Sequencing identifies multiple, early introductions of SARS-CoV2 to New York City Region

Matthew T. Maurano[1,2,*], Sitharam Ramaswami[3], Gael Westby[3], Paul Zappile[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 S. 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.snuderl@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 approach to link seemingly unrelated cases, and large-scale sequencing surveillance can inform on critical epidemiological parameters. Here, we report the analysis of 236 SARS-CoV2 sequences from cases in the New York City metropolitan area during the initial stages of the 2020 COVID-19 outbreak. The majority of cases throughout the region had no recent travel history or known exposure, and genetically linked cases were spread throughout the region. Comparison to global viral sequences showed that the majority were most related to cases from Europe. Our data are consistent with numerous seed transmissions from multiple sources and a prolonged period of unrecognized community spreading. This work highlights the complementary role of real-time genomic surveillance in addition 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 new pathogen as SARS-CoV2, a betacoronavirus related to the etiological agent of the 2002 SARS outbreak, SARS-CoV and of possible bat origin[1,2]. 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[3,4]. Rapid development of targeted sequencing protocols[5,6], open sharing of sequences through the GISAID (Global Initiative on Sharing All Influenza Data) repository[7], and rapid analysis and visualization of viral phylogenies using Nextstrain[8] have provided unprecedented and timely insight into the spread of the pandemic. 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[9].

The New York City metropolitan region has rapidly become an epicenter of the pandemic since the first community acquired case was detected on March 3, 2020 (a resident of New Rochelle in nearby Westchester County who worked in Manhattan). As of April 17, New York State had 222,284 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 (123,146 cases), followed by Nassau and Suffolk counties to the east on Long Island (51,954 cases)[10]. 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). Since even early COVID-19 cases diagnosed in our healthcare system had no travel history to countries with existing epidemics, determining the extent of asymptomatic community spread and transmission paths became paramount. In parallel with increased clinical capacity for PCR based testing, we sought to trace the origin of NYULMC-treated COVID-19 cases using phylogenetic analysis to compare to previously deposited

COVID-19 viral sequences. We further aimed to characterize the urban landscape of early COVID-19 transmission as well as to identify the provenance of the circulating strains. Lastly, we highlight the importance of sharing viral genomic data in real time given the critical need to match public health measures to viral spread[11].

Results

To assess the spread of SARS-CoV2 within the NYU Langone Health (NYULH) COVID-19 inpatient and outpatient population, we deployed and optimized a viral sequencing and analysis pipeline. Illumina RNA-seq libraries were generated from rRNA-depleted total RNA. We experimented with several library construction approaches, ultimately selecting an automated library construction approach used for later batches (Methods). Hybridization capture using custom biotinylated baits was used to enrich RNA-seq libraries for viral cDNA for Illumina sequencing (**Extended Data Fig. 1**). We extended our existing informatics processing and analysis pipeline which performs 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 (**Extended Data 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 by skipping the capture step. Up to 18 samples at a time were multiplexed in a given capture pool (**Extended Data Fig. 2c-d**). Samples with similar viral loads were grouped to minimize the range of target cDNA in any given capture pool (**Extended Data Fig. 2e-f**).

We analyzed a set of 236 sequences passing quality control. Samples had been randomly selected from cases confirmed positive between March 12 and April 6, 2020 and included a range of ages (**Fig. 1a-b**). Analysis of the clinical database found no recorded exposures for more than 60% of the cases (**Fig. 1c**). In fact, travel history was present in only 7% of analyzed cases. Cases originated 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 cases were residents of Brooklyn and Manhattan, followed by Nassau County (**Fig. 1e**). Analysis of residential ZIP codes showed that most were concentrated in the hospital catchment area throughout the New York metropolitan region (**Fig. 1f**). Notably, our dataset included no cases from Westchester County to the north of the city, outside of the NYULH catchment area, where the earliest detected regional outbreak was concentrated.

We performed a phylogenetic analysis to assess relatedness among cases (**Fig. 2a**). The diversity of mutational variation within cases suggested multiple introductions followed by substantial community spreading. We detected 282 nucleotide and 181 amino acid mutations across all cases, and identify those distinguishing relevant clades (**Extended Data Fig. 3**). Coloring cases

by county of residence within the New York region confirmed substantial mixing within the region from the onset of our sampling in the first week in March (**Fig. 2a**). Cases from April were generally linked to prior cases sampled in the New York area.

We then assessed relatedness of our cases to 9,918 sequences contributed from across the world to the GISAID EpiCov repository. Using a global phylogeny generated via Nextstrain (**Extended Data Fig. 4**), each case was assigned the location of the most similar sequence not originating from the region (**Fig. 2a**). The cases most related to our cohort were from Europe for 38% of our cases, while for 46% of cases the most related case had been sampled in the US or Canada (**Fig. 2b**). We also assessed the collection date of the most recent common ancestor broken down by different global regions (**Fig. 2c**). By this analysis, 67% of our cases could be linked to cases collected and analyzed in Europe as early as February 24. Given the paucity of early sequences relative to the scale of the pandemic, the genomic data alone do not implicate individual transmission events. However, the high relatedness of New York cases to European cases, in addition to the earlier spread of the pandemic to Europe, strongly suggest that the New York outbreak was seeded largely by way of Europe. Notably, detection of novel variants within our sample set continued apace as additional cases were sequenced (**Extended Data Fig. 5**), suggesting that further surveillance by regional, national, and international groups will be needed to monitor pandemic spread and public responses (**Extended Data Fig. 6**).

