Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City Region

Matthew T. Maurano[1,2,*], Sitharam Ramaswami[3], Paul Zappile[3], Dacia Dimartino[3], Ludovic Boytard[4], André M. Ribeiro-dos-Santos[1,2], Nicholas A. Vulpescu[1,2], Gael Westby[3], Guomiao Shen[2], Xiaojun Feng[2], Megan S. Hogan[1,2], Manon Ragonnet-Cronin[5], Lily Geidelberg[5], 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], Carolina Arguelles-Grande[4], Kimon V. Argyropoulos[2], Margaret Black[2], Antonio Serrano[2], Melissa E. Call[6], Min Jae Kim[6], Brendan Belovarac[2], Tatyana Gindin[2], Andrew Lytle[2], Jared Pinnell[2], Theodore Vougiouklakis[2], John Chen[9], Lawrence H. Lin[2], Amy Rapkiewicz[2], Vanessa Raabe[7], Marie I. Samanovic[8], George Jour[2,6], Iman Osman[4,6], Maria Aguero-Rosenfeld[2], Mark J. Mulligan[7], Erik M. Volz[5], 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] MRC Centre for Global Infectious Disease Analysis and Department of Infectious Disease Epidemiology, Imperial College London.
[6] Department of Dermatology, NYU School of Medicine, New York, USA.
[7] Division of Infectious Diseases and Immunology, Department of Medicine and NYU Langone Vaccine Center, NYU School of Medicine, New York, USA.
[8] Department of Medicine, NYU School of Medicine, New York, USA.
[9] Medical Center IT, NYU Langone Health, 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 864 SARS-CoV-2 sequences from cases in the New York City metropolitan area during the COVID-19 outbreak in Spring 2020. 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 early transmission was most linked 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.

Main text

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-CoV-2, a betacoronavirus related to the etiological agent of the 2002 SARS outbreak, SARS-CoV and of possible bat origin[1,2]. Building on infrastructure from past outbreaks[3,4], genomic epidemiology has been applied to track the worldwide spread of SARS-CoV-2 using mutations in viral genomes to link otherwise unrelated infections[5,6]. Collaborative development of targeted sequencing protocols[7,8], open sharing of sequences through the GISAID (Global Initiative on Sharing All Influenza Data) repository[9], and rapid analysis and visualization of viral phylogenies using Nextstrain[10] have provided unprecedented and timely insights into the spread of the pandemic. Notably, community transmission was identified in time to implement preventative measures using surveillance sequencing in the Seattle area[11,12].

The New York City metropolitan region rapidly became an epicenter of the pandemic following the identification of the first community-acquired case on March 3, 2020 (a resident of New Rochelle in nearby Westchester County who worked in Manhattan). As of May 10, 2020, New York State had 337,055 cases – the highest in the United States, and 8% of the worldwide total. Fully 55% of NY State cases lay within the five boroughs of New York City (185,357 cases), followed by Nassau and Suffolk counties to the east on Long Island (75,248 cases)[13]. The outlying boroughs and suburban counties reported markedly higher infection rates than Manhattan. The outbreak overlaps with the catchment area of the NYU Langone Health (NYULH) hospital system, including hospitals on the east side of Manhattan (Tisch/Kimmel), one in Brooklyn (Lutheran), and one in Nassau County (Winthrop). Since even early COVID-19 cases presented mostly without travel history to countries with existing outbreaks, determining the extent of asymptomatic community spread and transmission paths became paramount. In parallel with increased clinical capacity for diagnostic PCR based testing, we sought to trace the origin of NYULH-treated

COVID-19 cases using phylogenetic analysis to compare to previously deposited COVID-19 viral sequences. We further aimed to develop an approach to integrate sequencing as a complementary epidemiological indicator of outbreak trajectory.

To assess the spread of SARS-CoV-2 within the NYU Langone Health COVID-19 inpatient and outpatient population, we deployed and optimized a viral sequencing, quality control, and analysis pipeline by repurposing existing genomics infrastructure. Cases were selected for sequencing from those confirmed positive between March 12 and May 10, 2020, during which period, positive tests within the NYULH system mirrored New York City and nearby counties (**Extended Data Fig. 1**)[14]. Illumina RNA-seq libraries were generated from using a ribodepletion strategy starting from total RNA from nasal swabs. Hybridization capture with custom biotinylated baits was used to enrich RNA-seq libraries for viral cDNA for sequencing (**Extended Data Fig. 2**, **Online Methods**). Of samples extracted for sequencing, fully 78% yielded a successful sequence, although success rates were lower for samples with qPCR Ct values > 30 (**Extended Data Fig. 3a-b**). We observed that high-quality sequences could be generated directly from shotgun libraries for qPCR Ct values < 30, thereby simplifying pooling and logistical constraints by skipping the capture step. Up to 23 samples were multiplexed in a single capture pool (**Extended Data Fig. 3c-d**). Samples with similar Ct values were grouped to minimize the range of target cDNA representation across a single capture pool (**Extended Data Fig. 3e-f**). This resulted in 864 sequences passing quality control, representing 10% of COVID-19 positive cases in NYULH over that time period (**Extended Data Fig. 1, Supplementary Table 1)**.

The cohort of 864 sequenced cases included a range of ages (**Fig. 1a**). 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. 1b**). 66% of cases resided within New York City, and 86% within NY State (**Fig. 1c**). Analysis of residential ZIP codes showed that cases reflected the hospital catchment area within the New York metropolitan region (**Fig. 1d**). Notably, our dataset included few 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 compiled a database for 820 of these cases from electronic medical records, including potential exposure information for health care worker status, travel history, and close contact with a COVID-19 individual (**Online Methods**). We found no recorded exposures for 43% of cases (**Fig. 1e**). Travel history was present in only 5% of cases, and these cases were concentrated in March (**Fig. 1f**). Of 14 cases where travel destination information was available, 9 destinations were within the US, 4 were in Europe, and 1 was in South Asia. This assessment relies upon clinical notes during a period where clinical capacity was stretched, thus likely underestimates potential exposures. Conversely, we only assessed only potential exposures, so unrelated transmission routes are possible given the uncontrolled community spread at the time.

We inferred a maximum likelihood phylogeny to assess relatedness among cases (**Fig. 2**). Coloring cases by county of residence within the New York region showed identical or related viral sequences found across multiple counties from the onset of our sampling (**Fig. 2**). We detected 890 nucleotide and 547 amino acid mutations across all cases (**Extended Data Fig. 4**). Mutation of D614G in the spike protein, which has been suggested to affect transmission or virulence[15], was present in >95% of sequences. Further sequencing will be necessary to detect signals of positive selection purely from sequence analysis, and 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.

We then assessed relatedness of our cases to 5,004 sequences from across the world from the GISAID EpiCov repository **(Extended Data Fig. 5)**. A maximum likelihood tree showed that cases from the NY region demonstrated broader diversity than initially reported in Seattle[11], the only other US region with a comparable level of viral sequences **(Extended Data Fig. 6**). To investigate the timing of introductions and transmission to New York City, we inferred a rooted time-scaled phylogeny (**Fig. 3a, Extended Data Fig. 7a)**. This analysis identified 88 transmission chains introduced to the New York City region (**Fig. 3b, Supplementary Table 3**). Analysis of the estimated time of nodes representing divergence from the source placed most introductions broadly in late February and early March, while the first detected transmissions within New York City occurred slightly later (**Fig. 3c**). The timing of these introductions did not change for alternative nucleotide substitution models (**Extended Data Fig. 7b).** The number of NYULH samples in each transmission chain varied widely, and several early chains comprised 50 cases or more. Analysis of the geographical location of samples within early transmission chains showed most were collected in Europe, while chains seeded later contained more samples collected from elsewhere in the US (**Fig. 3d**).

While fine-scale delineation of individual transmissions is limited by viral mutation rate (many sequences demonstrate identical genotypes), incomplete sampling, and incomplete availability of exposure history[16], these estimates of introduction were consistent with the outbreak timeline (**Extended Data Fig. 1**) and with other reports of the initial stages of the New York City Region outbreak[17,18]. While it is possible that further analysis[12] and sequencing of archival samples may clarify the initial spread, in the context of the earlier spread of the pandemic to Europe, and reduced travel flows from Asia, the genetic data suggest that the New York outbreak was seeded largely by way of Europe.

