10 μL of extracted total RNA was used as input material to prepare Illumina sequencing libraries. Two methods for RNA-seq (cDNA) library preps were used, both based on a ribodepletion approach:

1. KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche Kapa catalog number KK8561). We followed the manufacturer's protocol, with the following modifications: for the adapter ligation step, we prepared a plate of IDT DNA UMI dual adapters at a concentration of 500 nM, and for the final PCR amplification of libraries, we ran 15 cycles.

2. Nugen Trio with human rRNA depletion (Tecan Genomics, cat No. 0606-96). Briefly the RNA library prep consists of the following steps: DNase treatment to remove any DNA, cDNA synthesis from the input RNA, single primer isothermal amplification (SPIA) of the resultant cDNAs, enzymatic fragmentation and construction of unique barcoded libraries followed by PCR library amplification (6 or 8 cycles were used, depending on input amount), and an AnyDeplete step to remove host rRNA transcripts. Later libraries were generated through an automated protocol on a Biomek FX^P Liquid handler integrated with a Biometra TRobot 96-well thermal cycler (Beckman Coulter). We found that, especially for lower viral inputs, the Nugen Trio kit had a lower rate of duplicate reads. Based on these results, we selected this approach as the baseline.

Purified libraries were quantified using qPCR (Kapa Biosystems, Kapa Biosystems, KK4824). Library size distribution was checked using the Agilent TapeStation 2200 system.

We used a combination of shotgun metagenomics and hybrid capture approaches according to the viral load. RNA-seq libraries from samples with viral load >100,000 copies/μL were sequenced without further enrichment for the viral sequences. Libraries from samples with <100,000 viral copies/μL were enriched for SARS-CoV2 genomic sequences using custom biotinylated DNA probe pools either from Twist Biosciences or Integrated DNA Technologies. In general, we pooled samples with viral loads within the same order of magnitude and accounting for variations in parent library concentration, multiplexing up to 18 libraries per reaction.

We followed the manufacturer's protocol for capture using the xGen COVID Capture Panel (Integrated DNA Technologies, Product# 10006764): https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/protocol/xgen-hybridization-capture-of-dna-libraries.pdf?sfvrsn=ab880a07_16. Hybridization of 500 ng to 1 μg of combined libraries with 4 μL of XGen Lockdown probes was carried out at 65 °C from 4-16 h. Post-capture PCR amplification cycles ranged from 6-10 for the Kapa libraries and 6-9 for the Nugen libraries, depending on the final captured library amount.

Samples were sequenced as paired end 100 or 150 reads on the NextSeq 500 or NovaSeq 6000 (using SP or S1 flowcells).

Sequenced Read Processing

All RNA-seq and Capture-seq data were processed using a uniform mapping and peak calling pipeline. Illumina sequencing adapters were trimmed with Trimmomatic v0.39 (Bolger et al. 2014). Reads were aligned to a custom index containing human genome reference (GRCh38/hg38) including unscaffolded contigs and alternate references plus the reference SARS-CoV2 genome (NC_045512v2, wuhCor1) using BWA (Li and Durbin 2009). PCR duplicates were marked using samblaster (Faust and Hall 2014). Variants were called using `bcftools mpileup --redo-BAQ --adjust-MQ 50 --gap-frac 0.05 --max-depth 10000 --max-idepth 200000` followed by `bcftools call --ploidy 1 --keep-

`alts --multiallelic-caller --variants-only`. Viral sequences were generated from VCF files based on the wuhCor1 reference using `bcftools consensus`; regions below 6x (for low-coverage) or 30x (for high-coverage) samples were masked with Ns.

Phylogenetic Analysis

Sequences were downloaded from GISAID EpiCov on April 13, 2020 and were analyzed using augur v7.0.2 with the Nextstrain pipeline as follows (Hadfield et al. 2018): The data were filtered to remove duplicate samples, incomplete sequences, samples with improperly formatted metadata, and samples known to be highly divergent. Sequences were then aligned to the reference genome using MAFFT v7.453 (Katoh and Standley 2013), and the resulting alignment was masked to remove the sequence ends and uninformative point mutations across all samples.

Phylogenetic tree reconstruction was initially performed with IQ-TREE v1.6.12 (Nguyen et al. 2015) using a generalized time reversible model, and refined by TreeTime v0.7.4 (Sagulenko et al. 2018) to produce a timetree rooted at the reference sequence. This tree was used to tabulate nucleotide mutations specific to lineages and samples, as well as the corresponding amino acid changes. In reporting mutations, gaps with respect to the reference were reported as deletions rather than missing data.

Using the ape R package (Paradis et al. 2004), we identified for each of our samples the nearest sequence from outside the New York region by looking for the minimal path length from our samples to all others. For analysis of ancestral collection dates, we recursively identified the most recent ancestor from a given geographical location, excluding those for which the branch length exceeds a limit of 0.1.

