Supplemental Material for Maurano et al.
"Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City Region"

Table of Contents

Supplemental Figures are available in-line on the following pages; Supplemental Tables S2-S4 and Supplemental Code are available as separate files.

Supplemental Table S1. Summary of control samples.

Shown are three classes of control samples:

i. Spike-in positive control corresponding to a 10,000,000 copies/mL sample concentration was made using 100,000 copies of synthetic viral RNA (Genbank MT007544.1, Twist Bioscience) spiked into 5 ng of human total RNA (ThermoFisher cat. 4307281).

ii. No-sample controls, where RNA extraction was performed on buffer only.

iii. Negative-sample controls, human nasopharyngeal swab testing PCR-negative for SARS-CoV-2.

Arbitrary Batch IDs are listed to group samples processed on same plate. Mean_Viral_Coverage, mean coverage

Batch	Control Type	Sample Type	BS	Mean_Viral_Coverage	QC Result
1	Positive control	Capture	BS04652A	7,065	PASS
1	Positive control	Capture	BS04653A	3,373	PASS
1	Positive control	Capture	BS04654A	3,347	PASS
2	No sample control	Shotgun	No library detected by TapeStation		
2	No sample control	Shotgun	No library detected by TapeStation		
2	PCR negative control	Capture	BS05329A	2	FAIL
3	PCR negative control	Shotgun	BS05638A	3	FAIL
4	PCR negative control	Shotgun	BS05720A	3	FAIL
4	PCR negative control	Shotgun	BS05721A	5	FAIL
5	PCR negative control	Shotgun	BS06206A	13	FAIL
5	PCR negative control	Shotgun	BS06207A	8	FAIL
6	No sample control	Capture	BS06832A	22	FAIL

Please note that Supplemental Tables S2-S4 and Supplemental Code are available as separate files

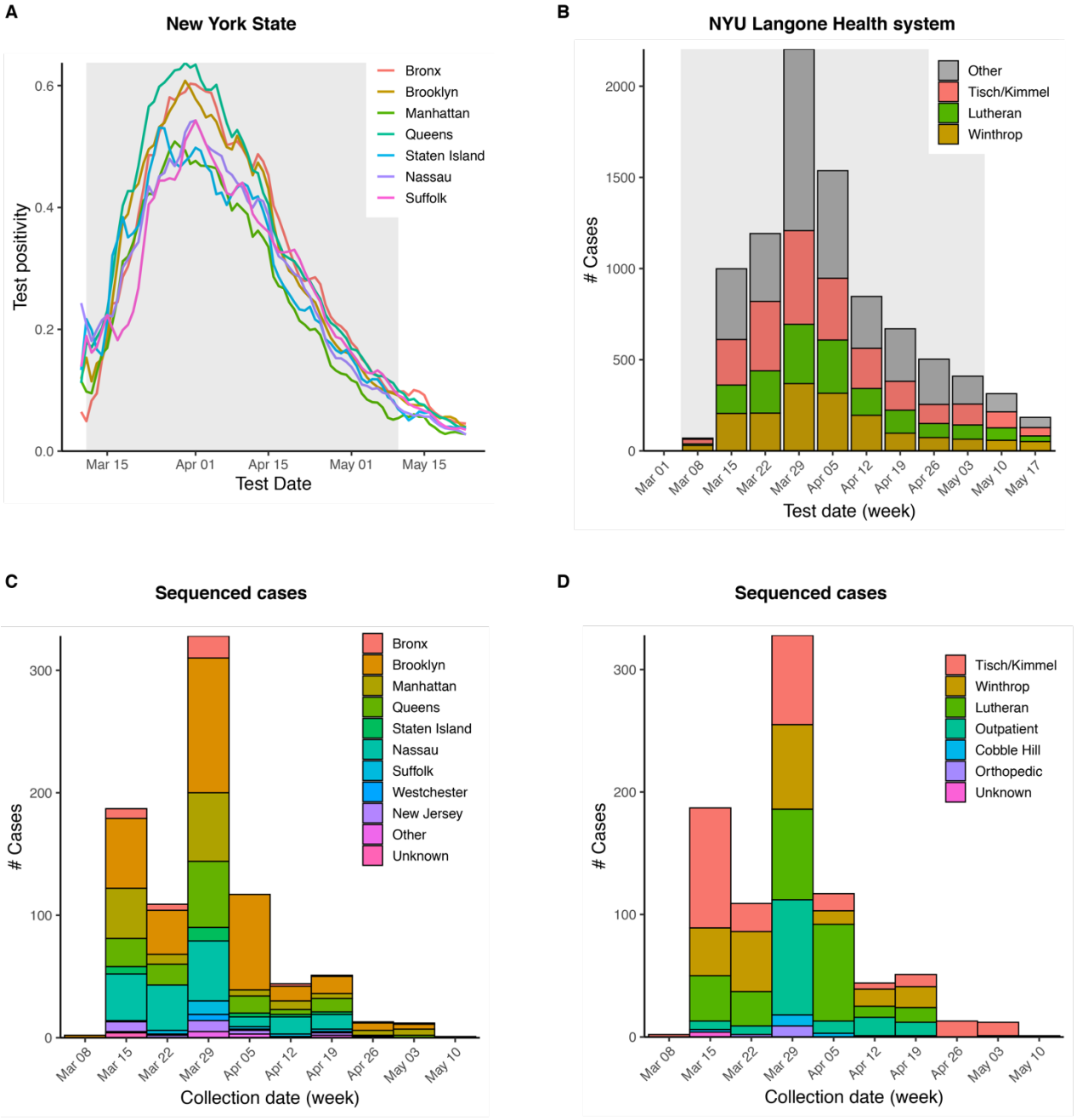
Supplemental Table S2. Summary of sequencing data.
Per sample summary of sequencing data for 864 cases and 10 controls. BS, biological sample ID. PropViralReads, proportion of nonredundant reads mapping to SARS-CoV-2 genome; analyzedViralReads, number of reads mapping to SARS-CoV-2 genome and passing 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.