To estimate the trajectory of the outbreak, we applied phylodynamic analysis to estimate viral effective population size from a subsample of sequences[19] (**Online Methods**). Under moderate assumptions, effective population size will be proportional to epidemic prevalence and growth rates of effective population size will correspond to epidemic growth[20]. This analysis identified a period of rapid growth, followed by return nearly to the start point (**Fig. 4a**). We estimate that the peak effective population size occured on March 29 [95% CI: March 19-April 5]. The growth rate decreased steadily after March 1 and was negative with high confidence by mid-April (**Fig. 4b**),

consistent with the epidemic curve of confirmed infections in the New York City Region (**Extended Data Fig. 1a**).

Our demonstration of rapid sample processing, deposition, and analysis underscores the potential for near real-time genomic epidemiology to provide an independent estimate of disease transmission, and its potential to recognize impending resurgence of a regional outbreak. Further surveillance by medical centers, regional public health departments, and national efforts will be needed to monitor genomic epidemiology, pandemic spread, and public responses (**Extended Data Fig. 5**). Given the logistical, regulatory, and methodological challenges to establishing such surveillance during an outbreak, it is critical to have this infrastructure already in place[21] for future waves of COVID-19 or other future pandemics.

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. Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161,** 1516–1526 (2015).
4. Carroll, M. W. *et al.* Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature* **524,** 97–101 (2015).
5. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4,** 10–19 (2019).
6. 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
7. 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).
8. *ARTIC SARS-CoV-2 Protocol*. Available at: https://artic.network/ncov-2019. (Accessed: 15 April 2020)
9. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22,** (2017).
10. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34,** 4121–4123 (2018).
11. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. (2020). doi:10.1101/2020.04.02.20051417
12. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and the US. doi:10.1101/2020.05.21.109322
13. *New York State Statewide COVID-19 Testing*. Available at: https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e. (Accessed: 18 June 2020)
14. Petrilli, C. M. *et al.* Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. *BMJ* **369,** m1966 (2020).
15. Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. doi:10.1101/2020.06.12.148726
16. Villabona-Arenas, C. J., Hanage, W. P. & Tully, D. C. Phylogenetic interpretation during outbreaks requires caution. *Nat Microbiol* (2020). doi:10.1038/s41564-020-0738-5
17. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* (2020). doi:10.1126/science.abc1917
18. Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181,** 990–996.e5 (2020).
19. Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10,** 540–550 (2009).
20. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput Biol* **9,** e1002947 (2013).
21. 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
22. Osman, I. *et al.* The urgency of utilizing COVID-19 biospecimens for research in the heart

of the global pandemic. *Journal of Translational Medicine* 1–4 (2020). doi:10.1186/s12967-020-02388-8
23. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
25. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30,** 2503–2505 (2014).
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)
27. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
28. 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).
29. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4,** 741 (2018).
30. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35,** 526–528 (2019).
31. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).
32. Volz, E. M. & Frost, S. D. W. Scalable relaxed clock phylogenetic dating. *Virus Evolution* **3,** 289 (2017).
33. Volz, E. M. & Didelot, X. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Systematic Biology* **67,** 719–728 (2018).

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, travel, exposure and demographic data[22]. Electronic medical records were reviewed to compile a clinical database for 820 cases listing health care worker status, travel history, and close contact with a known COVID-19 case. For cases where a given exposure was not directly stated in the clinical record, we recorded that field as missing data but included other exposures in our analysis. A summary field of exposure history per case was generated by the presence of COVID-19 contact, travel history, or health care worker status, in that order.

Sample collection

All samples were collected as part of clinical diagnostics. Nasopharyngeal swabs were collected and placed in 3 mL of Viral Transport Medium (VTM, Copan Universal Transport Medium) following clinical protocols. Samples were transported to the clinical microbiology laboratory at room temperature and tested for SARS-CoV-2 the same day. Remnant samples were stored at -70 °C.

Clinical testing

All initial detection of COVID-19 cases was performed as part of the clinical care. Clinical testing was performed using the following three FDA Emergency Use Certification (EUA) approved COVID-19 PCR based tests:

i. NYULH-validated PCR test using the US CDC primer design, targeting three regions of the virus nucleocapsid (N) gene, and an internal control primer targeting the human RNAse P gene (RP) (https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html) with PCR carried out on an ABI7500 Dx system. The limit of detection is 10,000 copies/mL.
ii. The Roche Cobas 6800 RT-PCR platform targeting the Orf1/a and E sequences, per manufacturer instructions. The limit of detection is 180 copies/mL.
iii. The Cepheid Xpert Xpress RT-PCR platform targeting the N2 and E viral sequences, per manufacturer instructions. The limit of detection is 250 copies/mL.

RNA extraction

RNA extraction was performed using two platforms for parallel sample processing:

i. Using the Maxwell RSC instrument (Promega, cat. AS4500), total RNA was extracted from 300 μL of viral transport medium 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 deacti-

vation, 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.

ii. Using the KingFisher Flex System (ThermoFisher Scientific) system, RNA was extracted from heat inactivated nasal swab samples in batches of 96 samples, following the manufacturer's instructions and the MagMax *mir*Vana Total RNA isolation Kit (ThermoFisher Scientific, A27828). Briefly, 250 uL of nasal swab collection was lysed in lysis buffer and β-mercaptoethanol and subsequently bound to magnetic beads and loaded into the KingFisher Flex instrument. A DNase I treatment step was performed as part of the instrument protocol and RNA samples were eluted in 50 uL of Elution Buffer and immediately stored at -80C.

Library preparation and sequencing

Ilumina sequencing libraries were prepared from 10 μL of total RNA. Two methods for cDNA RNA-seq library preps were used, both based on a ribodepletion approach:

i. KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche Kapa cat. KK8561). We followed the manufacturer's protocol, with the following modifications: for the adapter ligation step, we prepared a plate of barcoded adapters (IDT) at a concentration of 500 nM, and performed 15 cycles of PCR amplification of the final library.

ii. Nugen Trio with human rRNA depletion (Tecan Genomics, cat. 0606-96), including DNase I treatment, cDNA synthesis, single primer isothermal amplification (SPIA), enzymatic fragmentation, library construction, final PCR amplification (12-16 cycles), and an AnyDeplete step to remove host rRNA transcripts. An automated protocol was implemented on a Biomek FX^P liquid handler integrated with a Biometra TRobot 96-well thermal cycler (Beckman Coulter).

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

Libraries from samples with qPCR Ct values > 30 were enriched for SARS-CoV-2 genomic sequences using custom biotinylated DNA probe pools either from Twist Biosciences or Integrated DNA Technologies:

i. For capture using the IDT xGen COVID capture panel (Integrated DNA Technologies, cat. 10006764), we followed the manufacturer's protocol. Briefly, hybridization of 500 ng to 1 μg of combined library DNA with 4 μL of XGen Lockdown probes was carried out at 65 °C for 4-16 h, followed by PCR amplification for 6-10 cycles.

ii. For capture using the Twist Bioscience custom panel (Twist Design ID: TE-95888003), we followed the manufacturer's protocol using the Twist Hybridization and Wash Kit (Twist Biosciences, cat. 101025). Hybridization of 1-2 μg combined library DNA was carried out at 70 °C for 16-20 h. Post-capture PCR amplification cycles ranged from 12-14 cycles.

In general, we pooled samples with similar Ct values and accounted for variations in parent library concentration, multiplexing up to 23 libraries per reaction.