Geoplotting

The sample occurrence heat map was generated using R v.3.6.2 using the packages ggplot2 v.3.3.0 for plotting, and sf v.0.8.0 for geospatial data manipulation. Maps were generated based on the 2018 Zip code tabulated area geographical boundaries obtained from the United States Census Bureau (United States Census Bureau).

Figures

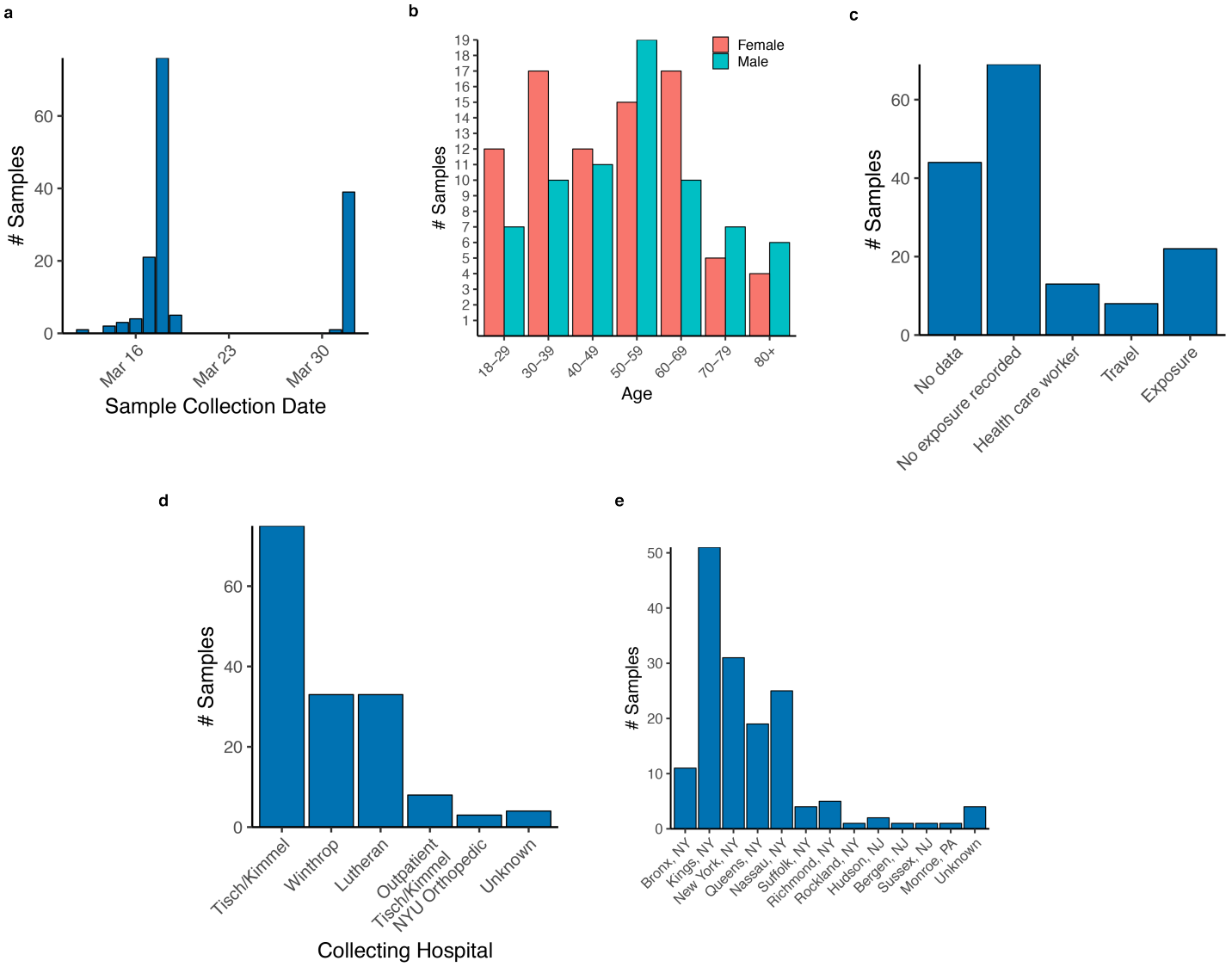

Fig. 1. Summary of study population.

Sample count by a. nasopharyngeal swab collection date; b. age (binned by decade) and sex distribution; c. potential exposure status, including occupation as health care worker, travel history, and contact with a COVID-19 positive individual; d. collecting hospital; and, e. borough/county of residence. Kings County (Brooklyn); New York County (Manhattan); Nassau County (Long Island); Richmond, NY (Staten Island).

