A Unique Clade of SARS-CoV-2 Viruses is Associated with Lower Viral Loads in Patient Upper Airways

Ramon Lorenzo-Redondo[1,*], Hannah H. Nam[1,*], Scott C. Roberts[1,*], Lacy M. Simons[1,*], Lawrence J. Jennings[2], Chao Qi[2], Chad J. Achenbach[1], Alan R. Hauser[1,3], Michael G. Ison[1], Judd F. Hultquist[1,+], Egon A. Ozer[1,+]

[1] Department of Medicine, Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, 60611, USA.

[2] Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA.

[3] Department of Microbiology-Immunology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA.

[*] These authors contributed equally to this work.

[+] These authors contributed equally to this work.

Correspondence to: e-ozer@northwestern.edu, judd.hultquist@northwestern.edu

ABSTRACT

The rapid spread of SARS-CoV-2, the causative agent of Coronavirus disease 2019 (COVID-19), has been accompanied by the emergence of distinct viral clades, though their clinical significance remains unclear. Here, we examined the genome sequences of 88 SARS-CoV-2 viruses from COVID-19 patients in Chicago, USA and identified three distinct phylogenetic clades. Clade 1 was most closely related to clades centered in New York, and showed evidence of rapid expansion across the USA, while Clade 3 was most closely related to those in Washington. Clade 2 was localized primarily to the Chicago area with limited evidence of expansion elsewhere. Average viral loads in the airways of patients infected with the rapidly spreading Clade 1 viruses were significantly higher than those of the poorly spreading Clade 2. These results show that multiple variants of SARS-CoV-2 are circulating in the USA that differ in their relative airway viral loads and potential for expansion.

INTRODUCTION

On December 31, 2019, the Wuhan Health Commission first reported clusters of people with severe respiratory infections and pneumonia from Wuhan, China (1-3). With more than 3 million confirmed cases worldwide and more than 200,000 attributable deaths during the first four months of 2020, the novel single-stranded positive-sense RNA betacoronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has emerged as the cause of an unprecedented worldwide healthcare crisis (4, 5). The spread of the novel coronavirus disease (COVID-19) has proceeded at different rates across the globe, but ongoing community transmission is widespread.

In the United States, the first confirmed case of COVID-19 was reported on January 20th in Washington state with a second confirmed case reported shortly thereafter in Chicago, Illinois on January 23rd (6). Both patients had recently traveled to Wuhan, China, where the transmission event is suspected to have taken place. The Chicago patient later passed the virus to her husband in the first reported case of local transmission on January 30th (6). The state of New York quickly developed into a hotspot after its first case was reported on March 1st. By the end of March, the United States had confirmed cases in all 50 states and reported the most cases of any country in the world. Stay-at-home orders went into effect in an attempt to mitigate virus spread in Illinois, New York, and Washington on March 21st, 22nd, and 23rd, respectively.

Based on publicly available data, prevalence and mortality rates differ on a state-by-state basis [as of April 27th, 2020: New York, 282,143 cases (5.8% mortality); Washington state, 13,319 cases (5.5% mortality); Illinois, 41,777 cases (4.5% mortality)] (7, 8). Specifically, in Illinois the upswing in cases occurred later than in Washington, despite estimated similar times of arrival, and the rate of increase has been slower than observed in New York (**Supplemental Figure 1**) (7). These differences can be attributed to a multitude of factors including: geography, population density, demographics, the timing and extent of community mitigation measures, testing availability, healthcare infrastructure, and public health reporting practices (9). Whether variability in viral genotype may also contribute to these observed differences in transmission and disease burden remains unknown.

In parallel with the unprecedented speed and scale of the pandemic, the rapid sharing and integration of clinical and epidemiological data has also occurred at an exceptional pace, with over 10,000 viral genome sequences of SARS-CoV-2 from across the globe currently available via online databases such as GISAID (10). Rapid whole-genome sequencing (WGS) of emerging viruses allows for informed development of molecular diagnostic assays, as well as

tracing of patterns of spread across multiple epidemiological scales (11). Understanding the molecular epidemiology of the virus through evolutionary models and phylogenetic analysis can further aid in estimating the genetic variability and evolutionary rate, which can have significant implications for disease progression as well as drug and vaccine development.

While extensive WGS has been reported in Washington state and New York, there are relatively few sequences available from Illinois. Here, we present a phylogenetic analysis of 88 SARS-CoV-2 viruses from the airways of COVID-19 patients seeking care in a single healthcare system in Chicago, Illinois from March 14, 2020 through March 21, 2020. These analyses find three distinct clades of virus circulating in the Chicago area, most closely resembling viruses circulating in New York (clade 1), China (clade 2), and Washington state (clade 3). Notably, the viral loads in patients infected with clade 1 viruses were significantly higher than clade 2. These results suggest that genetic differences in SARS-CoV-2 may be linked to differences in viral load and, potentially, with transmissibility of the virus.

METHODS

Specimen collection and testing

Individuals suspected of COVID-19 within two hospitals in Northwestern Medicine (NM) were screened for SARS-CoV-2 infection as per the CDC protocol (12). The NM a healthcare system comprises 10 hospitals spread across the Chicagoland area spanning approximately 50 square miles. Samples were collected between March 14, 2020 through March 21, 2020. Participants included individuals presenting to emergency departments at Northwestern Memorial Hospital (NMH) or Lake Forest Hospital (LFH), patients admitted to the hospital at NMH, and outpatient or corporate health clinic patients. Nasopharyngeal swab and/or bronchoalveolar lavage fluid specimens were collected from screened individuals and stored in either phosphate-buffered saline (PBS) or viral transport medium. Viral RNA was extracted from clinical specimens utilizing the QIAamp Viral RNA Minikit (Qiagen). Clinical testing for SARS-CoV-2 presence was performed by quantitative reverse transcription and PCR (qRT-PCR) with the CDC 2019-nCoV RT-PCR Diagnostic Panel utilizing N1 and N2 probes as previously described (12). All specimens with a cycle threshold (Ct) less than or equal to 35 were considered positive and included in this study. Specimens and clinical data of patients with positive tests had patient-identifying information removed and were assigned a study code number.

cDNA Synthesis and Viral Genome Amplification

cDNA synthesis was performed with SuperScript IV First Strand Synthesis Kit (Thermo) using 11 µl of extracted viral nucleic acids and random hexamers according to manufacturer's specifications. Direct amplification of the viral genome cDNA was performed using primer sets generating overlapping ~ 400 bp amplicons tiled across the genome. The multiplex primer set, comprised of two non-overlapping primer pools, was created using Primal Scheme and provided by the Artic Network (13). The version 3 release of this primer set was used for these studies. PCR amplification was carried out using Q5 Hot Start HF Taq Polymerase (NEB) with 5 µl of cDNA in a 25 µl reaction volume. A two-step PCR program was used with an initial step of 98°C for 30 seconds, then 35 cycles of 98°C for 15 seconds followed by five minutes at 65°C. Separate reactions were carried out for each primer pool and validated by agarose gel electrophoresis.

Sequencing Library Preparation and Nanopore Sequencing

Sequencing library preparation approach was adapted from the ARTIC Network protocol and Oxford Nanopore protocol "PCR tiling of COVID-19 virus" (13, 14). Amplicons from both primer pools were combined and purified with a 1x volume of AmpureXP beads (Beckman Coulter). A total of 50 ng of DNA was treated with NEB Ultra II End Prep Enzyme mix (NEB). Up to 24 specimen libraries were barcoded using Nanopore Native Barcoding Expansion kits (Oxford Nanopore) and NEBNext Ultra II Ligase (NEB) for simultaneous sequencing. Uniquely barcoded samples were pooled and cleaned with a 0.4x volume of AmpureXP beads. Adapter ligation and cleanup was performed using the Oxford Nanopore Genomic DNA by Ligation kit according to manufacturer specifications. Libraries were sequenced on the Nanopore MinION device using FLO-MIN106D Type R9.4.1 flow cells.