Discussion

The COVID-19 pandemic has quickly engulfed most of the world. The rapid growth of viral genome sequencing, enabled by real-time worldwide sharing of genomic data and scalable analysis platforms, supports an important role of genomic epidemiology in the continuing management of the COVID-19 pandemic. Here we show that the COVID-19 outbreak in NYC region originated from multiple independent origins and most of the viral sequences in our dataset show relationship to samples previously detected in Europe. Furthermore, a large majority of our samples had no travel history, confirming the major contribution of undetected community spread to the pandemic.

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

Real-time deposition of genomic sequences from multiple sources in the public domain provides valuable information about the source of infections and asymptomatic spread.

In the case of seasonal flu, virus mutations play a key role in virulence, vaccine efficacy, and oseltamivir resistance[12-14]. Given the rapid spread of SARS-CoV2, it is presumed 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 functional changes can be ascribed to any of these mutations, and what role mutations might play in shaping the ongoing pandemic.

On a regional level, genomic epidemiology provides 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[9]. 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[15]. Two other groups have reported sequencing surveillance of nearby or overlapping regions[16,17], and monitoring of a hospital system could be complemented by broader population studies[18]. 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 or other future pandemics. Real-time genomic tracking of the viral spread should also help inform policy decisions in future outbreaks.

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020).
2. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26,** 450–452 (2020).
3. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4,** 10–19 (2019).
4. Zhang, Y.-Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* (2020). doi:10.1016/j.cell.2020.03.035
5. Quick, J. *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).
6. *SARS-CoV-2 Protocol*. Available at: https://artic.network/ncov-2019. (Accessed: 15 April 2020)
7. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22,** (2017).
8. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34,** 4121–4123 (2018).
9. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. (2020). doi:10.1101/2020.04.02.20051417
10. *COVID-19 Tracker*. Available at: https://covid19tracker.health.ny.gov/views/NYS-COVID19-Tracker/NYSDOHCOVID-19Tracker-Map?%3Aembed=yes&%3Atoolbar=no&%3Atabs=n. (Accessed: 15 April 2020)
11. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* eabb4218 (2020). doi:10.18637/jss.v027.i08
12. Miller, M. S. & Palese, P. Peering into the crystal ball: influenza pandemics and vaccine efficacy. *Cell* **157,** 294–299 (2014).
13. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328,** 1272–1275 (2010).
14. Lee, J. M. *et al.* Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife* **8,** (2019).
15. Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* eabb5793 (2020). doi:10.1111/j.1467-9868.2005.00503.x
16. Fauver, J. R. *et al.* Coast-to-coast spread of SARS-CoV-2 in the United States revealed by genomic epidemiology. (2020). doi:10.1101/2020.03.25.20043828
17. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. (2020). doi:10.1101/2020.04.08.20056929
18. Kim, A. E. *et al.* 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. (2020). doi:10.1101/2020.03.04.20031211
19. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
21. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30,** 2503–2505 (2014).
22. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: im-

provements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
23. Nguyen, L.-T., Schmidt, H. A., Haeseler, von, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
24. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4,** 741 (2018).
25. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
26. *2018 Cartographic Boundary Files*. Available at: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html. (Accessed: 15 April 2020)

Materials and Methods

Bioethics statement.

The collection of COVID-19 human biospecimens for research has been approved by NYU Langone Health (NYULH) Institutional Review Board under the S16-00122 Universal Mechanism of human bio-specimen collection and storage for research. The approved IRB protocol allows for the collection and analysis of clinical and demographic data. A clinical database was compiled for 110 of the cases.

Sample collection.

Nasopharyngeal swabs were collected and placed in 3 mL of Viral Transport Medium (VTM, Copan Universal Transport Medium). Samples were transported to the clinical laboratory at room temperature and tested for SARS-CoV2 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. The 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.

For selected positive samples, we extracted total RNA from 300 μL of viral transport medium using the Maxwell RSC instrument (Promega, cat. AS4500) with the buccal swab DNA kit (Promega, cat. AS1640). 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 cDNA RNA-seq 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 flow cells).

Sequenced read processing.
Reads were demultiplexed with Illumina bcl2fastq v2.20 requiring a perfect match to indexing BC sequences. 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[19]. 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 v0.7.17[20]. PCR duplicates were

marked using samblaster v0.1.24[21]. Variants were called using bcftools v1.9 using the command `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 19, 2020 and were analyzed using augur v7.0.2 following the Nextstrain pipeline as follows[8]: the data were filtered to remove duplicate individuals, incomplete sequences, samples with improperly formatted metadata, and highly divergent samples. Sequences were then aligned to the reference genome using MAFFT v7.453[22], and the resulting alignment was masked to remove the sequence ends and uninformative point mutations across all sequences.

Phylogenetic tree reconstruction was initially performed with IQ-TREE v1.6.12[23] using a generalized time reversible model, and refined by TreeTime v0.7.4[24] to produce a timetree rooted at the reference sequence. This tree was used to tabulate nucleotide mutations specific to lineages and cases, 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. Trees were plotted with the ggtree R package.