Samples were sequenced as paired end 100- or 150-cycle 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 barcode sequences. All RNA-seq and Capture-seq data were processed using a uniform mapping pipeline. Illumina sequencing adapters were trimmed with Trimmomatic v0.39[23]. Reads were aligned using BWA v0.7.17[24] to a custom index containing human genome reference (GRCh38/hg38) including unscaffolded contigs and alternate references plus the reference SARS-CoV-2 genome (NC_045512.2, wuhCor1). Presumed PCR duplicates were marked using samblaster v0.1.24[25]. Only sequences with >23000 bp unmasked sequence were analyzed. Duplicate sequences from the same case were excluded, resulting in 864 final sequences (**Supplementary Table 1**). Variants were called across all samples using bcftools v1.9:

```
bcftools mpileup --redo-BAQ --adjust-MQ 50 --gap-frac 0.05 --max-
depth 10000 --max-idepth 200000 --output-type u |
bcftools call --ploidy 1 --keep-alts --multiallelic-caller -f GQ
```

Raw pileups were filtered using

```
bcftools norm --check-ref w --output-type u |
bcftools filter -i "INFO/DP>=10 & QUAL>=10 & GQ>=99 & FORMAT/DP>=10"
--SnpGap 3 --IndelGap 10 --set-GTs . --output-type u |
bcftools view -i 'GT="alt"' --trim-alt-alleles
```

Viral sequences were generated by applying VCF files to the reference sequence using `bcftools consensus` with -m to mask sites below 20x with Ns, and -m N to mask sites of ambiguous genotypes with N.

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].

Phylogenetic analysis

Sequences for non-NYULH cases were downloaded from GISAID EpiCov on June 14, 2020 and filtered to sequences collected on or before May 10, 2020. Sequences from non-human hosts, annotated by Nextstrain as duplicate individuals or highly divergent, with <27,000 non-ambiguous nucleotides, or with improperly formatted dates or location were excluded. Sequences from outside New York state were subsampled to a maximum of 20 samples per admin division (US) or country (outside US) per month, prioritizing sequences most similar to the focal set of 864 NYULH samples. This priority was penalized if many non-US samples were most similar to the same US sample, and mutations were weighted 333x more heavily than masked sites. Global sequences were then combined with the sequences from this study.

Sequences were analyzed using the augur v7.0.2 pipeline[10]. Sequences were aligned along with the reference genome using MAFFT v7.453[27], and the resulting alignment was masked to re-

move 100 bp from the beginning, 50 from the end, and uninformative point mutations (positions 11083, 13402, 21575, 24389, 24390).

Maximum likelihood phylogenetic reconstruction was performed with IQ-TREE v1.6.12[28] using a GTR substitution model. Support values were generated with the ultrafast bootstrapping option with 1000 replicates. This tree was used to tabulate nucleotide and amino acid changes specific to lineages and cases; gaps with respect to the reference were reported as deletions. TreeTime v0.7.4[29] was used to generate a timetree rooted at the reference sequence under a strict mutational clock under a skyline coalescent prior with a rate of 8×10^{-4} mutations per site per year and a standard deviation of 4×10^{-4}.

The first New York City transmission was identified as the most ancestral node or tip with >70% of sequences originating in the northeast (NY, CT, NJ, PA) using the ape[30] and phangorn[31] R packages. The transmission source was identified as the first upstream node defined by a unique mutation and ancestral to a sequence originating outside the northeast. Chains with identical source nodes and New York City genotypes were combined to yield transmission chains. Trees were plotted with the tidygraph and ggraph R packages.

Phylodynamic Analysis

To minimize ascertainment and sampling bias, analysis was performed on a subset of sequenced cases residing in New York City and outlying Westchester, Nassau and Suffolk counties, and excluded outpatients and known health care workers. Sequence data were aligned to reference (accession NC_045512.2) and ends trimmed using MAFFT 7.450[27]. A maximum likelihood tree was estimated using IQ-TREE 1.6.1 using a HKY substitution model[28]. A further 20 phylogenies were derived by randomly resolving polytomies and enforcing a small minimum branch length of 7×10^{-6} substitutions per site using the ape R package[30]. Rooted time scaled phylogenies were estimated using the treedater R package version 0.5.1[32] using a strict molecular clock. The skygrowth R package version 0.3.1[33] was used to estimate effective population size through time with an exponential prior for the smoothing parameter with rate 10^{-4}. The final estimates were generated by averaging results over the 20 estimated time trees. A script for reproducing these results is available at: https://gist.github.com/emvolz/d58cce01c3310a01df09faf615b77070.

Acknowledgements

We are indebted to the NYULH clinicians and laboratory personnel involved in the care and testing of the patients in this study. We would like to thank all the laboratories who have contributed sequences to GISAID (**Supplementary Table 2**). We thank Lea Starita and the Seattle Flu Study for technical assistance and sharing their bait design. This work was partially funded by NIH grants R35GM119703 (M.T.M.), P50CA016087 (I.O. and J.G.), P30CA016087 (I.O., P.C., and A.H.), UM1AI148574 (M.J.M), the NYULH Office for Science and Research, and

MR/R015600/1 from the MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London (M-R.C., E.M.V.)

Author Contributions

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

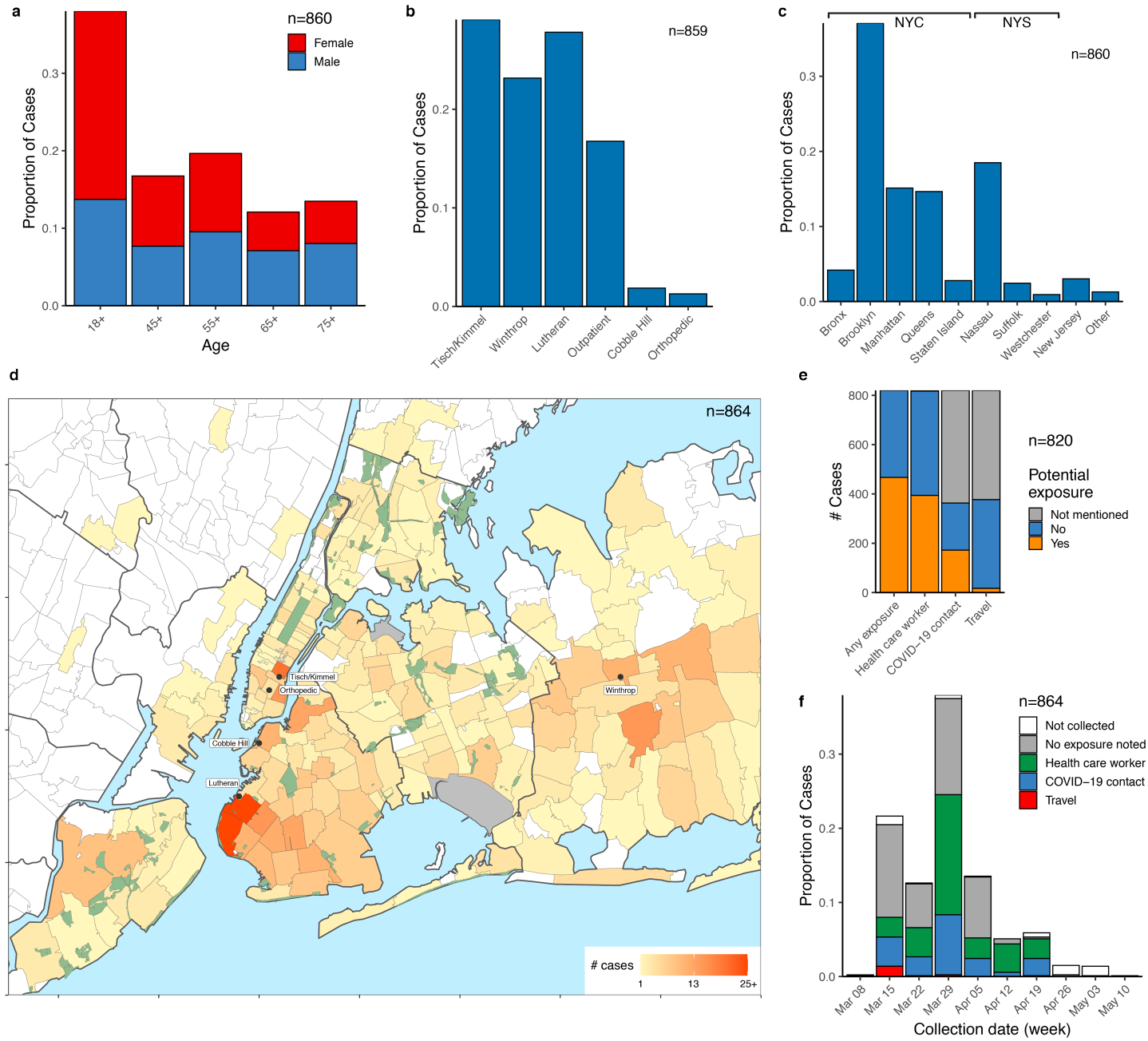


Fig. 1. Summary demographic parameters of sequenced SARS-CoV-2 cases in the NYULH system. Cases are broken down by: a. Age and sex. b. Collecting hospital. c. Residential location, grouped by borough and outlying counties; "Other" includes counties with few cases. d. Localization of case residences within the New York City Region. The color scale indicates numbers of cases per zip code. Collecting hospitals are indicated in rounded boxes. e. Potential exposure status, categorized by occupation as healthcare worker, travel history, and contact with a COVID-19 positive individual. f. Potential exposure status by collection date.

