science.sciencemag.org/cgi/content/full/science.abe3261/DC1

Supplementary Material for

Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events

Jacob E. Lemieux*, Katherine J. Siddle, Bennett M. Shaw, Christine Loreth, Stephen F. Schaffner, Adrianne Gladden-Young, Gordon Adams, Timelia Fink, Christopher H. Tomkins-Tinch, Lydia A. Krasilnikova, Katherine C. DeRuff, Melissa Rudy, Matthew R. Bauer, Kim A. Lagerborg, Erica Normandin, Sinéad B. Chapman, Steven K. Reilly, Melis N. Anahtar, Aaron E. Lin, Amber Carter, Cameron Myhrvold, Molly E. Kemball, Sushma Chaluvadi, Caroline Cusick, Katelyn Flowers, Anna Neumann, Felecia Cerrato, Maha Farhat, Damien Slater, Jason B. Harris, John A. Branda, David Hooper, Jessie M. Gaeta, Travis P. Baggett, James O'Connell , Andreas Gnirke, Tami D. Lieberman, Anthony Philippakis, Meagan Burns, Catherine M. Brown, Jeremy Luban, Edward T. Ryan, Sarah E. Turbett, Regina C. LaRocque, William P. Hanage, Glen R. Gallagher, Lawrence C. Madoff, Sandra Smole, Virginia M. Pierce, Eric Rosenberg, Pardis C. Sabeti*, Daniel J. Park, Bronwyn L. MacInnis*

†Corresponding author. Email: lemieux@broadinstitute.org (J.E.L.); pardis@broadinstitute.org (P.C.S.); bronwyn@broadinstitute.org (B.L.M.)

Published 10 December 2020 as *Science* First Release
DOI: 10.1126/science.abe3261

This PDF file includes:

Materials and Methods
Figs. S1 to S17
References

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/science.abe3261/DC1)

Tables S1 to S3 as separate .csv files
MDAR Reproducibility Checklist

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 (23), 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 a cluster in western MA from prior to March 10 2020 and 2) all available samples submitted to MADPH from Boston Healthcare for the Homeless Program (BHCHP) from Mar 19 2020 through April 18 2020, a period that included universal screening (5).

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 (26, 36), 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 (26, 36). 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 Registry Service (dockstore.org/organizations/BroadInstitute/collections/pgs). Code for analysis of assembled sequences is available at https://github.com/jacoblemieux/sarscov2pub. 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 constructed a set of 200 phylogenetic trees consisting of 5796 samples (772 genomes from this study plus 5024 comparator genomes) by repeatedly sub-sampling the much-larger GISAID database (downloaded on September 29, 2020). The quantiles of this distribution were used to construct confidence intervals for importation events. To perform this subsampling, we used the script *subsample_by_metadata_with_focal* with 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 a sensitivity analysis to investigate the effect of regional sampling size on the results of the importation analysis. We repeated this subsampling procedure for up to 100 samples per region, resulting in a set of 60 maximum-likelihood trees from independent draws from the GISAID database.

We also conducted additional analysis of the genomes sequenced in this study. We aligned the set of 772 genomes using MAFFT v7.471(*37*) and trimmed 5' and 3' (first 265 and last 228 bases) UTRs from the alignment in R (*27*). To estimate the root-to-tip distance, we constructed ML phylogenetic trees using PhyML(*38*) v3.3.20190909 with default parameters using the MAFFT alignment of 772 genomes. We used TempEst (*39*) 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 (*40*) 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 exponential 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. To ensure convergence, we inspected the MCMC traces and marginal posterior distribution of all model parameters. To confirm that we were not obtaining improper posterior distributions as a result of prior specification, in addition to inspection of marginal posteriors, we also tested several priors specifications, including including a normal[0.001,0.0002] truncated at 0 for clock rate, a unif[0, 100] for exponential population size, and a laplace[-50, 50]. These chains were found to yield equivalent results, confirming that the use of improper priors was not resulting in improper posterior distributions, and, once this was established, were terminated after 20 million states.

Detection of respiratory virus co-infection

We used Kraken2 (*31*) 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.

Ancestral State Reconstruction

To reconstruct the ancestral geographic location of unsampled nodes, we used three approaches to ancestral state reconstruction.

1) Maximum-likelihood reconstruction: We subsampled the GISAID database, using the sub-sampling procedure described in “Phylogenetic Tree Construction” above and inferred maximum-likelihood trees on this collection of trees as described. We then fit a two-state

markov chain consisting of states of “MA” vs “non-MA” using treetime(41), as implemented in the augur pipeline (42), to each of the maximum likelihood trees. We calculated the marginal probability of being in a given state at each node using the augur pipeline. We then iterated through the nodes in the tree and considered importation events as nodes whose marginal probability was MA > 0.9 and whose parent node was non-MA with marginal probability > 0.9. Once a node was considered to have been imported into MA, we did not count any further reintroductions of descendents during the study period as these were considered implausible. We repeated this procedure for a seven-state Markov chain consisting of six continents plus MA. We computed confidence intervals for importation counts by calculating the quantiles of the distribution over maximum-likelihood trees computeted from distinct GISAID subsamples.

2) Maximum parsimony: We applied parsimony-based reconstruction using the Narushima and Hanazawa method as implemented in the MPR function of the ape package in R. A collection of maximum-likelihood trees, as described above for maximum likelihood ancestral state reconstruction, was inferred using the augur pipeline. We used the output of the iqtree step. This method requires an outgroup, and given the difficulty of selecting an appropriate outgroup for SARS-CoV-2, we midpoint rooted trees and attached an arbitrary non-MA branch to the midpoint to serve as an outgroup. We then inferred the most parsimonious reconstruction of binary ancestral characters corresponding to MA and non-MA. State switches were counted similar to maximum-likelihood ancestral state reconstruction except that in place of marginal probability of 0.9 at each node, we considered a node to be in a given ancestral state only if both the upper and lower values of the reconstructed sets corresponded to Non-MA or MA. We excluded reimportation events during the study period.

3) Bayesian reconstruction of ancestral state: We conducted Bayesian discrete trait ancestral reconstruction in BEAST1.10.4. For global samples along with MA samples, given the large size of the trees, we were unable to obtain convergence by allowing BEAST to iterate over the space of tree topologies, so we provided a starting tree (the maximum-likelihood tree from Fig 2a) and ran BEAST with a fixed tree by removing the tree operators, using a binary (MA vs non-MA) markov chain model with asymmetric substitution model and an uncorrelated relaxed clock with a lognormal distribution. Priors for rate variables were left at their defaults and a prior normal(0.008, 0.002) truncated at 0 was used for the clock rate. We added indicator variables for state switches. The analysis was run until ESS of the rate parameters and state switch variables all exceeded 200, and the posterior distributions count of importations was summarized in R. For counting of importations into the SNF and homeless populations, we fit two-state continuous time Markov chains (with states corresponding to non-SNF and SNF, and non-Homeless and Homeless) with asymmetric substitution model and a relaxed clock. Unlike the global analysis, which required a fixed tree for computational reasons, we were able to obtain convergence while allowing tree topology to vary because the trees were smaller (including only the 772 samples in MA). For these models, we used an uninformative prior, CTMC rate reference, for the clock rate, ran BEAST1.10.4 for 100 million states, and summarized the posterior probability of state change transitions using R.

Haplotype Network Reconstruction

Haplotype networks were visualized using the software tool PopART v1.7 (32). 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 https://github.com/broadinstitute/sc2-variation-scripts/blob/master/scripts/msa_fasta_to_popart_nexus.py. A minimum spanning network of the sequences was constructed and 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 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 *(33)*.

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 *(35)* and a mean substitution rate of 1.04 x 10^{-3}/bp/year (Fig. S6C). To account for uncertainty in the substitution rate (which had a 95% highest posterior probability density interval of 0.91 - 1.17 x 10^{-3} substitutions/bp/year), we modeled the substitution rate as a normally distributed random variable with σ = 0.066 x 10^{-3}, drew 1 million sample rates from the distribution and generated one Poisson-distributed number of mutations for each, based on the 74 transmissions. The fraction of these draws that yielded 18 or fewer mutations constituted the reported p-value.

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 159,103 complete GISAID genomes was downloaded on November 2, 2020 and the percentage of ancestral and derived alleles was extracted from the alignment and plotted by geographic category.

To estimate the number of cases linked to the conference in each state, we estimated the proportion of genomes reported from that state in GISAID through 11/2/2020 with C2416T and C2416T/G26233T and multiplied the estimated proportions by case counts by state. We obtained tabulated case counts through November 1, 2020 from the NY times COVID data repository (https://github.com/nytimes/covid-19-data) and international data was obtained from the Johns Hopkins COVID tracking website (*43*). Genomes from patients in this study with known epidemiological linkages to the conference were removed from this analysis. We estimated confidence intervals for the proportion using the binomial distribution and multiplied those intervals by the total reported cases to obtain a confidence interval by state. To sum across states and account for variability in sampling at the state level, we used Monte Carlo simulation. We conducted 10,000 simulations in which the total number of cases with the variant of interest (e.g. C2416T) in state i, V_i, was simulated from a distribution constructed from $p_i \sim \mathrm{Beta}(\alpha_i + 0.5, \beta_i + 0.5)$, where α_i and β_i are counts of successes (T alleles) and failures (C alleles in the case of C2416T, or G alleles in the case of G26233T), which gives the binomial confidence interval using Jeffrey's formula. We then multiplied this by the total number of reported cases in each state, T_i. For C2416T, we multiplied the total case count in each state by a percentage attributable to the initial importation event (associated with the conference), $c \sim N^*(0.9, 0.05)$, based on the relative rate of reimportation of the European clades. For C2416T/G26233T, all cases were assumed to result from the conference, i.e. c = 1. The total number of cases across states was then estimated as $\Sigma_i T_i * V_i * c$. States reporting < 10 genomes total were removed from the analysis. We also only included states in which the allele of interest had been reported to account for the possibility that the distribution of cases in states is zero-inflated (i.e. the allele of interest has not entered a given state or country at all). We conducted 10,000 simulations and obtained confidence intervals for the total number of cases from the quantiles of this sample.

