Supplementary Materials

MATERIALS AND METHODS

Sample collections

This study was approved by the Partners Institutional Review Board under protocol 2019P003305 and MDPH IRB 00000701. We obtained samples and selected metadata from the MGH Microbiology Laboratory and MADPH under a waiver of consent for viral genomic sequencing. Samples were tested for SARS-CoV-2 by RT-qPCR. Samples that tested positive were eligible to be included.

Archived samples obtained from the MGH Microbiology Laboratory included nasopharyngeal (NP) swabs from five sources 1) all available cases prior to March 8, 2020, 2) all available samples from a skilled nursing facility in the Boston area (*29*), 3) samples from April 1 through April 14 from the MGH Respiratory Illness Clinic (RIC), established in Chelsea, MA, 4) samples from MGH Infection Control Unit investigations, and 5) samples drawn from the general pool of available samples tested by the MGH Microbiology Laboratory during the period from March 4 through May 9, 2020. Archived samples obtained from MADPH included NP swabs from 1) all available samples representing the first two known travel-associated introductions and the Berkshire County cluster and 2) all available samples submitted to MADPH from Boston Healthcare for the Homeless Program (BHCHP) from Mar 19 through April 18, a period that included universal screening (*16*).

Annotation of Cases

Epidemiological data on exposure and geography were obtained from medical record review (MGH) or collected by the DPH laboratory in the process of clinical testing. Zip code and county-level data were available for most samples from MGH. County-level data was available from DPH samples. Individuals who participated in the conference or who had known direct contact with attendees of the conference were deemed conference-associated (n = 28). One additional patient reported staying at the conference hotel but was diagnosed with COVID-19 over 1 month later;their exposure was considered unlikely to be due to the conference.

Viral sequencing

Samples were received at the Broad Institute as viral transport medium, universal transport medium, or molecular transport medium from NP swabs. In accordance with institutional biosafety committee approvals, samples were inactivated with Buffer AVL (Qiagen) or other chaotropic salt solution prior to extraction. RNA was extracted from 200uL of transport medium using either the QiAmp Viral RNA Mini Kit (Qiagen), or the MagMAX mirVana Total RNA Isolation kit on a KingFisher Flex automated extraction instrument (Thermo Fisher Scientific). Residual DNA was removed from the extracted material using TURBO DNase (Thermo Fisher Scientific).

Human ribosomal RNA was depleted using a ssDNA probe-based RNase H depletion method as previously described (*39*, *40*), or with the Ribo-Zero Plus rRNA Depletion Kit (Illumina). Unique ERCC RNA spike-ins were added to each sample as a quality control measure to track and mitigate potential cross contamination or downstream sample preparation issues. First and second strand cDNA was synthesized using either SuperScript III or IV Reverse Transcriptase (Thermo Fisher Scientific), and sequencing libraries were prepared with the Nextera XT or TruSeq RNA Library Prep kits as previously described (*39*, *40*). Libraries were sequenced using Illumina MiSeq, HiSeq, NextSeq, or NovaSeq machines with 100-nucleotide paired-end reads.

samples were extracted, prepared, and sequenced at the Broad Institute, Cambridge, MA, USA. The rRNA depletion, cDNA synthesis, and library construction protocols used in this study are publicly available on Benchling and can be found here: https://benchling.com/sabetilab/f_/gaLGu5X9-sabeti_group_sars-cov-2_metagenomic_sequencing_protocols/.

Genomic data analysis

We conducted all analyses using viral-ngs 2.0.21 on the Terra platform (app.terra.bio). All of the workflows named below are publicly available via the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). We demultiplexed individual libraries using the *demux_only* workflow for each lane of each flowcell, removed reads mapping to the human genome and to other known technical contaminants (e.g. sequencing adapters) using *deplete_only* (with bwaDbs=["gs://pathogen-public-dbs/v0/hg19.bwa_idx.tar.zst"] and blastDbs=["gs://pathogen-public-dbs/v0/GRCh37.68_ncRNA.fasta.zst", "gs://pathogen-public-dbs/v0/hybsel_probe_adapters.fasta"]), and performed reference-based assembly using *assemble_refbased* (once per sample, with all sequencing replicates merged in the read_unmapped_bams input and with a reference_fasta taken from https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta). We ran *assemble_refbased* on 1970 read set inputs spanning 1535 distinct samples (inclusive of controls).

We used the following stringent criteria to excluded any sample where i) fewer than 50,000 cleaned reads were obtained; ii) the proportion of reads mapping to the internal control (IC) sequence (ERCC spike-in) was >3 standard deviations from the mean observed for that IC sequence across all sequencing batches; iii) replicate genomes—where available—had 2 or more discordant SNPs or 1 or more discordant indels; iv) the number of normalized reads mapping to the SARS-CoV-2 genome was less than that observed in the highest negative control from the same sequencing batch. From the 1196 patient samples after filtering we obtained 850 assemblies with unambiguous consensus calls over at least 80% of the SARS-CoV-2 genome, and 778 with over unambiguous consensus calls over at least 98% of the SARS-CoV-2 genome, of which 772 were from unique individuals. We submitted 633 read sets to NCBI SRA and 837 genomes with at least 80% completeness to NCBI Genbank (using the *genbank* workflow). We used the 772 high-quality assemblies from unique individuals for the phylogenetic analyses described.

Failure to produce a SARS-CoV-2 genome from a PCR-positive sample may have been due to low viral titer, RNA degradation due to lack of sufficient cold chain, or technical sample handling issues (e.g. improper swab technique). Samples which failed to produce a genome at the first attempt were not further investigated at this time. To confirm the quality of our assemblies and mitigate any potential contamination we performed replicate library preparation and sequencing from RNA for 10% of samples. Among those samples that assembled a complete genome in both replicates, consensus-level genomes were identical.

Allele frequency was estimated as the proportion of derived / (derived + ancestral) versions of the allele. A 95% confidence interval was estimated for the proportion using the binomial distribution. The frequency of the iSNV for MA_MGH_00427 was calculated from 2 libraries; 50 reads contained the derived T allele and 146 reads contained the ancestral G allele based on the aligned reads from the viral-ngs pipeline (as described above).

Phylogenetic tree reconstruction

We constructed phylogenetic maximum likelihood (ML) and time trees with associated visualizations using the Augur pipeline (*augur_with_assemblies*). We used SARS-CoV-2-specific procedures taken from github.com/nextstrain/ncov, specifically setting the clock rate to 0.0008 +/- 0.0004, rooting the tree using the reference genome, and using the nextstrain site-masking and clade-definition files. In addition to our 772 genomes from unique individuals from Massachusetts, we included a global comparison set of 4,011 genomes subsampled from a download from the GISAID database on 15 June, 2020. These 4,011 genomes contain at most 50 representatives from each state or province in North America plus at most 50 representatives from each country outside of North America. Random subsampling was biased towards genomes genetically close to our focal set of genomes, using the distance matrix calculator at github.com/nextstrain/ncov/blob/master/scripts/priorities.py. The resulting augur output is visualizable on auspice.us or can be incorporated in custom deployments using Google Cloud Run using our template (github.com/dpark01/auspice-private-template); this template is used to showcase our data at auspice.broadinstitute.org.

We also conducted additional analysis of the genomes sequenced in this study. We aligned the set of 772 genomes using MAFFT v7.471(*41*) and trimmed 5' and 3' (first 265 and last 228 bases) UTRs from the alignment in R (*42*). To estimate the root-to-tip distance, we constructed ML phylogenetic trees using PhyML(*43*) v3.3.20190909 with default parameters using the MAFFT alignment of 772 genomes. We used TempEst (*44*) v.1.5.3 and selected the best-fitting root as identified using a heuristic residual mean squared function. To estimate branch support in maximum-likelihood phylogenies, we used IQ-Tree (*45*) with the ultrafast bootstrap and 10,000 bootstrap samples.

To construct Bayesian time-trees, we used BEAST 2.6.2 with a general time reversible substitution model with 4 rate categories drawn from a gamma distribution (GTR4G), a strict clock, coalescent exponential tree prior, a uniform [-inf, inf] prior for the clock rate, a 1/x [-inf, inf] prior for the coalescent effective population size; and a laplace [-inf, inf] prior for the growth rate. We ran the MCMC chain in BEAST2 for 100 million steps and thinned the chain by recording samples every 1000 steps. The first 30% of samples were discarded prior to calculating summary statistics from the posterior. We used TreeAnnotator v2.6.2 to construct maximum clade credibility trees with a burn-in percentage of 30%. We also compared a Hasegawa-Yoshino-Gawa substitution model with kappa = 2 and with 4 rate categories drawn from the gamma distribution (HKY4G) and ran this chain for 100 million steps using the same thinning and burn-in described for the GTR4G model.

Detection of respiratory virus co-infection

We used Kraken2 (*46*) to identify other viral taxa present in NP swab samples from COVID positive patients, excluding those removed by filters i and ii described above. To do so, we ran the *classify_single* workflow on all reads from all samples (with kraken2_db_tgz="gs://pathogen-public-dbs/v1/kraken2-broad-20200505.tar.zst", krona_taxonomy_db_kraken2_tgz="gs://pathogen-public-dbs/v1/krona.taxonomy-20200505.tab.zst", ncbi_taxdump_tgz="gs://pathogen-public-dbs/v1/taxdump-20200505.tar.gz", trim_clip_db="gs://pathogen-public-dbs/v0/contaminants.clip_db.fasta", spikein_db="gs://pathogen-public-dbs/v0/ERCC_96_nopolyA.fasta"). Our kraken2 database was constructed on 5 May, 2020, with the *kraken2_build* workflow (with standard_libraries=[“archaea”, “bacteria”, “plasmid”, “viral”, “human”, “fungi”, “protozoa”, “UniVec_Core”] and custom_libraries=[“gs://pathogen-public-dbs/v1/Hybsel_Viruses-20170523.2.fa.zst”, “gs://pathogen-public-dbs/v1/ercc_spike-ins-20170523.fa”]). The resulting per-sample outputs were run through the *merge_metagenomics* workflow and the resulting hits were filtered down to 20 common respiratory viruses of interest (adenovirus, HCoV-229E, HCoV-HKU1, HCoV-NL63, betacoronavirus 1, parainfluenza 1, parainfluenza 2, parainfluenza 3, Parainfluenza 4, enterovirus A, enterovirus B, enterovirus C, enterovirus D, influenza A, influenza B, human metapneumovirus, respiratory syncytial virus, SARS-CoV, MERS-CoV, human rhinovirus) using a threshold of 10 reads to identify a putative co-infection. We independently confirmed the presence of viral co-infections identified in the metagenomic sequencing data using the BioFire FilmAssay Respiratory Panel, performed at the MADPH or MGH Microbiology Laboratory. Three samples from early in the pandemic, for which no additional sample remained, were not tested.