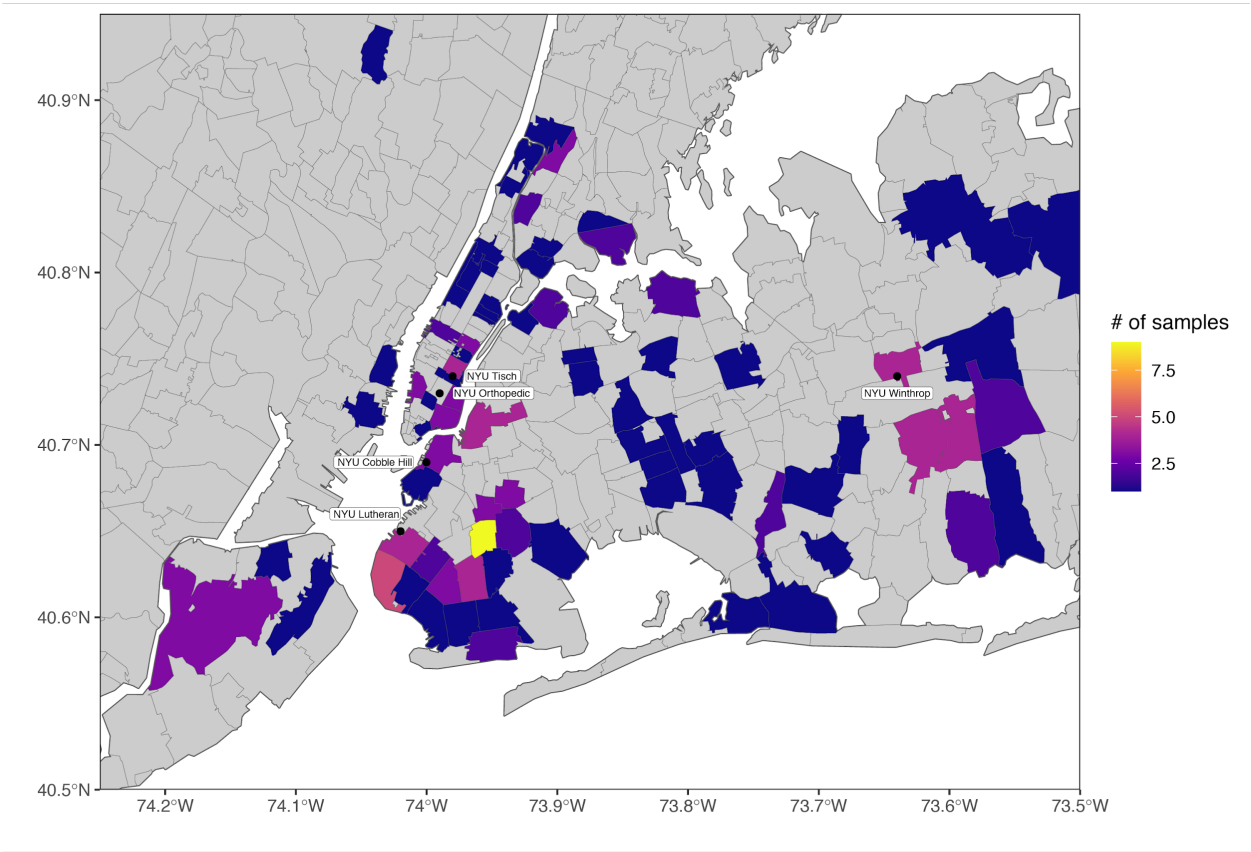

Fig. 2. Geographic summary of sequenced samples.
Samples are aggregated by residential ZIP code. Not shown are 8 individuals residing outside the plotted area, and 4 of unknown residence. Collecting hospitals are indicated in rounded boxes.

