Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

A clade of SARS-CoV-2 viruses 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

ARTICLE INFO

Article History:
Received 21 July 2020
Revised 20 October 2020
Accepted 21 October 2020
Available online xxx

Keywords:
SARS-CoV-2
COVID-19
Phylogenetics
Viral load
Viral genotype
Whole genome sequencing

ABSTRACT

Background: 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 aimed to investigate the phylogenetic characteristics of SARS-CoV-2 infections in Chicago, Illinois, and assess their relationship to clinical parameters.

Methods: We performed whole-genome sequencing of SARS-CoV-2 isolates collected from COVID-19 patients in Chicago in mid-March, 2020. Using these and other publicly available sequences, we performed phylogenetic, phylogeographic, and phylodynamic analyses. Patient data was assessed for correlations between demographic or clinical characteristics and virologic features.

Findings: The 88 SARS-CoV-2 genome sequences in our study separated into three distinct phylogenetic clades. Clades 1 and 3 were most closely related to viral sequences from New York and Washington state, respectively, with relatively broad distributions across the US. Clade 2 was primarily found in the Chicago area with limited distribution elsewhere. At the time of diagnosis, patients infected with Clade 1 viruses had significantly higher average viral loads in their upper airways relative to patients infected with Clade 2 viruses, independent of disease severity.

Interpretation: These results show that multiple variants of SARS-CoV-2 were circulating in the Chicago area in mid-March 2020 that differed in their relative viral loads in patient upper airways. These data suggest that differences in virus genotype can impact viral load and may influence viral spread.

Funding: Dixon Family Translational Research Award, Northwestern University Clinical and Translational Sciences Institute (NUCATS), National Institute of Allergy and Infectious Diseases (NIAID), Lurie Comprehensive Cancer Center, Northwestern University Emerging and Re-emerging Pathogens Program.

1. Introduction

In late 2019, the first clusters of people with severe respiratory infections and pneumonia were reported in the Hubei province of China [1–3]. The novel betacoronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has since emerged as the cause of a worldwide healthcare crisis with more than 27 million confirmed cases worldwide and more than 900,000 attributable deaths during the first eight months of 2020 [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 (US), the first confirmed case of COVID-19 occurred on January 15th in Washington state (WA) with a second confirmed case reported shortly thereafter in Chicago, Illinois on January 23rd. Both cases were linked with recent travel to China [6]. The Chicago patient later passed the virus to her husband in the first reported case of local transmission on January 30th [6]. By early March, Washington state was reporting sustained local transmission and developed into an early hotspot in the US with a peak of a little of over 500 confirmed cases per day in late March. Illinois likewise experienced a slow increase in case counts throughout March with case counts over 1000 per day by early April. Despite not reporting

[*] Corresponding authors.
E-mail addresses: judd.hultquist@northwestern.edu (J.F. Hultquist), e-ozer@northwestern.edu (E.A. Ozer).
[#] These authors contributed equally to this work.
[+] These authors contributed equally to this work.

https://doi.org/10.1016/j.ebiom.2020.103112

Research in context

Evidence before this study

The first case of coronavirus 2019 (COVID-19) caused by the novel coronavirus SARS-CoV-2 in the United States was reported in Washington state in January 2020. The disease has since spread to all 50 states, though the origins of the virus in different regions of the country are not always clear. Phylogenetic evidence of an Asian origin of the predominant virus in Washington state and a European origin of the predominant virus in New York state have been reported in the published and pre-print literature, but there has been little exploration of the origins of the virus in Chicago, Illinois, the third largest city in the US. We searched PubMed from database inception to September 10, 2020, for article titles and abstracts using the search terms “COVID-19” or “SARS-CoV-2” or “2019-nCoV” and “Illinois”. Our search yielded a few reports of specific transmission events, including that of the first known person-to-person transmission of SARS-CoV-2 in the US, but no broader analysis of SARS-CoV-2 origin or diversity in the Midwest. Recently, a few reports have shown correlations between viral genotype and viral load, with particular focus on the Spike mutation D614G. A search of the PubMed database for article titles and abstracts with the search terms “SARS-CoV-2”, “clade”, “mutation”, “variant”, “D614G”, “viral load”, or “cycle threshold” found a small number of published and pre-review reports linking the D614G mutation specifically to viral load in a clinical setting.

Added value of this study

This study is the first to investigate the population structure of SARS-CoV-2 viruses infecting patients in Chicago. This study demonstrates the presence of three distinct clades of SARS-CoV-2 viruses in the Chicago area: one that reflects the predominant clade on the East coast, one that reflects the predominant clade on the West coast, and one that is predominantly found in Illinois with close similarity to viruses found in China. Furthermore, this study supports a correlation between SARS-CoV-2 clade and relative viral load in the upper airways of infected patients.

Implication of all the available evidence

These results provide a snapshot of the population structure of SARS-CoV-2 viruses in the Chicago area relative to other regions in the US in March 2020. The finding that average viral loads in patient upper airways varied based on SARS-CoV-2 genotype may have implications for the design of future high specificity diagnostics and for future studies of virus transmissibility and pathogenicity.

its first confirmed case until March 1st, New York rapidly eclipsed both Washington and Illinois, reporting nearly 10,000 cases per day by the end of the month. By the end of March, the US 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 multiple states throughout March in an attempt to mitigate virus spread. In Illinois, New York, and Washington, stay-at-home orders began 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)] (Supplemental Fig. 1) [7,8]. 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 unclear.

In parallel with the unprecedented speed and scale of the pandemic, the sharing and integration of clinical and epidemiological data has also occurred at an exceptional pace, with tens of thousands of viral genome sequences of SARS-CoV-2 from across the globe available via online databases such as GISAID [10]. Whole genome sequencing (WGS) of SARS-CoV-2 viruses and subsequent phylogenetic analyses are most consistent with multiple independent introductions of the virus to the US [11]. While the predominant virus genotype in Washington state was most closely related to viruses in China in the early epidemic, the predominant virus genotype in New York was more closely related to viruses from Europe [12] .The latter viruses carry a mutation in their Spike protein (D614G) that has not only become more common worldwide over time, but that has also been associated with higher viral loads in COVID-19 patients. lsqb;13,14].