Sequence data analysis

Base calling of nanopore reads was performed on the Northwestern University's Quest High Performance Computing Cluster for Genomics using Guppy v3.4.5. Read quality filtering, reference alignment, and consensus sequence generation was performed using ARTIC Network software and protocols (15). Briefly, demultiplexed and quality filtered sequence reads were aligned to a SARS-CoV-2 reference genome sequence (NCBI accession number MN908947.3) using minimap2 v2.17 (16). Alignments were filtered to remove low-quality alignments, and

variants relative to the reference were determined from raw signal data using nanopolish v0.12.5 (17). Regions with less than 10-fold read coverage were masked. Validation of variant calls was performed using a combination of custom in-house software and manual examination of alignments using Tablet software (18). Consensus genome sequences were deposited in the GISAID database. Accession numbers are given in **Supplemental Table 1**. Publicly-available SARS-CoV-2 genome sequences were downloaded from the GISAID database on 04/10/2020 (10). Only full genome sequences with <1% Ns were selected for inclusion in the analyses.

Phylogenetic analysis and phylogeographic inference

Genome sequences were aligned using MAFFT v7.453 (19) software and manually edited using MEGA v6.06 (20). All Maximum Likelihood (ML) phylogenies were inferred under the generalized time-reversible (GTR) nucleotide substitution model using the double-precision FastTree v 2.1.11 (21), assessing the tree topology with the Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT) (22) with 1000 replicates. TreeTime v0.7.4 (23) was used for estimation of time scaled phylogenies and ancestral reconstruction of most likely sequences of internal nodes of the tree and transitions between geographical locations along branches. We also used TreeTime to infer the history of the effective population size (Ne), *i.e.* inverse coalescence rate, using Kingman coalescent as a tree prior when estimating the time trees for each phylogeny.

Bayesian time-scaled phylogenetic analyses were performed with BEAST v2.5.2 (24) to estimate the date and location of the most recent common ancestors (MRCA) as well as to estimate the rate of evolution of the virus. BEAST priors were introduced with BEAUTI v2.5.2 including an uncorrelated relaxed molecular clock model with a lognormal distribution, using a GTR substitution model with invariant sites, and a Coalescence Bayesian Skyline (25) to model the population size changes through time. Markov chain Monte Carlo (MCMC) runs of at least 100 million states with sampling every 5,000 steps were computed. The convergence of MCMC chains was monitored using Tracer v.1.7.1, ensuring that the effective sample size (ESS) values were greater than 200 for each parameter estimated.

Phylogeographic patterns were estimated using a discrete-state continuous time Markov chain to reconstruct the spatial dynamics between geographical locations (26), assuming an asymmetric transition model with separated rate parameters for each possible transition. MCMC was run for over 100 million steps with a burn-in of 20%. Parameters were sampled every 5,000 steps and trees sampled every 10,000 steps. Pairwise migration rate estimates had an ESS of more than 1000 in every case. Statistics from the trees were extracted using PACT v0.9.5 (27)

and SpreaD3 v0.9.6 (28) was used to visualize the phylogeographic patterns associated with the phylogenies. In each case, the maximum clade credibility (MCC) trees were obtained from the tree posterior distribution using TreeAnnotator v2.5.2 after 20% burn-in.

Statistical Analysis

All statistical analyses were performed in R version 4.0.0. To evaluate for differences in Ct values between clades, specimens were grouped according to the estimated clade of the detected virus. All specimens that did not cluster in any of the 3 major clades were placed into a separate group. A linear model was fitted after testing for normal distribution of the data. Clade membership was included in the model, and possible confounders such as the date of the qPCR, age, sex, race, maximal severity of illness, and the specimen collection source were controlled for. All possible contrasts within the model were performed and corrected for multiple comparisons using Tukey's test. Corrected p-values of less than 0.05 were considered statistically significant.

To test for association between disease severity and viral clade, we employed multinomial logistic regression, using maximal severity of COVID-19 per patient as the outcome variable [binned as mild (outpatient or ED only), moderate (inpatient hospitalization), or severe (ICU admission and/or death)], and clade, Ct values, and other demographic variables as the predictors. To analyze the association between clade and all available clinical and demographic variables, we fitted a logistic regression model including age, sex, race, severity of illness, and specimen source to test for any possible association between these variables and the clade observed. Finally, we also tested for differences in viral load by specimen source (nasopharyngeal swab vs bronchoalveolar lavage) by fitting a linear model that included the same confounders as before, extended to contain the interaction between clade and specimen source. Within this model, we tested for all possible combinations of clade, source, and their interaction and corrected for multiple comparisons as before. For these last three analyses we only used data from patients with viruses from our two major clades, Clades 1 and 2.

Ethics Statement

This study was reviewed and approved by the Institutional Review Board of Northwestern University.

RESULTS

Collection characteristics and sequencing results

Between March 14 and March 21, 2020, a total of 127 clinical tests performed by the Diagnostic Molecular Biology Laboratory at Northwestern Memorial Hospital (NMH) were positive for SARS-CoV-2 as determined by qRT-PCR. The residual RNA from these diagnostic tests were assigned unique sample identifiers and collected for sequence analysis. Of these, 37 specimens failed to achieve sufficient viral genome amplification required for sequencing or failed to achieve required coverage after alignment, defined in this study as at least 10x coverage across at least 90% of the genome. The remaining 90 specimens were considered successfully sequenced (**Supplemental Figure 2A**). Almost half of these 90 specimens were collected from patients evaluated in emergency departments at NMH or Lake Forest Hospital, 25% from non-ICU hospitalized patients at NMH, and 12% from patients in the NMH ICU (**Supplemental Figure 2B**). Two sequenced specimens (NM-nCoV-0082 and NM-nCoV-0087) were found to be redundant and excluded from this analysis, leaving 88 sequenced specimens from 88 individuals. The Ct values of these specimens ranged from 14.52 to 32.37. The average of median coverage across non-overlapping sequences was 1513.6 reads with a minimum median genome coverage of 65 and a maximum median genome coverage of 4059. After applying alignment quality filters and normalizing segment coverage to a maximum of 200 reads in each direction, the average median genome coverage was 348 reads per specimen (min 56, max 395).

Three main clades are detected through phylogenetic inference

We performed a maximum likelihood (ML) phylogenetic analysis of the 88 SARS-CoV-2 specimen genomes using the GTR nucleotide substitution model. To measure branch support, we used SH-aLRT with 1000 replicates. This analysis showed that most of the sequences analyzed (93%) clustered in three main clades with strong statistical support (>85% aLRT) (**Figure 1A**). Due to the lack of a consensus in current nomenclature, we will refer to these clades throughout this report, ordered by relative abundance in the sequence set, as Clades 1, 2, and 3. The intra- and inter-clade variability of the sequences showed a mean intra-group variability of 1.16×10^{-4} base differences per site [SEM= 3.16×10^{-5}] versus a mean inter-group variability of 4.80×10^{-4} base differences per site [SEM= 1.17×10^{-4}], confirming the higher within-

clade sequence similarity in the observed phylogenetic clustering. Among our sequences, 57.9% clustered in Clade 1, 27.3% in Clade 2, and 7.9% in Clade 3.