Using the ape R package[25], we identified for each cases the nearest sequence from outside the New York region, defined as minimal path length. 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 regional case heat map was generated using R v3.6.2 using the packages ggplot2 v3.3.0 for plotting, and sf v0.8 for geospatial data manipulation. Maps were generated based on the 2018 Zip code tabulated area geographical boundaries obtained from the United States Census Bureau[26].

Acknowledgements
We are indebted to the NYULH clinicians and laboratory personnel 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) and the NYS DOH Wadsworth Center for contributing sequences from the New York City area; a full list of is provided as supplementary data. This work was partially funded by NIH grants P30CA016087 (NYU Langone Genome Technology

Center), UM1AI148574 (M.J.M), and R35GM119703 to (M.T.M.). This work was partially supported by the NYULH Office for Science and Research.

Author Contributions

M.T.M., M.J.M., P.C., M.S. and A.H. conceived and supervised the study. G.S., X.F., M.B., A.S., B.B., T.G., A.L., J.P., T.V., L.B., L.H.L., A.R., V.R., M.I.S., G.J., I.O., M.A., M.J.M, P.C. and M.S. collected clinical samples and data. S.R., G.W., P.Z., D.D., M.S.H., P.M., Y.Z. and A.H. generated sequencing data. C.M., J.C., E.G. and J.C. contributed informatics tools. M.T.M, A.M.R., N.A.V., M.S.H, R.O., R.L., E.H. and A.H. performed the data analysis. M.T.M, M.S. and A.H. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

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

Figures

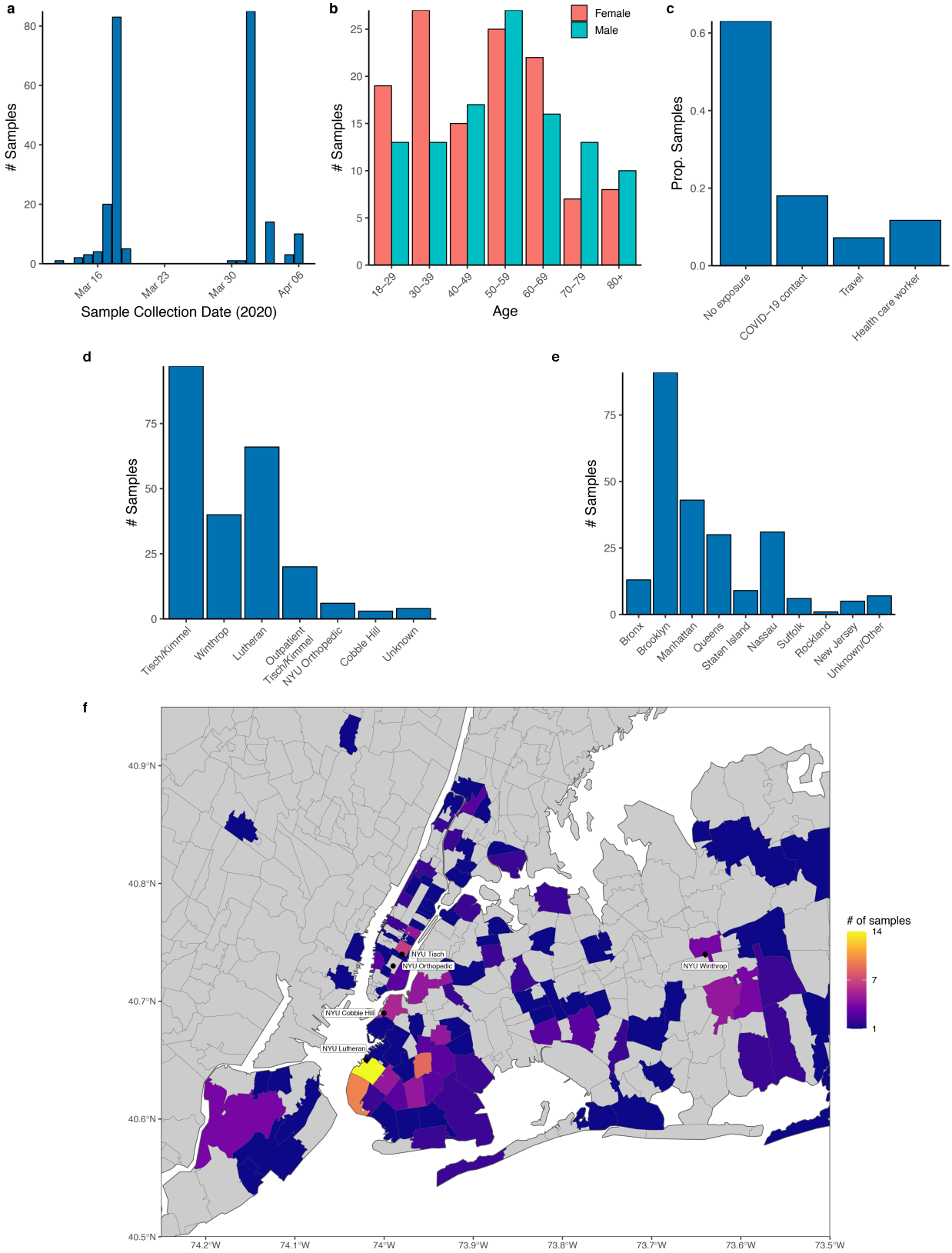


Fig. 1. Summary of study population. Case count by: a. Nasopharyngeal swab collection date. b. Age (binned by decade) and sex distribution. c. Potential exposure status, categorized by occupation as health care worker, travel history, and contact with a COVID-19 positive individual; n=110 cases. d. Collecting hospital. e. Borough, county, or state of residence. f. Localization of case residences within the Greater New York Region. Collecting hospitals are indicated in rounded boxes.