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. We further assessed these viral genomic data for correlations with clinical and demographic features of our cohort.

2. Methods

2.1. Specimen collection and testing

Individuals suspected of COVID-19 within two hospitals of the Northwestern Medicine healthcare system were screened for SARS-CoV-2 infection as per the CDC protocol [15]. 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 (VTM, ULab Scientific, cat. no. 2118-0413-99). Viral RNA was extracted from clinical specimens utilizing the QIAamp Viral RNA Minikit (Qiagen, cat. no. 52906). 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 in SARS-CoV-2 and RP probes for sample quality control as previously described (IDT, cat. no. 10006713) [15]. All specimens that failed to amplify the RP housekeeping gene were excluded from this study. All specimens with an N1 probe cycle threshold (Ct) less than or equal to 35 were considered positive and included in this study. RT-PCR was repeated in a random selection of specimens to validate Ct values obtained by the clinical diagnostic laboratory. Ct values from the N1 probes were used in all subsequent analyses.

2.2. cDNA synthesis and viral genome amplification

cDNA synthesis was performed with SuperScript IV First Strand Synthesis Kit (ThermoFisher, cat. no. 18091050) 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 in multiplexed PCR reactions to generate

~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 (version 3 release) [16]. PCR amplification was carried out using Q5 Hot Start HF Taq Polymerase (NEB, cat. no. M0493L) 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 s, then 35 cycles of 98 °C for 15 s followed by five minutes at 65 °C. Separate reactions were carried out for each primer pool and validated by agarose gel electrophoresis alongside negative controls. Each reaction set included positive and negative amplification controls and was performed in a space physically separated for pre- and post-PCR processing steps to reduce contamination.

2.3. 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” [16-18]. Amplicons from both primer pools were combined and purified with a 1x volume of AmpureXP beads (Beckman Coulter, cat. no. A63881). A total of 50 ng of DNA was treated with NEB Ultra II End Prep Enzyme mix (NEB, cat. no. E7546L). Up to 24 specimen libraries were barcoded using Nanopore Native Barcoding Expansion kits (Oxford Nanopore, cat. no. EXP-NBD-104 and EXP-NBD114) and NEBNext Ultra II Ligase (NEB, cat. no. E6056L) 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 (Oxford Nanopore, cat. no. SQK-LSK109) according to manufacturer specifications. Libraries were sequenced on the Nanopore MinION device using FLO-MIN106D Type R9.4.1 flow cells.

2.4. 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 with the DNA R9.4.1 450bps High Accuracy configuration, default quality score filtering (−qscore_filtering), and exclusion of reads without matching barcodes on ends (−require_barcodes_both_ends) to reduce the risk of potential barcode cross-contamination during library prep. Read quality filtering, reference alignment, and consensus sequence generation was performed using ARTIC Network software and protocols [18,19]. 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 [20]. 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 [21]. Regions with less than 10-fold read coverage were masked. We chose to use a lower read coverage cutoff than the default of 20 in the ARTIC Network software to allow for consensus sequence calls in regions with relatively poor amplification due to inefficient primer binding. Manual validation of variant calls in regions with 10−19x coverage was performed using Tablet software [22]. Targeted repeat sequencing of 11% of the final dataset (n = 10 specimens) was conducted with no differential variants detected in any sample. Final sequences from each flow cell were checked for batch effects as a final quality control check. Isolate lineages were determined with PANGOLIN using the 2020-05-19 definitions and with Nextstrain using the assign_clades.py program [23−25]. Consensus genome sequences were deposited in the GISAID database with accession numbers provided in Supplemental Table 1. All publicly-available SARS-CoV-2 genome sequences were downloaded from the GISAID database on April 10th, 2020 and included genome sequences from isolates collected as late as April 4th, 2020 [10]. Only full genome sequences in the database with <1% Ns were selected for inclusion in the analyses.

2.5. Phylogenetic analysis and phylogeographic inference

Genome sequences were aligned using MAFFT v7.453 software and manually edited using MEGA v6.06 [26,27]. The US analyses included 899 SARS-CoV-2 sequences from the US deposited in GISAID as of April 4, 2020 and the global analysis included 3099 SARS-CoV-2 sequences deposited in GISAID from 64 different countries up to April 4, 2020. All Maximum Likelihood (ML) phylogenies were inferred with IQ-Tree v 2.0.5 using its ModelFinder function before each analysis to estimate the nucleotide substitution model best-fitted for each dataset by means of Bayesian information criterion (BIC) [28,29]. We assessed the tree topology for each phylogeny both with the Shimodaira−Hasegawa approximate likelihood-ratio test (SH-aLRT) and with ultrafast bootstrap (UFboot) with 1000 replicates each [30,31]. TreeTime v0.7.4 was used for the assessment of root-to-tip correlation, the 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 [32]. TreeTime was run using an autocorrelated molecular clock under a skyline coalescent tree prior. We used the sampling dates of the sequences to estimate the evolutionary rates and determine the best rooting of the tree using root-to-tip regression with least-squares method. The substitution rate estimated by TreeTime was 1.2×10^{-3} for the USA tree and 8.6×10^{-4} for the global tree, consistent with other studies [12,33]. We used the population proportion per location in each of the datasets as weights in the analysis to correct for strong sampling biases in the datasets. We performed additional temporal phylogenies to confirm the clades observed in Chicago by downsampling the most overrepresented region of the USA (Washington state) using the R package FastaUtils and running the same pipeline described above.