Supplemental Table S3. Acknowledgements of GISAID sequences used.
List of sequences and contributors from GISAID.

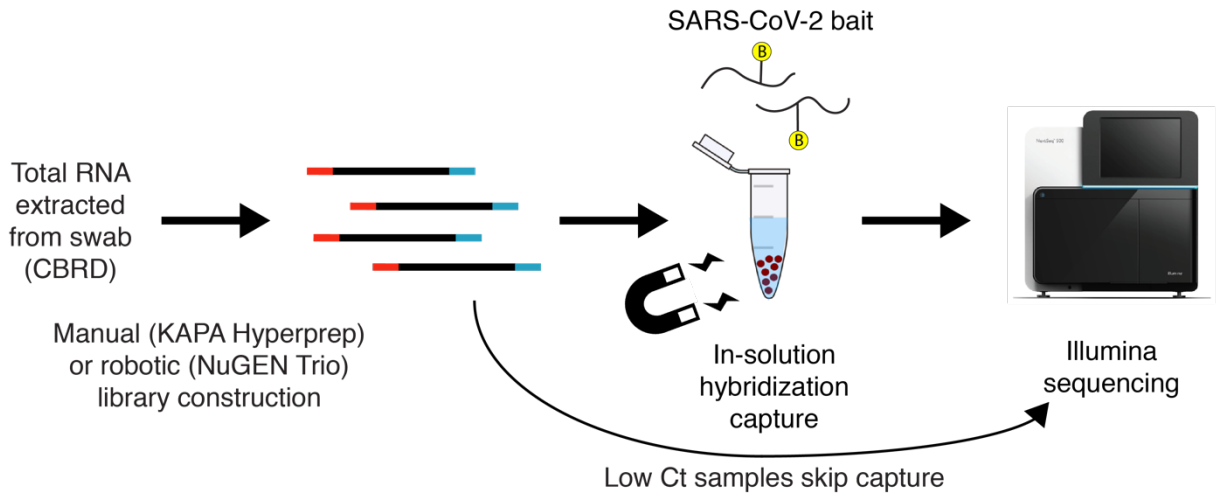
Supplemental Table S4. New York City Region transmission chains.
Summary of 109 transmission chains, including strain genotypes and date of nodes representing divergence from source and first NYC transmission.