Identifying viral importation events

For ancestral state inference, we inferred a state of “MA” vs “non-MA” using the augur pipeline(*47*). Cases whose ancestral state was inferred as non-MA with high confidence (>0.95) were considered imported cases. Conference-associated and nursing facility samples were excluded from the importation analysis.

Haplotype Network Reconstruction

Haplotype networks were visualized using the software tool PopART v1.7 (*48*). The assembled sequences were aligned against NC_045512.2 and the first 268bp at the 5' end and 230bp at the 3' end (UTR regions) were removed from the alignment. A nexus-format input file for PopART was created using a Python script to consolidate sequence information with metadata classifications. This script is available at http://www.github.com/broadinstitute/[repository]. A TCS network of the sequences (*49*) was constructed in PopART. Regions where any sequence had ambiguous bases were masked. For the construction of haplotype networks in Figure 4, one sample, MA_MGH_00090, was removed to prevent masking of the G3892T variant. For the displayed haplotype networks, the area of the circle corresponds to how many verbatim-identical sequences (after masking) bin together as the same haplotype. The hash marks on the edges indicate the SNP distance between sequence haplotypes (1 mark=1 SNP apart). Gene graphs were constructed using pairwise distance matrices computed on aligned SARS-CoV-2 genomes and clustered using the R package adegenet (*50*).

SNF genetic diversity analysis

For this analysis, the main SNF cluster was restricted to samples collected before April 15, 2020, and the conference cluster to samples collected before March 8, 2020. We assumed that the number of transmissions was the minimum possible (one fewer than the number of samples in the cluster). The p-value for the comparison between the clusters assessed the probability that the observed numbers of mutations were produced by Poisson processes with the same value of λ, using the R function poisson.test (in the *stats* package v3.6.2). For the expected number of mutations, we assumed that substitutions occur predominantly during the transmission bottleneck and calculated the expected rate based on a generation time of 5.0

days (*51*) and a mutation rate of 1.0 x 10^{-3}/bp/year (Fig. S5C), which together yield an expectation of 0.41 substitutions/transmission.

Epidemiological and demographic data analysis

We downloaded publicly available daily and weekly data on cases of SARS-CoV-2 in MA for the period January 1 - August 1 from the website of the MADPH (https://www.mass.gov/info-details/covid-19-response-reporting). This data included cases by day, cases by county over time, and cases involving congregate living facilities and staff. We compiled detailed case statistics by exposure category using the press releases reporting early case totals and exposure available on the MADPH website. During the study period, an additional case from February 6, 2020, was added to MADPH tallies. This case was missing detailed case information such as exposure category and was not included in early press releases from MADPH; it was therefore excluded from the tallies of cases by exposure category and estimates of the sampling proportion, but included in total case counts over time as reported in the main text to incorporate the most recent tallies. To calculate the cumulative proportion of alleles by county, conference-associated and SNF-associated individuals were removed and the cumulative allele frequency through the end of the study period was calculated for each of the four counties with the largest numbers of genomes (Suffolk, Middlesex, Essex, and Norfolk). To calculate the proportion of domestic and global sequences from the GISAID database, a multiple sequence alignment of 58,043 complete GISAID genomes was downloaded on July 14 2020 and the percentage of ancestral and derived alleles was extracted from the alignment and plotted by geographic category.

We used R (*42*), Bioconductor (*52*), ggplot2, tidyverse (*53*), and ggtree (*54*) to clean and plot data and trees, and choroplethr to draw maps.

Fig S1. A. Counts of complete genomes reported in this study, by county. **B.** Case counts by county reported by MADPH through July 1, 2020. **C.** Scatterplot of counts of complete genomes in this study vs. MADPH-reported cases through July 1, 2020. **D.** Sampling proportion by county (fraction of complete genomes sequenced in this study out of total cases by county reported to MADPH through July 1, 2020).

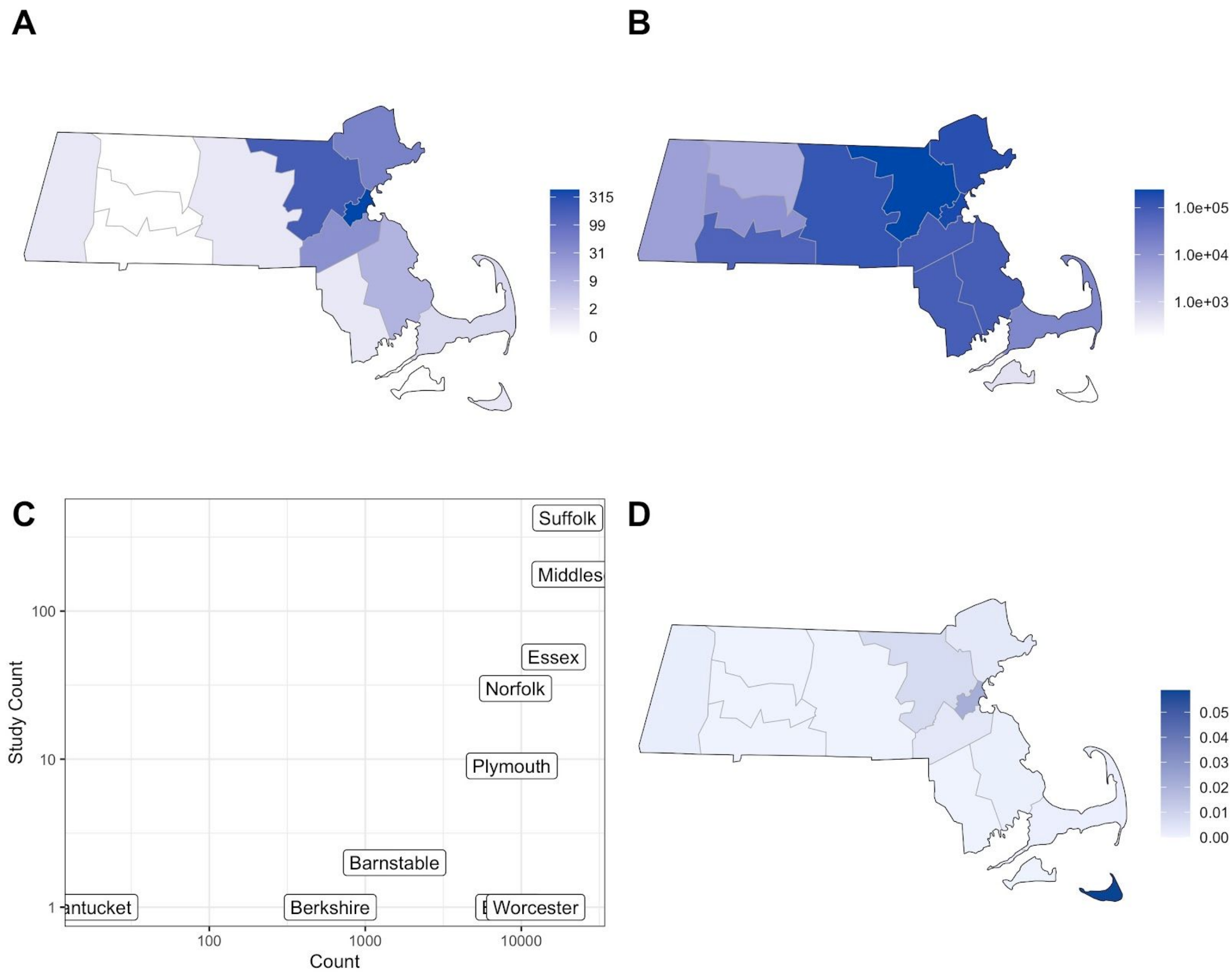


Fig S2. A. Mean coverage (on a log scale) vs. viral C_t for all samples included in the study. A linear regression fit is shown in blue. **B.** Fraction of the genome that is complete is shown vs. viral C_t. A $C_t < 28$ was strongly associated with recovery of a complete virus genome. Fit from a logistic regression model is shown in blue. **C.** The numbers of genomes at given thresholds of completeness are displayed. **D.** Histogram of the numbers of genomes at different thresholds of completeness. **E.** Combined coverage across sequenced SARS-CoV-2 genomes. [*next page]*

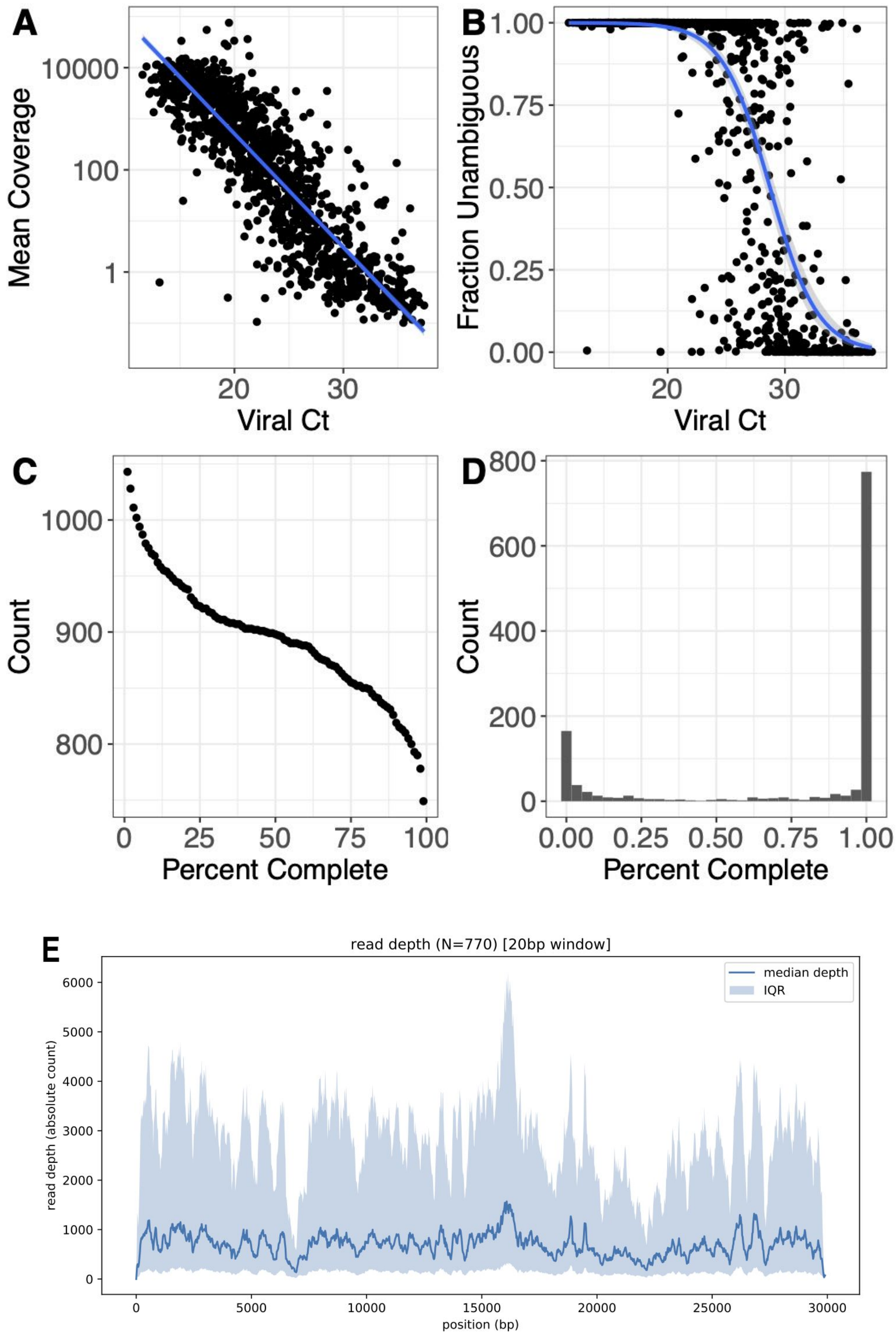

A
Mean Coverage
10000
100
1
20
30
Viral Ct
B
Fraction Unambiguous
1.00
0.75
0.50
0.25
0.00
20
30
Viral Ct
C
Count
1000
900
800
0
25
50
75
100
Percent Complete
D
Count
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00
Percent Complete
E
read depth (N=770) [20bp window]
median depth
IQR
read depth (absolute count)
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
position (bp)