Clade 2 is uniquely prominent in the Chicago area

To place the three observed clades in a broader geographic context, we expanded our analysis to include 901 SARS-CoV-2 sequences from the USA deposited in GISAID as of April 4, 2020. We included the reference sequence obtained from Wuhan, China at the beginning of the pandemic to root the trees. With this dataset, we performed ML phylogenetic reconstruction of the sequences, generated a temporal phylogeny, and reconstructed the ancestral states of the tree to identify clade-defining mutations (**Figure 1B**) and possible geographical location transitions. Finally, we performed phylodynamic analyses using a Bayesian approach to confirm the clades observed in our sequences and understand their origin and dynamics. These analyses confirmed our previous clustering, with a bulk of the Chicago sequences falling into three distinct clades (**Figure 2A**). Overall, the local epidemic in the Chicago area is broadly reflective of the epidemic in the USA as a whole, with most major and minor clades represented within this one hospital system. These results are consistent with three main introductions of SARS-CoV-2 into the Chicago area seeding the three predominant clades, while also suggesting several other minor introductions with more limited expansion (**Figure 2A**).

Clade 1 viruses cluster most similarly with those predominant in New York, while Clade 3 viruses cluster most similarly with those circulating on the West Coast, particularly in Washington state. These two clades show a much broader spread across the USA than Clade 2, which is primarily localized to the Chicago area. In fact, Clade 2 was comprised almost entirely of our newly-sequenced specimens, with only a few representatives from other parts of the USA, including a couple sequences each from Georgia and Wisconsin (**Figure 2A**, **Supplemental Figure 3**). These sequences were closely related to the SARS-CoV-2 sequences from the first COVID-19 patients identified in Illinois by the Illinois Department of Public Health (IDPH) at the end of January 2020 (6).

Ancestral state reconstruction analyses identified a number of mutations uniquely circulating among different strains (**Figure 1B**). Clade 1 was defined by at least three mutations, a synonymous change in Orf1a (C3037T), a nonsynonymous change in orf1ab (C14408T; Nsp12 P322L), and a nonsynonymous change in S (A23403G; Spike D614G). Two mutations were shared by Clade 2 sequences, both of which are non-synonymous mutations in the viral orf1a gene (T490A and C3177T; Nsp1 D75E and Nsp3 P153L, respectively). Clade 3 carried

two defining mutations, a synonymous change in orf1ab (C18060T) and a nonsynonymous change in orf1ab (A17858G; Nsp13 Y541C).

We next studied the phylodynamic and phylogenomic characteristics of these three clades to infer possible differences in their population dynamics. These analyses suggest very limited spread of Clade 2, with an average proportion of branches residing in Illinois of 0.81 [95% CI: 0.66-0.88] (**Supplemental Figure 3**). The most probable USA origin of Clade 2 was in Illinois, and the time of the most recent common ancestor (TMRCA) was around January 18, 2020 [95% Highest Posterior Density (HPD) interval: January 9 - January 21, 2020] (Figure 4D). This timing, location, and the sequences at the inferred root of this tree (USA_IL_1 and USA_IL_2) correspond with the first reported COVID-19 cases in Illinois: an individual who returned from Wuhan on January 13 and her husband. These results suggest that Clade 2 was introduced to the Chicago area early in the USA epidemic with limited spread beyond Illinois (**Figure 2B**, middle).

A similar analysis of Clade 1 suggested multiple introductions, a high degree of diversity, and the existence of several probable sub-clades. Although the highest proportion of residing branches of this clade were in New York, this proportion was only 0.34 [95% CI: 0.31-0.37], indicating broad geographic spread of this clade (**Supplemental Figure 3**). These findings are most consistent with Clade 1 being centered in New York, but rapidly expanding throughout the New York region and to many other locations across the USA over the course of the epidemic (**Figure 2B,** bottom). The TMRCA of this clade in the USA was calculated to be February 10, 2020 [95% HPD interval: January 28 - February 20, 2020], an estimated date in close alignment with previous reports from New York City. These results are most consistent with this clade having had multiple introductions into Illinois with some occurring early during the epidemic in February and March of 2020.

Finally, Clade 3 is estimated to have originated in Washington state and remained centered there throughout the investigated time course (proportion of branches residing in Washington of 0.83 [95% CI: 0.81-0.85]) (**Supplemental Figure 3**). This clade was also introduced into the USA early in the epidemic (TMRCA = January 16, 2020 [95% HPD interval January 11 - January 17, 2020]), and subsequently expanded at a modest rate to multiple locations (**Figure 2B**, top). This timing corresponds with the first confirmed case was reported in Washington state, which was a patient who returned from Wuhan on January 15 (29). These data are consistent with two main introductions of Clade 3 into Illinois, followed by limited spread. Consequently, only 8% of the sequences in our collection are Clade 3.

The three clades also differ considerably in their inferred Effective Population Sizes (Ne) (**Figure 2C**). Clade 1 had a later introduction into the USA, but a subsequent exponential expansion made it the predominant clade. Clade 3 had two phases of expansion to reach a slightly lower peak population size. Cluster 2, on the other hand, has had a constant and limited expansion despite its early introduction (**Figure 2C**).

Overall, our results suggest that as of mid-March 2020, three main clades of SARS-CoV-2 were circulating in the Chicago area. Clade 2 appears to have been introduced in Illinois early in the epidemic and seems to have undergone more limited spread when compared to the other two clades, Clade 1 centered in New York and Clade 3 centered in Washington.

Clade 2 has limited global distribution and most likely originated in China

To further investigate the origin and population dynamics of the Chicago-centered Clade 2, we performed a phylogenetic analysis at the global pandemic level. To this end, we included an additional 3099 SARS-CoV-2 sequences deposited in GISAID from throughout the world up through April 4, 2020. We conducted similar ML analyses as those performed on the USA sequences above (**Supplemental Figure 4**). These results further confirmed the strong statistical support for Clade 2 (>90%). Except for some sequences from Australia, the spread of this clade has been exceptionally limited. Ancestral reconstruction of these global sequences showed a close relationship between Clade 2 with sequences obtained from China, mainly Shanghai, with low divergence from the first sequenced samples in Wuhan, China (**Figure 3A**).

To better discern the most likely origin of Clade 2, we performed a phylogeographical analysis using all available Clade 2 sequences and related global sequences. As suggested above, these results determined the most likely origin of this clade at our current state of global sampling to be Wuhan, China (**Figure 3B**). Clade 2-like sequences from other areas outside the USA most likely represent introductions of similar viruses into those areas rather than secondary seeding from Illinois. One exception to this is a cluster of sequences in Australia that appear to be more closely related to the Illinois sequences than those from China (**Figure 3B**). Together, these findings are most consistent with a single introduction of this clade into Illinois directly from Wuhan, China around mid-January 2020 as the source of the observed Clade 2 sequences in Chicago. This result coincides with the history of the first case reported in Illinois, which was a traveler from Wuhan. The sequence of the virus from this individual is very close to the node that defines this clade (6).

Additionally, global analysis with the other two clades (**Supplemental Figure 4**) suggest that the New York-centered Clade 1 was likely to have been introduced into the USA from China

through Europe, consistent with other reports. On the other hand, Clade 3 was most likely introduced into the USA directly from China, thereafter remaining and spreading almost exclusively in the USA.

Together, these analyses show very distinct global phylodynamics of the three main clades observed in the Chicago area: Clade 1 was introduced by way of New York through Europe, Clade 2 was introduced directly from China, and Clade 3 was introduced through Washington state from China. While Clade 1 is well-represented on the East Coast and Clade 3 is well-represented on the West Coast, Clade 2 is almost exclusively found in the Chicago area.

Clade 2 viruses are associated with lower viral titers at the time of diagnosis

The limited spread of Clade 2 SARS-CoV-2 viruses both in the USA and worldwide are suggestive of phenotypic variability that may impact transmission rate. To further investigate this, we compared qRT-PCR cycle threshold (Ct) values, a proxy for viral load in the patients' airways, calculated for each sample at the time of diagnosis. We first grouped the viruses according to their assigned clade, creating an additional group for the viruses that did not fall into any of the three major clades. We fitted a linear model to test for differences in Ct values between the clades, controlling for the date of measurement, age, sex, race, severity of illness, and specimen collection source (**Supplemental Table 2**). We detected significantly higher Ct values corresponding to lower viral loads in Clade 2 (Chicago-centered) compared to Clade 1 (New York-centered) (q=0.0080 after multiple comparison correction) (**Figure 4A**). Similarly, Clade 3 (Washington-centered) had lower viral loads on average compared to Clade 1 (q= 0.0270), although we had fewer samples from this Clade. These results suggest that Chicago-centered Clade 2 and Washington-centered Clade 3 viruses have lower overall viral loads in the airways of infected patients relative to Clade 1 viruses.