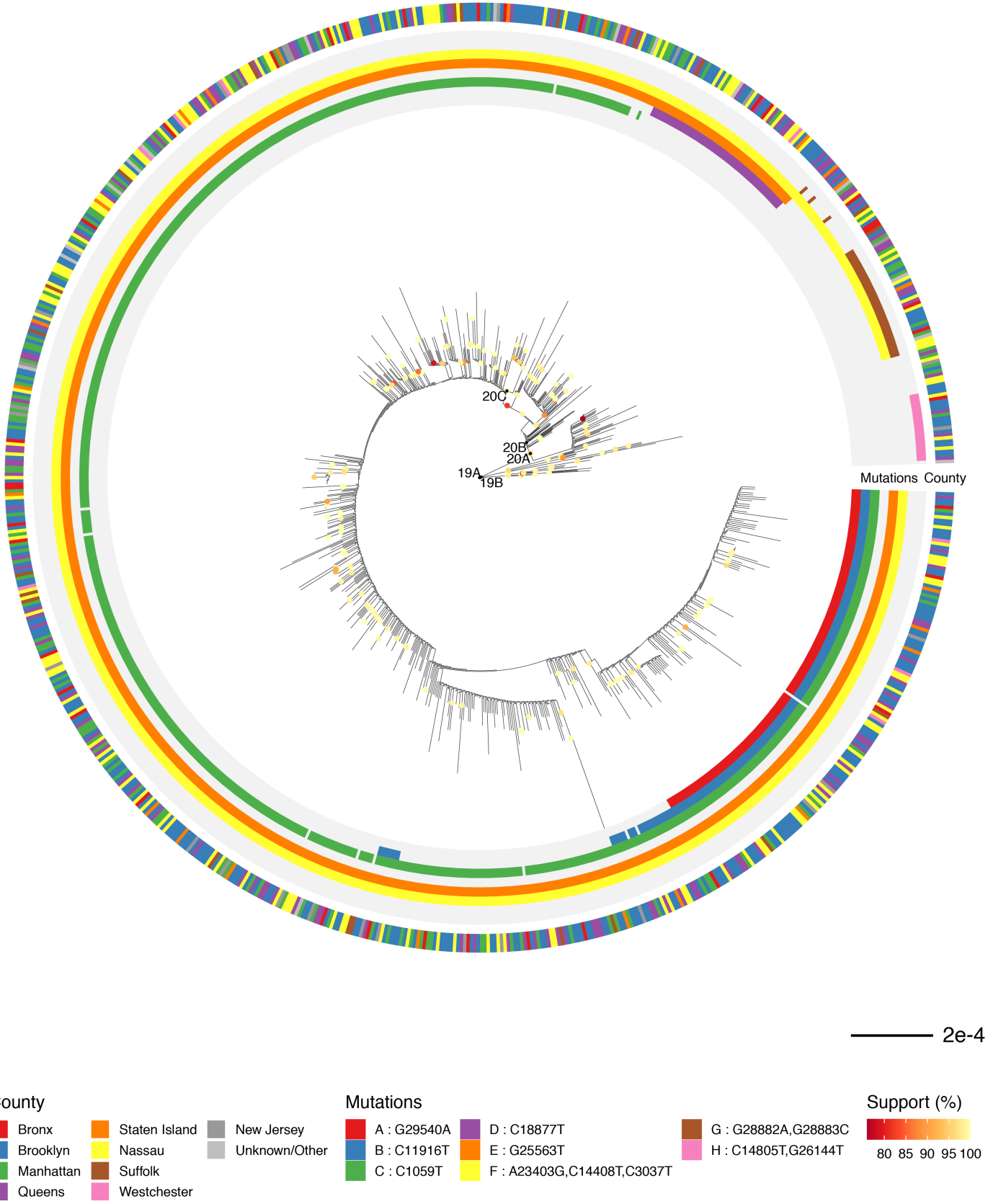


Fig. 2. Phylogenetic relationship of regional viral sequences. a. Maximum likelihood phylogeny inferred from 864 cases. Nodes with bootstrap support values above 75 are noted. Letters indicate groups of clade-defining mutations**.** From innermost to outermost: inner ring indicates county of residence. Scale bar represents nucleotide substitutions per site.

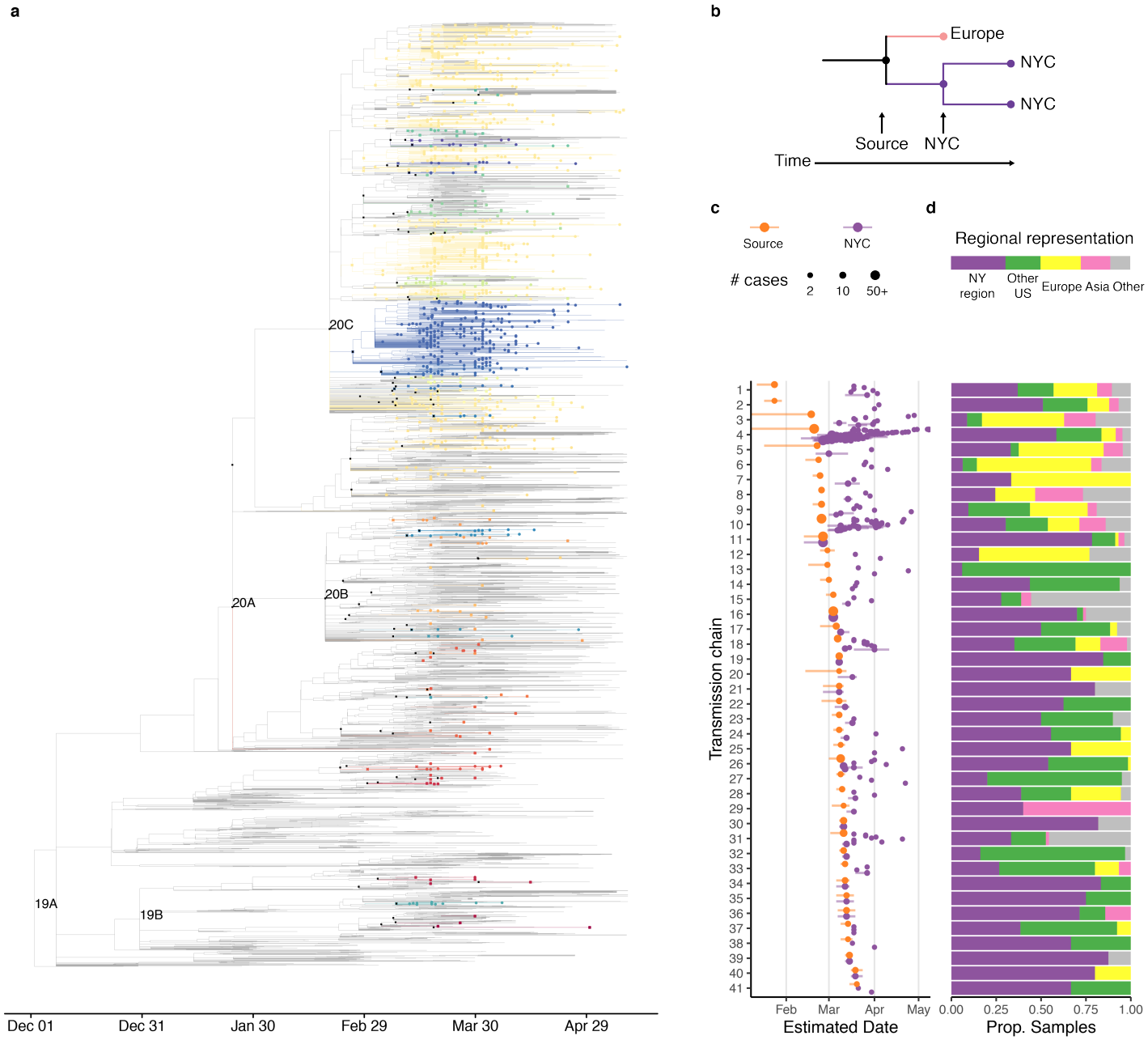


Fig. 3. Time-scaled phylogeny showing global sequence context.