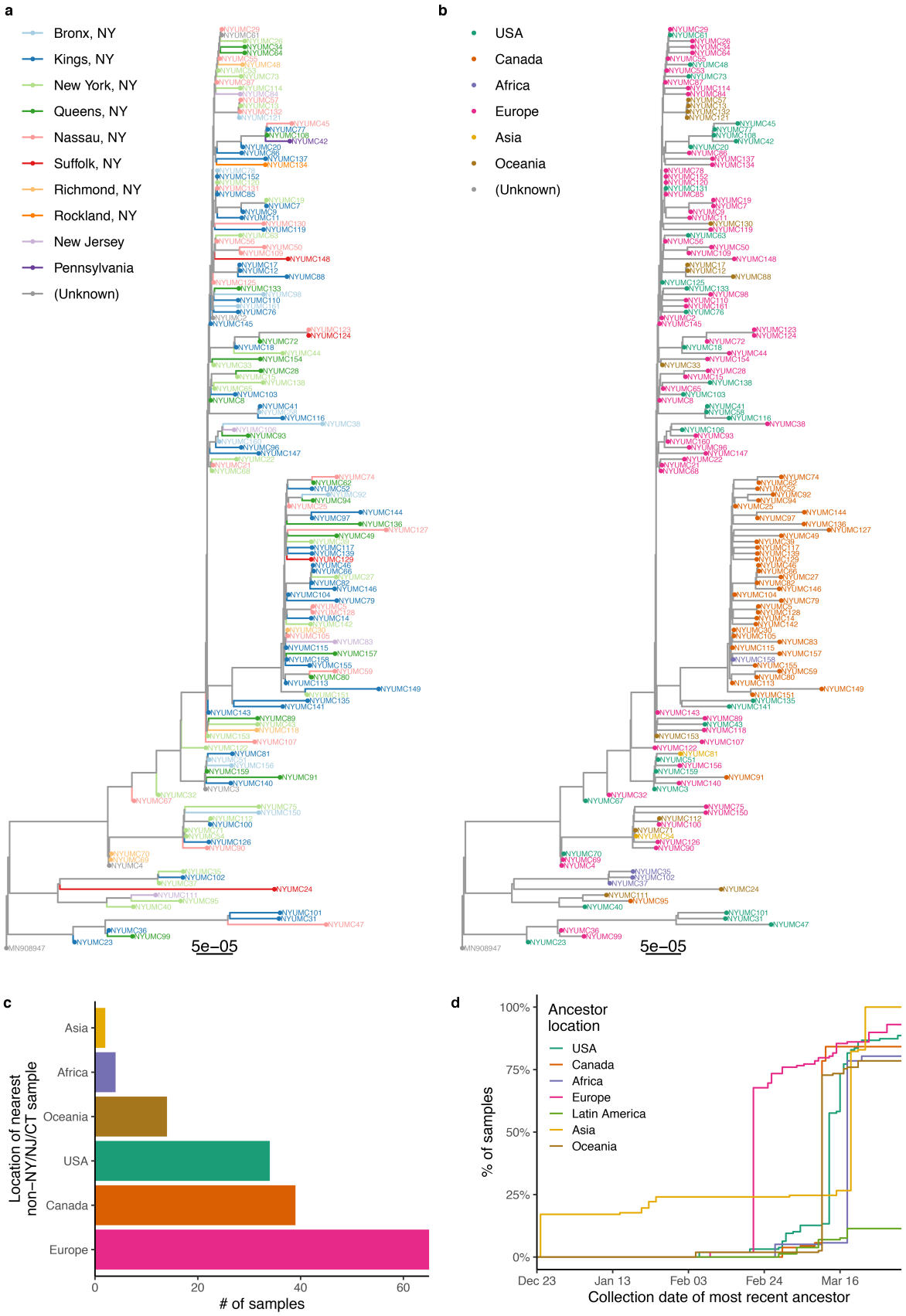

Fig. 3 Phylogenetic relationship of regional viral sequences. a-b. Phylogeny constructed from 156 samples, colored by (a) residence or (b) the location of the nearest sample from outside the New York, New Jersey and Connecticut area. c. Counts of samples matched to each location in (b). d. Cumulative plot showing collection date of most recent ancestral sample from a given region.