Next, we tested whether the source of the sample (nasopharyngeal swab [NP] vs bronchoalveolar lavage fluid [BAL]) influenced the viral load, as measured by the Ct value, and/or could account for the differences between clades. Due to the low number of BAL samples represented in our dataset, we narrowed our analysis to just Clades 1 and 2 and tested for differences between sample source, clade, and both, accounting for their potential interaction. Overall, this analysis showed a significant difference between NP and BAL samples (q= 0.0484), with lower Ct values in the BAL samples (**Figure 4B**). When considering only NP samples, the Ct values of Clade 1 viruses were still significantly lower than Clade 2 viruses as observed previously (q=0.0062). However, no significant difference in Ct value was detected by clade in the BAL samples. While the number of BAL samples are low, this result suggests that

the viral load differences between the two clades are primarily observed in the upper airways of infected patients. That being said, most of our BAL samples came from patients in the ICU while only two NP samples where obtained from the ICU, making it difficult to disentangle whether the differences observed between BAL and NP are due to the severity of the disease or the anatomical compartment from which the samples were obtained (**Supplemental Table 2**).

Given early reports that COVID-19 disease severity is correlated with SARS-CoV-2 viral load (30, 31), we next looked for associations between viral clade and additional clinical characteristics. To test this, we fit a logistic regression model including all available demographic and clinical variables to evaluate for associations with these variables with maximal disease severity and/or clade. We found no significant association between any available variable and disease severity in our limited dataset. Consistent with our previous results, Ct value was the only variable associated with viral clade and vice versa. These data suggest that Ct value in the upper airway at the time of diagnosis may not be associated with disease severity, though it is associated with the viral genotype. Given the rapid rates of phylogeographic and population size expansion of Clade 1 versus Clades 2 and 3, and the observed higher viral loads in the upper airways of Clade 1-infected patients at the time of diagnosis, it is possible that Clade 1 viruses exhibit higher transmissibility, though this has yet to be formally tested.

DISCUSSION

In this study, we report the phylogenetic and phylodynamic analyses of 88 SARS-CoV-2 genomes from COVID-19 patients in Chicago, Illinois, USA. These results reveal at least three major clades circulating in the Chicago region as of mid-March 2020. Two of the three clades are broadly representative of the major circulating clades that expanded and disseminated rapidly across the USA through early April 2020 with points of origins in New York (Clade 1) and Washington state (Clade 3). In addition, we identified a third major clade (Clade 2) that is relatively unique in the USA and likely originated in Illinois. Despite an early estimated arrival date, viruses in this clade lack the rapid and wide dispersion evidenced with the other viral clades. Interestingly, patients infected with viruses in Clade 1 presented with significantly higher viral loads in the upper airways of patients at the time of diagnosis than patients infected with viruses in Clades 2 or 3. These data suggest that differences in virus genotype may have direct impacts on viral load, which in turn may influence transmission and viral spread.

Chicago, Illinois is the third-largest city in the USA and is located centrally in the country with direct flights to all continents except Antarctica. Despite recording the second confirmed case of COVID-19 and the first confirmed case of person-to-person transmission in the USA, relatively little SARS-CoV-2 sequencing data had been available from the Chicago area with only 11 sequences available on GISAID [as of May 11, 2020]. The data presented here add an additional 88 sequences to this collection and constitute the first systematic analysis of SARS-CoV-2 whole genome sequences from COVID-19 patients in the Chicago area.

Phylogenetic analyses of these sequences demonstrate that the major clades circulating in Chicago in mid-March 2020 represent separate introduction events that can be traced back to distinct sources. Although the Chicago-centered Clade 2 and the Washington-centered Clade 3 are phylogenetically related, this analysis shows that they were the result of geographically and temporally separate introductions into the USA from China. The first positive test for SARS-CoV-2 in Illinois was from an individual who had returned to Chicago directly from Wuhan, China six days prior to becoming ill (6). This individual would later pass this virus to her husband in the first confirmed case of local transmission in the USA. The sequences of these patients' viruses are closely related to viruses from Wuhan, China as well as viruses in Clade 2. Together, these data suggest that Clade 2 viruses were introduced into the Chicago area from China and began community transmission around mid-January 2020. While Clade 3 viruses are also related, they are more consistent with an independent introduction event through an intermediate in Washington state. We additionally find that that Clade 1 viruses are most likely to have been introduced to Chicago via New York, through Europe, a model which is also supported by a recent pre-print report suggesting that most strains of SARS-CoV-2 circulating in New York City are phylogenetically more closely related to strains from the European continent than other potential sources (32).

Both the USA and global sequence data suggest relatively little dispersion of Clade 2 strains beyond Illinois as of mid- to late-March 2020. Indeed, though the clade appears to have originated from China as early as mid-January 2020, there is little evidence to date of spread beyond a few likely cases in other Midwest states and a secondary cluster in Australia. This is in contrast to Clade 3 and especially Clade 1 viruses, which have larger estimated population sizes and have spread over a larger geographic area in less time. Though a large number of isolates collected in the US and globally were available to this study through the GISAID, under-sampling of some regions and/or patient groups could have resulted in imprecise estimates of transmission patterns or potentially have missed a broader distribution of Clade 2 specimens. That being said, 24 of 88 sampled viruses (27.3%) from our system were Clade 2 compared to

only 10 of 901 viruses (1.1%) from across the US and only 34 of 3099 viruses (1.1%) from non-US countries, suggesting that Clade 2 is more abundant in the Chicago area than elsewhere in the US and world as of late March 2020. As more genomes are sequenced and deposited both from Illinois as well as across the US and globally, we will continue to assess changes in the population structure of SARS-CoV-2.

Notably, we found that Clade 2 and Clade 3 viruses had significantly lower viral loads in patient upper airways relative to Clade 1 viruses at the time of diagnosis. This is among the first reports of SARS-CoV-2 phylogenetic lineage correlating with phenotypic differences. Lower viral loads in the airways of infected patients may result in decreased potential for transmission due to correspondingly lower viral titers in respiratory secretions and droplets. This could account for some of the observed differences in the population sizes and spread of the different viral clades. Clade 1 viruses are defined by a couple of nonsynonymous mutations in the RNA-dependent RNA polymerase (Nsp12 P322L) and Spike (S D614G) proteins. Other studies have likewise linked the Spike D614G mutation to a broader global spread and to higher viral loads, though the mechanism of action is unclear (33, 34). Similarly, other reports have noted the rapid spread of the Nsp12 P322L mutation (35). Our study was not designed to determine causal relationships between mutations and phenotypes, and it is possible that the virus is adapting multiple, independent changes that aid in its propagation. Further laboratory and clinical studies will be required to test the impact of these variants on viral replication, disease severity, and transmissibility. The viral diversity within Chicago enabled us to capture this variation in a single health care system and control for sampling date, patient demographics, and hospital location in these analyses, but it is also possible that unmeasured confounders are also influencing these results.

In addition to the potential for lower viral loads, there are several other conceivable factors that could account for the muted spread of Clade 2 viruses outside the Chicago region. First, the state of Illinois instituted a relatively rapid and stringent lockdown by implementing restrictions on mass gatherings on March 13th, closing schools on March 17th, and issuing stay-at-home orders on March 21st. These restrictions may have contributed to a slower spread of this version of the virus beyond the state. Second, the virus may be more prevalent among members of a relatively more insular community within the state with comparatively less travel or socialization with non-community members. Third, these viruses may have distinct clinical manifestations that lessen the opportunity for asymptomatic transmission to multiple other individuals. Ongoing studies of clinical and demographic patient data associated with these specimens in our institution could potentially elucidate whether there is support for any of these

scenarios, though more specimens and sequences will be required to determine these associations.