a. Tips highlighted with dots are NYULH cases from this study; edges are with unique colors for each transmission chain. Black crosses (x) indicate source nodes, black squares indicate nodes of first New York City transmission.

b. Schematic of approach to infer transmission chains.

c-d. Transmission chains in the New York City region ordered by inferred divergence date from source. c. Dates estimated for source transmission (orange) and first detected local transmission (purple) inferred from sequenced cases; lines represent 90% confidence intervals. Shown are transmission chains with more than 1 NYULH case. d. Representation of global regions in each transmission chain. Bar at top shows overall representation of regions in phylogeny.

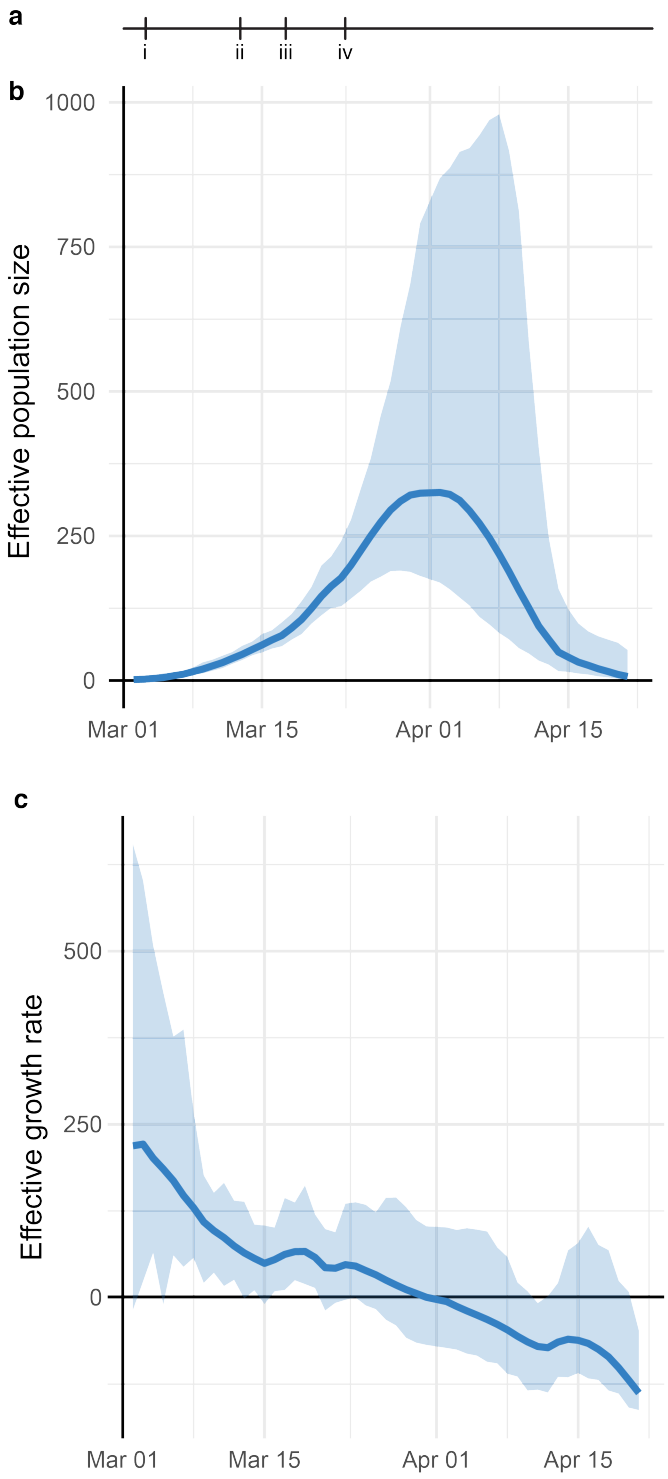
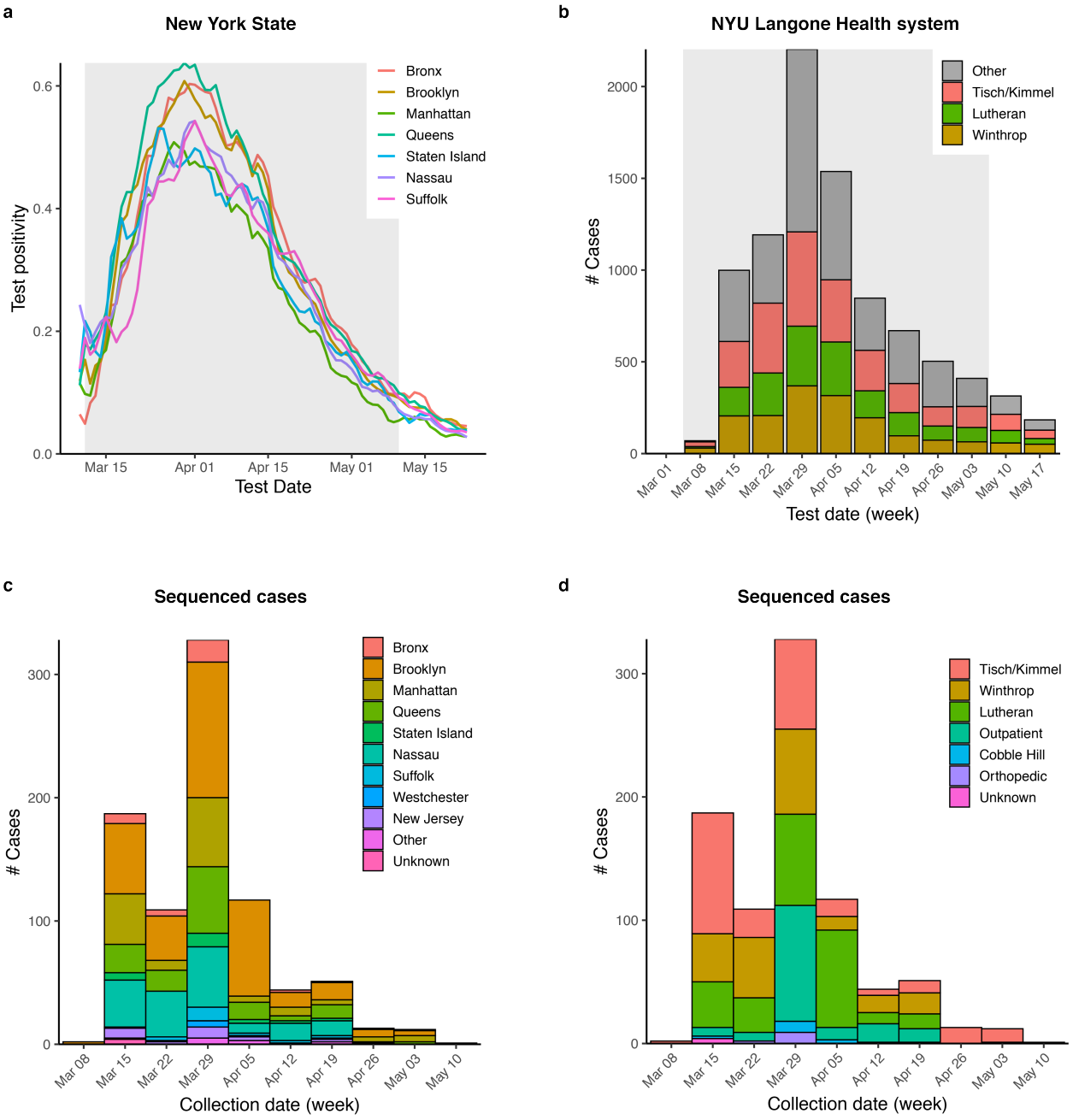
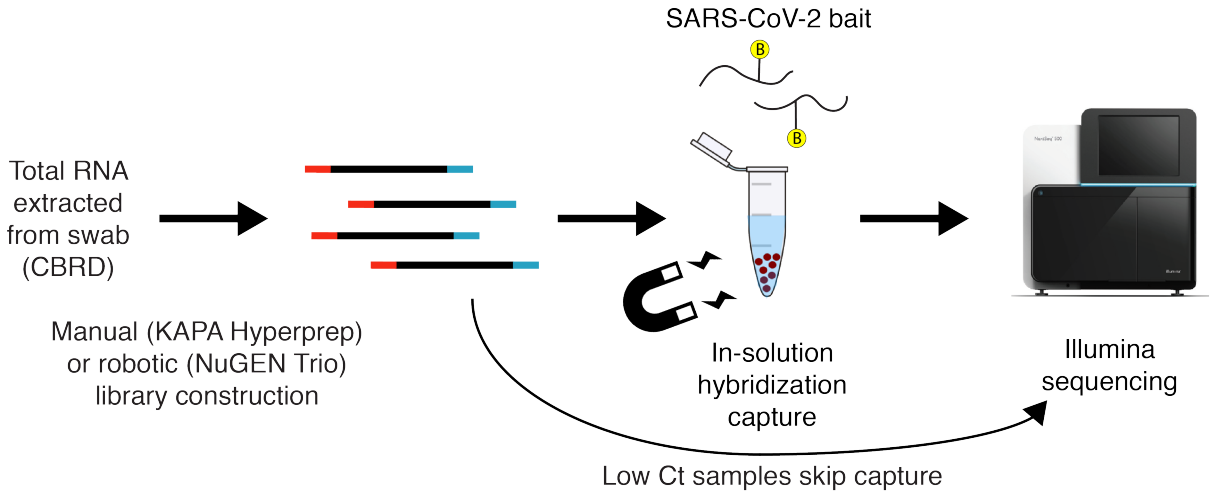