Bayesian time-scaled phylogenetic analyses were only performed for the phylogenies of each clade separately due to the high complexity of the full US and global phylogenies. We used BEAST v2.5.2 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 [34]. BEAST priors were introduced with BEAUTI v2.5.2 including an uncorrelated relaxed molecular clock model with a lognormal distribution of the evolutionary rate, the TreeTime estimate (1.2×10^{-3}) as the prior for the mean, and a standard deviation of 0.1 after optimization with preliminary runs. We assumed a GTR substitution model with invariant sites, as the best-fitted model obtained with ModelFinder, and a Coalescence Bayesian Skyline to model the population size changes through time. The posterior evolutionary rate was estimated to be 1.5×10^{-3} for Clade 1, 1.2×10^{-3} for Clade 2, and 1.1×10^{-3} for Clade 3. All the analyses indicated support for a relaxed clock as the average rate coefficient of variation was 0.78 for Clade 1, 0.25 for Clade 2, and 0.57 for Clade 3, which indicates a degree of rate autocorrelation among adjacent branches in the tree. Additionally, Bayesian phylogenetic reconstruction was performed with the global sequences most closely related to Clade 2. For this analysis we used both the previously estimated global (8.6×10^{-4}) and Clade 2 (1.2×10^{-3}) evolutionary rates as priors. Both analyses resulted in similar tree topologies and same geographical location of origin for Clade 2. Markov chain Monte Carlo (MCMC) runs of at least 100 million states with sampling every 5,000 steps were computed [35]. 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 [36], 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 and SpreaD3 v0.9.6 was used to visualize the phylogeographic patterns associated with the phylogenies [37,38]. 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.

2.6. Statistical analysis

All statistical analyses were performed in R version 4.0.0. The R package *lme4* was used for the linear and logistic regressions, *fitdistrplus* was used to test the best fitted model for the data, *lsmeans* was used to test for all possible contrasts and perform multiple comparison correction, *nnet* was used for the multinomial logistic regression, *finalfit* package was used to generate the descriptive table, and *ggplot2* and *ggpubr* were used to generate the statistical figures.

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, controlling for possible confounders including the date of sample collection, age, sex, race, maximal severity of illness, the specimen collection source, and immunosuppression status (defined by existence of an immunosuppressing condition such as stem cell or organ transplant or malignancy and/or concurrent treatment with immunosuppressing medications such as steroids). 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.

We also tested for differences in viral load by specimen source (nasopharyngeal swab or bronchoalveolar lavage) by fitting a linear model that included the same confounders as above, extended to contain the interaction between clade and specimen source. For this analysis we only used data from patients with viruses from the two most abundant clades, Clades 1 and 2 (n = 76 patients total). Within this model, we tested for differences between clades by source and corrected for multiple comparisons as before.

To determine if the difference in Ct values between clades was driven by differences in time since symptom onset or exposure, the number of days between symptom onset or exposure and specimen collection was determined by chart review. We tested for differences between Clade 1 and Clade 2 in the number of days between symptom onset and specimen collection (n = 76) or possible virus exposure (n = 29 patients) using a Wilcoxon Signed-Rank Test.

Finally, 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 1 and 2 and all available clinical and demographic variables, we fitted a logistic regression model including age, sex, race, severity of illness, immunosuppression status, time since symptoms onset, and specimen source to test for any possible association between these variables and the clade observed.

2.7. Ethics statement

This study was reviewed and approved by the Institutional Review Board of Northwestern University (STU00212260, STU00212267, and STU00206850). All specimens were retrospectively collected and analyzed; informed consent and HIPAA authorization were waived for all participants.

2.8. Role of the funding sources

The funding sources had no role in the design of the study, in the collection/analysis/ interpretation of the data, in the writing of the report, or in the decision to publish. The corresponding authors had access to all of the data in the study and had final responsibility for the decision to submit this work for publication.

3. Results

3.1. Collection characteristics and sequencing results

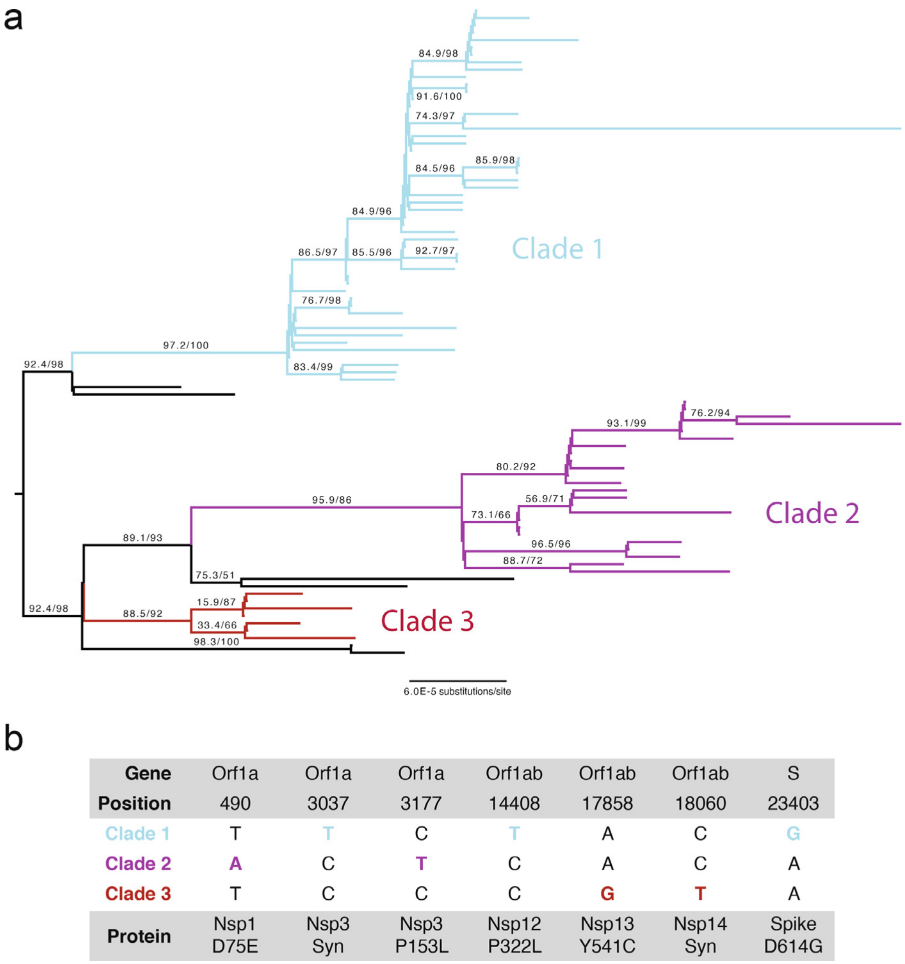
Between March 14 and March 21, 2020, a total of 127 qRT-PCR clinical tests performed at NMH were positive for SARS-CoV-2 (Supplemental Fig. 2A). Residual RNA from these diagnostic tests was collected for WGS. Of these, 39 specimens failed to achieve sufficient amplification required for sequencing, failed to yield sufficient purity after barcoding, had inadequate read coverage after sequencing, or were found to be repeat samples from the same patient; these samples were excluded from further analysis. Almost half of the 88 successfully sequenced specimens had been collected from patients evaluated in emergency departments at NMH or LFH, 25% from non-ICU hospitalized patients at NMH, and 12% from patients in the NMH ICU (Supplemental Figure 2B). The qRT-PCR cycle threshold (Ct) values of these specimens ranged from 14.52 to 32.37, consistent with similar reports [13]. After applying alignment quality filters and normalizing segment coverage to a maximum of 200 reads in each direction, the average median genome coverage across non-overlapping amplicon sequences was 348 reads per specimen (min 56, max 395). Library preparation and sequencing was repeated on a random selection of 10 isolates with no differences in variant calls detected.