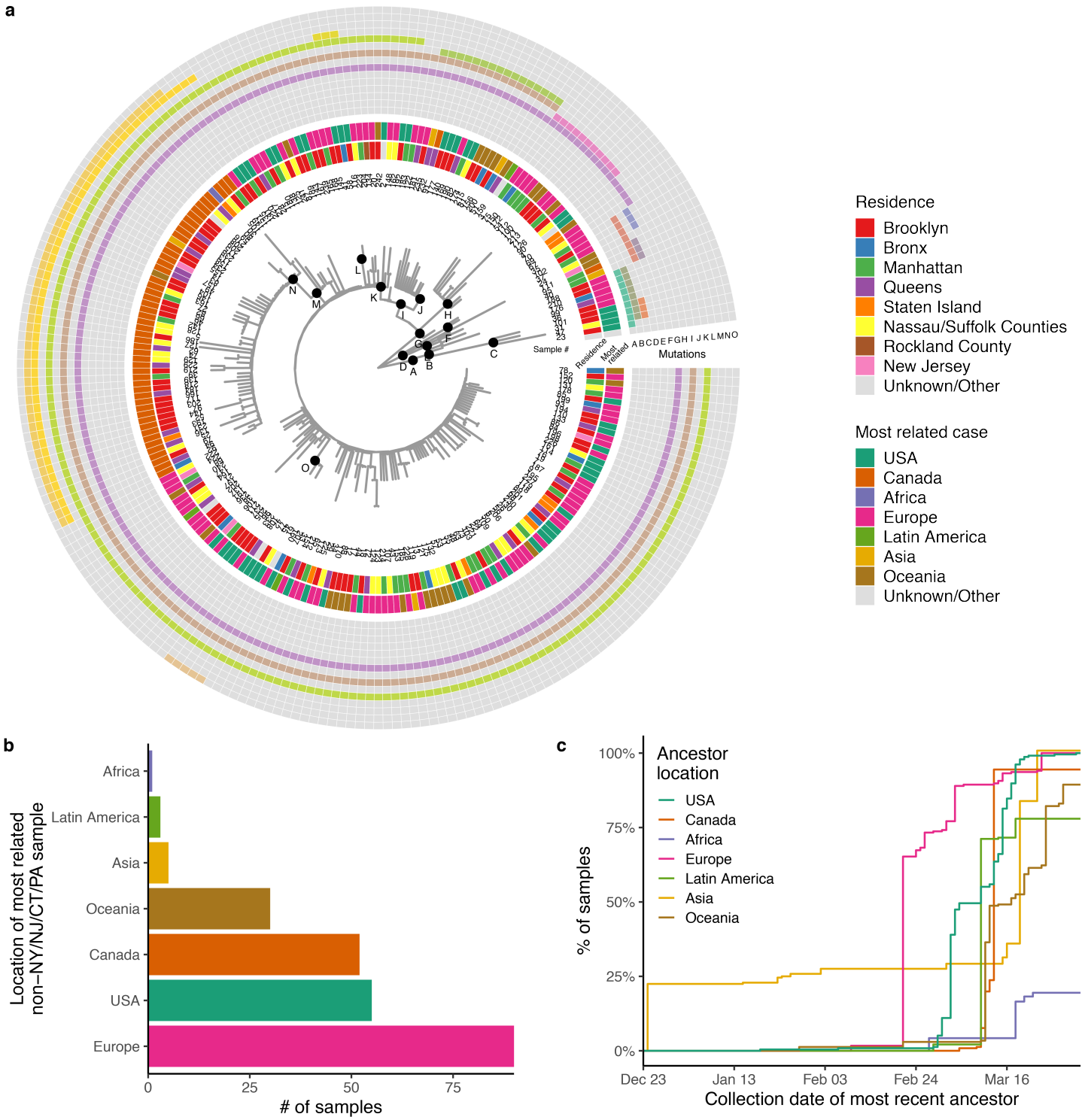
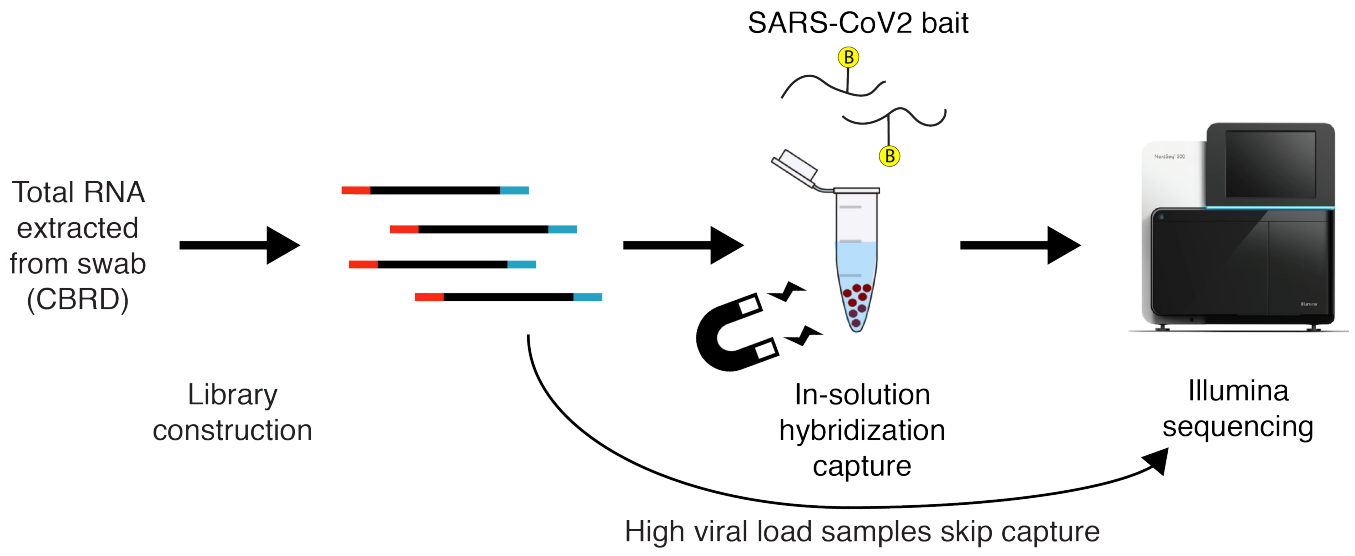