The clinical data we examined, including disease severity, did not correlate significantly with clade or Ct value. For example, a similar proportion of patients infected with either Clade 1 or Clade 2 viruses were located in the ICU at time of sampling (5 of 51 and 3 of 24, respectively). Previous reports suggested that disease severity is correlated with viral load as measured by Ct value during disease course, but our data suggest the same may not hold true for Ct values collected at diagnosis. Although patients in the ICU had lower Ct values than patients in other locations, supporting a link between viral load and severity, most ICU samples were BAL specimens, not NP specimens. This correlation may therefore simply be reflective of different Ct values in the lower vs. upper airways. Further studies with more specimens and associated clinical and demographic data will be needed to further elucidate these relationships.

As the specimens in this study were collected over the first 8 days during which testing was broadly available at our institution, the phylogenetic patterns and abundances observed represent a cross-sectional snapshot of our region. Longitudinal evaluations could reveal more accurate clade dynamics and population structure. Additionally, these specimens were collected from one health system primarily serving patients from the North and West sides of Chicago, potentially limiting generalizability to other regions in and around Chicago or other parts of Illinois. Nevertheless, these data suggest Chicago is an ideal location for observational and interventional studies to capture the breadth of SARS-CoV-2 genetic diversity. This ability to compare diverse viral genotypes in a single hospital setting allowed us to identify clade-specific differences in viral load at the time of diagnosis, which may in turn have consequences for transmissibility.

ACKNOWLEDGEMENTS

This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. Funding for this work was provided by: a Dixon Translational Research Grant made possible by the generous support of the Dixon family (E.A.O. and J.F.H.); a COVID-19 Supplemental Research award from the Northwestern Center for Advanced Technologies (NUCATS - M.G.I., C.J.A., and J.F.H.), the Gilead Sciences

Research Scholars Program in HIV (J.F.H.), and NIH grants K22 AI136691 (J.F.H.), U19 AI135964 (E.A.O., A.R.H.), K24 AI04831 (A.R.H.), and T32 AI095207 (H.H.N., S.C.R.).

AUTHOR CONTRIBUTIONS

Conceptualization: R.L.R., H.H.N., S.C.R., L.M.S., A.R.H., M.G.I., J.F.H., E.A.O. Investigation: R.L.R., H.H.N., S.C.R., L.M.S., J.F.H., E.A.O. Resources: H.H.N., S.C.R., C.Q., L.J.J., A.R.H., M.G.I., J.F.H., E.A.O. Formal analysis: R.L.R., H.H.N., S.C.R., L.M.S., E.A.O. Supervision: L.J.J., A.R.H., M.G.I., J.F.H., E.A.O. Funding acquisition: A.R.H., M.G.I., J.F.H., E.A.O. Writing: R.L.R., H.H.N., S.C.R., L.M.S., A.R.H., M.G.I., J.F.H., E.A.O.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. The Lancet. 2020;395(10223):470-3.
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33.
3. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New Engl J Med. 2020;382(13):1199-207.
4. Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. J Med Virol. 2020;92(5):522-8.
5. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020;395(10224):565-74.
6. Ghinai I, McPherson TD, Hunter JC, Kirking HL, Christiansen D, Joshi K, et al. First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA. Lancet. 2020;395(10230):1137-44.
7. The COVID Tracking Project [Available from: https://covidtracking.com/data.
8. IDoP H. Coronavirus Disease 2019. [cited 2020. Available from: https://dph.illinois.gov/covid19.
9. Geographic Differences in COVID-19 Cases, Deaths, and Incidence - United States, February 12-April 7, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(15):465-71.
10. GISAID: Over 10,000 viral genome sequences of hCoV-19 shared with unprecedented speed via GISAID GISAID2020 [Available from: gisaid.org.
11. Eden JS, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J, et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. Virus Evol. 2020;6(1):veaa027.
12. (CDC) CfDCaP. CDC 2019-Novel Coronavirus (2019-nCoV) Real Time RT-PCR Diagnostic Panel. Diseases DoV; 2020 3/30/2020. Report No.: CDC-006-00019, Revision 03.
13. Artic Network [Available from: https://artic.network/ncov-2019.
14. Oxford Nanopore Technologies [Available from: https://community.nanoporetech.com.
15. nCoV-2019 novel coronavirus bioinformatics protocol [Available from: https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.
16. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.

17. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12(8):733-5.

18. Milne I, Bayer M, Stephen G, Cardle L, Marshall D. Tablet: Visualizing Next-Generation Sequence Assemblies and Mappings. Methods Mol Biol. 2016;1374:253-68.

19. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80.

20. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725-9.

21. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.

22. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21.

23. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4(1):vex042.

24. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):e1003537.

25. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005;22(5):1185-92.

26. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.

27. Posterior Analysis of Coalescent Trees [April 27, 2020]. Available from: https://github.com/trvrb/PACT.

28. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol. 2016;33(8):2167-9.

29. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med. 2020;382(10):929-36.

30. Yu X, Sun S, Shi Y, Wang H, Zhao R, Sheng J. SARS-CoV-2 viral load in sputum correlates with risk of COVID-19 progression. Crit Care. 2020;24(1):170.

31. Zheng S, Fan J, Yu F, Feng B, Lou B, Zou Q, et al. Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January-March 2020: retrospective cohort study. BMJ. 2020;369:m1443.

32. Gonzalez-Reiche AS, Hernandez MM, Sullivan M, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. medRxiv. 2020:2020.04.08.20056929.

33. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020:2020.04.29.069054.

34. Bhattacharyya C, Das C, Ghosh A, Singh AK, Mukherjee S, Majumder PP, et al. Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of *TMPRSS2* and *MX1* Genes. bioRxiv. 2020:2020.05.04.075911.

35. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18(1):179.

FIGURE LEGENDS

Figure 1 | A) ML phylogenetic tree of 88 SARS-CoV-2 specimen genomes from Northwestern Memorial Hospital and Lake Forest Hospital. All non-zero statistical support values for each branch are indicated. B) Clade-defining mutations at the USA level. Positions are numbered according to the reference genome (NC_045512.2).

Figure 2 | A) ML phylogenetic reconstruction of full genome sequences from the United States. We included sequences from Northwestern and sequences from USA available in GISAID. Branches are colored by location and tips corresponding to Northwestern sequences are highlighted. Well supported Clades of the tree that include our defined Chicago Clades are indicated. B) Phylogeographic patterns of US isolates in three major clades represented in the Chicago collection under a discrete diffusion model. Westward movements are indicated by lines with an upward curvature, eastward movements are indicated by lines with a downward curvature, lines are colored according to the most probable geographical location of their descendent node, and circle sizes around a node are proportional to the number of lineages maintaining that location. C) Effective population sizes (Ne) of US isolates in three major clades represented in the Chicago collection. The estimate of the coalescent population size reconstructed by ML along with its confidence intervals is shown for the three clades. X-axes correspond to the inferred phylogenetic date.

Figure 3 | **a)** Phylodynamic tree of USA and global Clade 2 related genome sequences. We used global sequences phylogenetically related to Clade 2 available in the GISAID database and performed the analysis encompassing simultaneous estimation of sequence and discrete (geographic) trait data. The depicted phylogenetic tree corresponds to the maximum clade credibility tree. Branch colors represent the most probable geographical location of their descendent node inferred through Bayesian reconstruction of the ancestral state. The width of the branches represents their posterior probability. X-axis corresponds to the inferred date. **b)** Phylogeographic reconstruction of the origin of Clade 2 under a discrete diffusion model. Westward movements are indicated by lines with an upward curvature, eastward movements are indicated by lines with a downward curvature, lines are colored according to the most probable geographical location of their descendent node, and circle sizes around a node are proportional to the number of lineages maintaining that location.