Supplemental Code. Sequencing data processing pipeline.
Archive of sequencing processing pipeline (also available from https://github.com/mauranolab/mapping/tree/master/dnase).

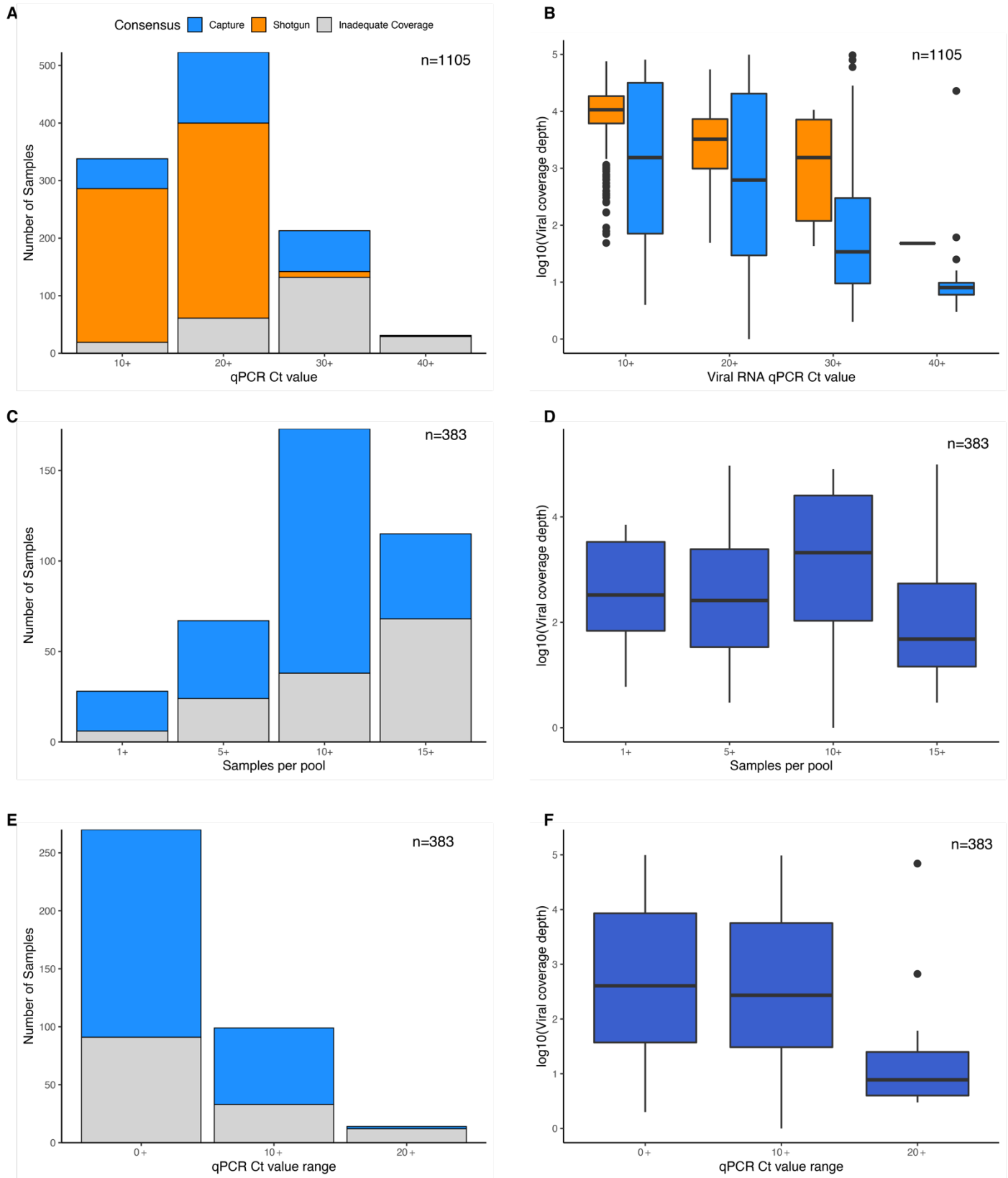
Supplemental Figures



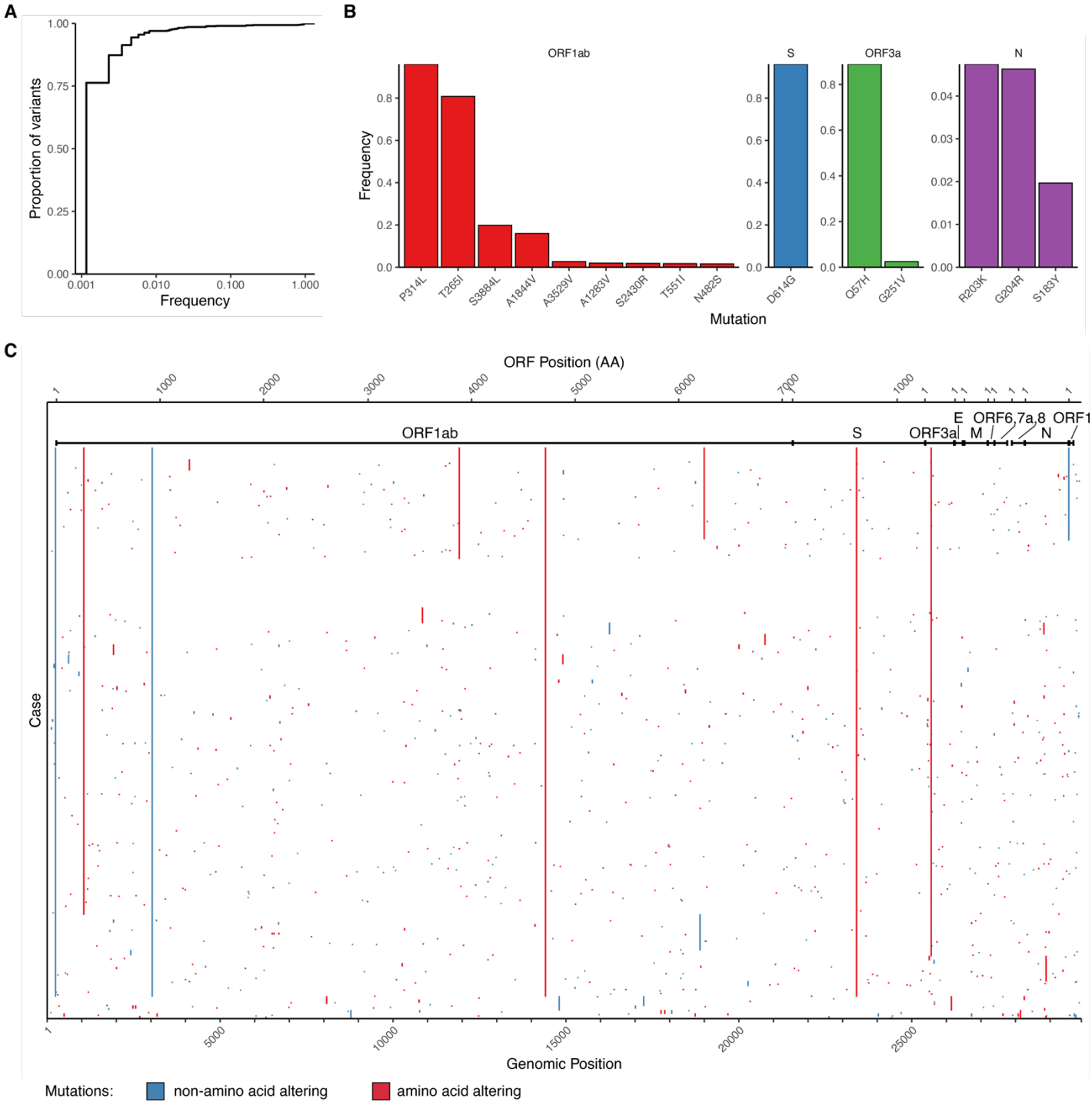
Supplemental Fig. S1. Outbreak trajectory and sampling of NYU Langone Health catchment area.
(A) SARS-CoV-2 positivity rate for New York City boroughs and outlying counties reported by New York State Department of Health (NYS Department of Health).
(B) Summary of weekly positive tests across NYU Langone 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).



Supplemental Fig. S2. 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.