Fig S3. A. Scatterplot of MGH Roche cobas 6800 instrument PCR Ct values for SARS-CoV-2 target vs. quantification prior to library construction. **B.** Scatterplot of MGH Roche cobas 6800 instrument PCR C_t values for Pan SARS target vs. quantification prior to library construction. **C.** Scatterplot of Roche cobas 6800 PCR C_t targets. **D**. Scatterplot of DPH N1 assay vs. quantification prior to library construction. **E**. Scatterplot of DPH N2 assay vs. quantification prior to library construction. **F.** Scatterplot of DPH N1 vs. N2 targets. **G.** Scatterplot of MGH Roche cobas 6800 instrument PCR C_t values for SARS-CoV-2 target vs. mean coverage (log 10 scale). **H.** Scatterplot of MGH Roche cobas 6800 instrument PCR C_t values for Pan SARS target vs. mean coverage (log 10 scale). **I.** Quantification prior to sequencing vs. mean coverage (log 10 scale) for MGH samples. **J.** Scatterplot of DPH N1 assay vs. mean coverage (log 10 scale). **K.** Scatterplot of DPH N2 assay vs. mean coverage (log 10 scale). **L.** Quantification prior to sequencing vs. mean coverage (log 10 scale) for DPH samples.

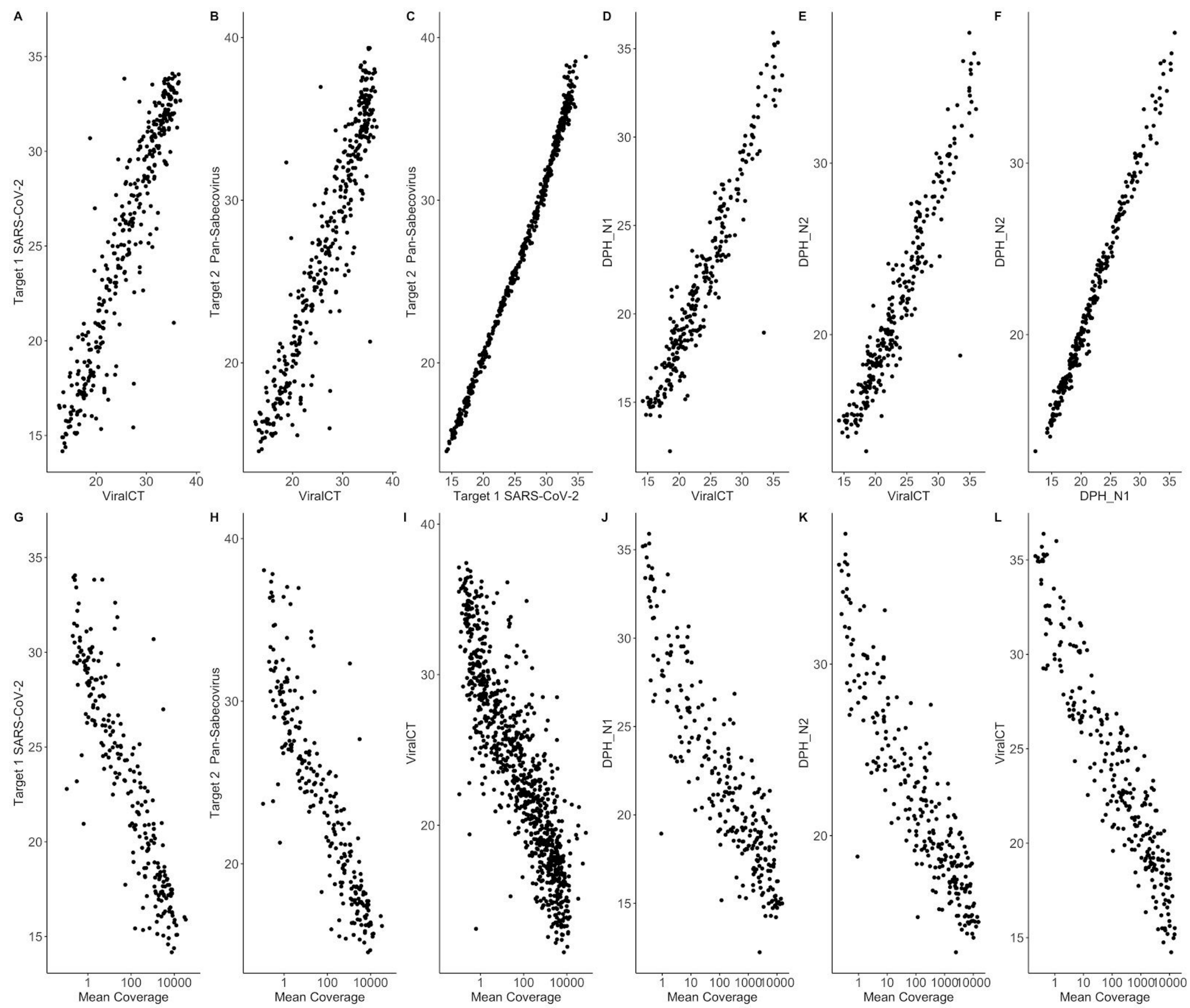


Fig S4. A. Distance matrix of pairwise distances for all complete genomes (>98% complete) from unique individuals in this study. **B.** Histogram of pairwise distances for all possible pairwise comparisons between complete genomes in the study. **C.** Tajimas's D values in 500-base-pair intervals across the genome.

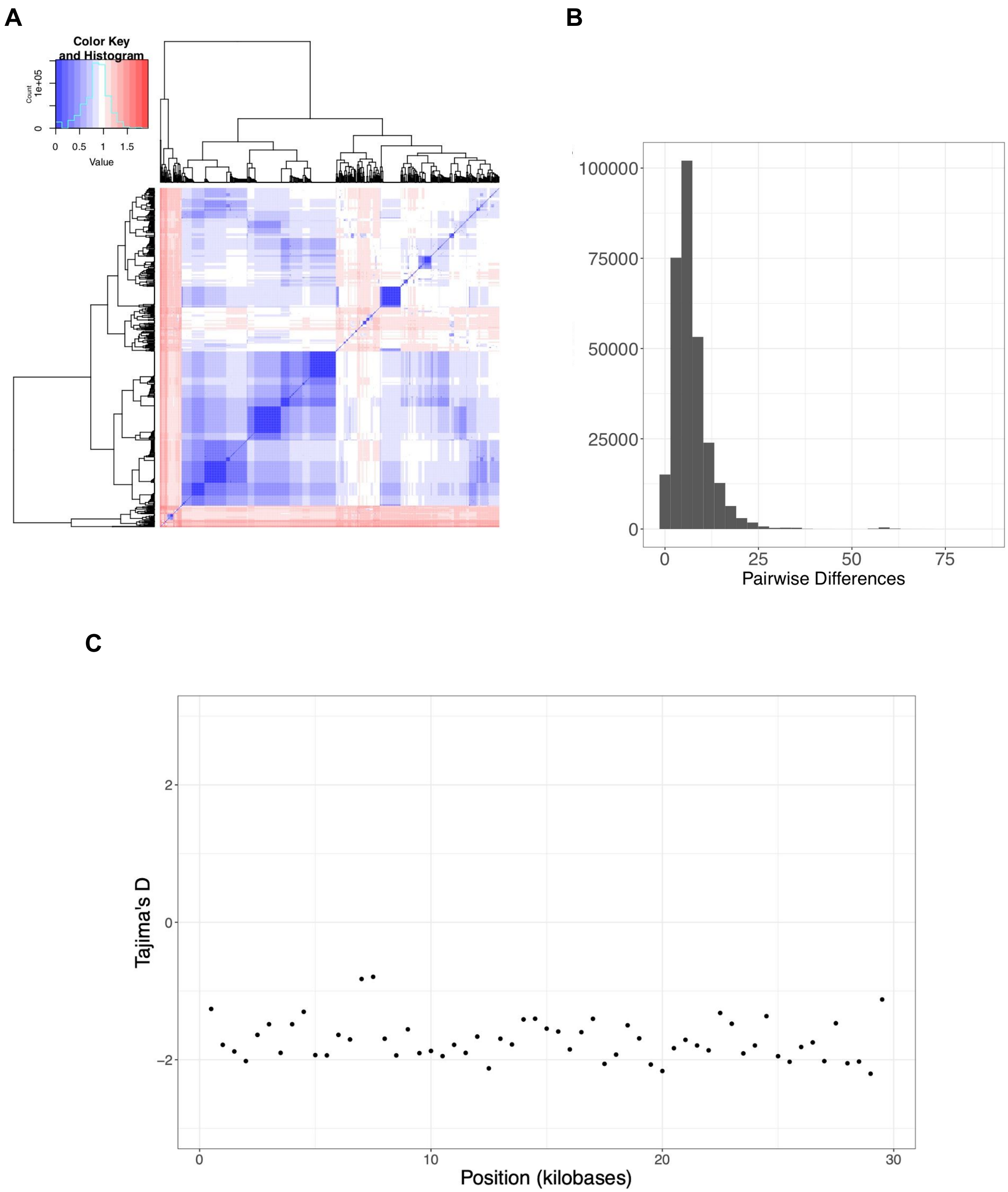


Fig S5. A. Linear regression of root-to-tip distance vs. date of sampling. Root-to-tip distance was calculated using TempEst (*44*) on maximum likelihood trees inferred using PhyML (*43*).