For the time-adjusted model, each time period was modeled separately, and the results were summed. When the total number of genomes in a given state in a given time period was less than 10, we used pooled counts across time periods for $\alpha_i + 0.5$, $\beta_i + 0.5$. We also implemented a version of the simulation using a normal approximation. In this case, $V_i \sim N^*(p_{i,hat}, sd_{i,hat})$ where $p_{i,hat}$ = number of successes / number of trials and $sd_{i,hat}$ = sqrt(pi,hat* (1-$p_{i,hat}^*$)/n+1), $p_{i,hat}^*$ = number of successes + 0.5 / number of trials + 1, i.e. standard formulas, but with a continuity correction to prevent the variance being estimated at 0 in the absence of any successes or failures. N* is a truncated normal distribution on the interval [0,1]. As reported (Figure S15M-N), both approaches yielded essentially identical estimates, but the latter approach allowed us to account for the possibility that the reported allele frequency in a given state is more variable than expected under independent binomial sampling, by inflating the standard deviation by a factor of 2 (termed "robust 1" model) and 3 ("robust 2"), a situation may occur, for example, if states are sequencing clusters of cases. Similar to the binomial model, for the time-dependent model, we used pooled estimates across both time periods for $p_{i,hat}$, $sd_{i,hat}$ if the number of genomes in a given time period for a given state was less than 10. For the estimation of countries with the G26233T allele, we used an identical approach with country in place of US state, and only considered countries in which the allele had been reported.

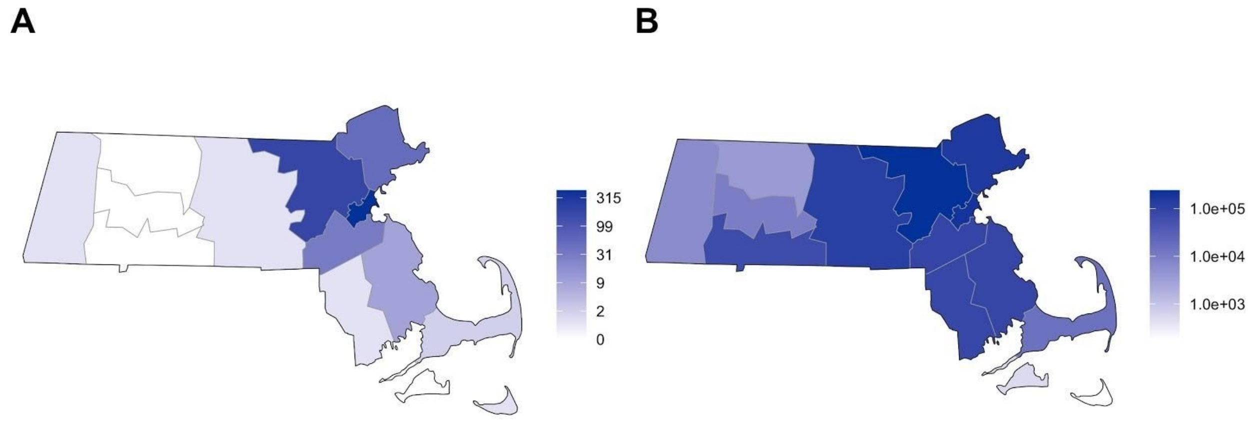


Fig S1.
A. Counts of complete genomes reported in this study, by county. **B.** Case counts by county reported by MADPH through July 1, 2020.

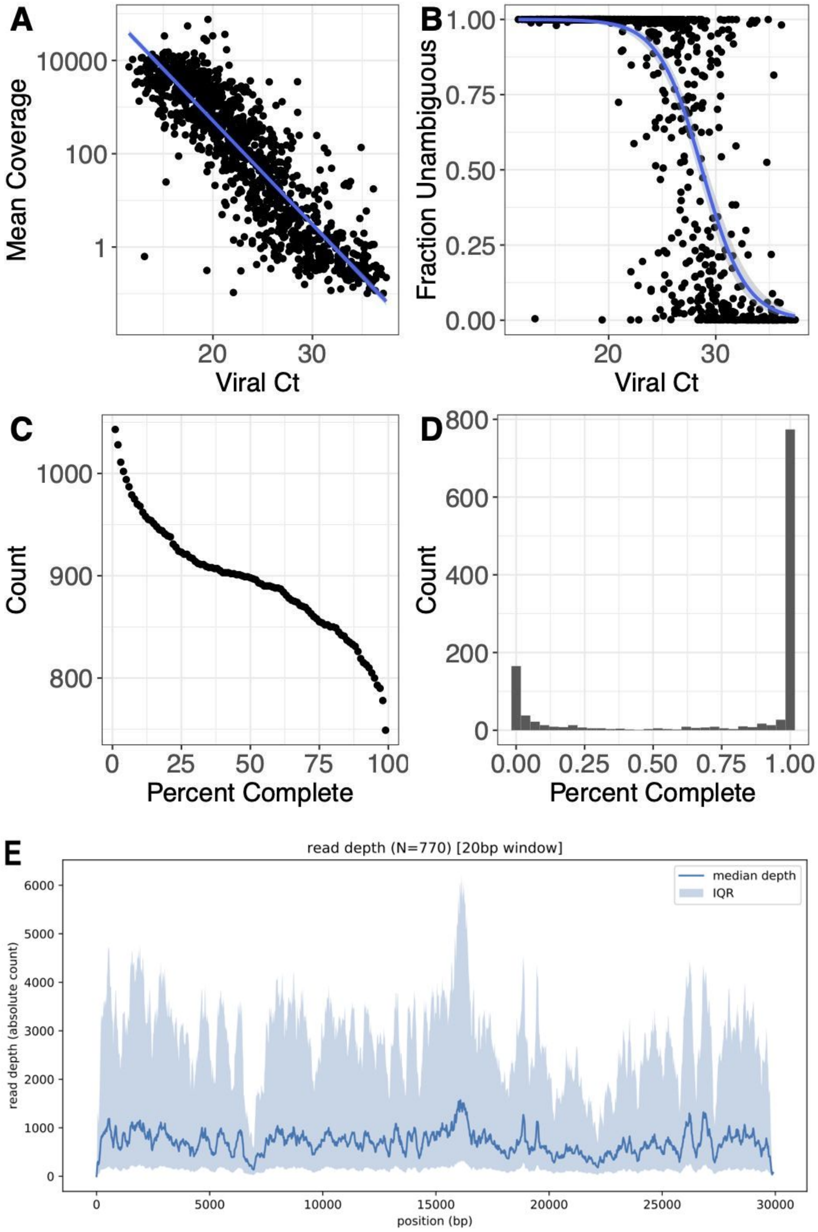


Fig S2.
A. Mean coverage (on a log scale) vs. viral C_t for all samples included in the study. A linear regression fit is shown in blue. **B.** Fraction of the genome that is complete is shown vs. viral C_t. A $C_t < 28$ was strongly associated with recovery of a complete virus genome. Fit from a logistic 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. [*previous page]*

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.

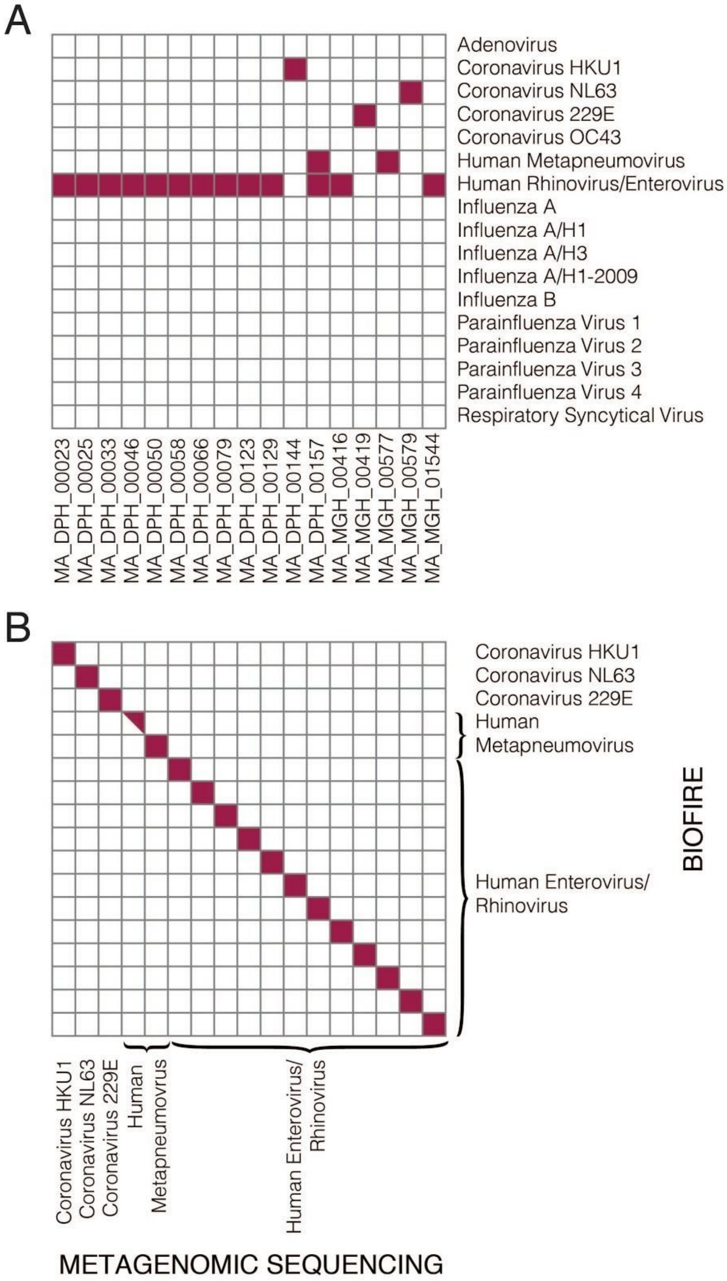


Fig S5.

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.