3.2. Three main clades are detected through phylogenetic inference

We first performed a maximum likelihood (ML) phylogenetic analysis of the 88 SARS-CoV-2 specimen genomes in this study. This analysis showed that most of the sequences (93%) clustered in three main clades (support: >85% aLRT and >85% UFboot) (Fig. 1A). For the purposes of this report we will refer to these clades, ordered by relative abundance in the sequence set, as Clades 1 (n=51, 57.9%), 2 (n=24, 27.3%), and 3 (n=7, 7.9%). Using the Pangolin and Nextstrain methods of lineage classification [23,24], isolates in Clade 1 were assigned mainly to Pangolin lineage B.1 (92% assigned to lineage B.1; 2% to B.1.1; and 6% to B.1.5) and Nextstrain clade 20C (61% to clade 20C, 37% to clade 20A, and 2% to clade 20B), the majority of isolates in Clade 2 belonged to Pangolin lineage A.3 (92% to A.3; 8% to A) and Nextstrain clade 19B (92% to clade 19B and 8% to clade 19A), and isolates in clade 3 belonged mainly to lineage A.1 (86% to A.1; 14% to A) and all belonged to Nextstrain clade 19B (Supplemental Table 1). 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.

3.3. Clade 2 viruses are prominent in the Chicago area

To place the three observed clades in a broader geographic context, we expanded our analysis to include all 899 SARS-CoV-2 sequences from the US deposited in GISAID as of April 4, 2020. With the complete US 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 (Fig. 1B) and possible geographical transitions. These analyses confirmed our previous clustering, with a bulk of the Chicago



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

Fig. 1. Phylogenetic Analysis of SARS-CoV-2 Isolates in Chicago. **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. Black branches represent sequenced isolates that do not belong to any of the three major clades. Midpoint rooting was used for representation purposes. **b)** Clade-defining mutations at the US level. Positions are numbered according to the reference genome (NC_045512.2).

sequences falling into three distinct clades (Fig. 2A). Overall, the local epidemic in the Chicago area was broadly reflective of the epidemic in the US as a whole. Clade 1 viruses clustered predominantly with isolates centered around New York, while Clade 3 viruses clustered with those circulating on the West Coast, particularly in Washington state. These two clades showed a broader spread across the US than Clade 2, which was comprised almost exclusively of our newly sequenced specimens in Illinois with only a few representatives from elsewhere in the US (Fig. 2A, Supplemental Figure 3). Clade 2 also contained a few previously sequenced SARS-CoV-2 specimens identified in Illinois at the end of January 2020 [6]. Because phylogeographic analyses can be strongly biased by the geographical sampling, an under-sampling of the Midwest compared to other possibly over-sampled US regions could bias the analyses showing limited spread of Clade 2. Therefore, we calculated the number of genome sequences used per number of positive SARS-CoV-2 tests by April 4, 2020 for each US region. This analysis showed that, as a function of number of SARS-CoV2 positive tests, the Midwest was the second-most sampled region in the US by early April (Supplemental Figure 4A). The Northwest region was the most sampled where Washington state was most highly-represented. Nevertheless, to exclude the possibility that the clustering of Clade 2 was induced by the over-sampling of Washington state where the other less widely-spread Clade 3 was mainly centered, we inferred additional phylogenetic trees after randomly downsampling Washington state sequences to include only 50 from the 403 sequences present in the initial dataset. This analysis further confirmed the distinct clustering of Clade 2 sequences (Supplemental Figure 4B).

We next studied the phylodynamic and phylogenomic characteristics of these three clades using a Bayesian approach to infer possible differences in their population dynamics. These analyses suggested 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 Fig. 3B). Illinois was found to be the most probable origin of Clade 2 in the US with an estimated most recent common ancestor (TMRCA) around January 18, 2020 [95% Highest Posterior Density (HPD) interval: January 9 - January 21, 2020]. The timing, location, and the sequences close to the inferred root of this tree (USA_IL_1 and USA_IL_2) coincide with the first reported COVID-19 cases in Illinois [6], though multiple or alternate introductions cannot be ruled out. These results suggest that Clade 2 may have been introduced to the Chicago area early in the US epidemic and subsequently had limited spread beyond Illinois (Fig. 2B, middle panel).