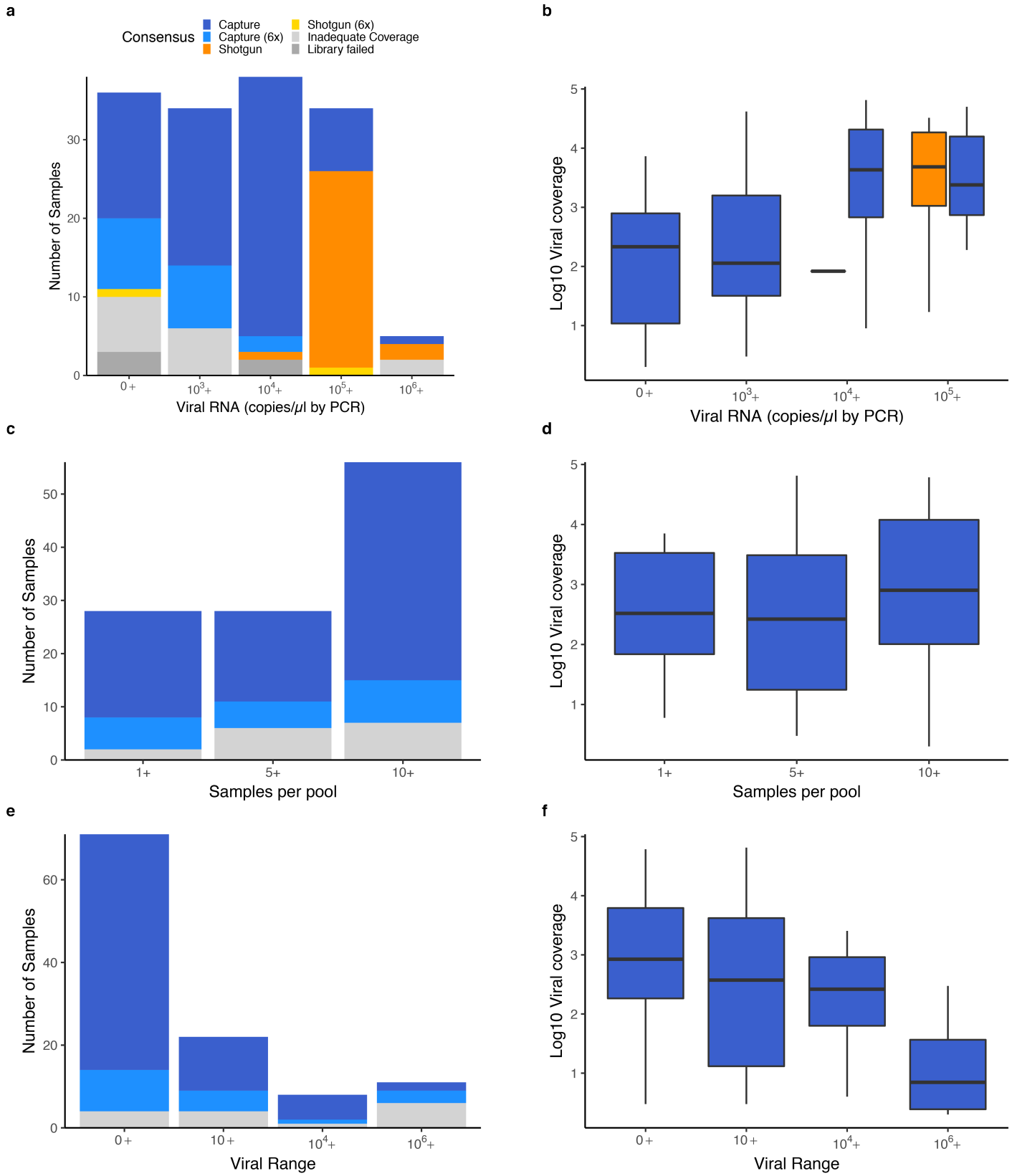