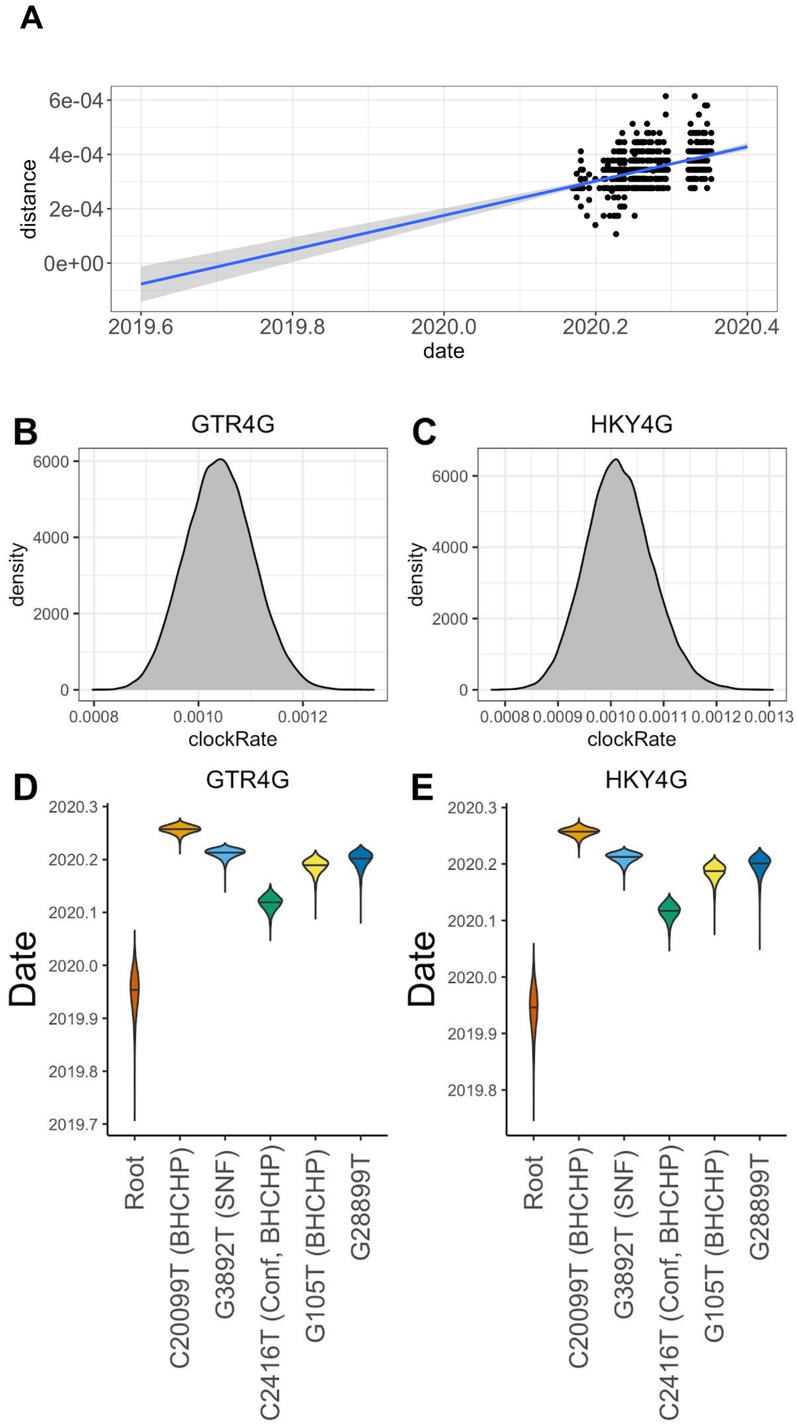


Fig S6. Probability of an importation event over time. Samples whose ancestral state was inferred as non-MA are coded as 1 and samples whose ancestral state is inferred as MA are coded as 0 (a small amount of noise is added to the y-coordinate to show the density of the data). A logistic regression (red curve) shows the probability of importation decreasing through the study period (β_1 = -0.04999 +/- 0.01056, p = 2.2e-06). A loess smoother is shown with a dashed line.

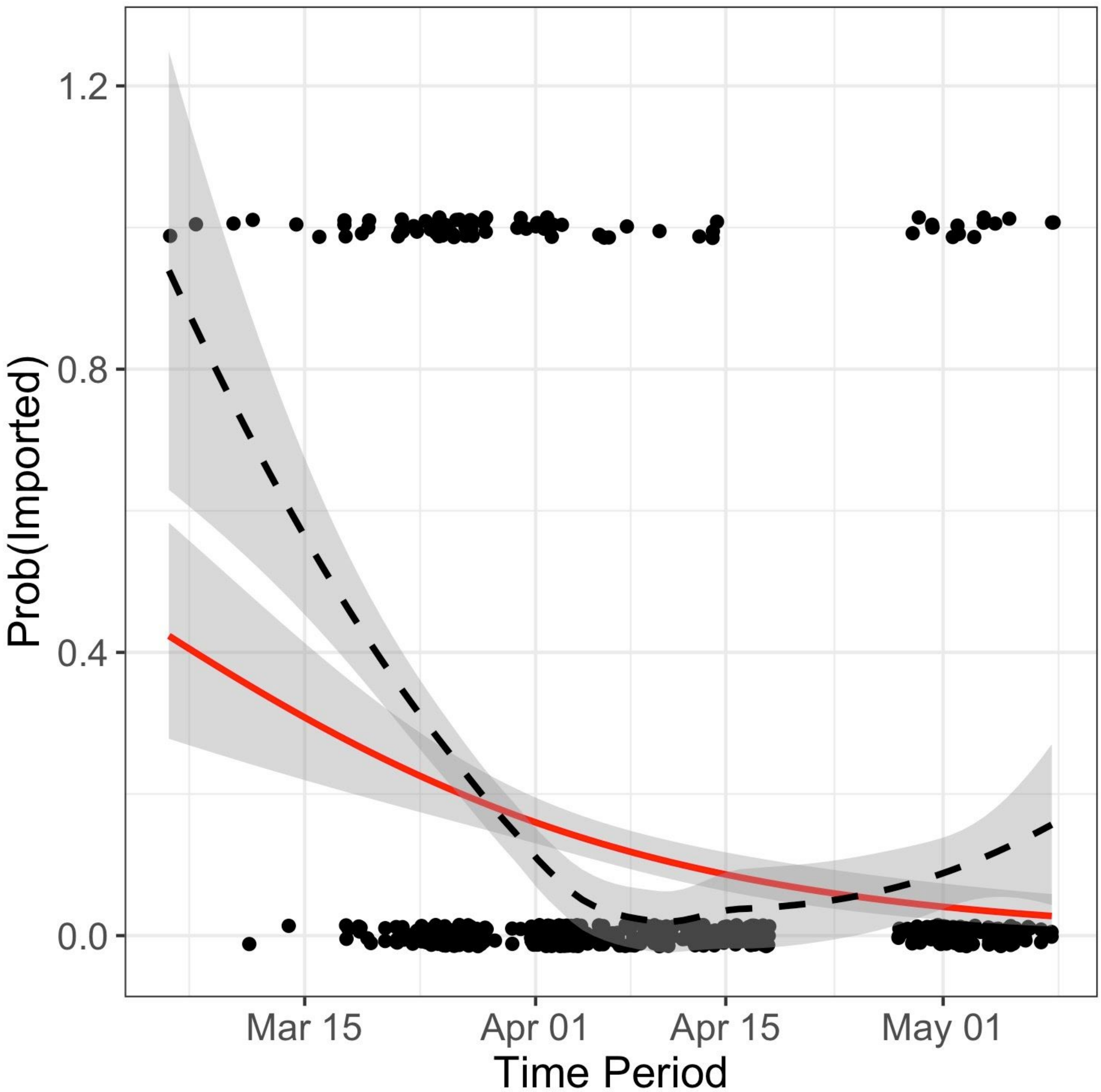


Fig S7. Portion of global time-stamped phylogeny (inferred using augur(*47*) with GISAID and MA genomes) containing MA-1 (red arrow).

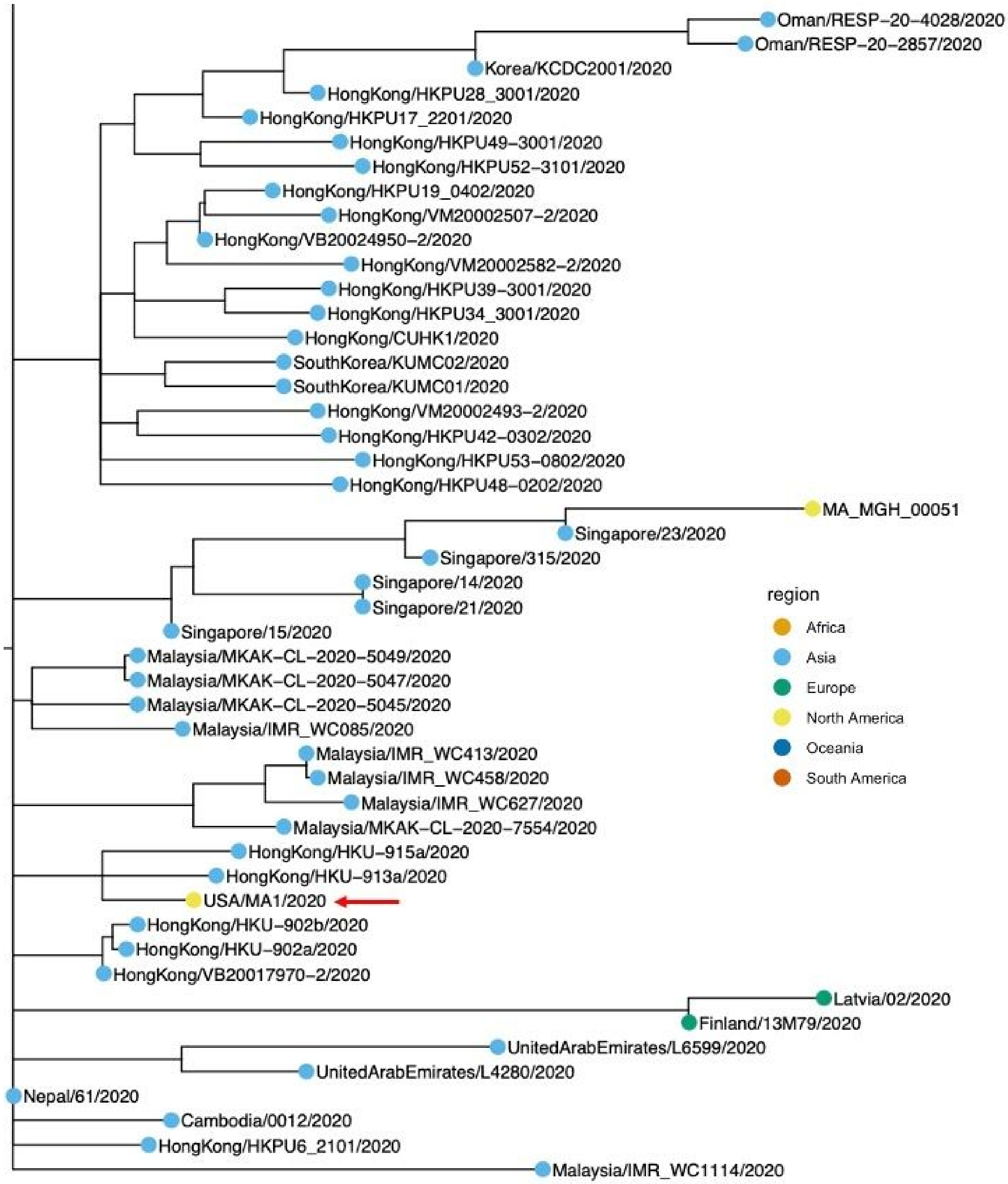


Fig S8. Portion of the global time-scaled phylogenetic tree (inferred using augur(*47*) with GISAID and MA genomes) containing DPH_00002 and DPH_00003 (marked with red arrows).