Supplemental Fig. S3. Technical factors related to sequencing quality control.
Shown are sample counts by final quality control (QC) outcome (left) and average coverage depth (right) stratified by qRT-PCR Ct values (A-B), size of capture pool (C-D), and qRT-PCR 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.

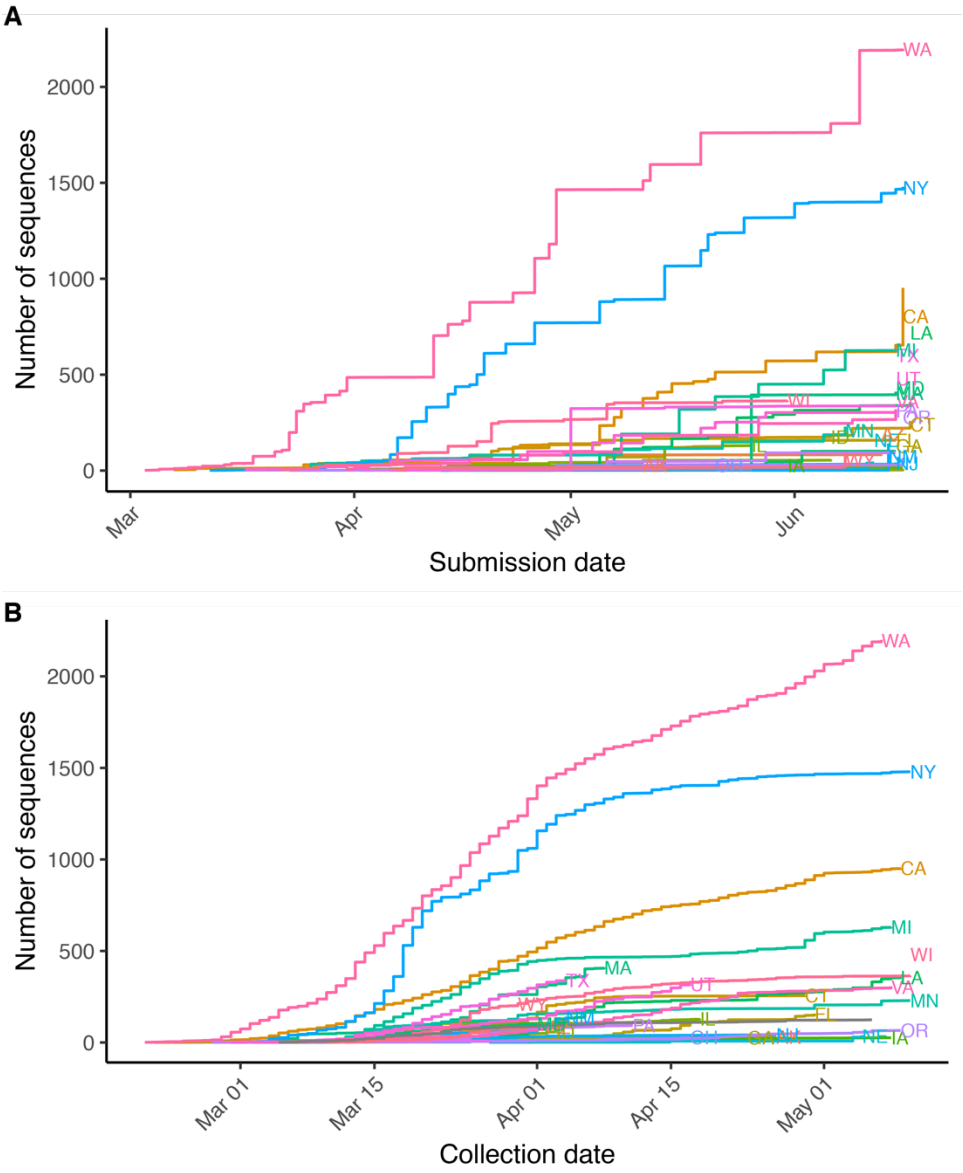


Supplemental Fig. S4. Overview of mutations identified.