Fig. 4. Phylodynamic analysis of outbreak trajectory.
a. Timeline of New York City outbreak, highlighting (i) announcement of first community-acquired case (Mar 3); (ii) ban on gatherings exceeding 500 people (Mar 12); (iii) closure of schools, restaurants, and bars, and other venues (Mar 16); (iv) closure of non-essential businesses (Mar 22). b-c Outbreak trajectory estimated from genetic data showing b. effective population size relative to March 1 and c. growth rate of effective population size (units of 1/years). Shaded regions represent 95% credible interval.

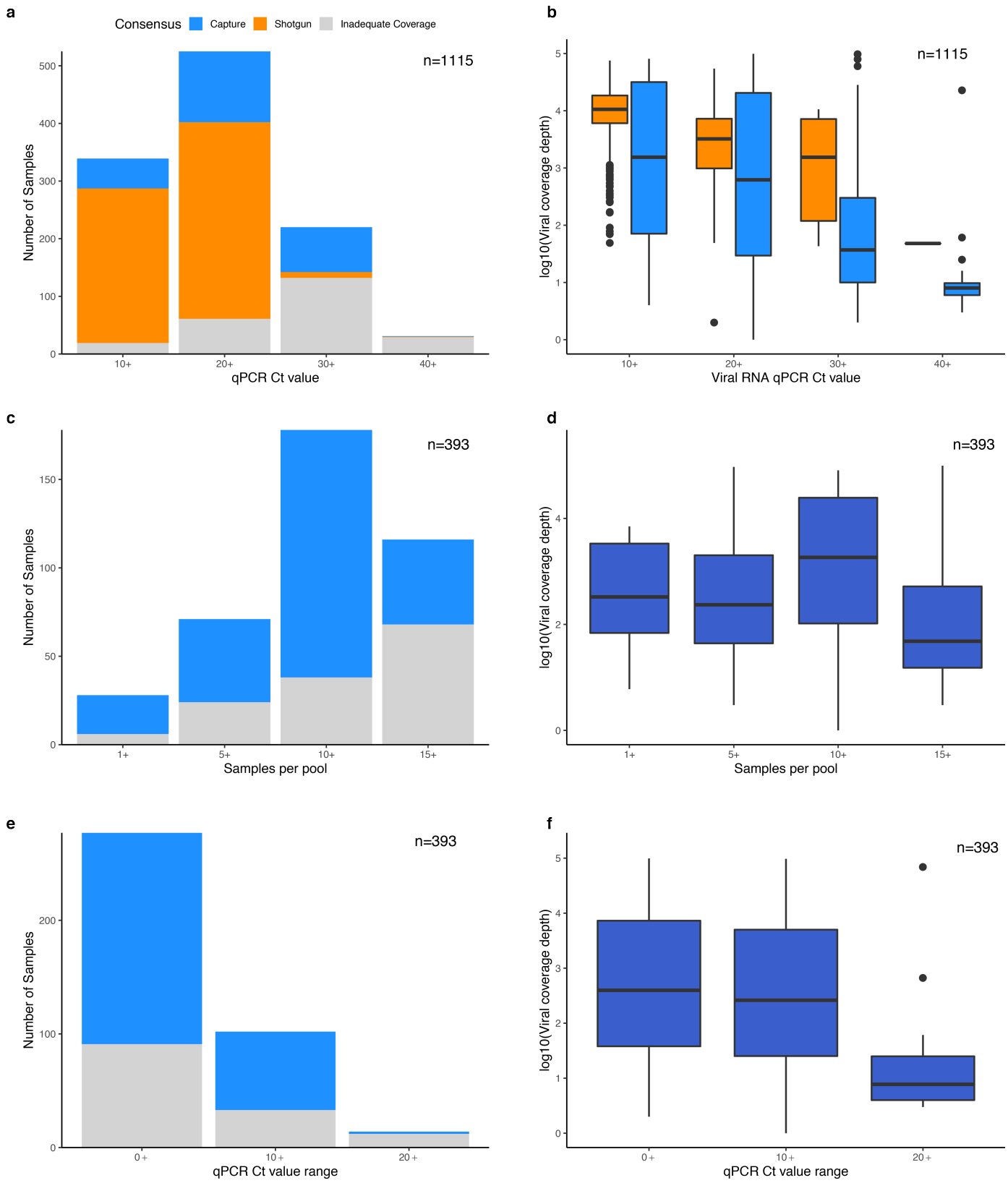
Supplemental Figures



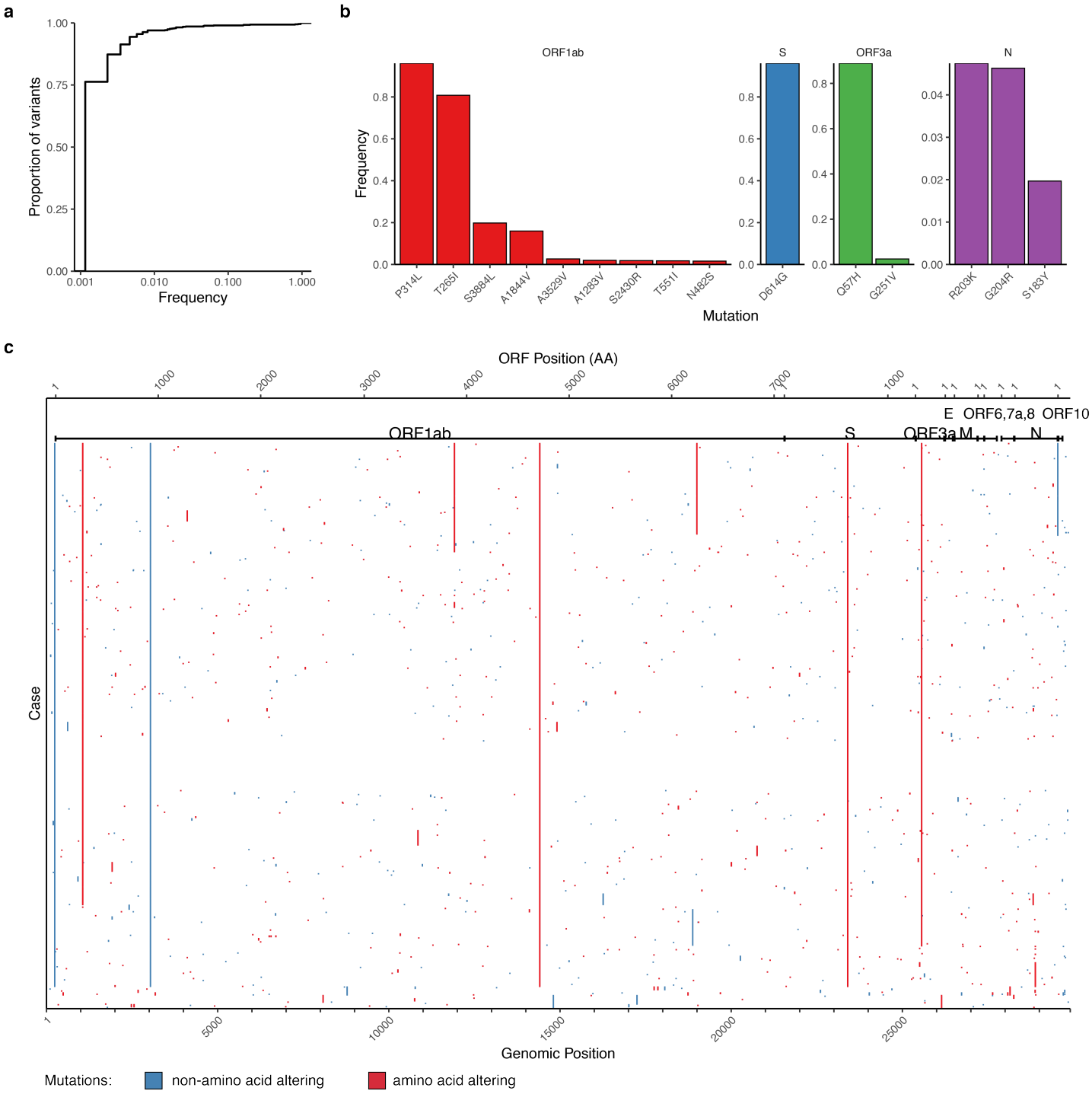
Extended Data Fig. 1. Outbreak trajectory and sampling of NYU Langone Health catchment area.
a. SARS-CoV-2 positivity rate for New York City boroughs and outlying counties reported by New York State Department of Health[13].
b. Summary of weekly positive tests across NYU Langone Health. Shaded region indicates time period sampled for sequenced cases.