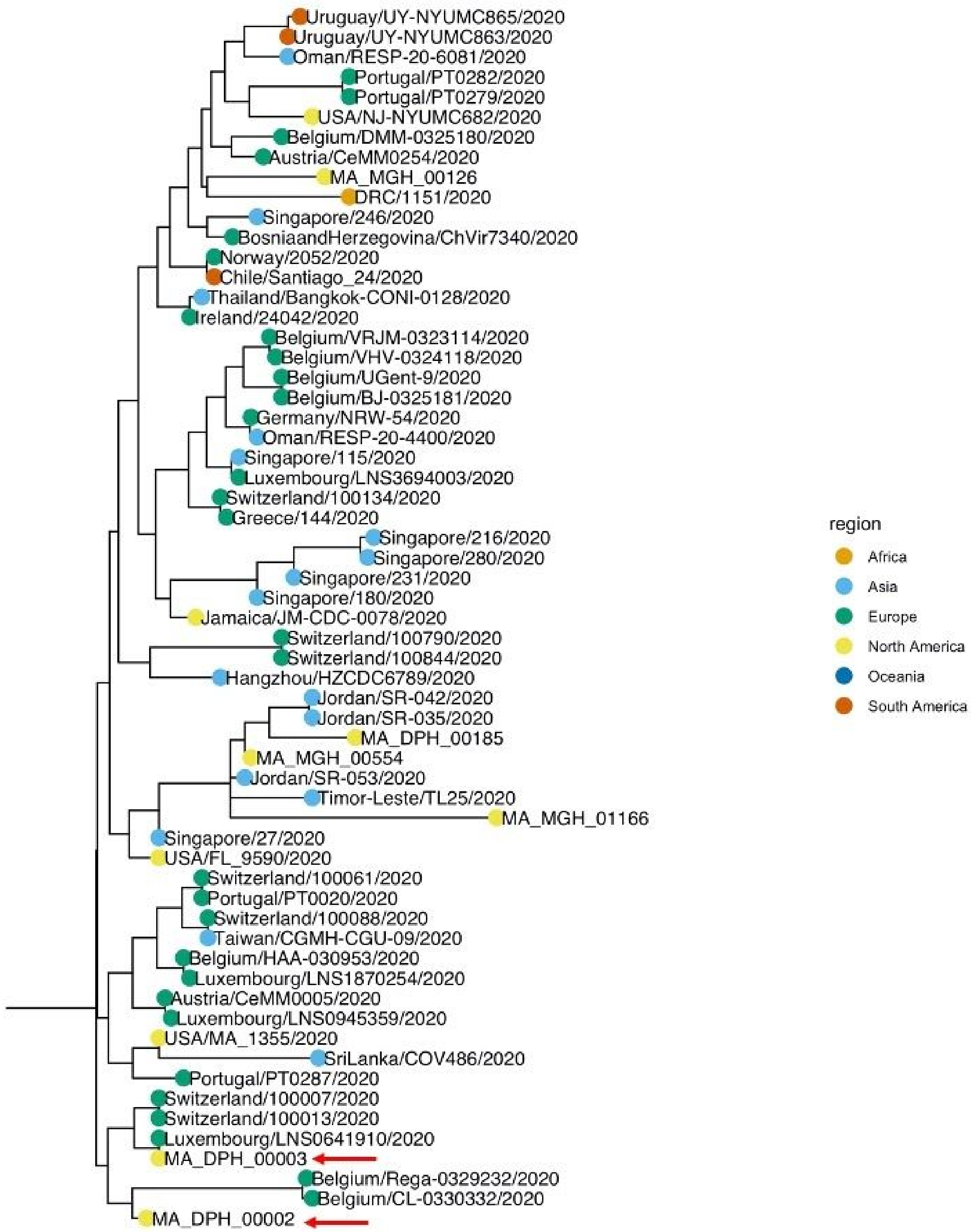


Fig S9. Phylogenetic ML tree of MA samples from this study plus partial genomes (>5kb) from Berkshire County Cluster. Ultrafast bootstrap support (when > 80) is shown at nodes. Tips corresponding to samples collected from the Berkshire County cluster are shown in orange.

Fig S10. Phylogenetic position of DPH_00011 (type genome for the Western MA cluster, sample labeled with red arrow).

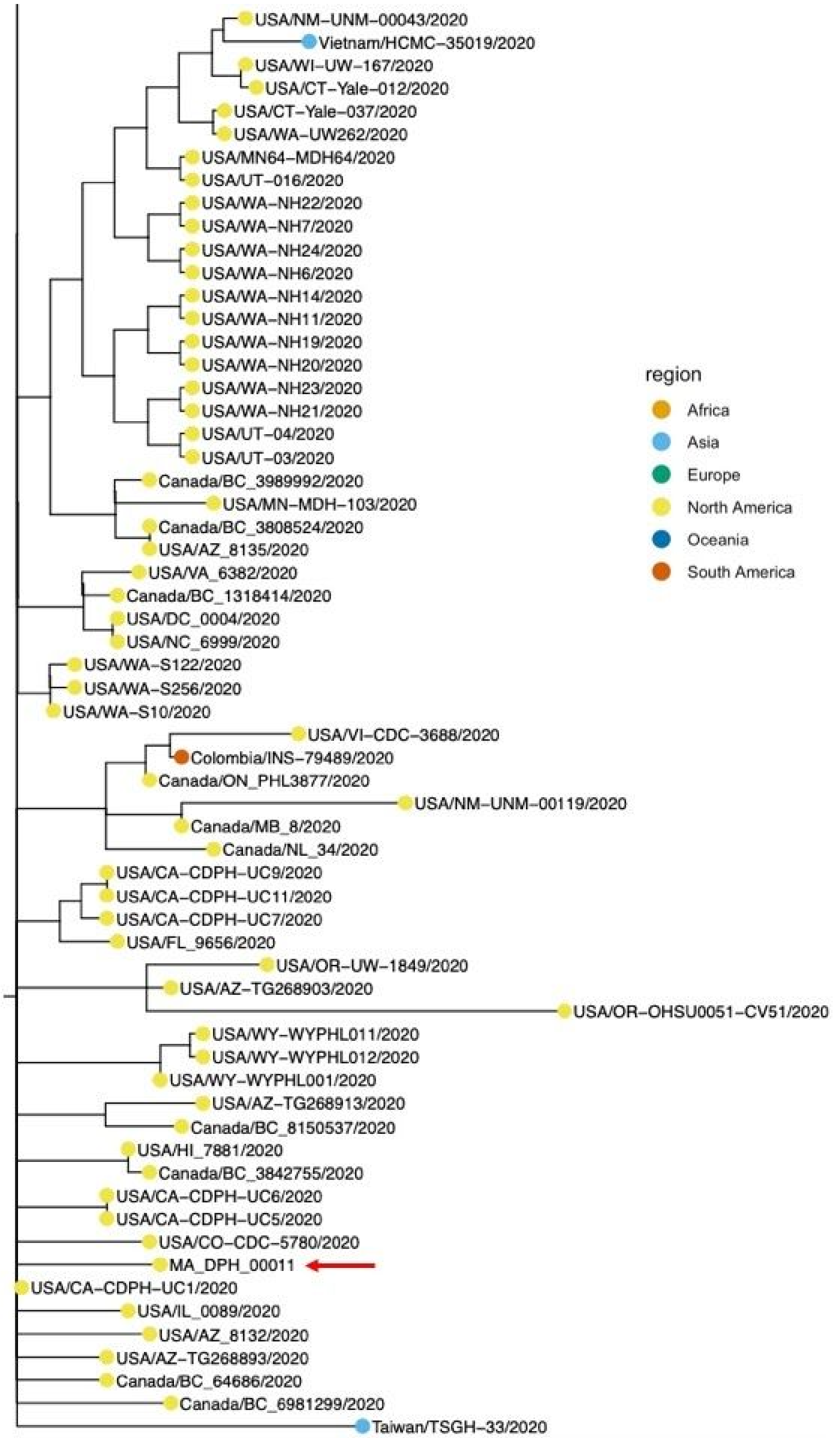


Fig S11. ML Tree of MA genomes computed using IQtree, with ultrafast bootstrap support shown at nodes with support > 80.