Fig S6.
A. Linear regression of root-to-tip distance vs. date of sampling. Root-to-tip distance was calculated using TempEst (38) on maximum likelihood trees inferred using PhyML (37). **B.** Posterior distribution of clock rate using a GTR substitution model with 4 gamma-distributed rate categories (GTRG4). **C**. Posterior distribution of clock rate using an HKY substitution model with 4 gamma-distributed rate categories (HKYG4). **D-G**. Posterior distributions of exponential population size and growth rate under both models. **H-I.** Posterior distributions of tMRCA for major Boston-area clades under GTRG4 and HKYG4 models.

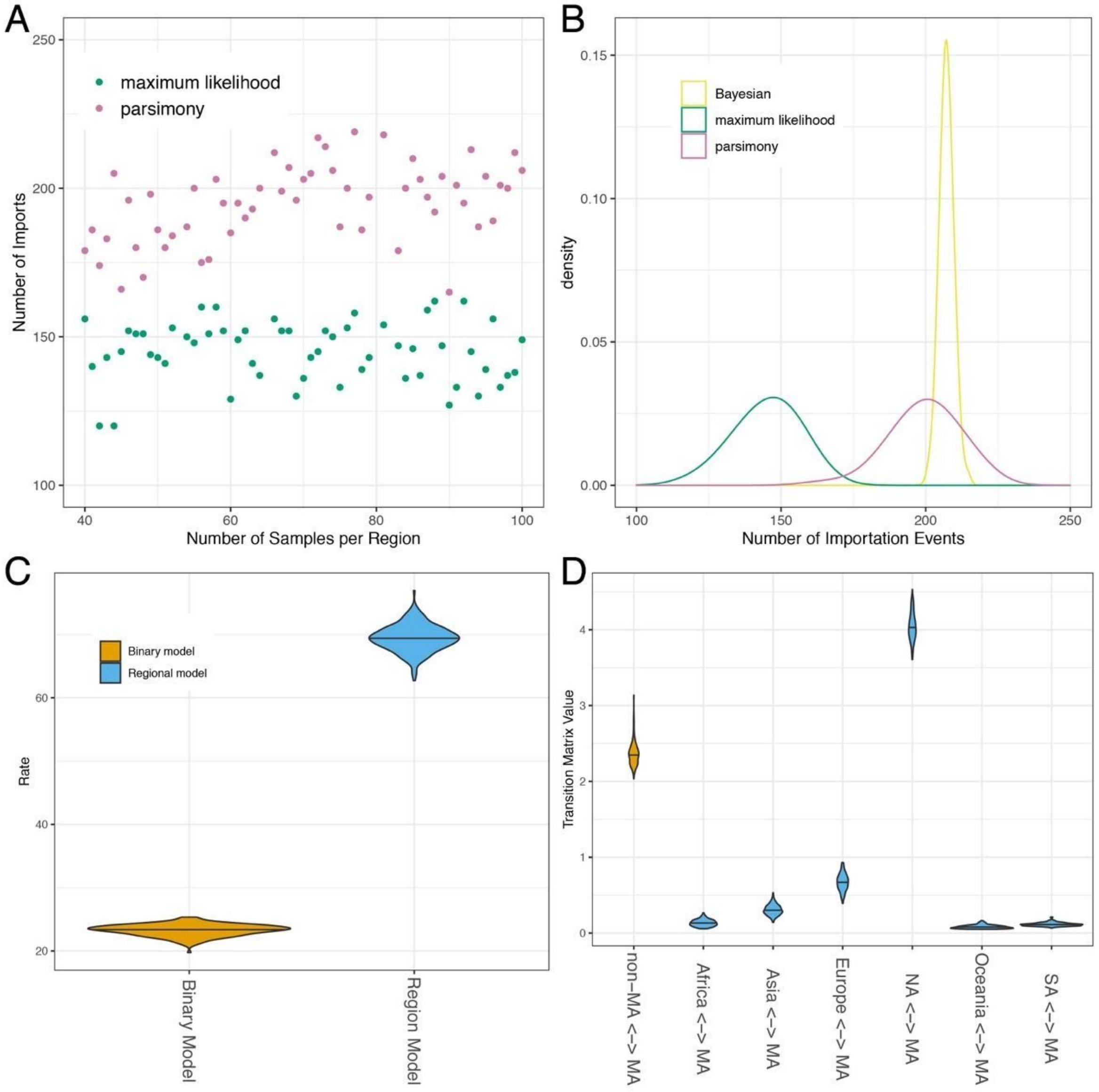


Fig S7.

A. Number of importation events as assessed by maximum likelihood ancestral state reconstruction and parsimony-based reconstruction, using subsamples of varying size from the GISAID database. **B.** Density estimates for the distribution of import sizes across the range of tree sizes for which the number of imports reached a plateau (complete range for maximum likelihood reconstruction, and above 60 for maximum parsimony-based reconstruction). The posterior distribution of imports into MA from a Bayesian ancestral inference model implemented in BEAST, using a single, fixed ML tree are also shown. **C.** Distribution of rate parameters for ML trees across subsamples. **D.** Distribution of fitted transition matrix elements for the ancestral category exchanges in the ML model.

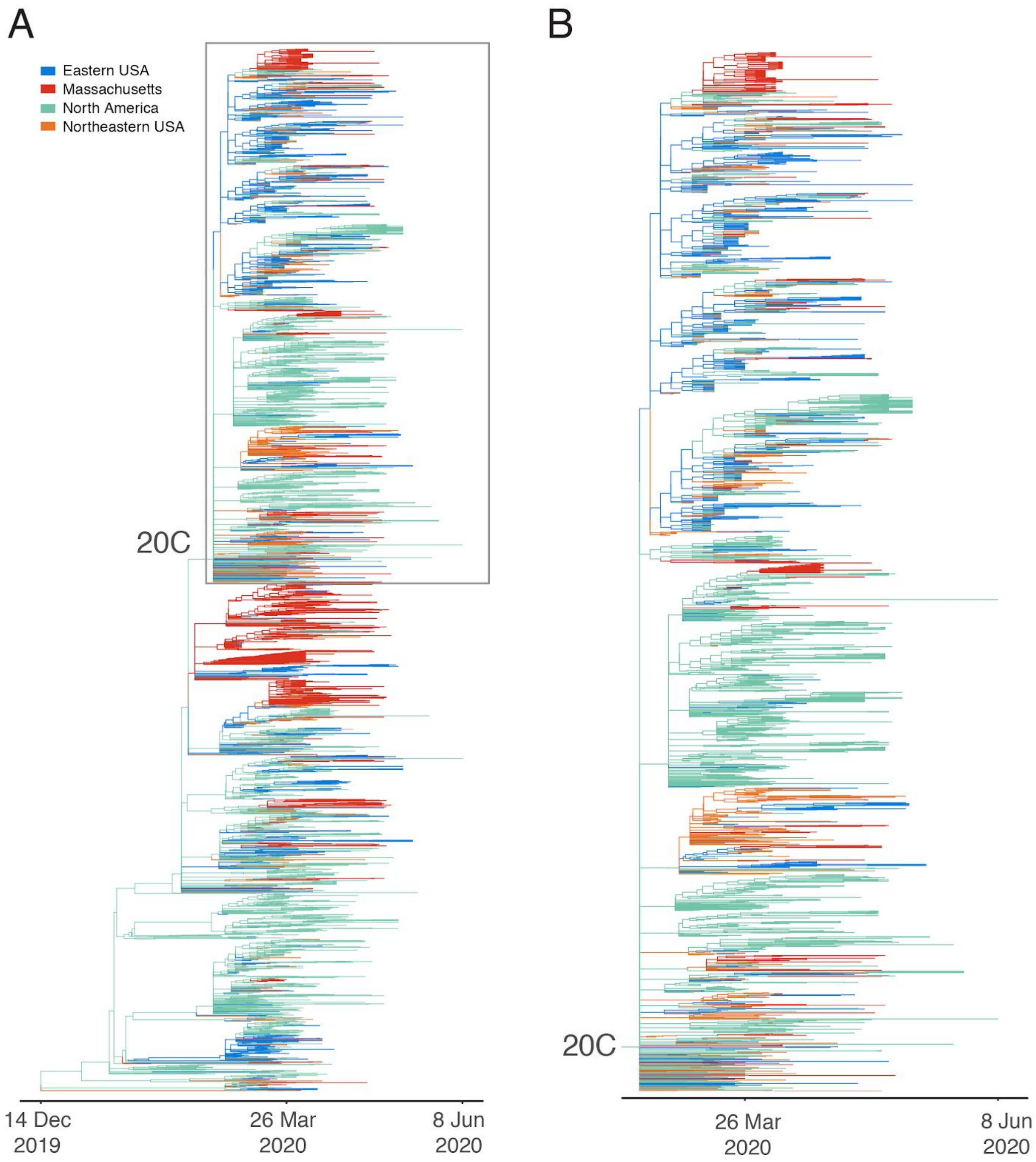


Fig S8.
A. Maximum Likelihood Time Trees of 4349 SARS-CoV-2 genomes from the United States collected between January and June 2020. Zoomed-in view of clade 20C, marked by a grey box in panel A.