Fig. 2. Phylogenetic relationship of regional viral sequences. a. Phylogeny constructed from 236 cases. From innermost to outermost: colored boxes indicate residence, the location of the nearest case from outside the region (Connecticut, New York, New Jersey, Pennsylvania). Letters indicate groups of clade-defining mutations (**Extended Data Fig. 3).** b. Counts of cases matched to each location in (b). c. Cumulative plot showing collection date of most recent ancestral case from a given region.

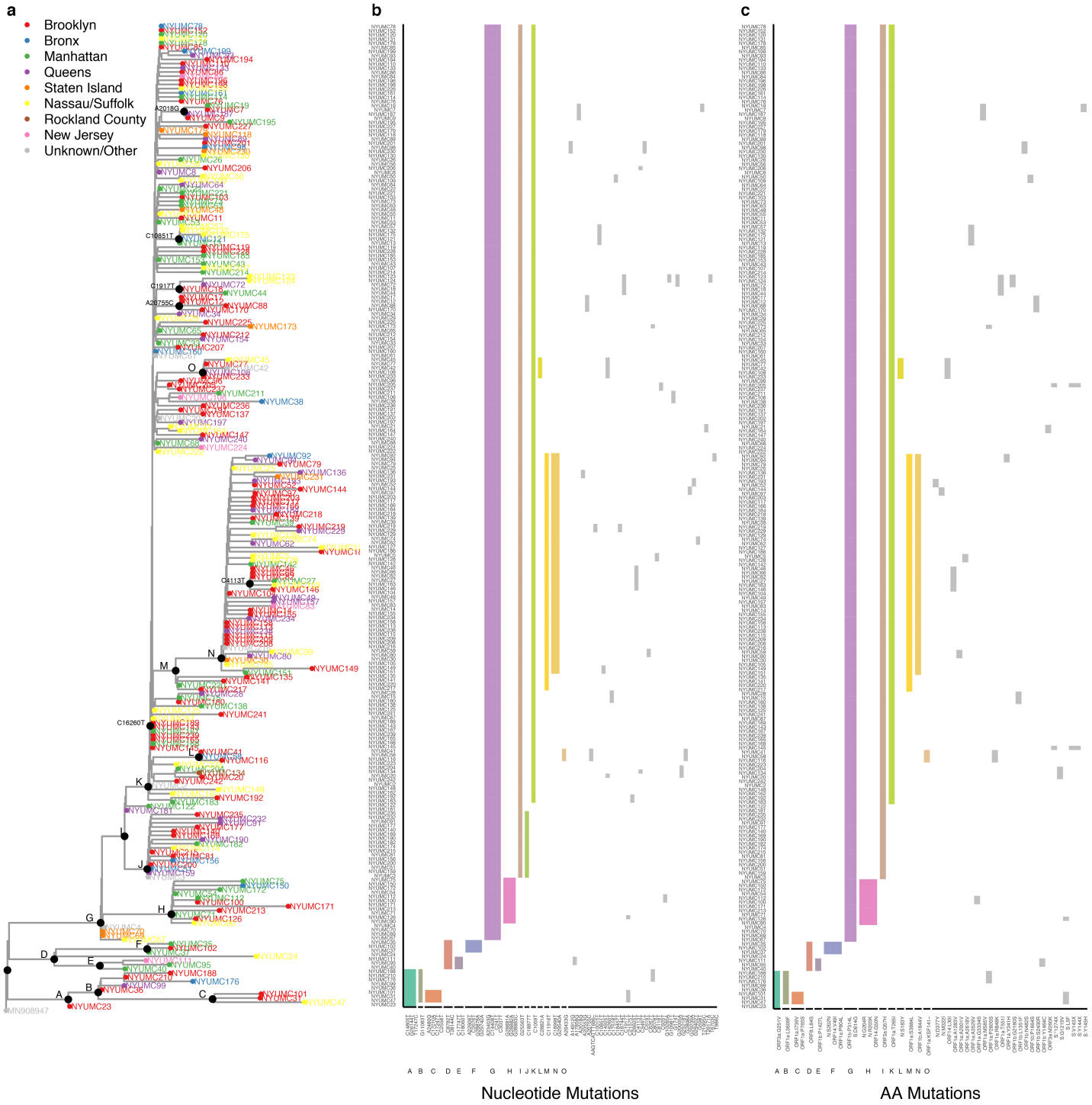
Supplemental Figures



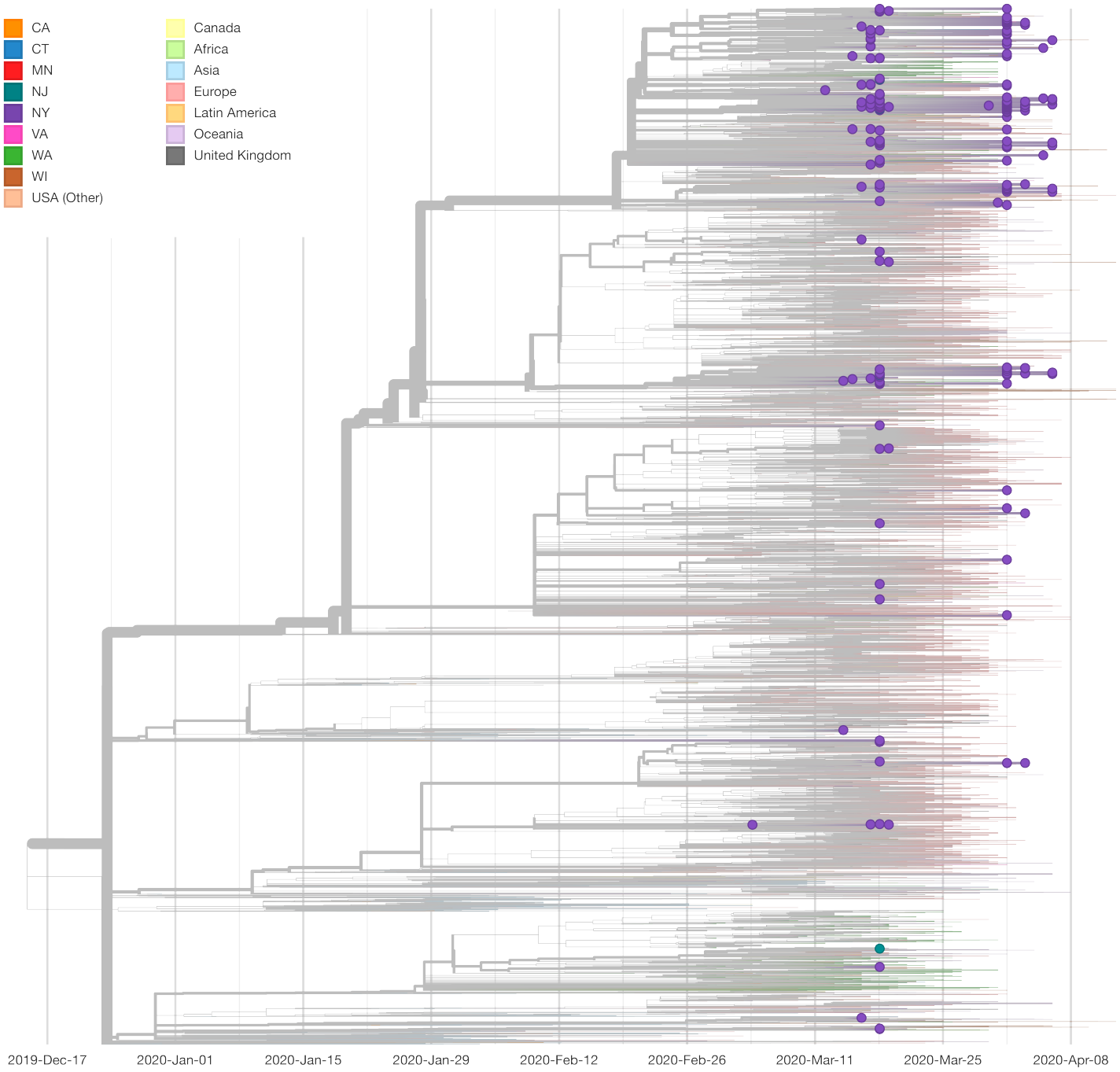
Extended Data Fig. 1. SARS-CoV2 sequencing workflow.
Total RNA was extracted using high throughput extractors, at the Center for Biorepository Specimen and Development (CBRD) at NYU Langone Health. Ribodepletion RNA-seq library preps were