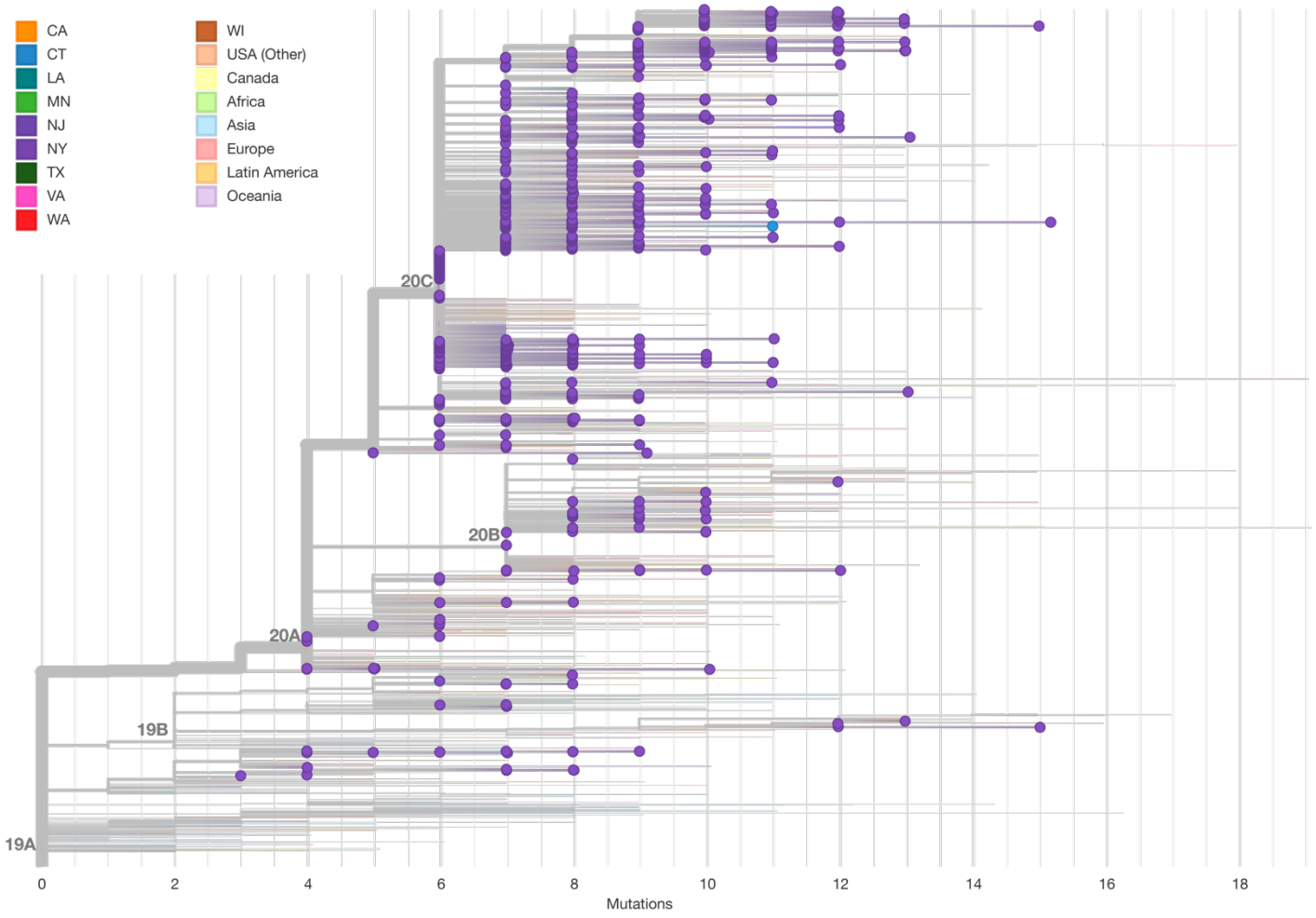
(A) Cumulative frequency distribution of variants identified across 864 cases.