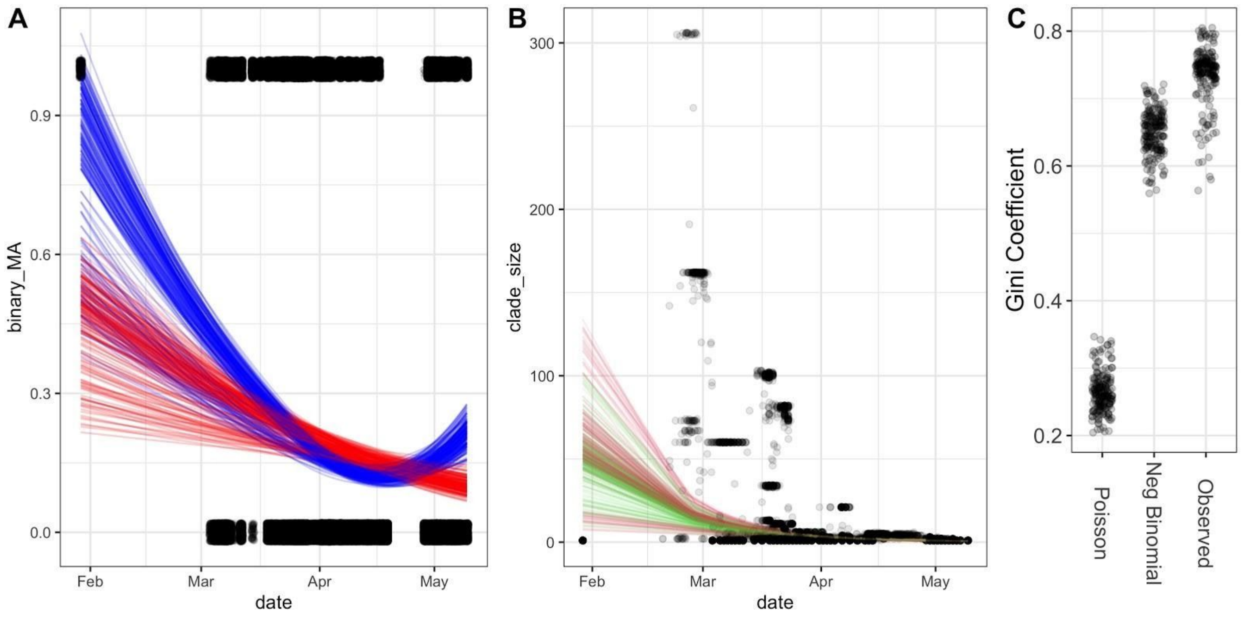


Fig S9.
A. Probability of an importation event over time. Ancestral state reconstruction was inferred for a population of trees. Samples whose ancestral state was inferred as non-MA are coded as 1 and samples whose ancestral state is inferred as MA are coded as 0 (a small amount of noise is added to the y-coordinate to show the density of the data). This was repeated for 200 subsamples from GISAID. For each subsample, a logistic regression (red curve; median β_1 = -3.48 95% CI [-4.42 , -1.53]) shows the probability of importation decreasing through the study period and a loess smoother (blue) shows the change in importation probability over time for a given ML tree. **B.** Scatterplot showing the size of each imported clade vs. time. Regression models have been fit to model imported counts vs. time. Poisson (red line, median β_1 = -16.66 95% CI[-19.96, -8.61]) and negative binomial (green line, β_1 = -15.69, 95% CI[-18.23, -10.63]) regression of clade size vs date for each ML tree. In all cases, negative binomial regression provided an improved fit for all regression lines ($p < 2 \times 10^{-16}$ for all trees, likelihood ratio test). **C.** Gini coefficient for observed import count data, as compared to simulated data from poisson and negative binomial distributions fit to the observed count data using maximum likelihood.

Fig S10.
Geographic distribution of select lineage-defining variants in Eastern Massachusetts by zip code. The scale is in log10(case counts + 1).

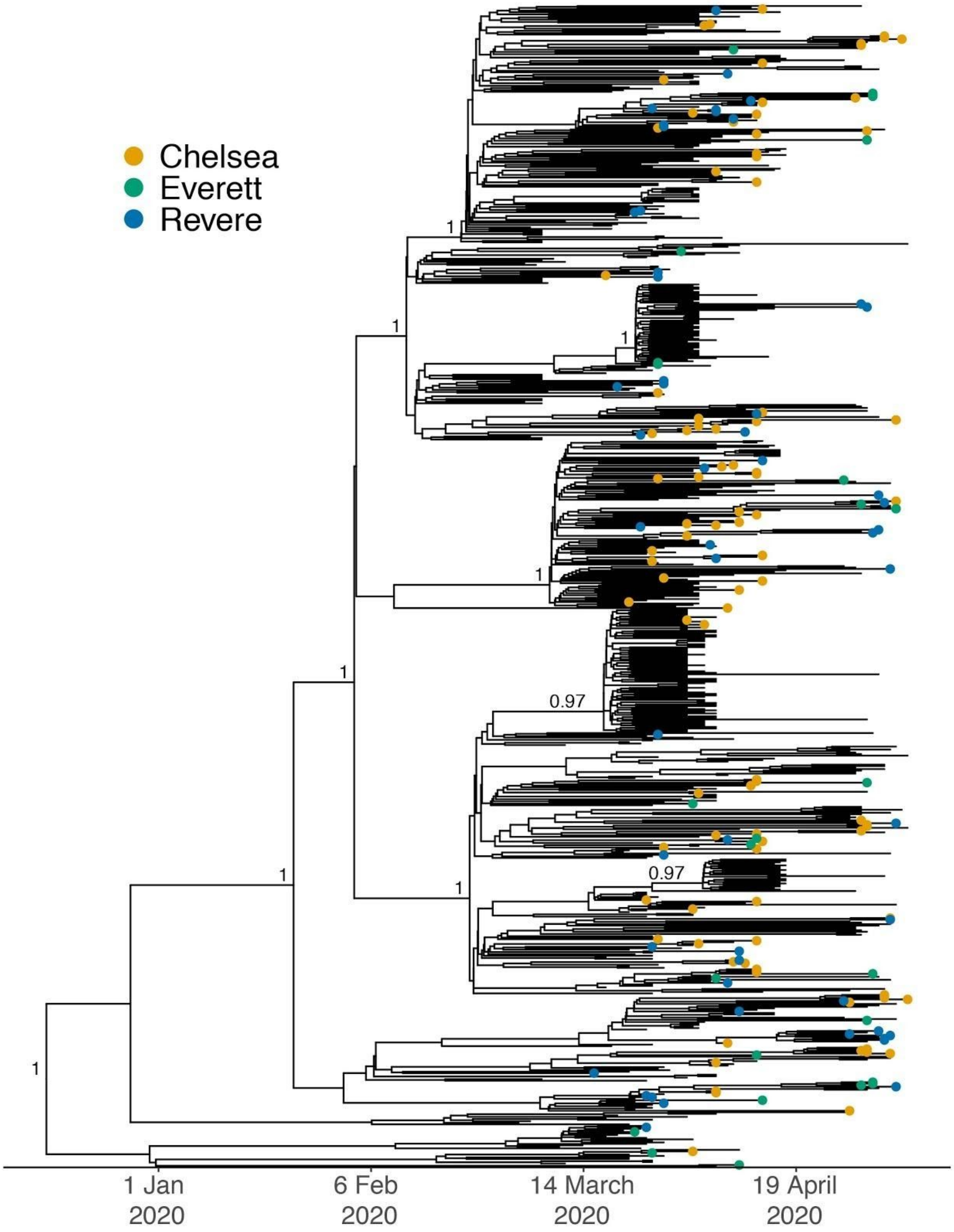


Fig S11.
Sequenced samples labeled by zip code of residence for the top three zip codes in the set of 772 genomes from unique patients.

Fig S12.
Maximum Likelihood tree of 772 Massachusetts genomes with tips labeled by exposure. Tree was computed using IQtree, with ultrafast bootstrap support shown at nodes with support > 80. Legend is in substitutions per site.

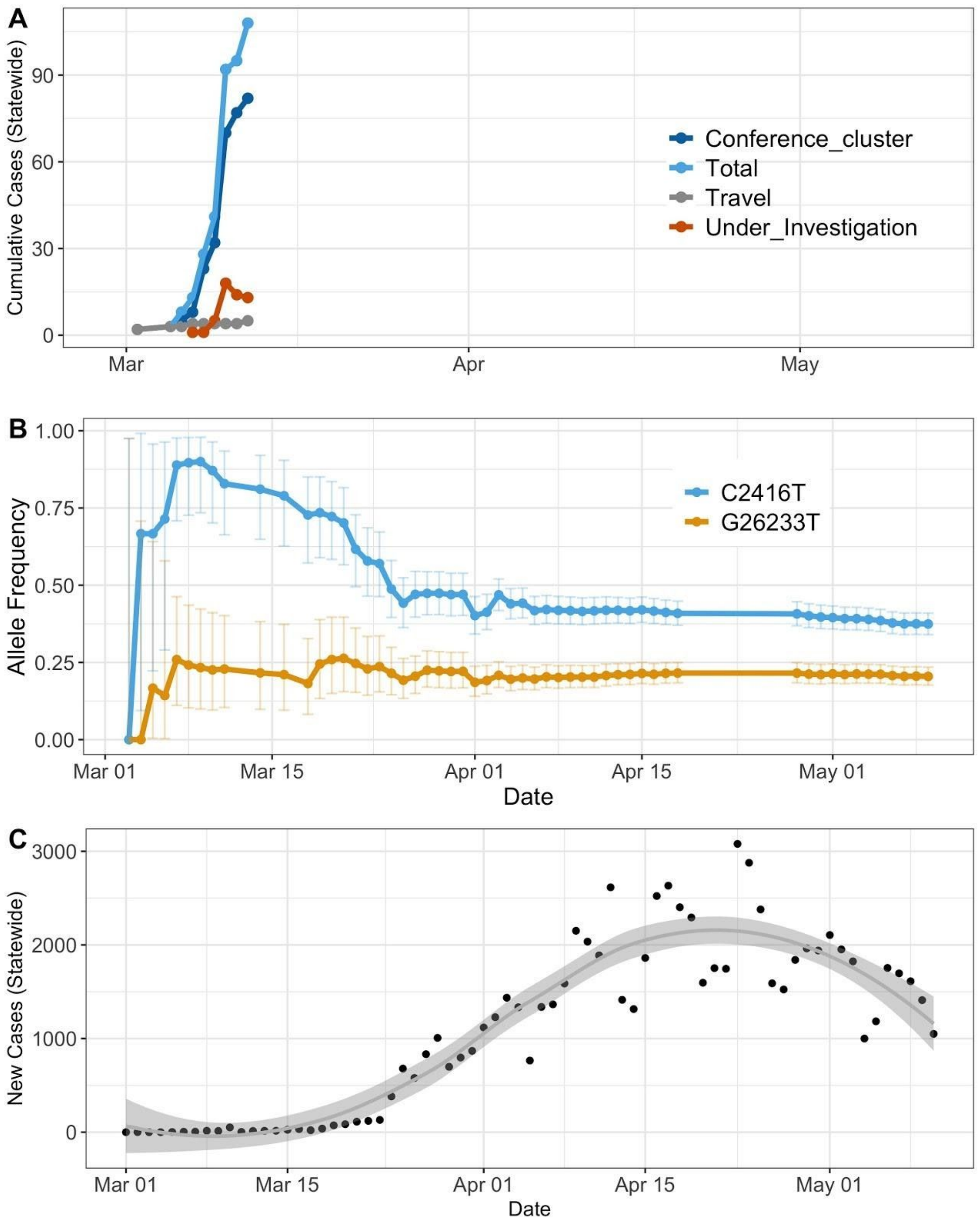


Fig S13.