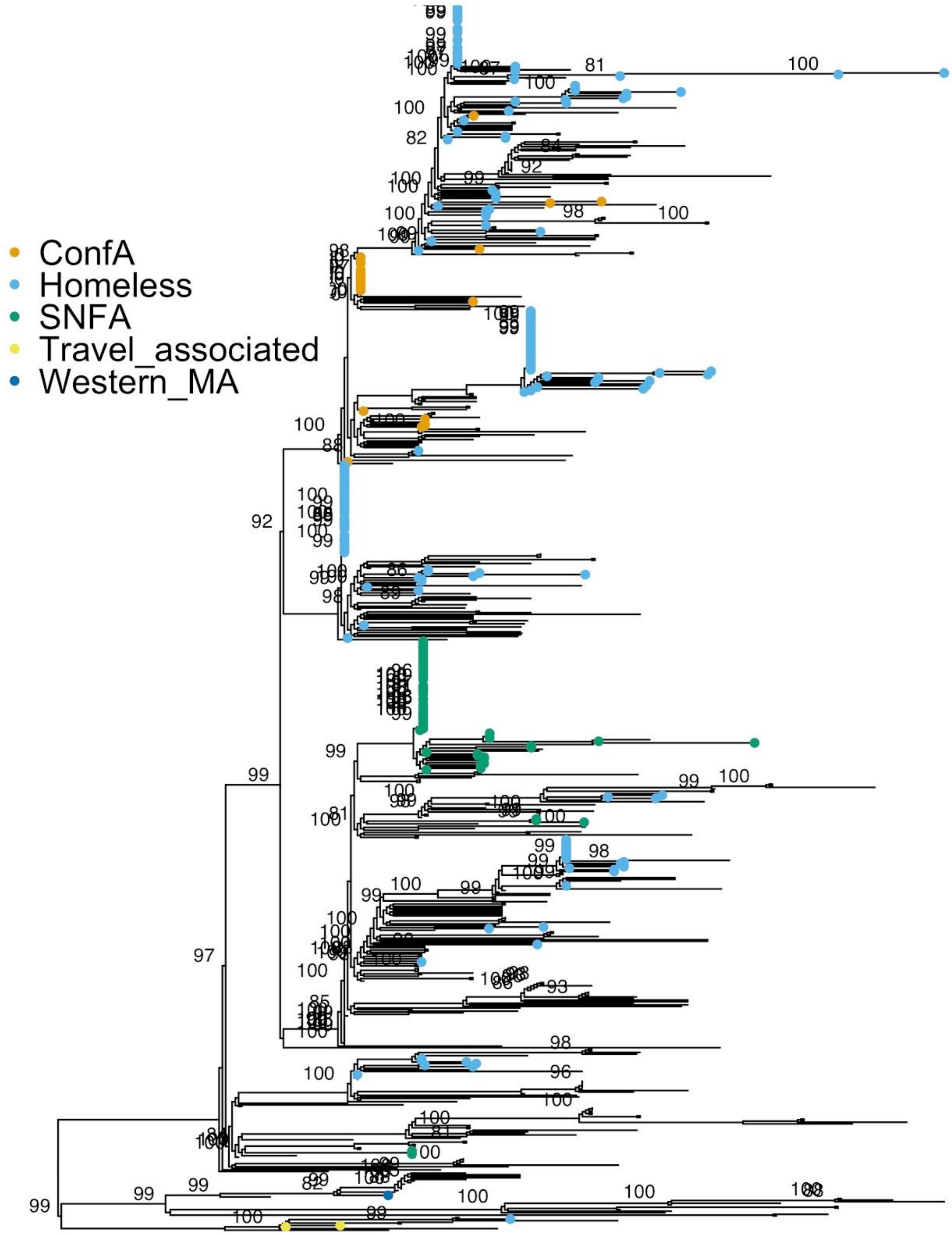
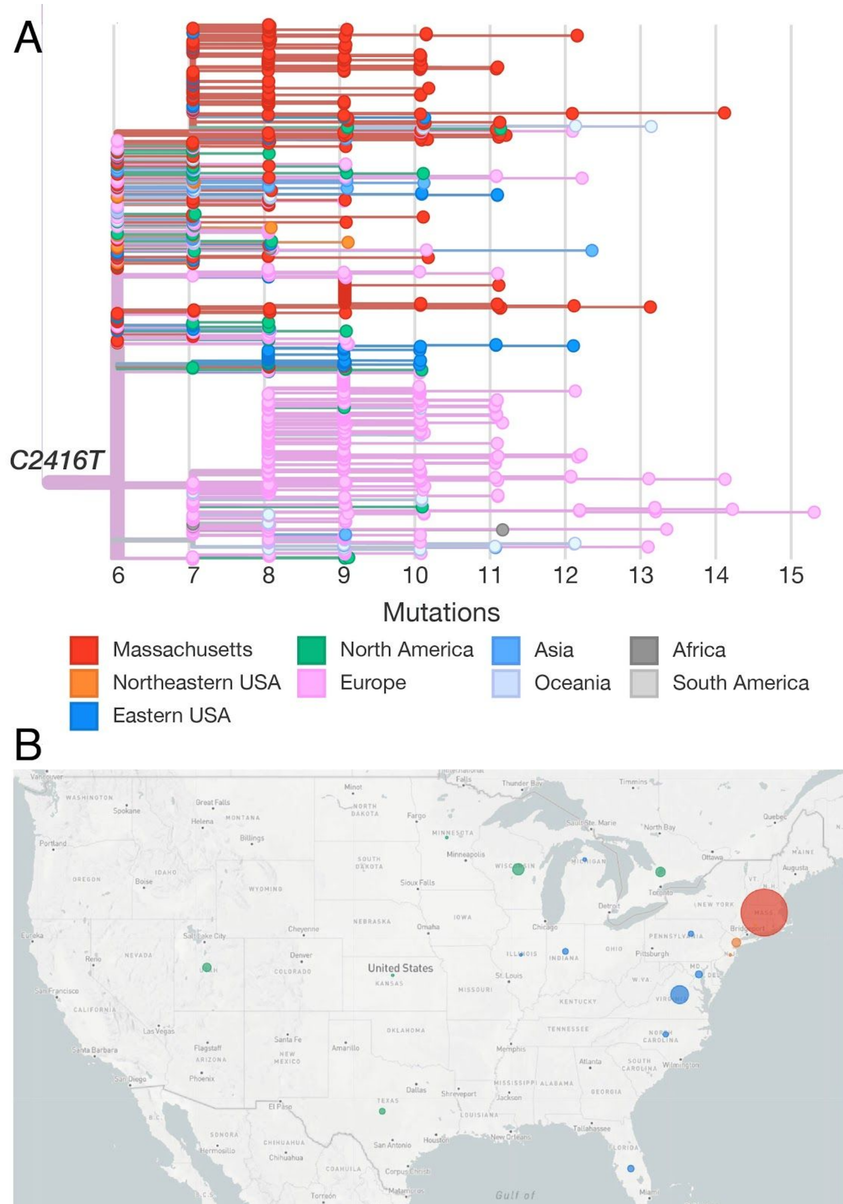


Fig S12. A. Divergence tree of the C2416T variant showing all global sequences (in GISAID through June 14, 2020) with the C2416T variant. **B.** Map showing the distribution of the C2416T variant across the United States. Circle size reflects the number of reported genomes per state. **C.** Phylogeny (left panel) and map showing global distribution of C2416T/G8371T. **D.** Phylogeny (left panel) and map showing global distribution of C2416T/G20578T.



C


Phylogeny
RESET LAYOUT
Category Geolocation
Massachusetts
Northeastern USA
Eastern USA
North America
Europe
Asia
Oceania
Africa
South America
New Hampshire
Mutations
Geography
RESET ZOOM
PLAY
RESET

D

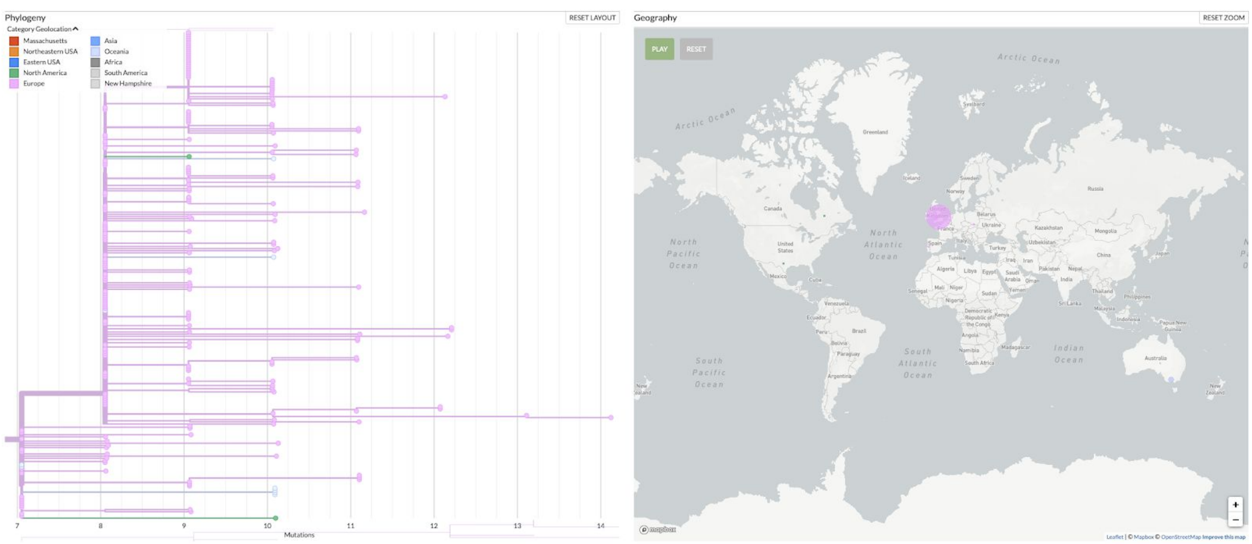

Phylogeny
RESET LAYOUT
Category Geolocation
Massachusetts
Northeastern USA
Eastern USA
North America
Europe
Asia
Oceania
Africa
South America
New Hampshire
Mutations
Geography
RESET ZOOM
PLAY
RESET

Fig S13. ML with samples from MGH suspected nosocomial clusters labeled. Posterior support for strongly supported (ultrafast bootstrap support > 80) nodes is shown.