A similar analysis of Clade 1 suggested multiple introductions, a high degree of diversity, and the existence of several probable subclades. Although the highest proportion of branches were found to reside in New York, this proportion was relatively low at 0.34 [95% CI: 0.31–0.37], in agreement with the broad geographic spread of this clade (Supplemental Fig. 3A). These findings are most consistent with Clade 1 being centered in New York but rapidly expanding throughout the region and to many other locations across the US (Fig. 2B, bottom panel). The TMRCA of this clade in the US was

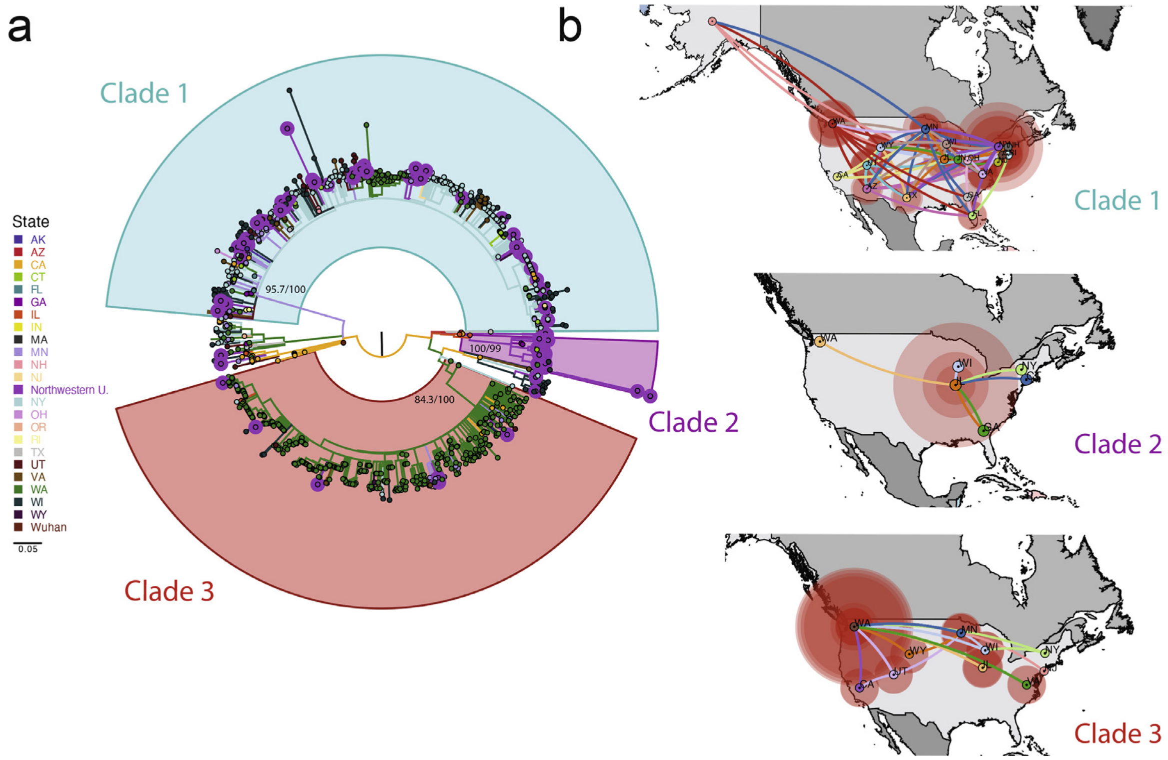


Fig. 2. Phylogenetic and Phylogeographic Analysis of Chicago Isolates compared to the US Epidemic. **a)** ML phylogenetic reconstruction of full genome sequences from the United States. We included sequences from Northwestern and all sequences from the US available in GISAID as of April 4, 2020. Branches 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; approximate likelihood-ratio test (aLRT) and bootstrap values of the main clade branches are indicated proximal to the branch points. **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.

calculated to be February 10, 2020 [95% HPD interval: January 28 - February 20, 2020], an estimated date in close alignment with prior reports [12]. These results are suggestive of this clade having had multiple introductions into the Chicago area with some occurring early during the epidemic in February and March of 2020.

Finally, US isolates in Clade 3 are 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 3C). This clade was also estimated to have been introduced into the US early in the epidemic (TMRCA = January 16, 2020 [95% HPD interval January 11 - January 17, 2020]), though we cannot rule out the possibility that this clade was independently introduced on more than one occasion [11]. These data are most consistent with subsequent expansion of Clade 3 from Washington state to multiple locations throughout the US, including Illinois, followed by limited spread elsewhere (Fig. 2B, top panel).

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 New York-centered Clade 1 and the Washington-centered Clade 3.

3.4. Clade 2 has limited global distribution

To further investigate the origin and population dynamics of Clade 2, we performed a phylogenetic analysis at the global level. In addition to our previous set of US and Chicago sequences described above, we included all 3099 SARS-CoV-2 sequences deposited in GISAID from outside the US up through April 4, 2020. We conducted ML analyses similar to those performed on the US sequences described above (Supplemental Figure 4). These results further confirmed strong statistical support for Clade 2 (91.6% aLRT and 92% UFboot). Again, the detected spread of this clade has been exceptionally limited, with a majority of sequences derived from this study, China, or Australia. Ancestral reconstruction of these global sequences showed low divergence of these samples from those first sequenced in Wuhan, China (Fig. 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 to be China at our current state of global sampling (Fig. 3B). This analysis was performed using the estimated global (8.6×10^{-4}) and Clade 2 (1.2×10^{-3}) evolutionary rates as priors to examine any influence of this parameter on the observed phylogeography. Both inferences indicated a strong support for China as the most probable geographical location of origin of Clade 2 with a posterior probability of 0.9912 when using the global evolutionary rate and 0.9958 when using the estimated Clade 2 evolutionary rate as priors. Based on this phylodynamic analysis, Clade 2-like sequences from other areas outside the US 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 those in Illinois than China (Fig. 3B). Together, these findings are most consistent with an introduction of Clade 2 into Illinois from China as early as January 2020. Additionally, global analysis with the other two clades (Supplemental Figure 5) suggest that the New York-centered Clade 1 was likely to have been introduced into the US from Europe while the Washington-centered Clade 3 was most likely introduced directly from China, consistent with other reports.[11]