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.

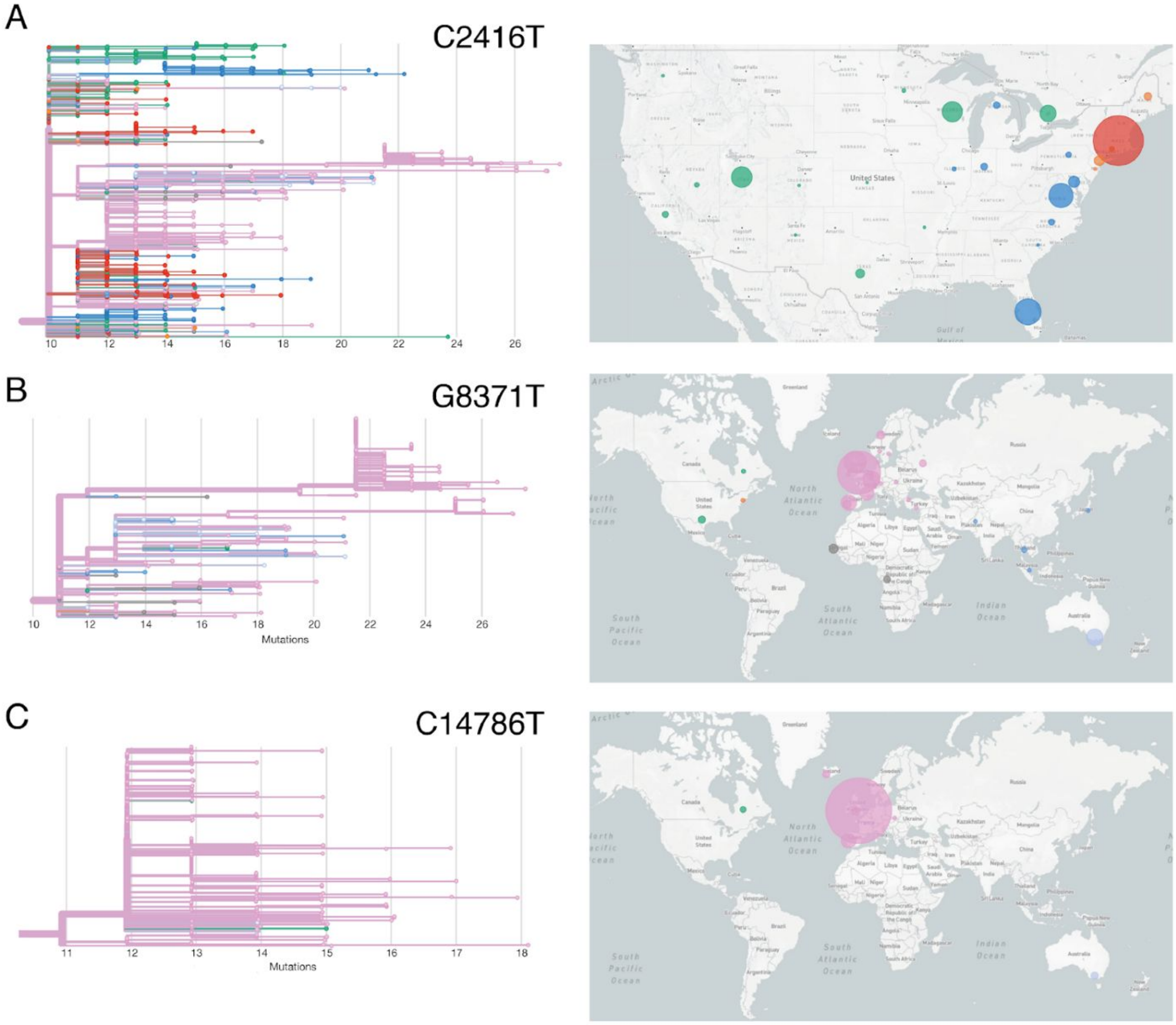
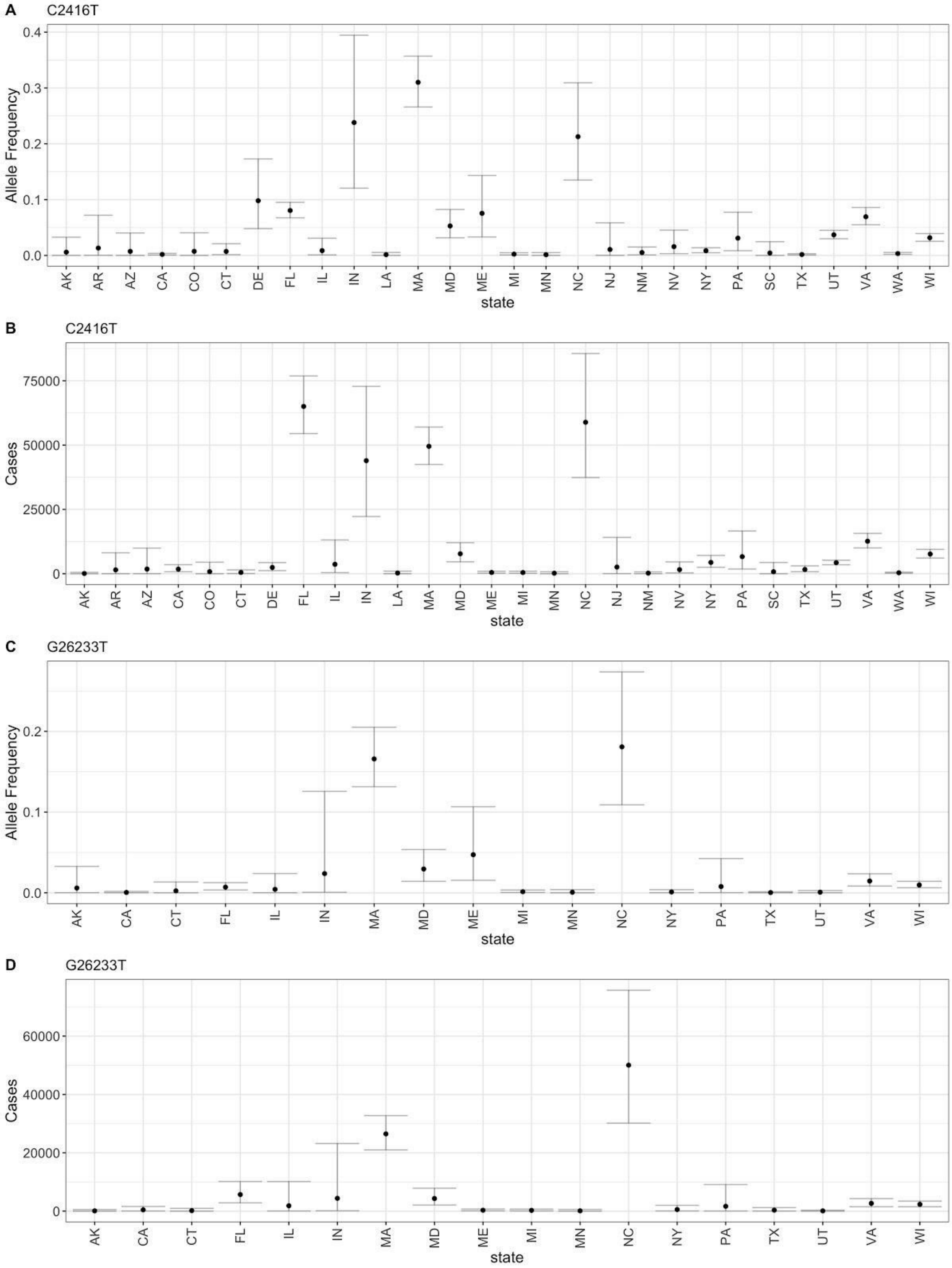
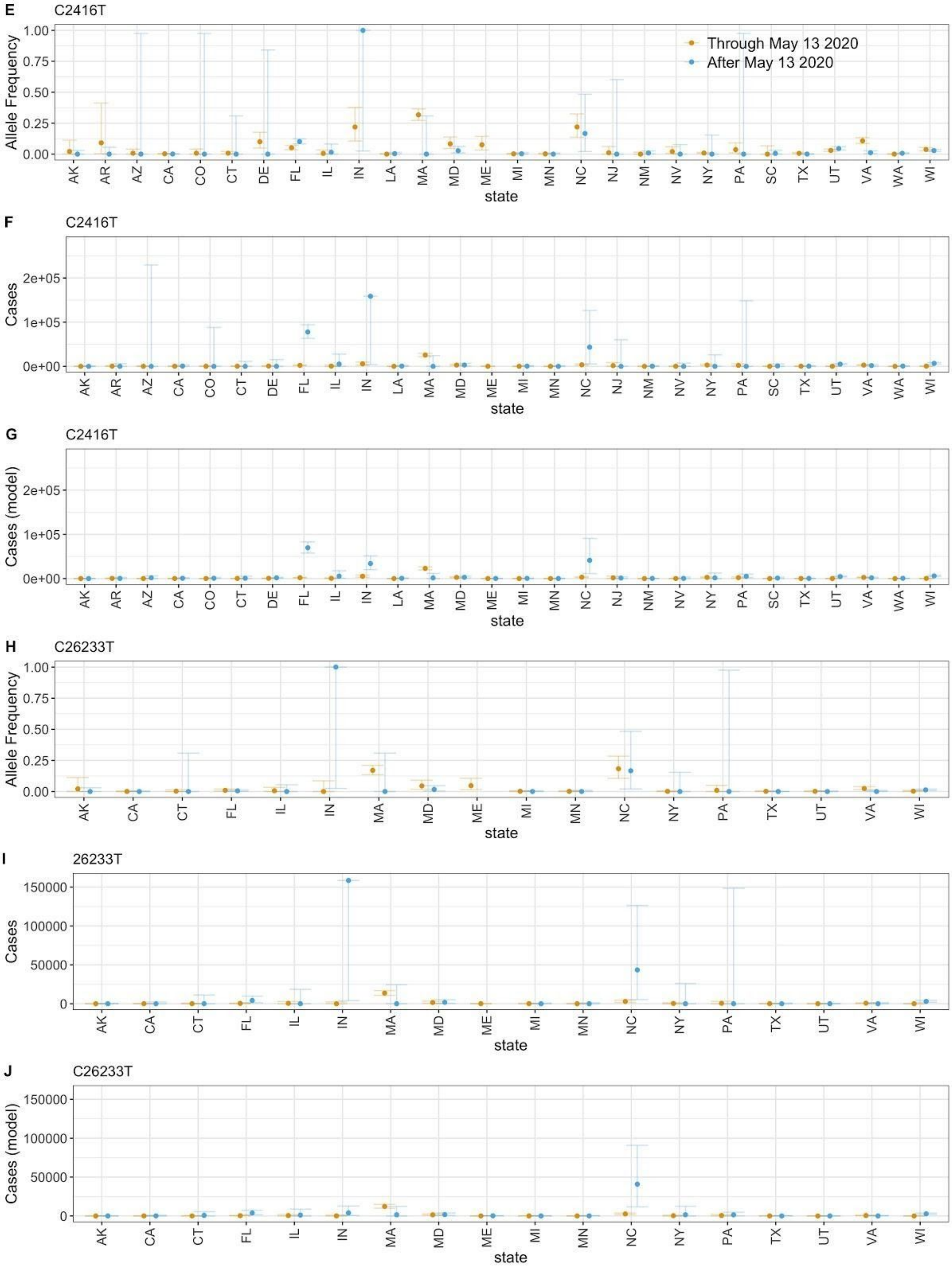