Figure 4 | **a)** PCR Cycle threshold (Ct) values of patient samples grouped by major Clade assignment. **b)** PCR Cycle threshold (Ct) values per sample source grouped by the two main Clades assigned. In both figures, significance is indicated for the comparisons performed within each fitted model (* = q-value<0.05; ** = q-value <0.01). Horizontal line in each box represents the median value and lower and upper error bars are the interquartile ranges.

Supplemental Figure 1 | Per-state SARS-CoV-2 infections in the United States from Feb 27 through April 27, 2020. Lines represent total infection volume over time in number of positive tests per 100,000 residents in each of the 50 US states and the District of Columbia. States highlighted in black are Illinois (IL, solid line), New York (NY, dotted line), and Washington (WA, dashed line).

Supplemental Figure 2 | Collection characteristics of sequenced SARS-CoV-2 specimens. **a)** Numbers of specimens collected per day. **b)** Clinical site of specimen collection. ED = Emergency department, Inpatient = Non-ICU hospital ward, ICU = Intensive Care Unit, Other / Unknown = Outpatient, corporate health, or source undetermined.

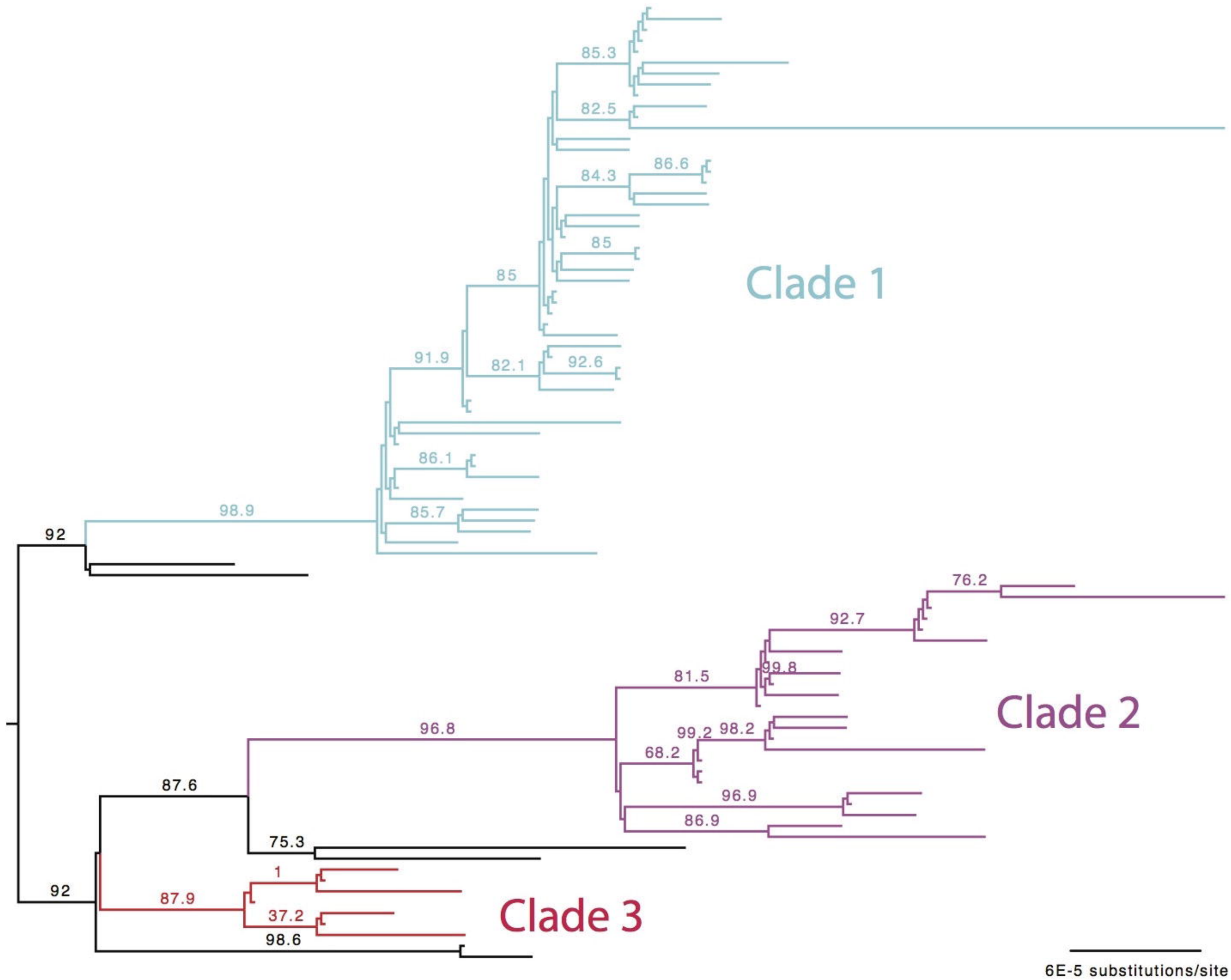
Supplemental Figure 3 | Phylodynamic trees of US isolates in three major clades represented in the Chicago collection. Maximum clade credibility tree where branch colors represent the most probable geographical location of their descendent node inferred through Bayesian reconstruction of the ancestral state. The width of the branches represents their posterior probability. X-axis corresponds to the date in decimal years. **a)** Clade 1, **b)** Clade 2, **c)** Clade 3.

Supplemental Figure 4 | ML phylogenetic tree of all available global SARS-CoV-2 genomes. Clades most closely related to the three major clades in the Chicago collection are highlighted. We included sequences from Northwestern and global sequences from GISAID. Branches are colored by location and tips corresponding to Northwestern sequences are highlighted. Well supported Clades of the tree that include our defined Chicago Clades are indicated.

Supplemental Table 1 | Accession numbers for the consensus SARS-CoV-2 genome sequences deposited in this study in the GISAID database.