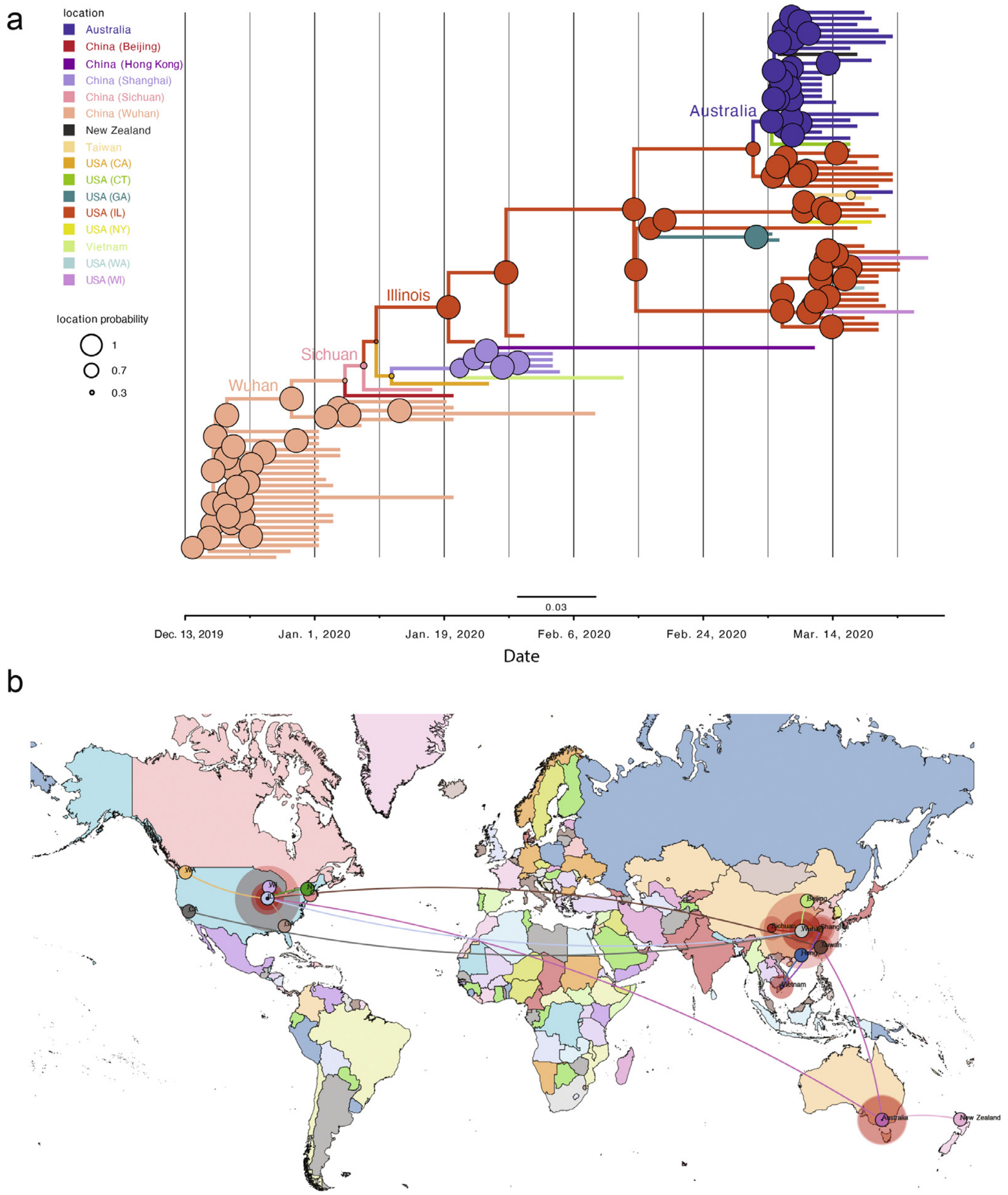


Fig. 3. Phylogenetic and Phylogeographic Analysis of Chicago Isolates compared to the Global Pandemic. **a)** Phylodynamic tree of US and global Clade 2 related genome sequences. We used global sequences phylogenetically related to Clade 2 available in the GISAID database and performed the analysis encompassing 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. 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.

Together, these analyses suggest distinct global phylodynamics of the three main clades observed in the Chicago area: Clade 1 was likely introduced by way of New York through Europe, Clade 2 may have been introduced directly from China, and Clade 3 was likely 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, as of early April 2020 Clade 2 was almost exclusively found in the Chicago area.

3.5. 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 US and worldwide could suggest phenotypic variability that may impact transmission rate. To further investigate this, we compared the Ct values for each sample, which serves as a proxy for viral load in the patients' airways at the time of diagnosis. We 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 fit a linear model to test for differences in Ct values between the clades, controlling for the date of isolation, hospital location, age, sex, race, severity of illness, immunosuppression status, and specimen collection source (Supplemental Table 2). We detected significantly higher Ct values corresponding to lower viral loads in Clade 2 (Chicago-centered; mean=21.4 sd=3.8) compared to Clade 1 (New York-centered; mean= 24.4 sd=3.75) with a corrected p-value=0.008, and Clade 3 (Washington-centered; mean= 25.6 sd=5.73) with a corrected *p*-value=0.028) (Figure 4A). Due to the limited sample sizes for Clade 3