Fig S14.

A. Divergence tree of the C2416T variant showing all global sequences (in GISAID through September 29, 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.


A
C2416T
Allele Frequency
0.4
0.3
0.2
0.1
0.0
AK AR AZ CA CO CT DE FL IL IN LA MA MD ME MI MN NC NJ NM NV NY PA SC TX UT VA WA WI
state
B
C2416T
Cases
75000
50000
25000
0
AK AR AZ CA CO CT DE FL IL IN LA MA MD ME MI MN NC NJ NM NV NY PA SC TX UT VA WA WI
state
C
G26233T
Allele Frequency
0.2
0.1
0.0
AK CA CT FL IL IN MA MD ME MI MN NC NY PA TX UT VA WI
state
D
G26233T
Cases
60000
40000
20000
0
AK CA CT FL IL IN MA MD ME MI MN NC NY PA TX UT VA WI
state


E
C2416T
Allele Frequency
1.00
0.75
0.50
0.25
0.00
Through May 13 2020
After May 13 2020
AK AR AZ CA CO CT DE FL IL IN LA MA MD ME MI MN NC NJ NM NV NY PA SC TX UT VA WA WI
state
F
C2416T
Cases
2e+05
1e+05
0e+00
AK AR AZ CA CO CT DE FL IL IN LA MA MD ME MI MN NC NJ NM NV NY PA SC TX UT VA WA WI
state
G
C2416T
Cases (model)
2e+05
1e+05
0e+00
AK AR AZ CA CO CT DE FL IL IN LA MA MD ME MI MN NC NJ NM NV NY PA SC TX UT VA WA WI
state
H
C26233T
Allele Frequency
1.00
0.75
0.50
0.25
0.00
AK CA CT FL IL IN MA MD ME MI MN NC NY PA TX UT VA WI
state
I
26233T
Cases
150000
100000
50000
0
AK CA CT FL IL IN MA MD ME MI MN NC NY PA TX UT VA WI
state
J
C26233T
Cases (model)
150000
100000
50000
0
AK CA CT FL IL IN MA MD ME MI MN NC NY PA TX UT VA WI
state

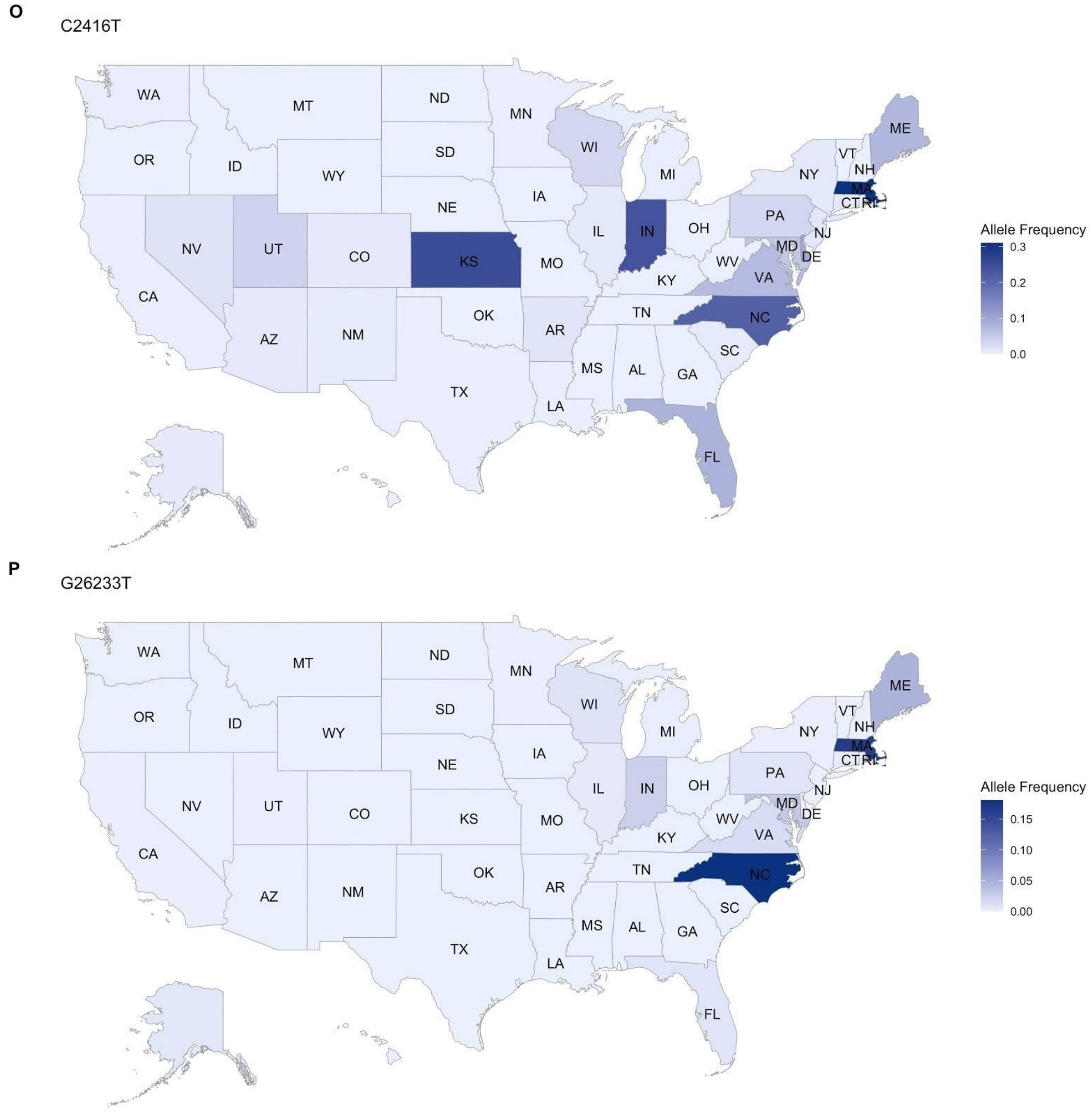


Fig S15.

A. Allele frequency of C2416T in 49 states reporting genome data in GISAID. **B.** Estimated case counts linked to C2416T based on total reported cases by state, through 11/01/2020, and allele frequency estimates. **C.** Allele frequency of G26233T by state. **D.** Estimated case counts linked to G26233T based on total reported cases by state, through 11/01/2020, and allele frequency estimates. **E.** Allele frequency estimates of C2416T by time period, for the time-dependent model. **F.** Crude estimates of case counts linked to C2416T by time period, for the time-dependent model (allele frequency * total reported cases). **G.** Adjusted estimates of case counts linked to C2416T. As described (Material and Methods), these estimates sum cases only in states reporting at least 1 copy of the T allele and > 10 genomes, account for the possibility of reimportation, and pool across time periods in a given state when < 10 genomes are reported from a given time period in a given state. **H.** Allele frequency estimates of G26233T by time period, for the time-dependent model. **I.** Crude estimates (allele frequency * total reported cases) of case counts linked to G26233T by time period, for the time-dependent model. **J.** Adjusted estimates, as in panel G, of case counts linked to G26233T. **K.** Allele frequency estimates of G26233T in countries reporting this allele in GISAID through 11/2/2020. **L.** Estimated cases of G26233T linked to the conference, by country. **M.** Total number of conference-linked cases of each allele, computed by Monte Carlo simulation (materials and methods), with confidence intervals calculated from the binomial. The two alleles are shown by color. Estimates in each time period are given for the time-dependent model. **N.** Total number of conference-linked cases of each allele, computed by Monte Carlo simulation (materials and methods), with confidence intervals calculated estimated from a normal distribution. r1 and r2 denote a robust model, as described (Materials and Methods), with inflated variance for the estimated allele frequency at the state level, to demonstrate the potential effect of clusters. **O.** Map of allele frequency of the C2416T allele by state. **P.** Map of allele frequency for the G26233T allele.