c-d. Sequenced cases by collection date, broken down by county of residence (c) and collecting hospital (d).



Extended Data Fig. 2. SARS-CoV-2 sequencing pipeline.
Total RNA was extracted using high throughput extractors, at the Center for Biorepository Specimen and Development (CBRD) at NYU Langone Health. RNA-seq libraries were prepared using two ribodepletion protocols. For samples with qPCR Ct values < 30, we skipped the hybridization capture enrichment and sequenced RNA-seq libraries directly; other libraries were enriched for the viral genome sequences using hybridization-based capture baits (IDT or Twist) designed against the Wuhan SARS-CoV-2 RefSeq. Libraries were sequenced on the Illumina NovaSeq 6000 or NextSeq 500.



Extended Data Fig. 3. Technical factors related to sequencing success.
Shown are sample counts by final QC outcome (left) and average coverage depth (right) stratified by qPCR Ct values (a-b), size of capture pool (c-d), and qPCR Ct range 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.

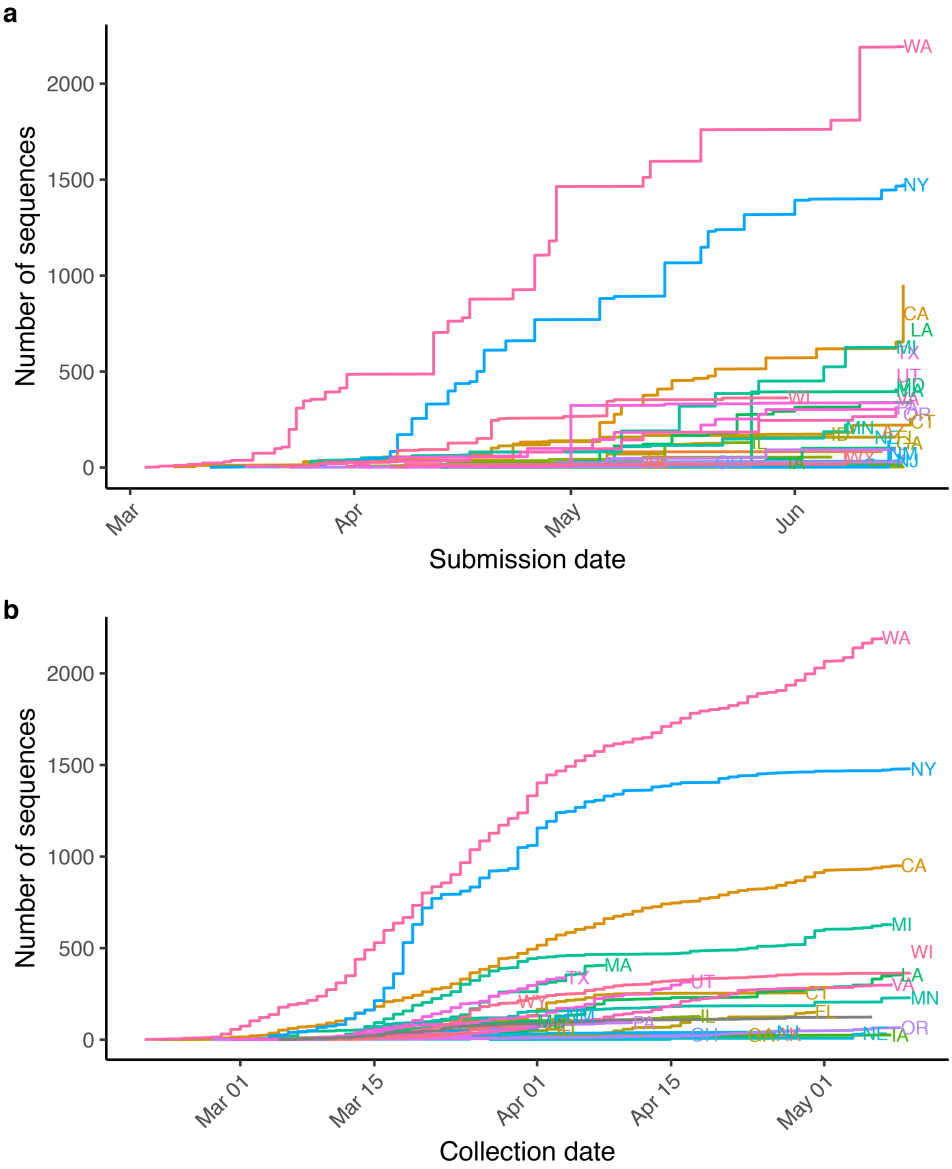


Extended Data Fig. 4. Overview of mutations identified.