used to prepare libraries (Kapa Hyper Riboerase or Nugen Trio) followed by enrichment for the viral genome sequences using hybridization-based capture baits (IDT or Twist) designed against the Wuhan SARS-CoV2 RefSeq. For samples with viral load >100,000 copies/μL, we skipped the hybridization capture enrichment and sequenced RNA-seq libraries directly. Libraries were sequenced on the Illumina NovaSeq 6000 or NextSeq 500.



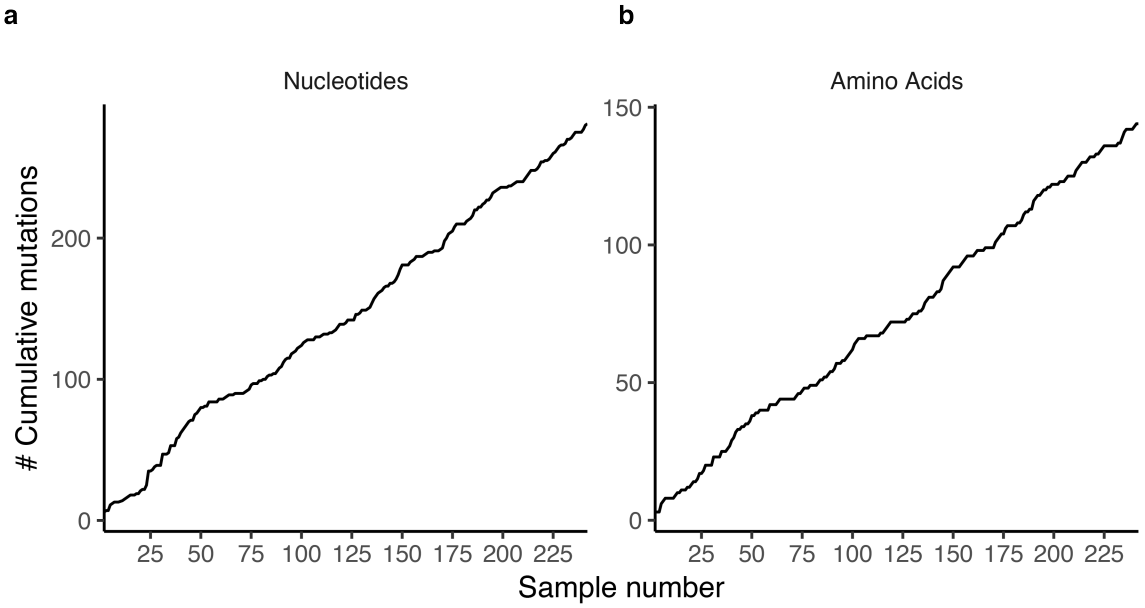
Extended Data Fig. 2. Technical factors related to sequencing success rate.
Shown are sample counts by final QC outcome (left) and average coverage (right) by viral load (a-b), size of capture pool (c-d), and range of viral load among samples in the same capture pool (e-f). Boxes in b, d, f indicate first and third quartiles, whiskers extend to 1.5 times interquartile range.