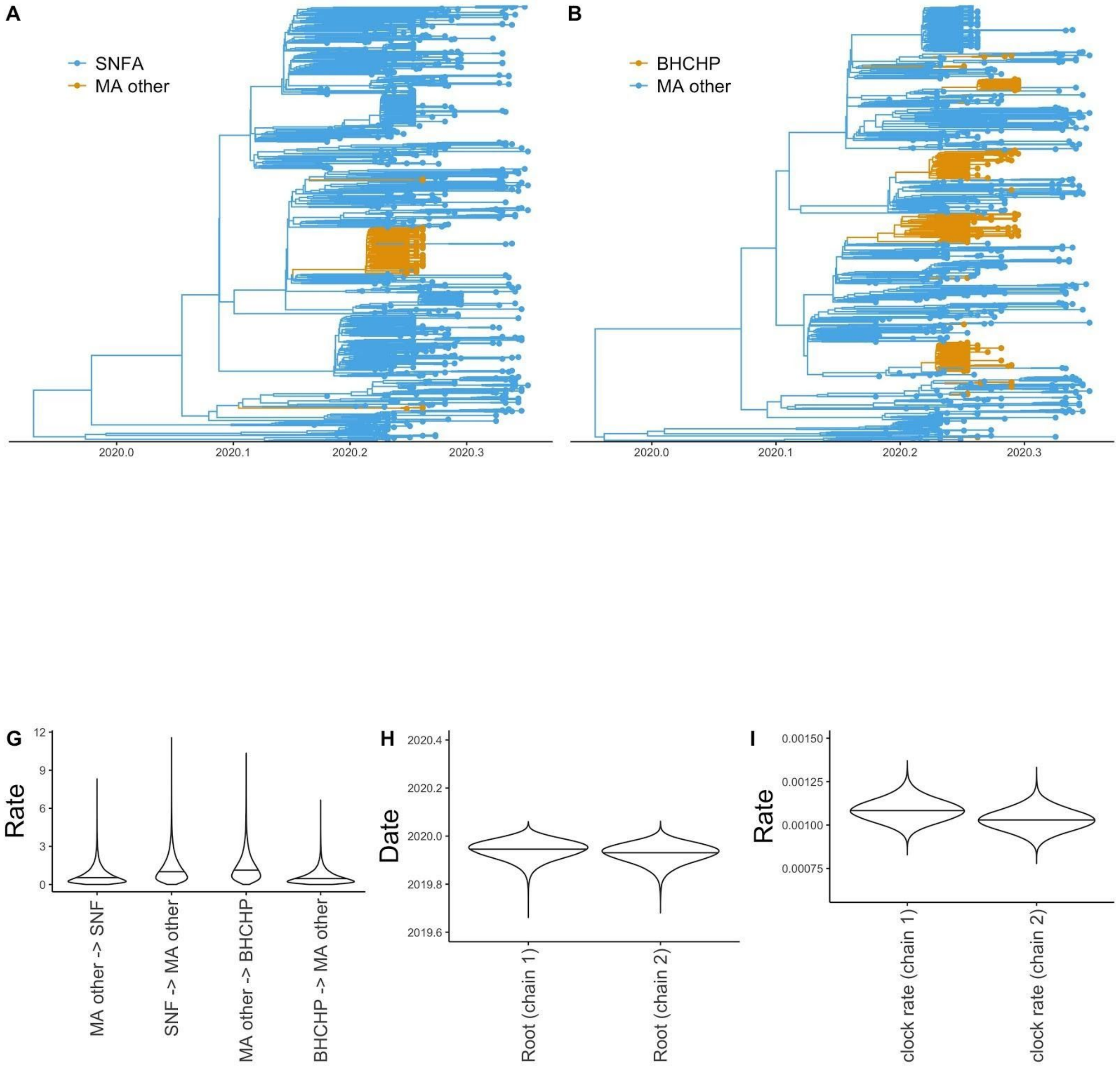


Fig S16.

Results of ancestral trait inference on skilled nursing facility (SNF) subjects and residents and staff from the Boston Health Care for Homeless Program (BHCHP). **A.** MCC tree for SNF subjects, annotated by ancestral state. **B.** MCC tree for BHCHP subjects, annotated by ancestral state. **C-D.** Histograms of counts for SNF imports [median 2, 95% HPD 2-3] and exports [median 2, 95% HPD 2-3] into the SNF. **E-F.** Histograms of counts for imports [median 16, 95% HPD 14-18] and exports [median 4, 95% HPD 2-5] into the BHCHP population. **G.** Histogram of samples from the marginal posterior distributions of rate parameters for the model. **H.** Marginal posterior for tMRCA of the root for the chain used for each analysis (chain 1, SNF model; chain 2, BHCHP model). **I.** Marginal posterior for clock rate.

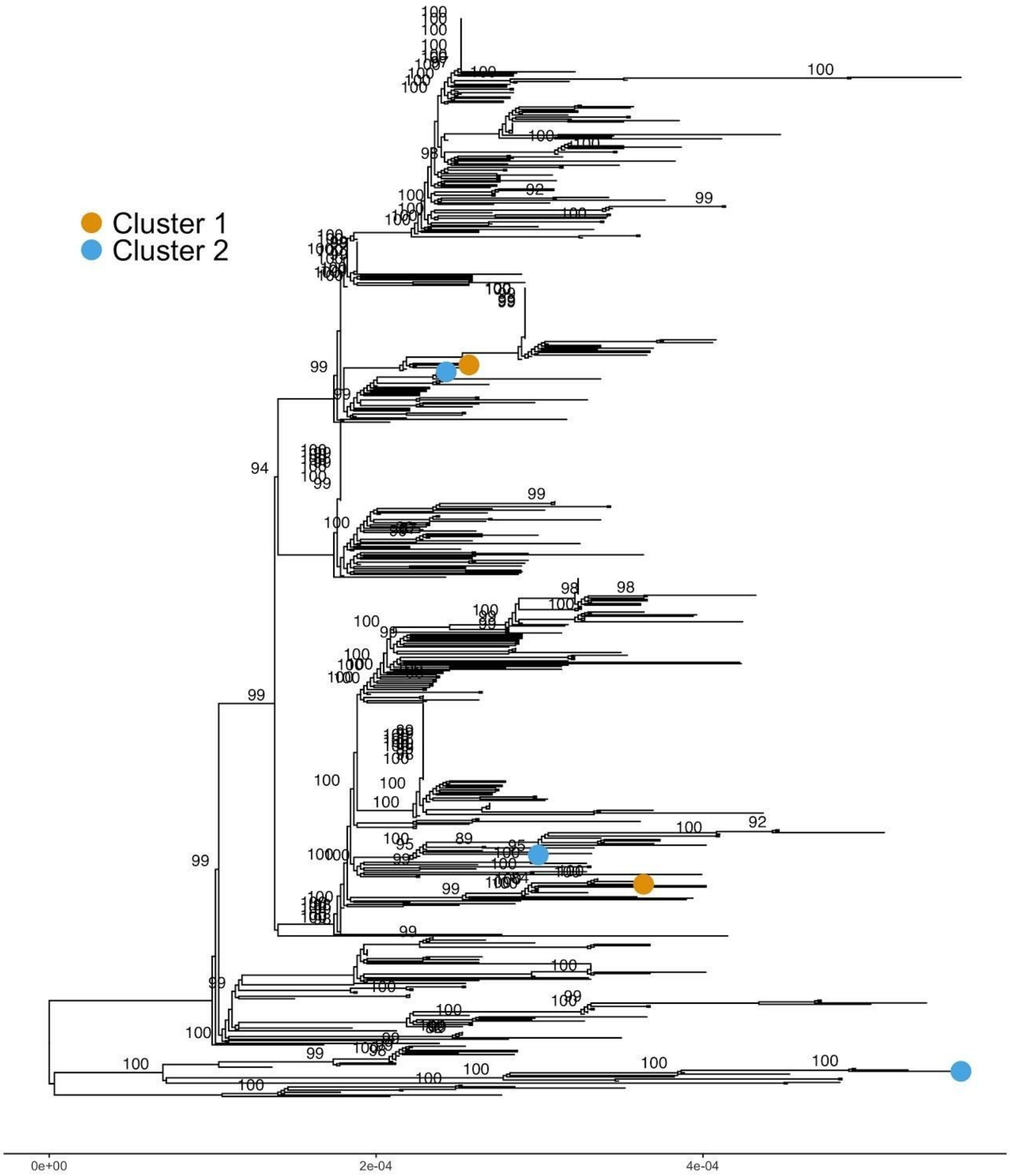


Fig S17.
Maximum likelihood tree of 772 Boston area SARS-CoV-2 genomes. Samples from two independent, suspected nosocomial clusters are labeled. Bootstrap values for strongly supported nodes (ultrafast bootstrap support > 80) are shown. Scale bar shows substitutions per site.

References and Notes

1. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, (JHU), COVID-19 Dashboard, (available at https://coronavirus.jhu.edu/map.html).

2. M. A. Waltenburg, C. E. Rose, T. Victoroff, M. Butterfield, J. A. Dillaha, A. Heinzerling, M. Chuey, M. Fierro, R. H. Jervis, K. M. Fedak, A. Leapley, J. A. Gabel, A. Feldpausch, E. M. Dunne, C. Austin, C. S. Pedati, F. S. Ahmed, S. Tubach, C. Rhea, J. Tonzel, A. Krueger, D. A. Crum, J. Vostok, M. J. Moore, H. Kempher, J. Scheftel, G. Turabelidze, D. Stover, M. Donahue, D. Thomas, K. Edge, B. Gutierrez, E. Berl, M. McLafferty, K. E. Kline, N. Martz, J. C. Rajotte, E. Julian, A. Diedhiou, R. Radcliffe, J. L. Clayton, D. Ortbahn, J. Cummins, B. Barbeau, S. Carpenter, J. C. Pringle, J. Murphy, B. Darby, N. R. Graff, T. K. H. Dostal, I. W. Pray, C. Tillman, D. A. Rose, M. A. Honein; CDC COVID-19 Emergency Response Team, Coronavirus Disease among Workers in Food Processing, Food Manufacturing, and Agriculture Workplaces. *Emerg. Infect. Dis.* **27**, (2020). 10.3201/eid2701.203821 Medline

3. W. E. Wei, Z. Li, C. J. Chiew, S. E. Yong, M. P. Toh, V. J. Lee, Presymptomatic Transmission of SARS-CoV-2 - Singapore, January 23-March 16, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 411–415 (2020). doi:10.15585/mmwr.mm6914e1 Medline