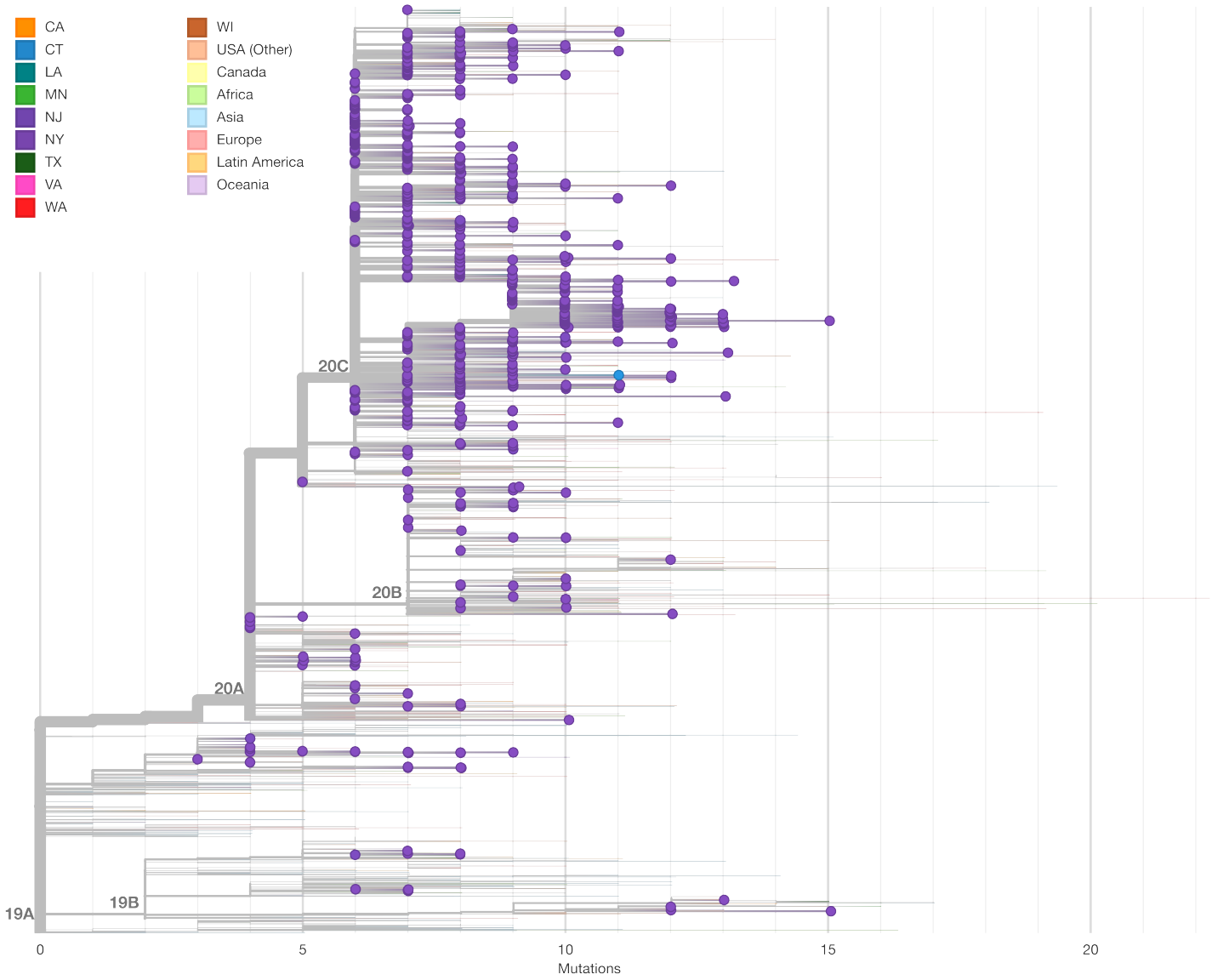
a. Cumulative frequency distribution of variants identified across 864 cases.

b. Frequency for top 15 amino acid-altering mutations by open reading frame (ORF).

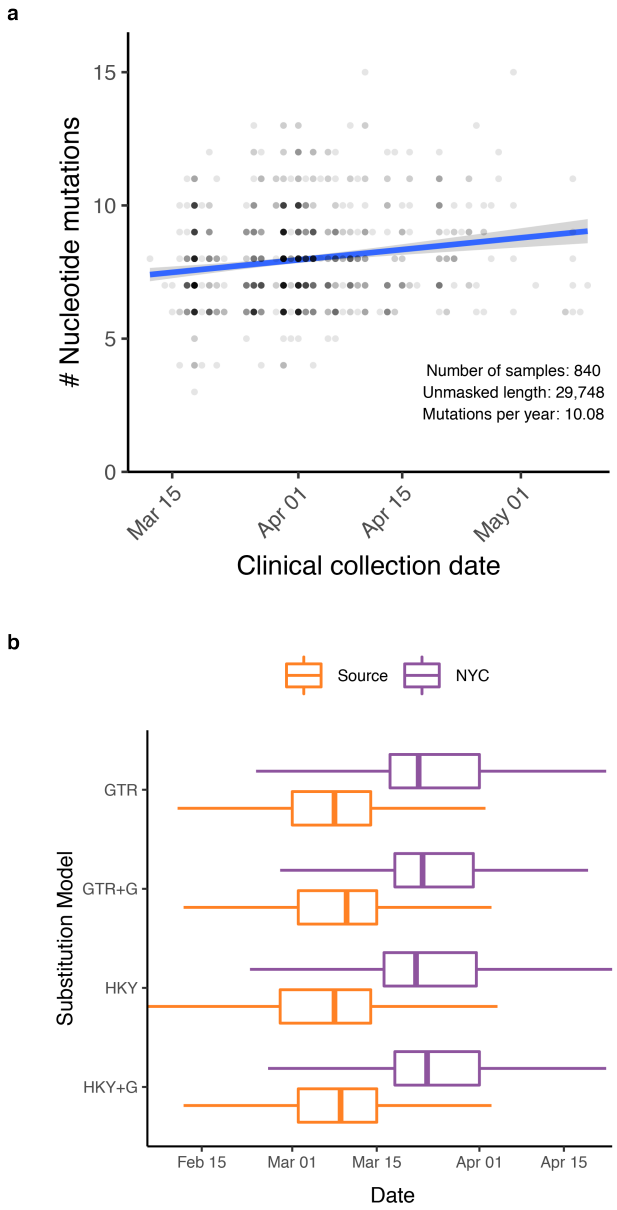
c. Heatmap of mutations identified per case, with x-axis being genomic coordinates, and rows in the same order as **Fig. 2a**. ORFs are annotated according to GenBank MN908947.3.



Extended Data Fig. 5. SARS-CoV-2 sequences in GISAID by US state.
Summary of US sequences in the GISAID EpiCov repository collected through May 10, 2020 showing all submitters per state by (a) submission date, and (b) collection date.



Extended Data Fig. 6. Maximum likelihood tree including 5,004 global sequences from GISAID. NYULH sequences are highlighted with dots, tip and edge coloring indicates geographical location.



Extended Data Fig. 7. Time-scaled phylogeny analysis.

a. Root to tip plot showing the number of point mutations per case by collection date. Sequences with >10% ambiguous nucleotides are excluded.

b. Effect of alternate substitution models on identification of transmission chain dating. Shown is the distribution of 90% confidence intervals across all transmission chains weighted by number of samples in each chain. Boxes indicate first and third quartiles, whiskers extend to 1.5 times interquartile range.

Supplementary Tables

Supplementary Tables are available as separate files.

Supplementary Table 1. Summary of sequencing data.

Per sample summary of sequencing data for 864 cases. PropViralReads, proportion of nonredundant reads mapping to SARS-CoV-2 genome; analyzedViralReads, number of reads mapping to SARS-CoV-2 genome and passing all filters; PropDupViralReads, proportion of analyzedViralReads marked as PCR duplicates; Mean_Viral_Coverage, mean coverage depth; Num_bp_20x, number of bp covered at ≥20x depth

Supplementary Table 2. Acknowledgements of GISAID sequences used.

List of sequences and contributors from GISAID.

Supplementary Table 3. New York City Region transmission chains.

Summary of 88 transmission chains, including estimated characteristic mutations and date of nodes representing divergence from source and first NYC transmission.