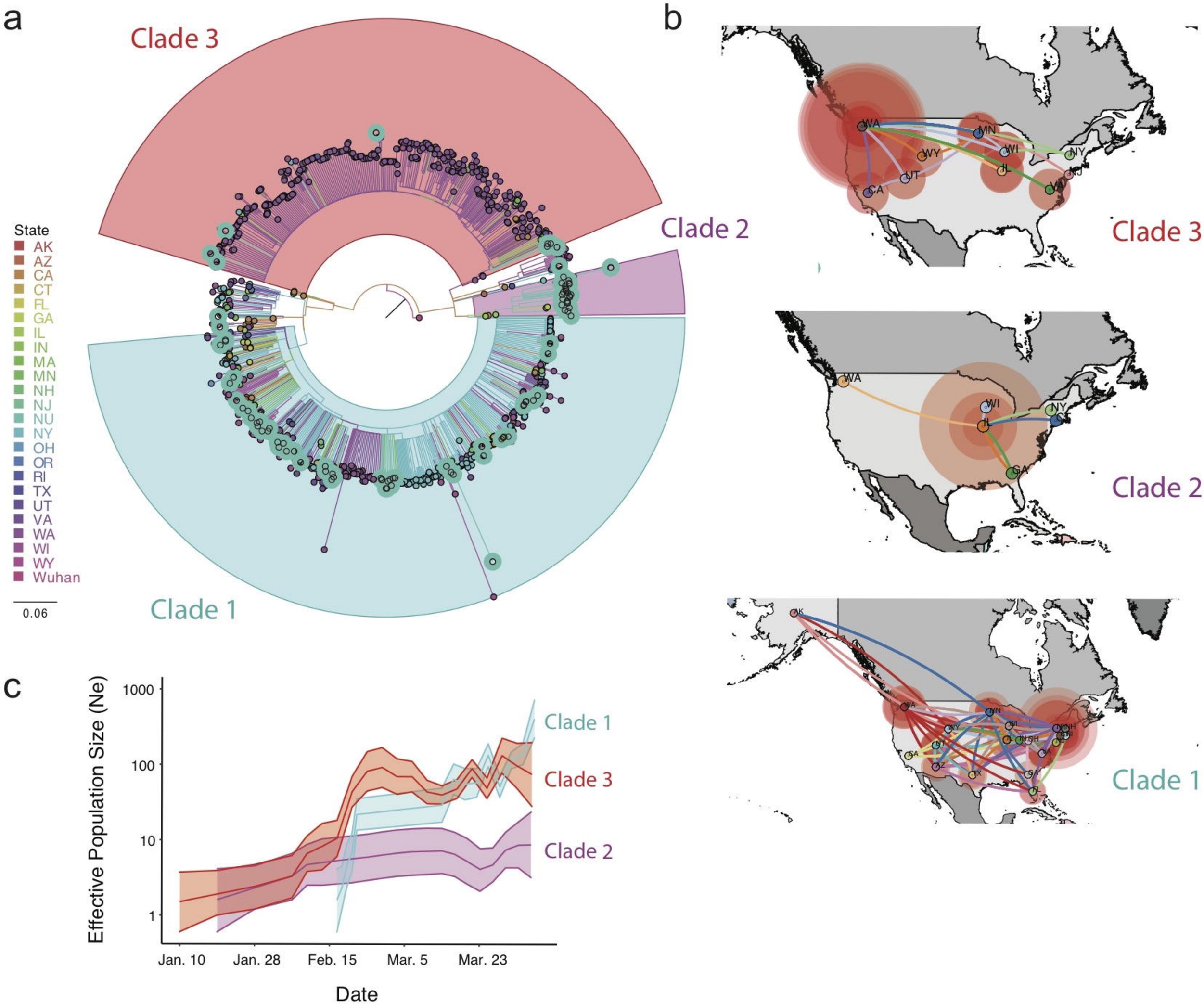
Supplemental Table 2 | Summary of the clinical and demographic data for all patient samples analyzed in this study. P-values represent the results of Chi-squared tests for categorical explanatory variables and F-tests for continuous variables.