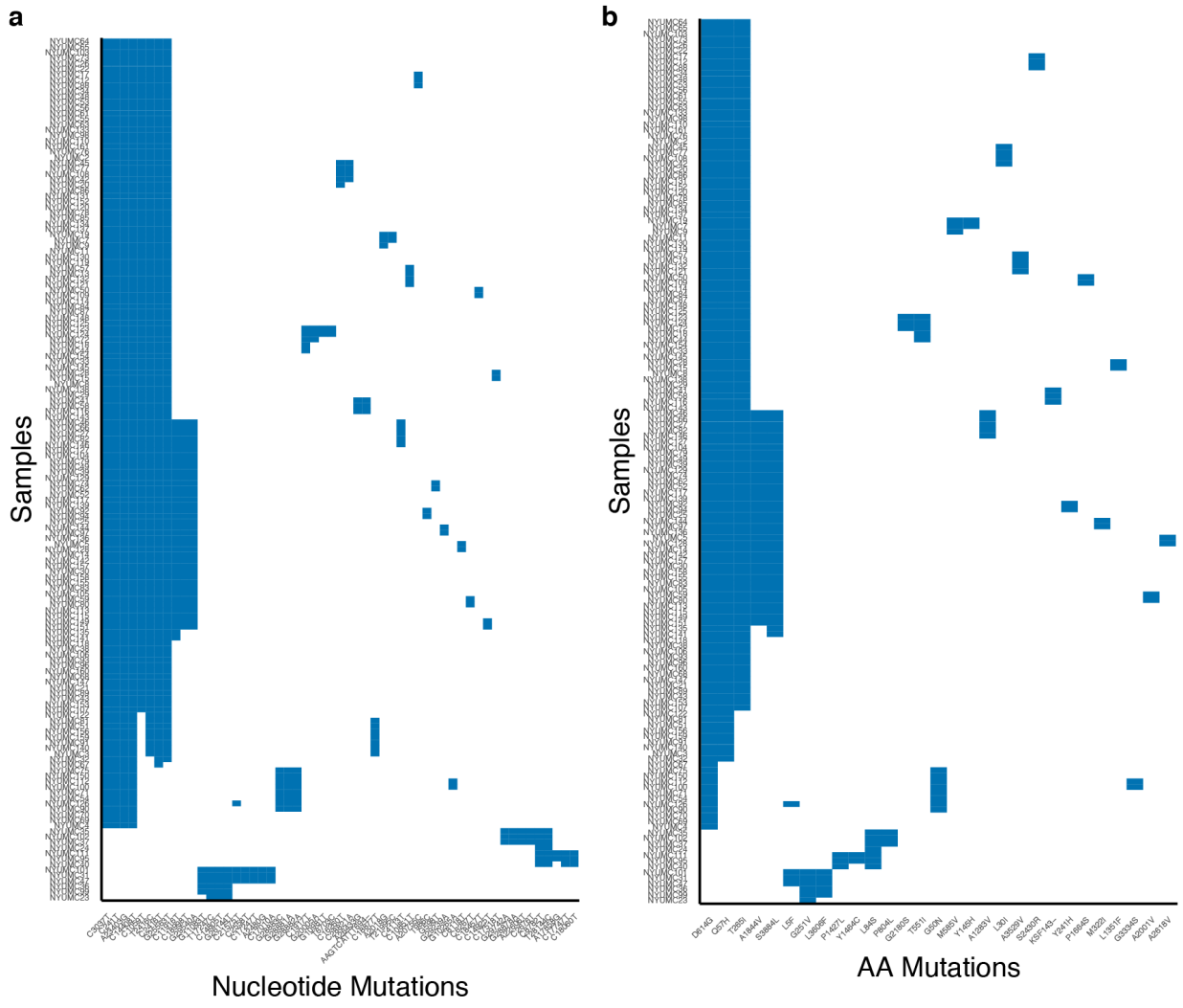

Fig. 4. Mutation heatmap.

Instances of a) nucleotide and b) amino acid mutations across all samples. Mutations found in only a single sample have been excluded for legibility.

Supplemental Figures

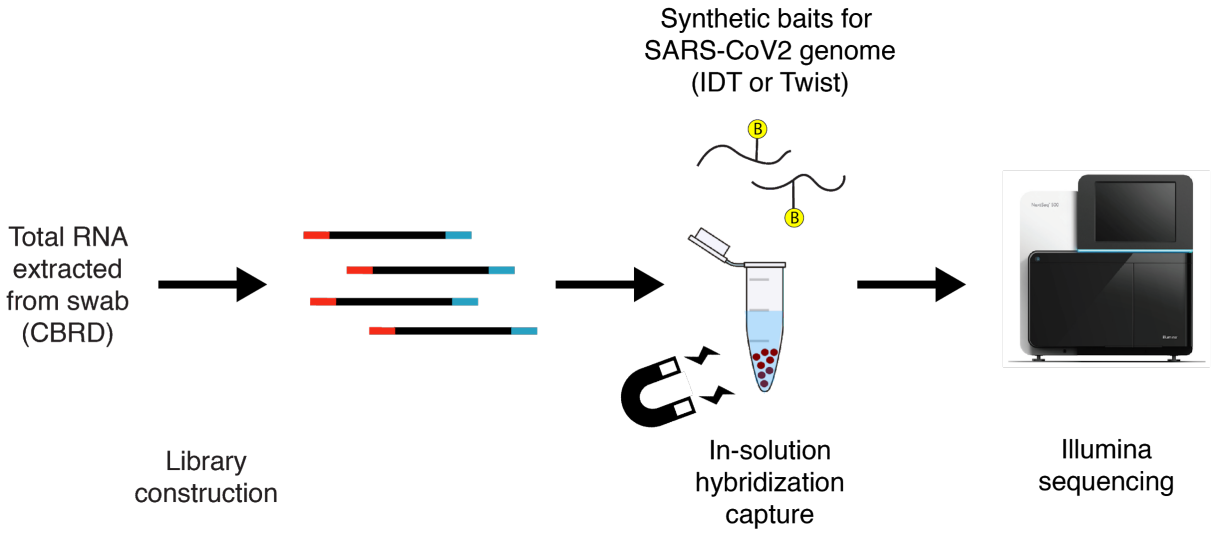

Fig. S1. SARS-CoV2 sequencing workflow.
Total RNA was extracted using high throughput extractors, at the Center for Biorepository Specimen and Development (CBRB) at NYU Langone Health. Ribodepletion RNA-seq library preps were used to prepare libraries (Kapa Hyper Riboerase or Nugen Tri) followed by enrichment for the viral genome sequences using hybridization-based capture baits designed against the Wuhan SARS-CoV2 RefSeq. For samples with viral load >100,000 copies/μL, we skipped the hybridization capture enrichment and sequenced the RNA-seq libraries prep directly. Samples were sequenced on the Illumina NovaSeq 6000 or NextSeq 500.