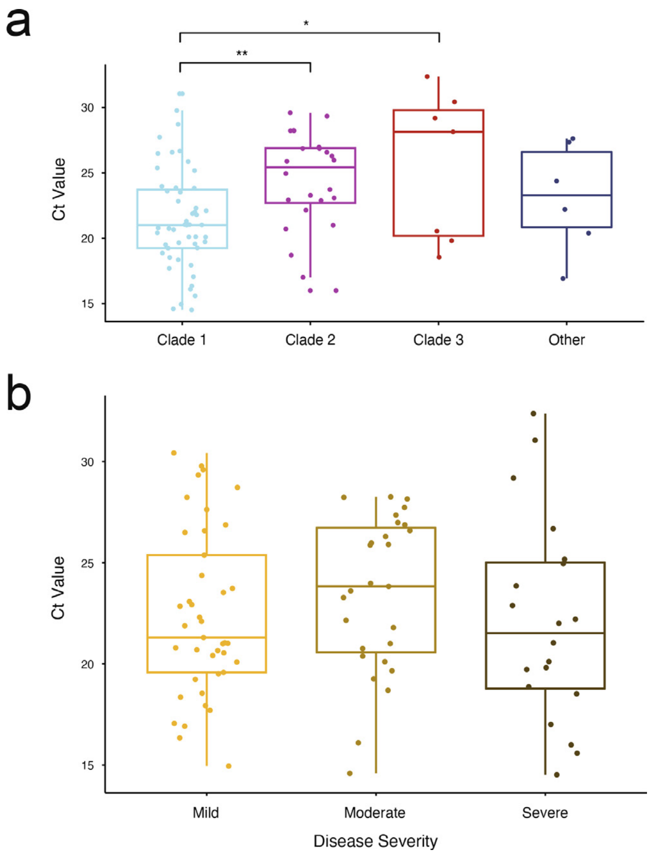


Fig. 4. Associations between Viral Clade and Ct Value and Disease Severity. **a)** PCR Cycle threshold (Ct) values of patient samples grouped by major Clade assignment. **b)** Specimen Ct values by maximum disease severity. Mild (blue) = no hospital admission; Moderate (red) = hospital admission, but no ICU stay; Severe (green) = ICU admission. In both panels, horizontal lines in each box represent the median value and the lower and upper error bars are the interquartile ranges. Significance is indicated for the comparisons performed within each fitted model (* = q-value<0.05; ** = q-value <0.01).

(n = 7) and ungrouped specimens (n = 6), our subsequent analyses focus solely on Clades 1 (n = 51) and 2 (n = 24).

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 1 and 2 (Supplemental Fig. 6A). We tested for differences between clades, within each sample source, accounting for the potential interaction between clade and source in the linear model. As BAL specimens were only collected from patients in the ICU and illness severity in our study was defined by ICU admission, we could not control for severity in our model. When comparing within NP samples, the Ct values of Clade 1 viruses were still significantly lower than Clade 2 viruses, as observed previously (corrected p-value = 0.0003). However, no significant difference in Ct value was detected by clade in the BAL samples. While the numbers 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.

It has been reported that viral loads in the upper airways generally decrease over the course of infection [39,40]. To determine whether the difference in average Ct values between samples from patients infected with Clade 1 or Clade 2 viruses corresponded with the time of specimen collection, the number of days between reported symptom onset and specimen collection was determined for each patient from available case files. No significant difference in time from symptom onset to specimen collection between patients infected with Clade 1 or Clade 2 viruses was observed (Supplemental Fig. 6B). Additionally, the date of possible virus exposure was able to be determined for 29 patients, and again no significant difference was found between clades (Supplemental Figure 6C). Consistent with prior reports, we did find that later dates of specimen collection after symptom onset significantly correlated with higher Ct values among patients infected with Clade 1 viruses, though a similar correlation was not seen among Clade 2 isolates (Supplemental Figure 6D). One potential explanation for why the observed Ct values did not seem to increase over time for Clade 2 viruses is that with the overall higher average Ct values among members of the Clade, there was reduced ability to detect further increases in Ct value before the specimens become undetectable and/or unable to be sequenced. Further studies containing more patients with Clade 2 isolates will be needed to better assess the viral dynamics. These results suggest that the average viral load in the upper airways of patients infected with Clade 2 viruses is lower relative to those infected with Clade 1 viruses at the time of diagnosis independent of time since symptom onset.

While reports linking disease severity and SARS-CoV-2 viral load have been mixed [41–45], we next looked for associations between viral Clades 1 and 2 and maximal disease severity. Disease severity was binned as mild (outpatient or ED only), moderate (inpatient hospitalization), or severe (ICU admission and/or death) and set as the outcome variable. We then fit a logistic regression model including all available demographic and clinical variables. Interestingly, we found no significant association between any available variable and disease severity in our dataset, including clade (Fig. 4B). Finally, we fit a logistic regression model including all available demographic and clinical variables with clade as the outcome variable (Supplemental Table 3). 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 clade.

4. Discussion

In this study, we report the phylogenetic and phylodynamic analyses of 88 SARS-CoV-2 genomes from COVID-19 patients in Chicago, Illinois, US, which largely fall into three major clades. Two of these three clades are broadly representative of the major circulating clades that expanded and disseminated rapidly across the US through early April 2020 with imputed points of origin 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 US and predominantly found in the Chicago area based on current sampling. Viruses in this clade lack the rapid and wide dispersion evidenced with the other viral clades. Notably, patients infected with Clade 1 viruses presented with significantly higher viral loads in the upper airways at the time of diagnosis compared to patients infected with Clade 2 viruses. These data suggest that differences in SARS-CoV-2 genotype may impact viral load; additional research will be required to determine if this can in turn influence transmission and viral spread.

Chicago, Illinois is the third-largest city in the US 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 US, relatively little SARS-CoV-2 sequencing data has been reported 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 likely the result of geographically and temporally separate introductions into the US from China. Our analysis suggests that Clade 2 viruses were introduced into the Chicago area from China and were being transmitted in the community as early as mid-January 2020, although multiple introductions cannot be ruled out. While Clade 3 viruses are 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 report describing the European origin of SARS-CoV-2 in New York City [12]. These data do not rule out the possibility of multiple introductions of each clade, which may have spread through asymptomatic transmission prior to detection [11,46].