4. T. M. McMichael, D. W. Currie, S. Clark, S. Pogosjans, M. Kay, N. G. Schwartz, J. Lewis, A. Baer, V. Kawakami, M. D. Lukoff, J. Ferro, C. Brostrom-Smith, T. D. Rea, M. R. Sayre, F. X. Riedo, D. Russell, B. Hiatt, P. Montgomery, A. K. Rao, E. J. Chow, F. Tobolowsky, M. J. Hughes, A. C. Bardossy, L. P. Oakley, J. R. Jacobs, N. D. Stone, S. C. Reddy, J. A. Jernigan, M. A. Honein, T. A. Clark, J. S. Duchin; Public Health–Seattle and King County, EvergreenHealth, and CDC COVID-19 Investigation Team, Epidemiology of Covid-19 in a Long-Term Care Facility in King County, Washington. *N. Engl. J. Med.* **382**, 2005–2011 (2020). doi:10.1056/NEJMoa2005412 Medline

5. T. P. Baggett, H. Keyes, N. Sporn, J. M. Gaeta, Prevalence of SARS-CoV-2 Infection in Residents of a Large Homeless Shelter in Boston. *JAMA* **323**, 2191–2192 (2020). doi:10.1001/jama.2020.6887 Medline

6. L. Hamner, P. Dubbel, I. Capron, A. Ross, A. Jordan, J. Lee, J. Lynn, A. Ball, S. Narwal, S. Russell, D. Patrick, H. Leibrand, High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice - Skagit County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 606–610 (2020). doi:10.15585/mmwr.mm6919e6 Medline

7. A. James, L. Eagle, C. Phillips, D. S. Hedges, C. Bodenhamer, R. Brown, J. G. Wheeler, H. Kirking, High COVID-19 Attack Rate Among Attendees at Events at a Church - Arkansas, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 632–635 (2020). doi:10.15585/mmwr.mm6920e2 Medline

8. A. Schuchat; CDC COVID-19 Response Team, Public Health Response to the Initiation and Spread of Pandemic COVID-19 in the United States, February 24-April 21, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 551–556 (2020). doi:10.15585/mmwr.mm6918e2 Medline

9. MA Department of Public Health, Man returning from Wuhan, China is first case of 2019 Novel Coronavirus confirmed in Massachusetts (2020), (available at

https://www.mass.gov/news/man-returning-from-wuhan-china-is-first-case-of-2019-novel-coronavirus-confirmed-in).

10. COVID-19 Response Reporting, *Massachusetts Department of Public Health* (2020), (available at https://www.mass.gov/info-details/covid-19-response-reporting).

11. D. J. Park, G. Dudas, S. Wohl, A. Goba, S. L. M. Whitmer, K. G. Andersen, R. S. Sealfon, J. T. Ladner, J. R. Kugelman, C. B. Matranga, S. M. Winnicki, J. Qu, S. K. Gire, A. Gladden-Young, S. Jalloh, D. Nosamiefan, N. L. Yozwiak, L. M. Moses, P.-P. Jiang, A. E. Lin, S. F. Schaffner, B. Bird, J. Towner, M. Mamoh, M. Gbakie, L. Kanneh, D. Kargbo, J. L. B. Massally, F. K. Kamara, E. Konuwa, J. Sellu, A. A. Jalloh, I. Mustapha, M. Foday, M. Yillah, B. R. Erickson, T. Sealy, D. Blau, C. Paddock, A. Brault, B. Amman, J. Basile, S. Bearden, J. Belser, E. Bergeron, S. Campbell, A. Chakrabarti, K. Dodd, M. Flint, A. Gibbons, C. Goodman, J. Klena, L. McMullan, L. Morgan, B. Russell, J. Salzer, A. Sanchez, D. Wang, I. Jungreis, C. Tomkins-Tinch, A. Kislyuk, M. F. Lin, S. Chapman, B. MacInnis, A. Matthews, J. Bochicchio, L. E. Hensley, J. H. Kuhn, C. Nusbaum, J. S. Schieffelin, B. W. Birren, M. Forget, S. T. Nichol, G. F. Palacios, D. Ndiaye, C. Happi, S. M. Gevao, M. A. Vandi, B. Kargbo, E. C. Holmes, T. Bedford, A. Gnirke, U. Ströher, A. Rambaut, R. F. Garry, P. C. Sabeti, Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161**, 1516–1526 (2015). doi:10.1016/j.cell.2015.06.007 Medline

12. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020). doi:10.1126/science.abc8169 Medline

13. L. Yurkovetskiy, K. E. Pascal, C. Tompkins-Tinch, T. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, K. Veinotte, S. B. Egri, S. F. Schaffner, J. E. Lemieux, J. Munro, P. C. Sabeti, C. Kyratsous, K. Shen, J. Luban, SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. *bioRxiv* (2020), p. 2020.07.04.187757.

14. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori; Sheffield COVID-19 Genomics Group, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020). 10.1016/j.cell.2020.06.043 Medline

15. MA Department of Public Health, Update and Interim Guidance on Outbreak of 2019 Novel Coronavirus (2019-nCoV) in Wuhan, China (2020), (available at https://www.mass.gov/clinical-advisory/update-and-interim-guidance-on-outbreak-of-2019-novel-coronavirus-2019-ncov-in).

16. MA Department of Public Health, Coronavirus Disease 2019 (COVID-19) Cases in MA, March 15 2020 (2020).

17. The CDC sequenced 19 MA genomes prior to March 8 2020. 17/19 cases (89%) contained C2416T. The CDC MA genomes are not annotated with exposure information, but given official MADPH data reporting that 23/28 cases as of March 8 2020 were linked to the

conference (19), and the five non-conference associated cases include the travel-associated cases from this time period (MA-1, DPH_00002, and DPH_00003) and one from the Berkshire county cluster (all of which lack C2416), it can be inferred that a minimum of 16/17 C2416T-containing samples sequenced by the CDC were conference-associated.

18. North Carolina Department of Health and Human Services, Five More People in North Carolina Test Positive for COVID-19 (2020), (available at https://www.ncdhhs.gov/news/press-releases/five-more-people-north-carolina-test-positive-covid-19).

19. Indiana State Department of Health, State Health Department Announces 2nd COVID-19 Case (2020), (available at https://calendar.in.gov/site/isdh/event/isdh-news-release-state-health-department-announces-2nd-covid-19-case/).

20. Tennesse Department of Health, TDH Releases Further Information Regarding COVID-19 Case, (2020), (available at https://www.tn.gov/health/news/2020/3/5/tdh-releases-further-information-regarding-covid-19-case.html).

21. Indiana State Department of Health, State Health Department Confirms 1st Case of COVID-19 in Hoosier with Recent Travel (2020).

22. H. Reese, A. D. Iuliano, N. N. Patel, S. Garg, L. Kim, B. J. Silk, A. J. Hall, A. Fry, C. Reed, Estimated incidence of COVID-19 illness and hospitalization - United States, February-September, 2020. *Clin. Infect. Dis.* ciaa1780 (2020). doi:10.1093/cid/ciaa1780 Medline

23. S. A. Goldberg, J. Lennerz, M. Klompas, E. Mark, V. M. Pierce, R. W. Thompson, C. T. Pu, L. L. Ritterhouse, A. Dighe, E. S. Rosenberg, D. C. Grabowski, Presymptomatic Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 Among Residents and Staff at a Skilled Nursing Facility: Results of Real-time Polymerase Chain Reaction and Serologic Testing. *Clin. Infect. Dis.* 10.1093/cid/ciaa991 (2020).

24. M. Bharel, Order of the Commisioner of Public Health, (2020), (available at https://www.mass.gov/doc/march-15-2020-assisted-living-visitor-restrictions-order/download).

25. A. Endo, S. Abbott, A. J. Kucharski, S. Funk; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020). doi:10.12688/wellcomeopenres.15842.3 Medline

26. C. B. Matranga, K. G. Andersen, S. Winnicki, M. Busby, A. D. Gladden, R. Tewhey, M. Stremlau, A. Berlin, S. K. Gire, E. England, L. M. Moses, T. S. Mikkelsen, I. Odia, P. E. Ehiane, O. Folarin, A. Goba, S. H. Kahn, D. S. Grant, A. Honko, L. Hensley, C. Happi, R. F. Garry, C. M. Malboeuf, B. W. Birren, A. Gnirke, J. Z. Levin, P. C. Sabeti, Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014). doi:10.1186/s13059-014-0519-7 Medline

27. R. Ihaka, R. Gentleman, R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).

28. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, J. Zhang, Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004). doi:10.1186/gb-2004-5-10-r80 Medline

29. H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019). doi:10.21105/joss.01686

30. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T. Lam, ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017). doi:10.1111/2041-210X.12628

31. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019). doi:10.1186/s13059-019-1891-0 Medline

32. J. W. Leigh, D. Bryant, POPART: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015). doi:10.1111/2041-210X.12410

33. T. Jombart, I. Ahmed, adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011). doi:10.1093/bioinformatics/btr521 Medline

34. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005). doi:10.1038/nature04153 Medline

35. L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser, Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020). doi:10.1126/science.abb6936 Medline

36. C. B. Matranga, A. Gladden-Young, J. Qu, S. Winnicki, D. Nosamiefan, J. Z. Levin, P. C. Sabeti, Unbiased Deep Sequencing of RNA Viruses from Clinical Samples, Unbiased Deep Sequencing of RNA Viruses from Clinical Samples. *J. Vis. Exp.* (113): (2016). 10.3791/54117 Medline

37. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

38. S. Guindon, J. F. Dufayard, W. Hordijk, V. Lefort, O. Gascuel, in *Infection Genetics and Evolution* (ELSEVIER SCIENCE BV PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS, 2009), vol. 9, pp. 384–385.

39. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016). doi:10.1093/ve/vew007 Medline

40. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015). doi:10.1093/molbev/msu300 Medline

41. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018). doi:10.1093/ve/vex042 Medline

42. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018). doi:10.1093/bioinformatics/bty407 Medline

43. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020). doi:10.1016/S1473-3099(20)30120-1 Medline