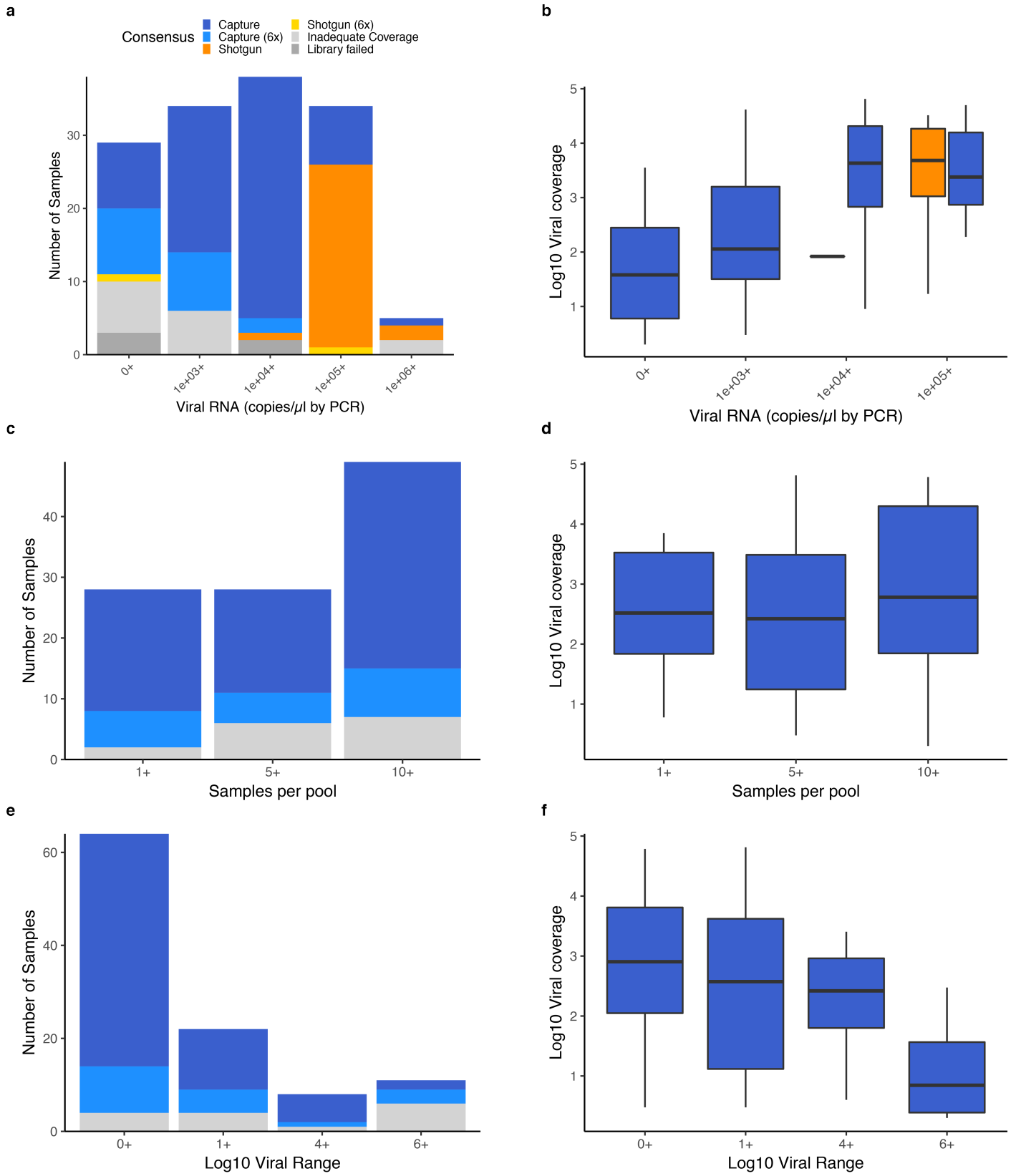

Fig. S2. Technical factors related to sequencing success rate.

Shown are sample counts by final QC outcome (left) and average coverage (right) by viral load (a-b), size of capture pool (c-d), and range of viral load among samples in the same capture pool (e-f).