Both the US 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 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 collective, 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. However, the proportion of viruses sampled from Chicago in this study that were Clade 2 (24 of 88, 27.3%) was far greater than the proportion in the database sampled across the US (10 of 899, 1.1%) or from non-US countries (34 of 3099, 1.1%), suggesting that Clade 2 was more abundant in the Chicago area relative to other sampled locations as of late March 2020. As more genomes both from Illinois and elsewhere are sequenced, we will continue to assess for changes in the population structure and distribution of SARS-CoV-2.

Notably, we found that Clade 2 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. Clade 1 viruses are defined by 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 [13,47]. Similarly, other reports have noted the rapid spread of the Nsp12 P322L mutation [48]. 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. Lower viral loads in the upper airways of infected patients may decrease viral transmissibility and impact the overall spread, but further laboratory and clinical studies will be required to test the impact of these variants on viral replication, disease severity, and transmissibility.

Other than the potential for lower viral loads, there are other 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. While these restrictions went into place concurrent with sample gathering for this study, they may have contributed to a slower spread of Clade 2 viruses outside 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 be found to have distinct clinical manifestations that lessen the opportunity for asymptomatic transmission to multiple other individuals. Although it could be postulated that the relative frequency of Clade 2 isolates among these samples is a reflection of their rapid expansion in the Chicago area prior to mid-March, such an explanation would seem to be in contrast with a lack of similar expansion in other regions in which Clade 2 isolates had previously been sampled in the GISAID data set. Ongoing studies of clinical and demographic patient data associated with continued specimen collection may provide support for these or other scenarios.

The other clinical and demographic data we examined, including disease severity, did not correlate significantly with clade or Ct value. Previous reports have similarly found that disease severity may not always correlate with viral load [43–45,49]. 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 in this specimen set. Finally, we did observe a correlation between Ct value and the time since symptom onset, confirming previous reports [39,40], though this did not change significantly by clade and did not explain the higher Ct values of Clade 2 viruses. Additional unmeasured confounders may play a role in these relationships, including antibody status and immune response, which may be incompletely accounted for in this limited specimen set. Further studies with more specimens and associated clinical data will be needed to 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 detailed clade dynamics and population structure. Additionally, these specimens were collected from one healthcare 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.

Contributors

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. All authors read and approved the final version of this manuscript.

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 Foundation (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.), a supplemental award from the Lurie Comprehensive Cancer Center at Northwestern University (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.).

Declaration of Competing Interests

The authors declare no competing financial interests.

Data sharing statement

Due to IRB restrictions and for protection of patient privacy individual participant data will not be made available. An aggregate summary of the patient data underlying the results of this study is presented in Supplemental Table 2. All SARS-CoV-2 genome sequences derived from participant specimens are available online at GISAID (https://www.gisaid.org/) with accession numbers reported in Supplemental Table 1.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2020.103112.

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, 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, et al. Early transmission dynamics in wuhan, China, of novel coronavirus–infected pneumonia. New Eng 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, 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, 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. https://covidtracking.com/data (accessed April 27, 2020).

[8] IDoP H. Coronavirus disease. https://dph.illinois.gov/covid19 (accessed April 27, 2020).

[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. gisaid.org (accessed 4/22/2020 2020).

[11] Worobey M, Pekar J, Larsen BB, et al. The emergence of SARS-CoV-2 in Europe and the US. bioRxiv 2020.

[12] Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 2020;369(6501):297–301.

[13] Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182(4):812–27 e19.

[14] Volz EM, Hill V, McCrone JT, et al. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. medRxiv 2020 2020.07.31.20166082.

[15] (CDC) CfDCaP. CDC 2019-novel coronavirus (2019-nCoV) real time RT-PCR diagnostic panel. 2020.

[16] Artic Network. https://artic.network/ncov-2019 (accessed March 23, 2020).

[17] Oxford Nanopore Technologies. https://community.nanoporetech.com (accessed April 27, 2020).

[18] Quick J, Grubaugh ND, Pullan ST, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 2017;12(6):1261–76.

[19] nCoV-2019 novel coronavirus bioinformatics protocol. https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html (accessed March 23, 2020).

[20] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–100.

[21] 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.

[22] 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.

[23] O'Toole Á, McCrone JT. pangolin (phylogenetic assignment of named global outbreak lineages). https://github.com/hCoV-2019/pangolin (accessed June 4, 2020).

[24] Rambaut A, Holmes EC, O'Toole A, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020.

[25] Nextstrain. ncov. https://github.com/nextstrain/ncov (accessed September 13, 2020).

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

[27] 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.

[28] Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 2020;37(5):1530–4.

[29] Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. Model-Finder: fast model selection for accurate phylogenetic estimates. Nat Methods 2017;14(6):587–9.

[30] 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.

[31] Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 2018;35(2):518–22.

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

[33] Boni MF, Lemey P, Jiang X, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol 2020.

[34] Bouckaert R, Heled J, Kuhnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2014;10(4):e1003537.

[35] 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.

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

[37] Hueck CJ. Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol Mol Biol Rev 1998;62(2):379–433.

[38] 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.

[39] M.S. H, Seong M-W, Kim N, et al. Viral RNA load in mildly symptomatic and asymptomatic children with COVID-19, Seoul. Emerging Infectious Diseases 2020;26(10).

[40] Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. New Eng J Med 2020;382(12):1177–9.

[41] 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.

[42] Zheng S, Fan J, Yu F, 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.

[43] To KK, Tsang OT, Leung WS, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. Lancet Infect Dis 2020;20(5):565–74.

[44] Lui G, Ling L, Lai CK, et al. Viral dynamics of SARS-CoV-2 across a spectrum of disease severity in COVID-19. J Infect 2020;81(2):318–56.

[45] Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. N Engl J Med 2020;382(22):2081–90.

[46] Bedford T, Greninger AL, Roychoudhury P, et al. Cryptic transmission of SARS-CoV-2 in Washington state. Science 2020 eabc0523.

[47] Bhattacharyya C, Das C, Ghosh A, 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.

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

[49] Joynt GM, Wu WK. Understanding COVID-19: what does viral RNA load really mean? Lancet Infect Dis 2020;20(6):635–6.