a

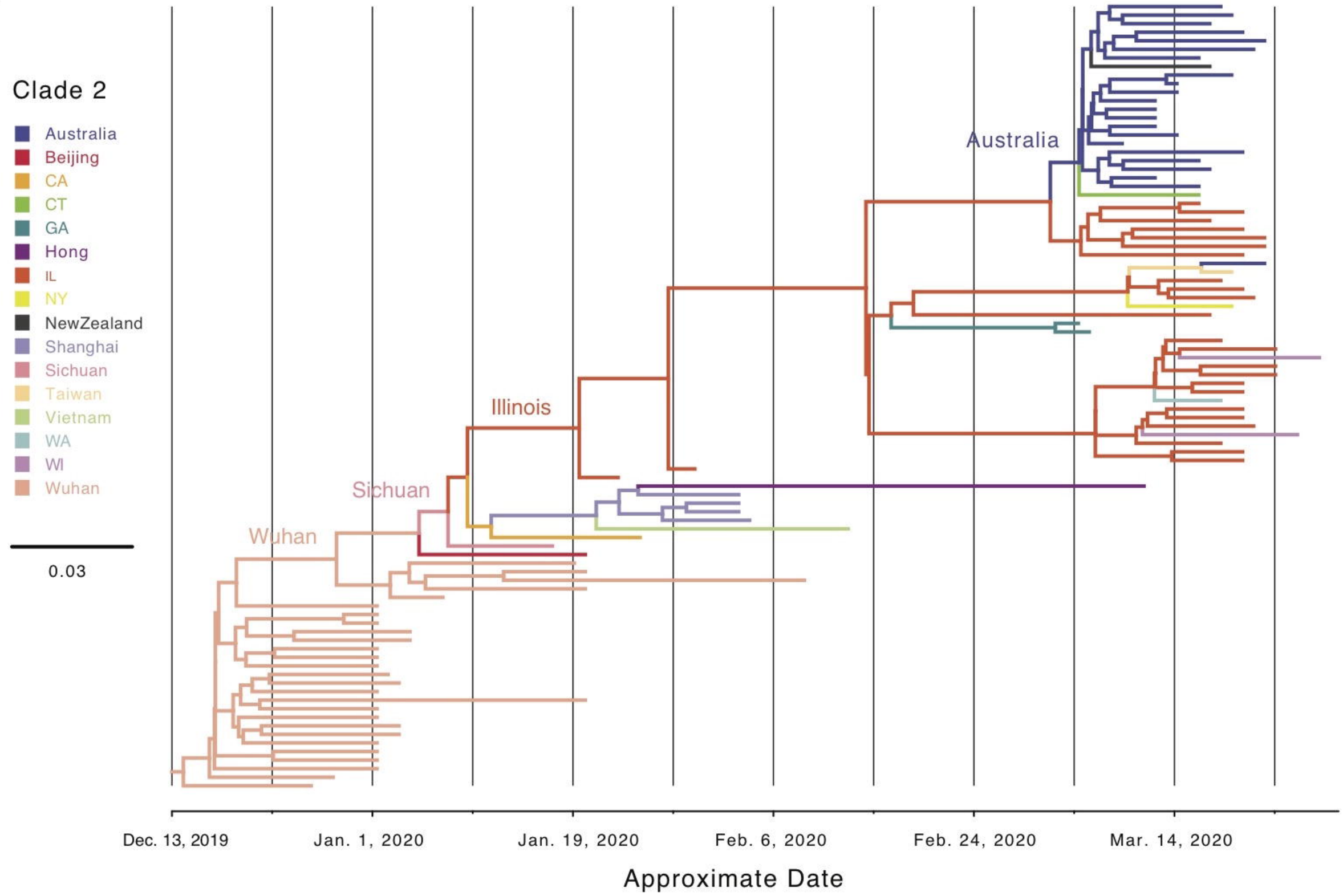


b

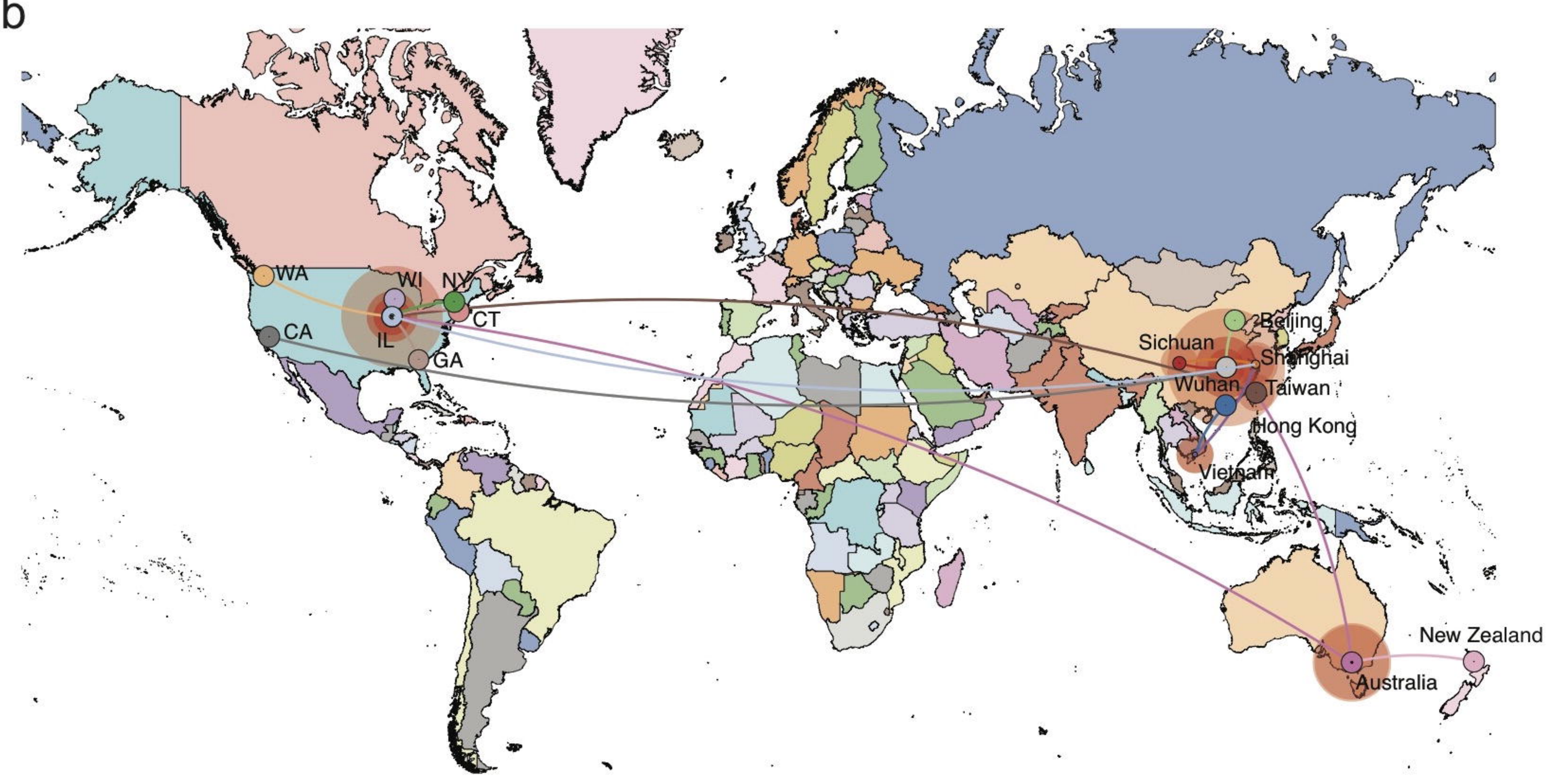
Gene	Orf1a	Orf1a	Orf1a	Orf1ab	Orf1ab	Orf1ab	S
Position	490	3037	3177	14408	17858	18060	23403
Clade 1	T	T	C	T	A	C	G
Clade 2	A	C	T	C	A	C	A
Clade 3	T	C	C	C	G	T	A
Protein	Nsp1 D75E	Nsp3 Syn	Nsp3 P153L	Nsp12 P322L	Nsp13 Y541C	Nsp14 Syn	Spike D614G

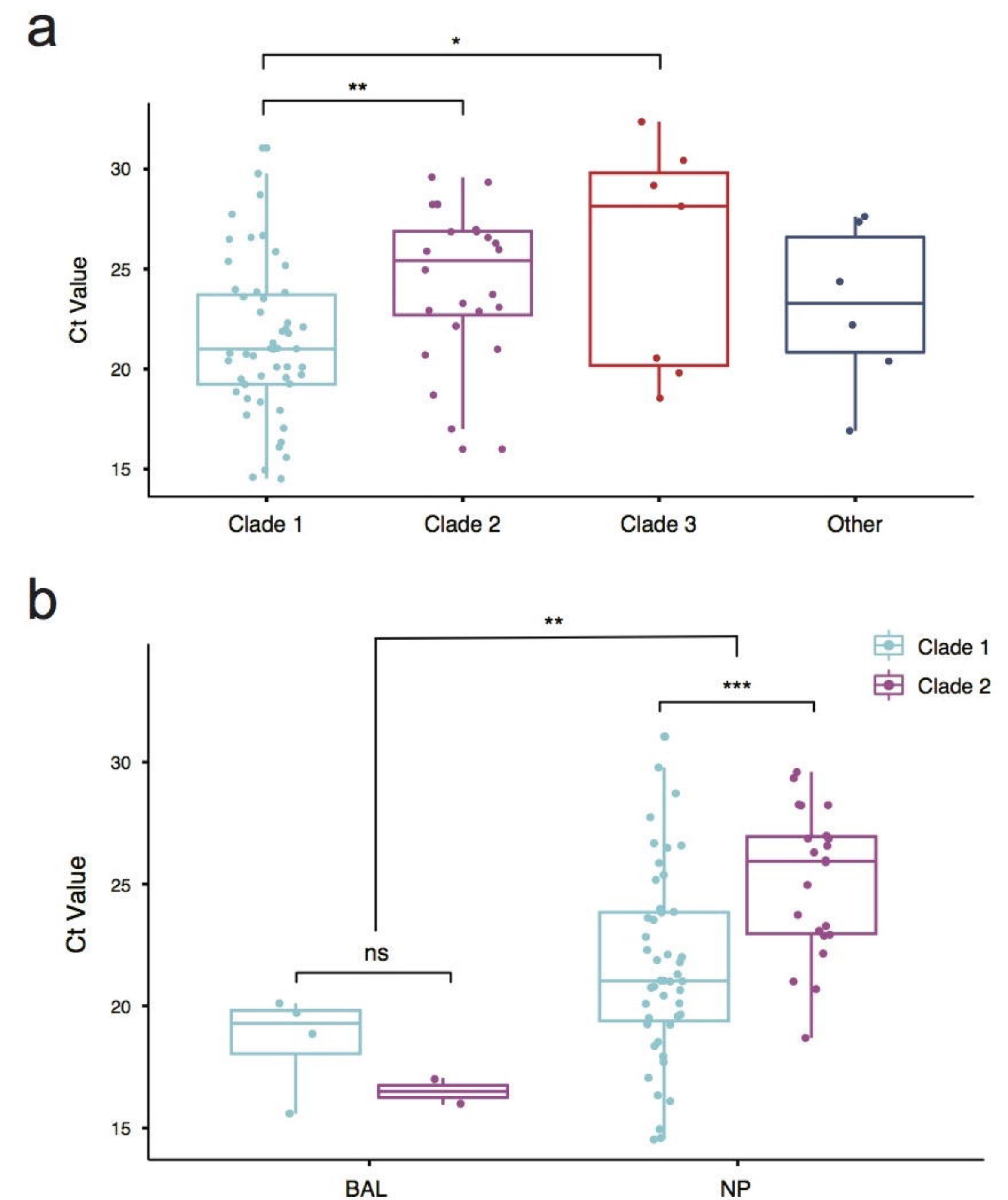

a
Clade 3
Clade 2
Clade 1
State
AK
AZ
CA
CT
FL
GA
IL
IN
MA
MN
NH
NJ
NU
NY
OH
OR
RI
TX
UT
VA
WA
WI
WY
Wuhan
0.06
b
Clade 3
Clade 2
Clade 1
c
Effective Population Size (Ne)
1000
100
10
1
Jan. 10
Jan. 28
Feb. 15
Mar. 5
Mar. 23
Date
Clade 1
Clade 3
Clade 2

a


Clade 2
Australia
Beijing
CA
CT
GA
Hong
IL
NY
NewZealand
Shanghai
Sichuan
Taiwan
Vietnam
WA
WI
Wuhan
0.03
Australia
Illinois
Sichuan
Wuhan
Dec. 13, 2019
Jan. 1, 2020
Jan. 19, 2020
Feb. 6, 2020
Feb. 24, 2020
Mar. 14, 2020
Approximate Date

b


WA
WI
NY
CT
CA
IL
GA
Beijing
Sichuan
Shanghai
Wuhan
Taiwan
Hong Kong
Vietnam
New Zealand
Australia


a
*
**
Ct Value
30
25
20
15
Clade 1
Clade 2
Clade 3
Other
b
**
***
ns
Clade 1
Clade 2
Ct Value
30
25
20
15
BAL
NP