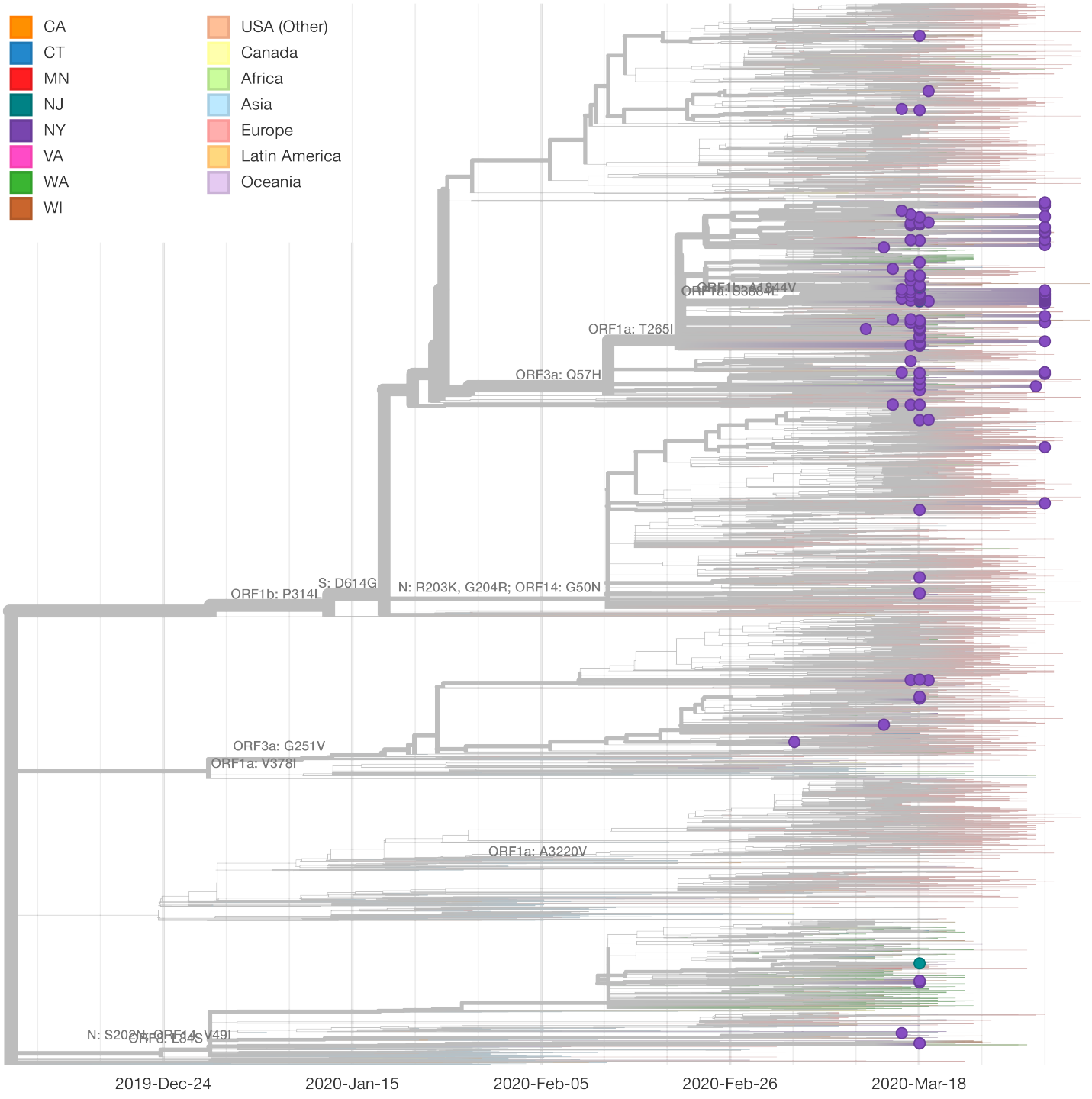

Fig. S3. Phylogenetic analysis of 7,627 SARS-CoV2 sequences from the GISAID EpiCov repository via Nextstrain. Tips highlighted with dots are samples from this study. Samples and edges are colored by geographical location.