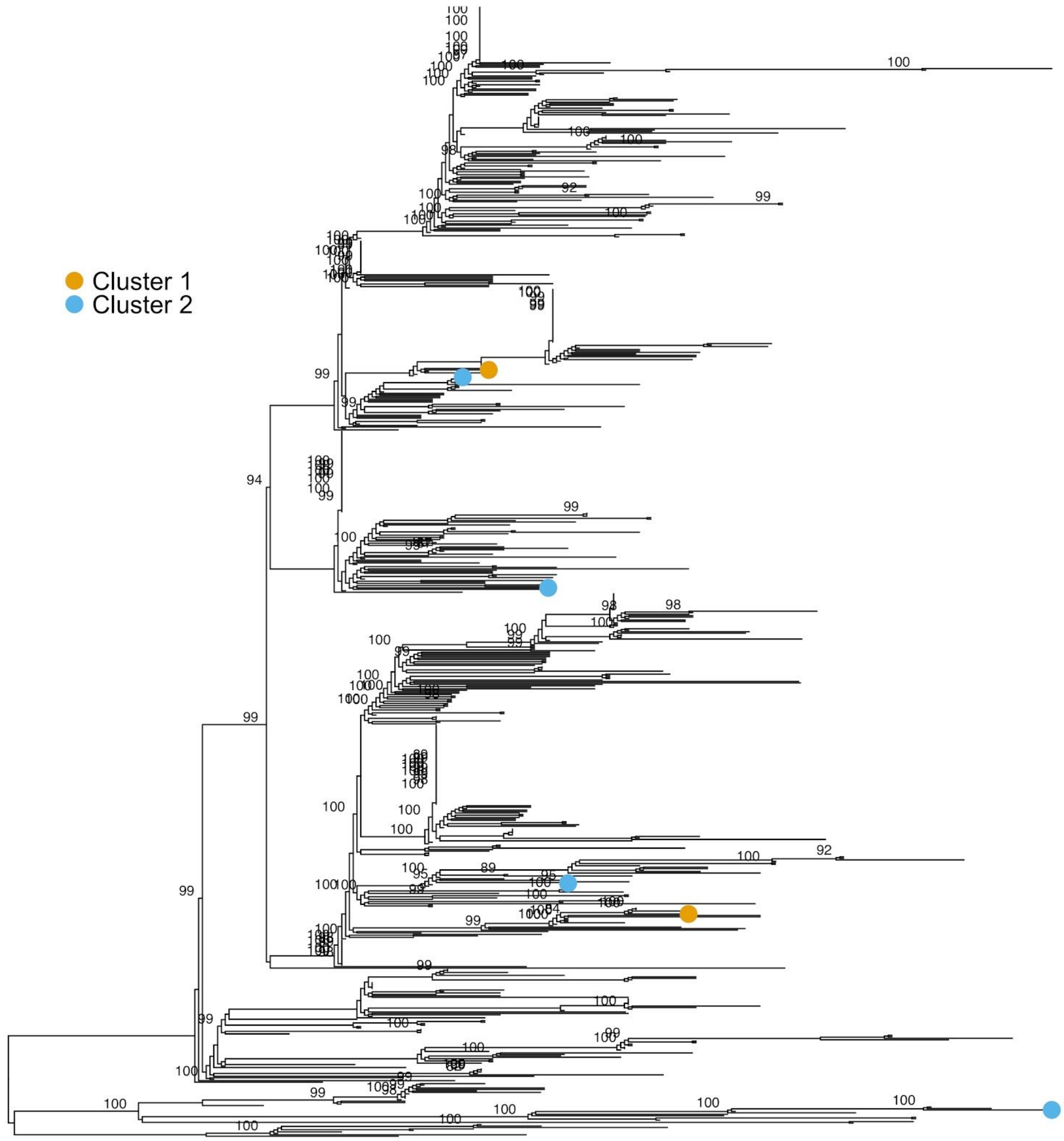


Fig S14. Sequenced samples labeled by zip code of residence for the top three zip codes in the set of 772 genomes from unique patients.

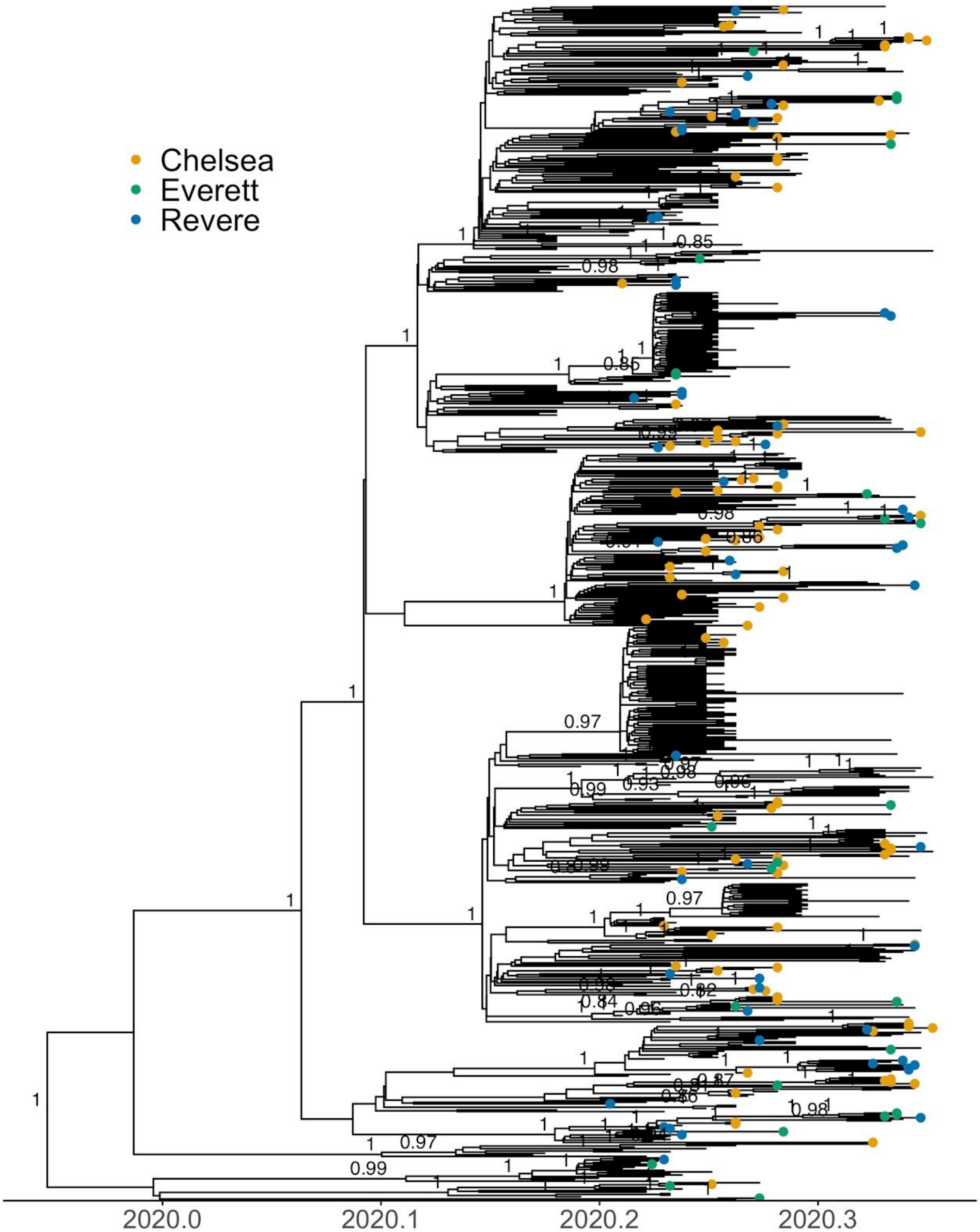


Fig S15. Geographic distribution of select lineage-defining variants in Eastern Massachusetts. The scale is in log10(case counts + 1).

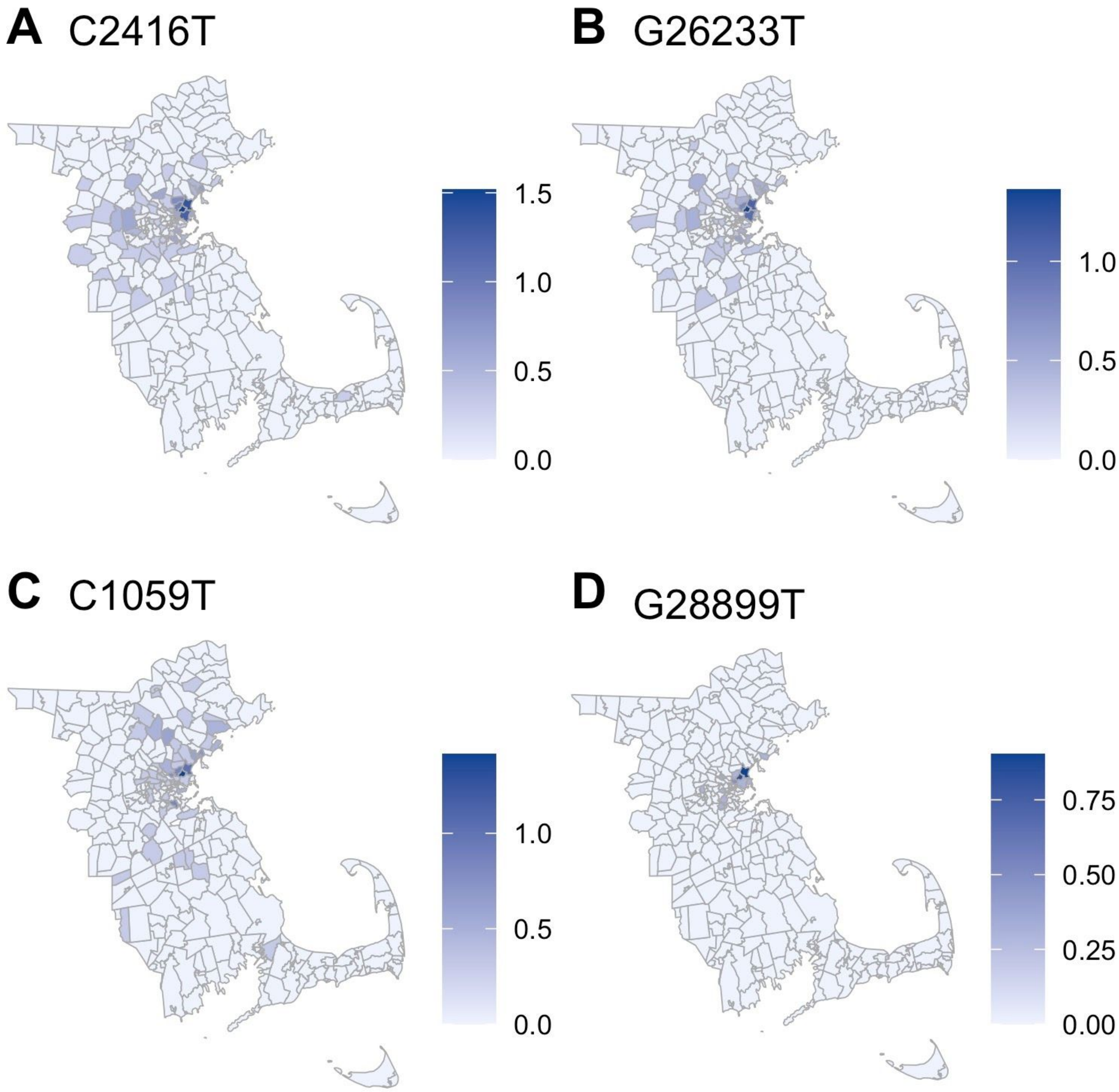


Fig S16. Prevalence of conference-associated variants by day among the Chelsea Respiratory Illness Clinic (RIC), among individuals experiencing homelessness sampled by BHCHP, and among samples available from the MGH Microbiology Laboratory that were not a part of known clusters (Conference, SNF) or from the Chelsea RIC. The solid line gives the cumulative allele frequency in each group.