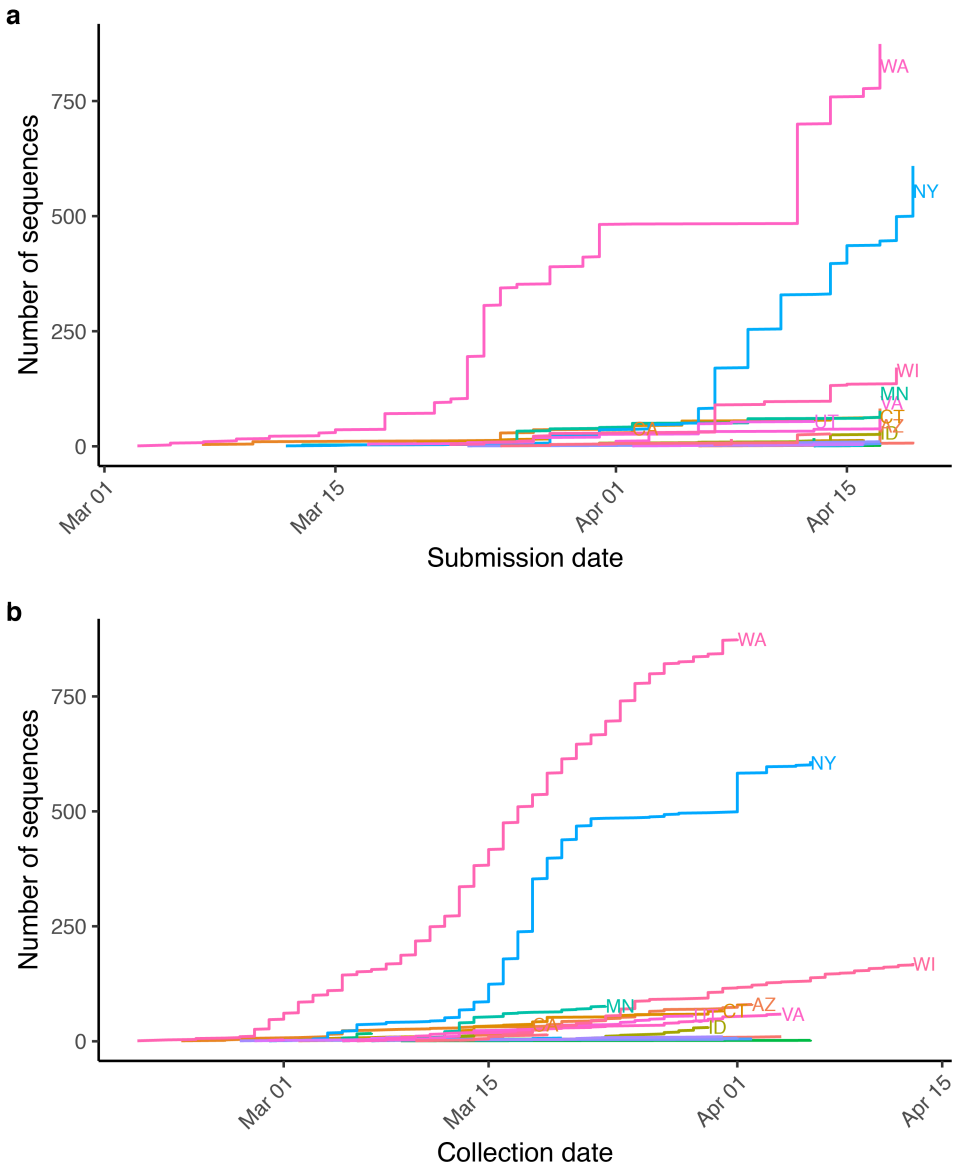
Extended Data Fig. 3. Phylogenetic relationship of regional viral sequences and mutations.
a. Colors indicate case residential location, capital letters indicate informative mutations b.-c. Key for b. nucleotide and c. amino acid mutations present in two or more cases.



Extended Data Fig. 4. Phylogenetic analysis of SARS-CoV2 sequences relative to the GISAID EpiCov repository via Nextstrain. Tips highlighted with dots are cases from this study. Tips and edges are colored by geographical origin.



Extended Data Fig. 5. 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.



Extended Data Fig. 6. US SARS-CoV2 sequences in GISAID by state.
Shown is a summary of the GISAID EpiCov repository showing all submitters per state by (a) submission date, and (b) collection date.