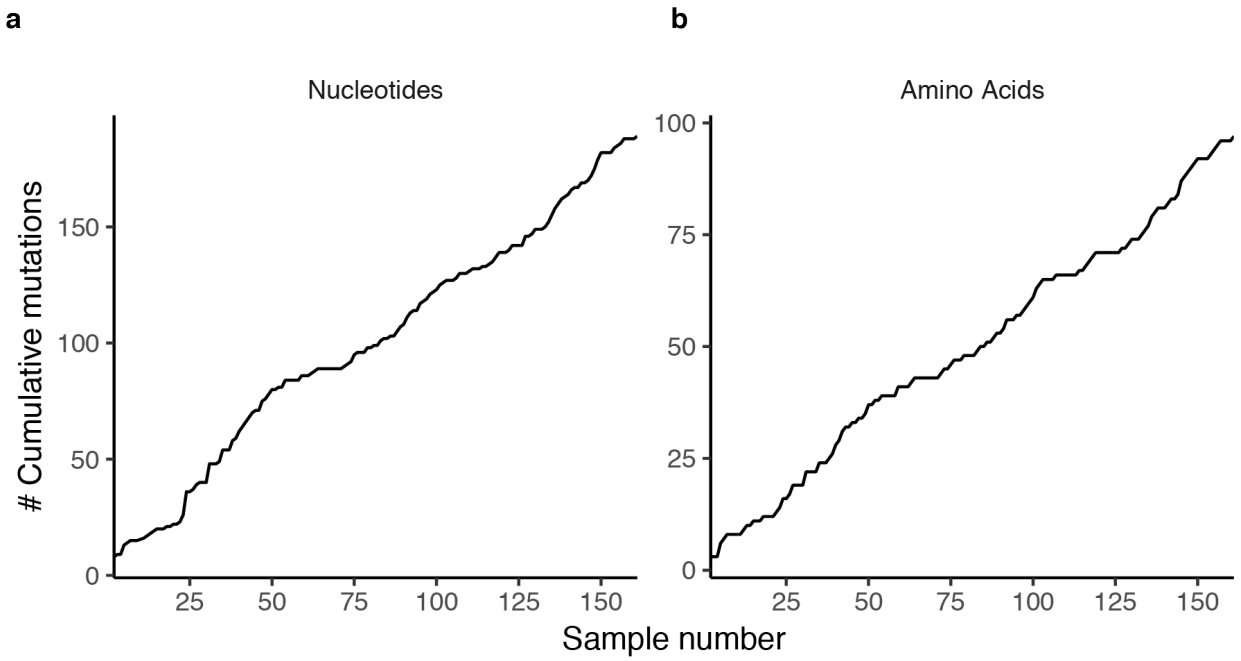

Fig. S4. Cumulative number variants detected per sample.
Number of unique (a) nucleotide and (b) amino acid mutations as a function of the total number of samples sequenced.

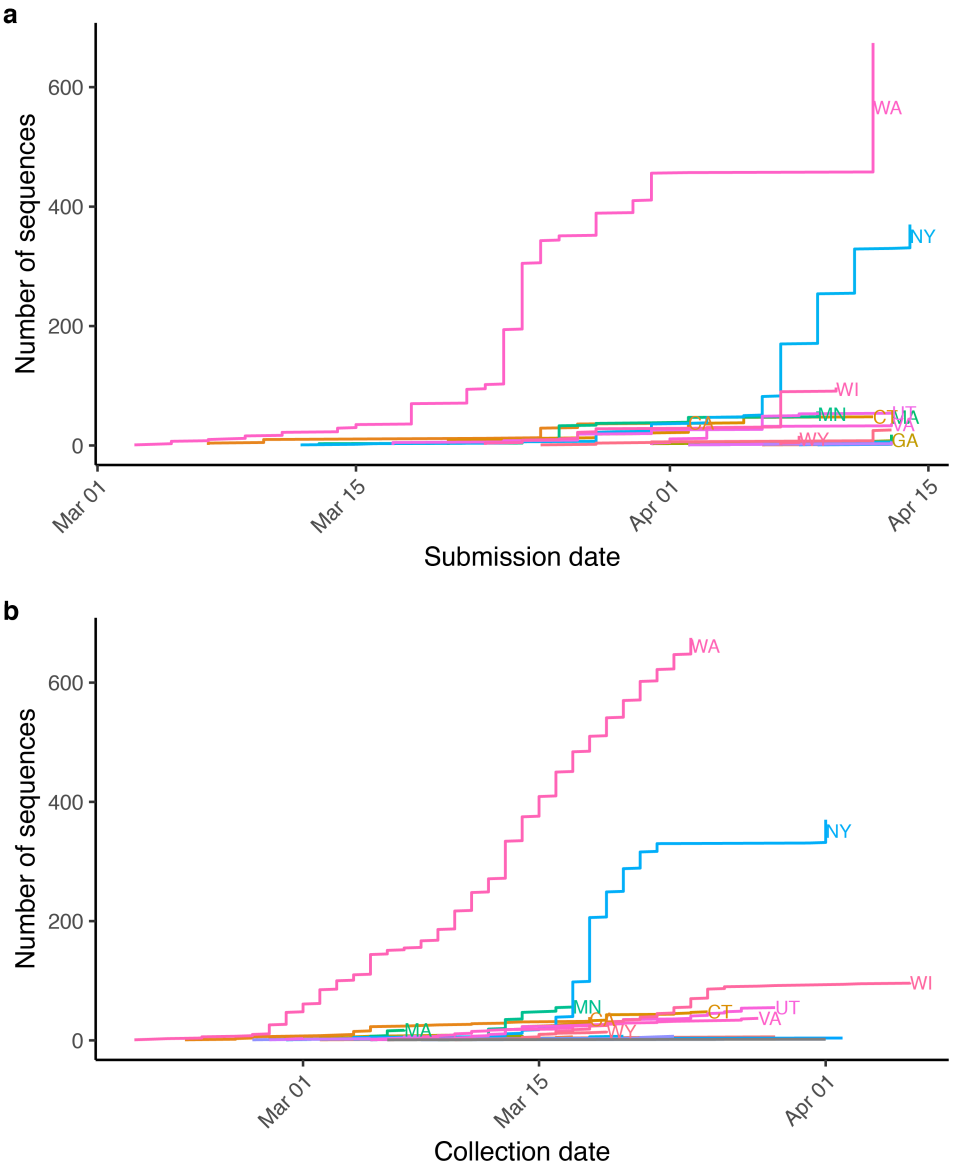

Fig. S5. US SARS-CoV2 sequences in GISAID by state.
Shown is a summary of the GISAID EpiCov repository by (a) submission date, and (b) collection date.