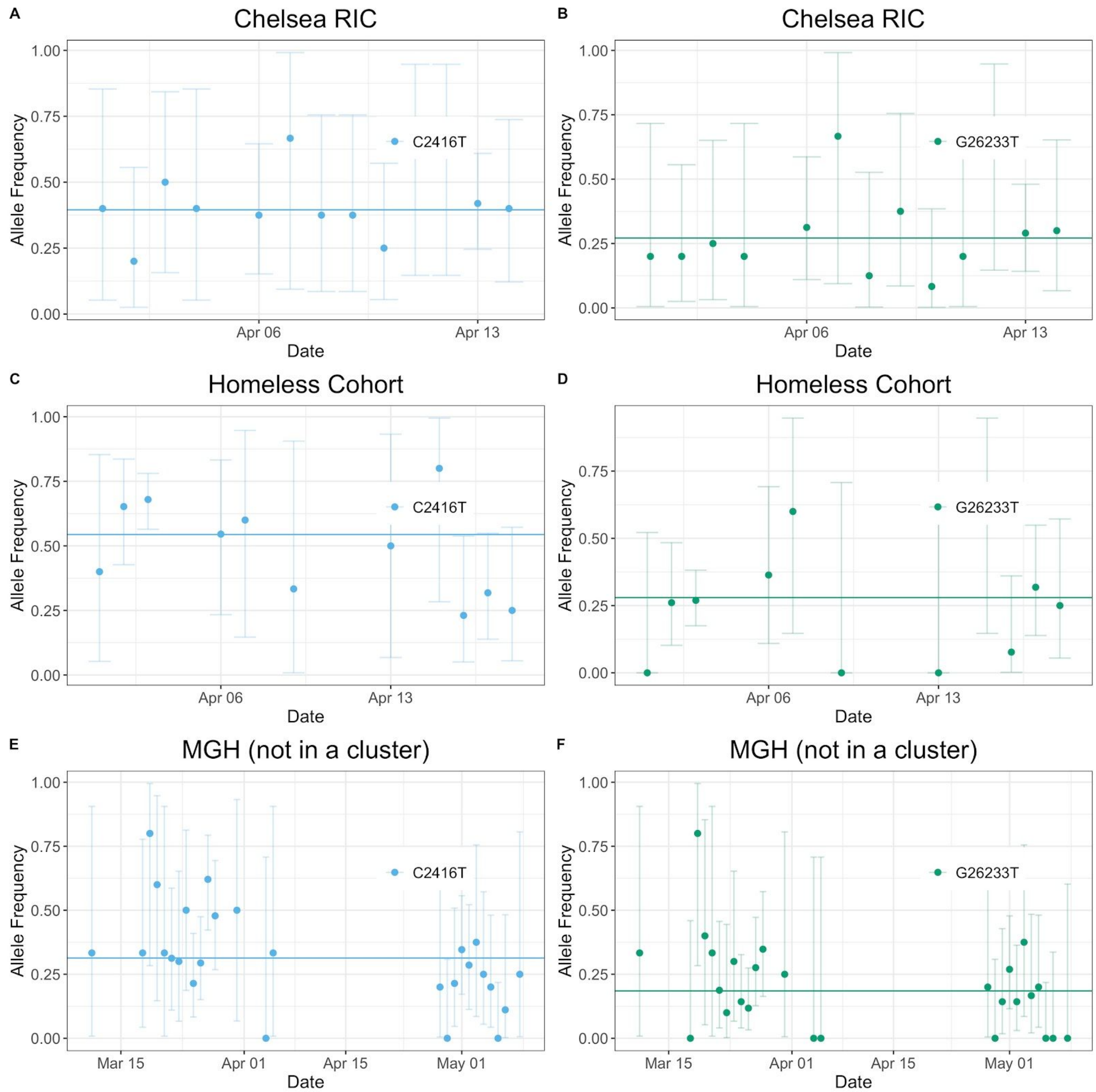


Fig S17. A. Cumulative case numbers by exposure group from March 9 through March 12 (period of data availability for the given exposures). **B.** Cumulative allele frequency of conference-associated alleles vs. time. **C.** Number of new infections reported by MADPH vs. time.

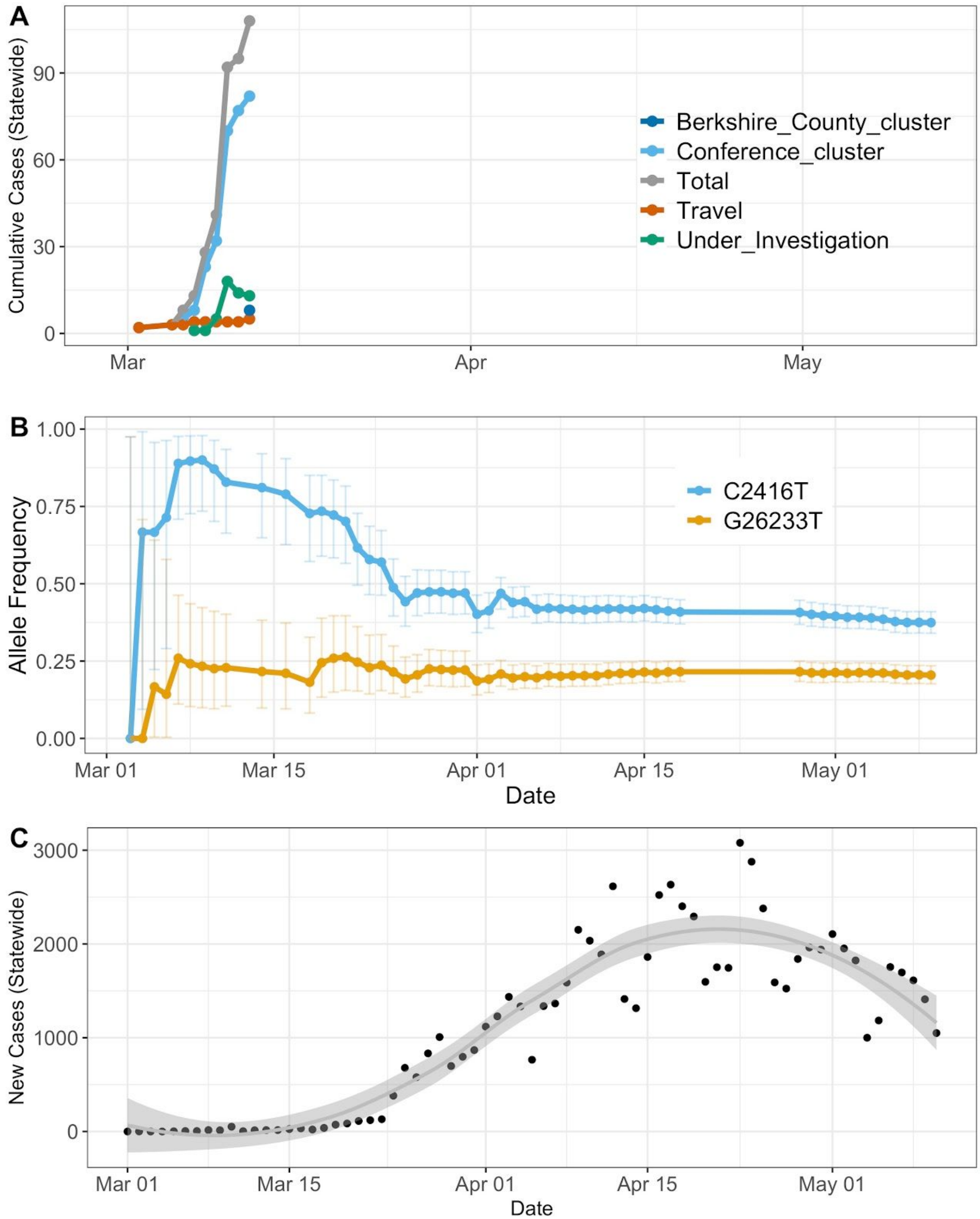


Fig S18. Confirmation of respiratory virus detection in metagenomic sequencing results. **A.** Results of the BioFire FilmArray Respiratory Virus Panel performed on the 17 available samples for which co-infections were detected by metagenomic sequencing. **B.** Concordance between BioFire and metagenomic sequencing results for respiratory viruses.

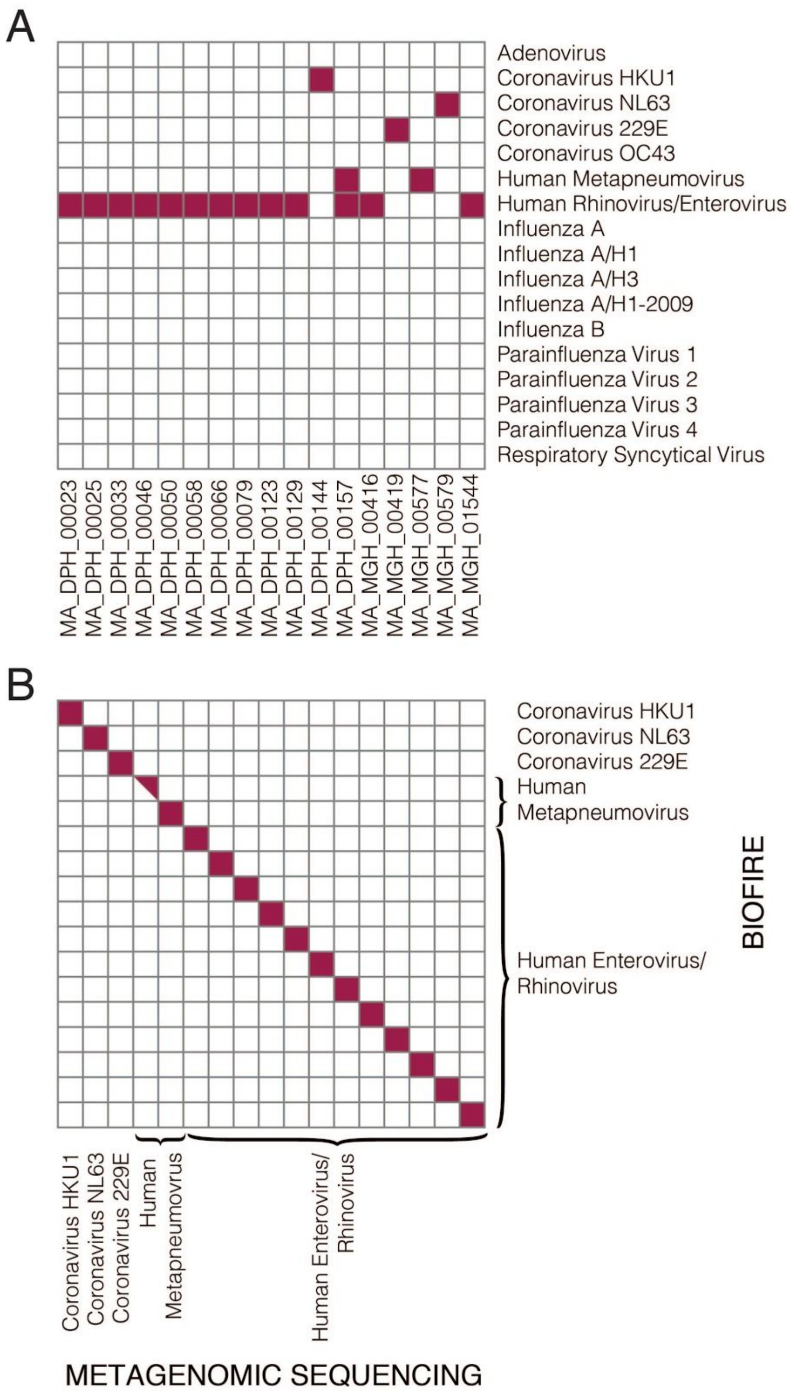


Table S1.
Download of sample_set table with summary assembly variables.

Table S3:

Table of geographic ancestral trait inferences.

Table S3.

Table of counts for all viral species classified by Kraken2.