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

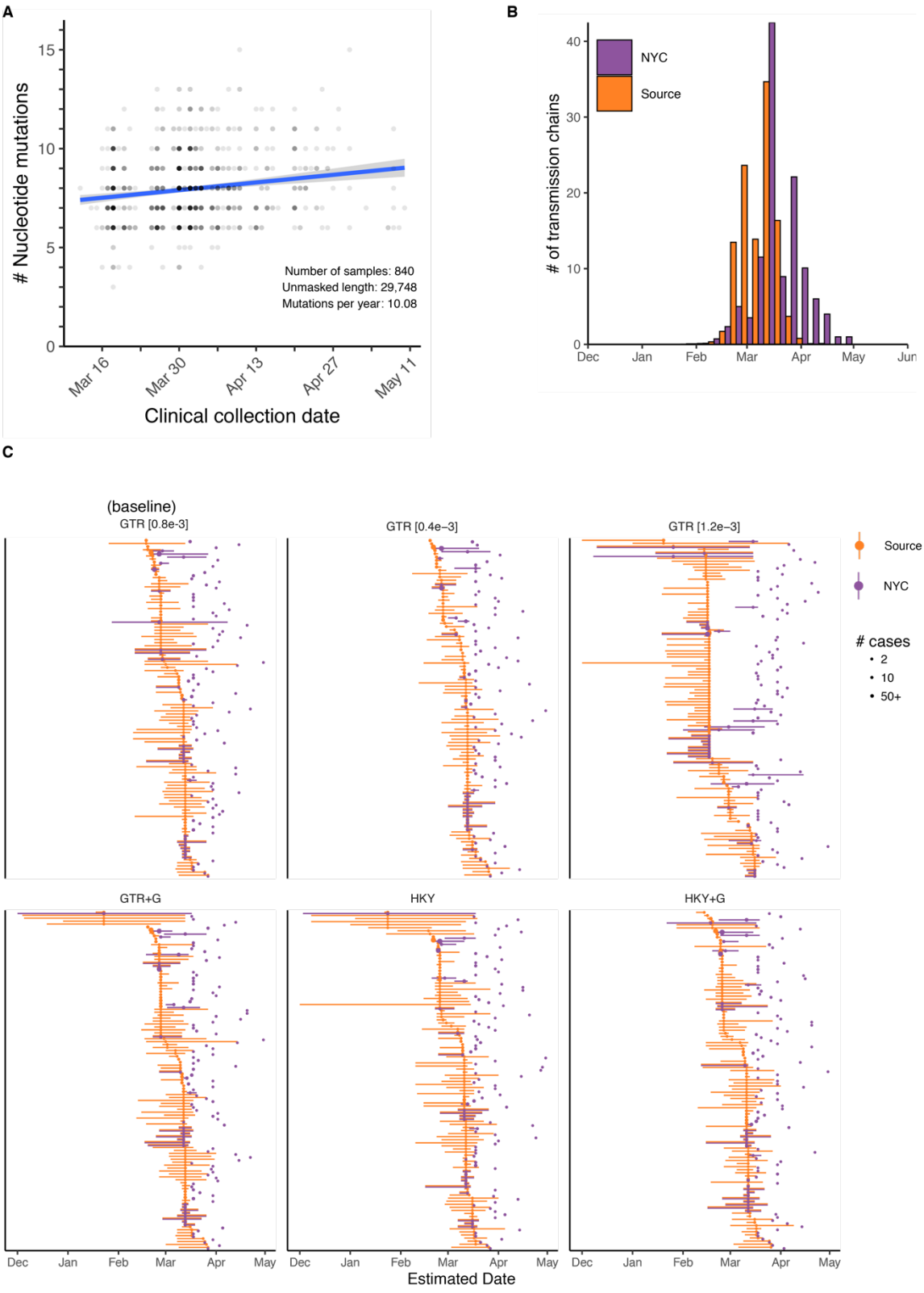
(C) Heatmap of mutations identified per case, with x-axis being genomic coordinates, and rows in the same order as **Fig. 2A**. ORFs are annotated according to GenBank MN908947.3.



Supplemental Fig. S5. 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.



Supplemental Fig. S6. Maximum likelihood tree including 5,004 global sequences from GISAID. NYULH sequences are highlighted with dots, tip and edge coloring indicates geographical location.



Supplemental Fig. S7. 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) Histogram of estimated dates for transmission chains identified in **Fig. 3**. (C) Effect of alternate substitution models (GTR+G, HKY and HKY+G) and substitution rates (0.4e-3 or 1.2e-3) on transmission chain dating. A time scaled phylogeny was inferred under alternate substitution models and substitution rates, and transmission chains re-identified for each.