Viral genome sequencing places White House COVID-19 outbreak into phylogenetic context

Trevor Bedford[1,2,3], Jennifer K. Logue[4], Peter D. Han[2,3], Caitlin R. Wolf[4], Chris D. Frazar[3], Benjamin Pelle[3], Erica Ryke[3], Jover Lee[1], Mark J. Rieder[2,3], Deborah A. Nickerson[2,3], Christina M. Lockwood[5], Lea M. Starita[2,3], Helen Y. Chu[2,4], Jay Shendure[2,3,6]

[1]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA*
[2]*Brotman Baty Institute for Precision Medicine, Seattle, WA, USA*
[3]*Department of Genome Sciences, University of Washington, Seattle, WA, USA*
[4]*Department of Medicine, Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, USA*
[5]*Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA*
[6]*Howard Hughes Medical Institute, Seattle, WA, USA*

Abstract

In October 2020, an outbreak of at least 50 COVID-19 cases was reported surrounding individuals employed at or visiting the White House. Here, we applied genomic epidemiology to investigate the origins of this outbreak. We enrolled two individuals with exposures linked to the White House COVID-19 outbreak into an IRB-approved research study and sequenced their SARS-CoV-2 infections. We find these viral sequences are highly genetically similar to each other, but are distinct from over 160,000 publicly available SARS-CoV-2 genomes, possessing 5 nucleotide mutations that differentiate this lineage from all other circulating lineages sequenced to date. We estimate this lineage has a common ancestor in the USA in April or May 2020, but its whereabouts for the past 5 to 6 months are not clear. Looking forwards, sequencing of additional community SARS-CoV-2 infections collected in the USA prior to October 2020 may reveal linked infections and shed light on its geographic ancestry. In sequencing of SARS-CoV-2 infections collected after October 2020, the relative rarity of this constellation of mutations may make it possible to identify infections that likely descend from the White House COVID-19 outbreak.

Introduction

After its emergence in late 2019, COVID-19 spread throughout the world, resulting globally in over 42 million cases and over 1.1 million deaths as of October 25, 2020 [1]. In the United States alone, there have been over 8.9 million confirmed cases and over 225,000 deaths reported [2]. COVID-19 has been repeatedly associated with localized outbreaks surrounding social settings like weddings and bars [3] as well as workplaces [4], including so-called “superspreader” events. Social events and workplaces, among other exposure settings, are thought to be driving ongoing transmission in the United States [5]. In October 2020, an outbreak of COVID-19 was reported surrounding individuals employed at the White House,

individuals attending an event at the White House Rose Garden announcing Amy Coney Barrett's nomination to the Supreme Court and individuals attending other events between September 26 and September 30, 2020 [6]. As of October 30, 2020, there are at least 50 individuals who have publicly announced cases of COVID-19 linked to this outbreak [7]. The origins of the White House outbreak have been characterized as “unknowable” [8].

Due to long incubation times and a spectrum of disease severity, contact tracing of COVID-19 spread is challenging. However, genomic sequencing of the SARS-CoV-2 virus causing individuals’ infections offers an alternative avenue for investigating SARS-CoV-2 transmission and spread [9–11]. This technology enables genomic epidemiology where genetic relationships among sequenced samples are used to make inferences about how infections are epidemiologically related. Because SARS-CoV-2 mutates approximately once every two weeks along a transmission chain [9,12], it is possible to use patterns of shared mutations to group viruses and discover transmission relationships.

We applied genomic epidemiology to shed light on the origins of the White House outbreak. We enrolled two individuals with exposures linked to the White House COVID-19 outbreak into an IRB-approved research study, collected nasal swabs and sequenced the SARS-CoV-2 virus from these specimens. These two individuals reported no direct contact with each other in the days preceding their COVID-19 diagnoses and are independently linked to the White House COVID-19 outbreak. We refer to these infections as WH1 and WH2. We report here on the genetic relationships between WH1, WH2 and publicly available SARS-CoV-2 genomic data.

Results

Samples WH1 and WH2 were each obtained as an anterior nasal swab from individuals who had previously tested positive for SARS-CoV-2 shortly after attending event(s) associated with the White House COVID-19 outbreak. After confirmation of the positive test (Ct = 25 for Orf1b and S for WH1 and Ct = 25 for Orf1 and S for WH2), we performed genome sequencing on both samples either using hybrid capture (WH1 + WH2) or shotgun metagenomics (WH2) approaches to prepare the libraries. This sequencing resulted in 550X average depth-of-coverage of WH1, yielding a consensus genome of 29,857 resolved bases or 99.8% of reference. It possesses 14 mutations relative to the reference strain Wuhan/Hu-1/2019 (241T, 1059T, 1977G, 3037T, 7936T, 14250T, 14408T, 16260T, 18417C, 19524T, 20402T, 23403G, 25563T, 28821A). WH2 yielded a partial genome with 50X average coverage and 2643 resolved bases or 8.9% of reference.

Given WH2 is currently represented as a partial genome, we sought to characterize the genetic match between WH1 and WH2 at positions with sequencing coverage in WH2. Of the 14 sites that distinguish WH1 from reference, 5 of these sites had read coverage in WH2. At all 5 of these sites, reads from WH2 fully agreed with the mutation call in WH1. These were 3037T with 783 Ts out of 783 reads, 7936T with 78 Ts out 78 reads, 16260T with 2 Ts out of 2 reads, 20402T with 1 T in 1 read, and 28821A with 1 A in 1 read. In other words, no reads from WH2 supported the reference sequence at these 5 bases. Mutations 7936T and 20402T are rare in

circulating viruses appearing in 0.04% and 0.02% of sequenced genomes; mutations 16260T and 28821A are uncommon, appearing in 0.9% and 0.6% of sequenced genomes; and mutation 3037T is common, appearing in 83.2% of sequenced genomes. This allelic constellation strongly supports WH1 and WH2 as belonging to the same viral lineage. Of 160,291 publicly available viral sequences, only WH1 and WH2 possess all 5 of these mutations. Given confidence in base call for 7936T in particular, we would estimate the upper bound for chance of random collision at approximately $P = 0.0004$.

These mutations place WH1 and WH2 squarely within circulating genetic diversity in the United States (**Fig. 1**). WH1 and WH2 belong to Nextstrain clade 20C, which is the clade that predominates in the USA epidemic [13] and additionally belong to Pangolin lineage B.1.26 [14]. Pangolin lineage B.1.26 is demarcated by mutations C16260T and C28821A and contains viruses sampled from the USA, Canada and New Zealand (cov-lineages.org/lineages/lineage_B.1.26.html).

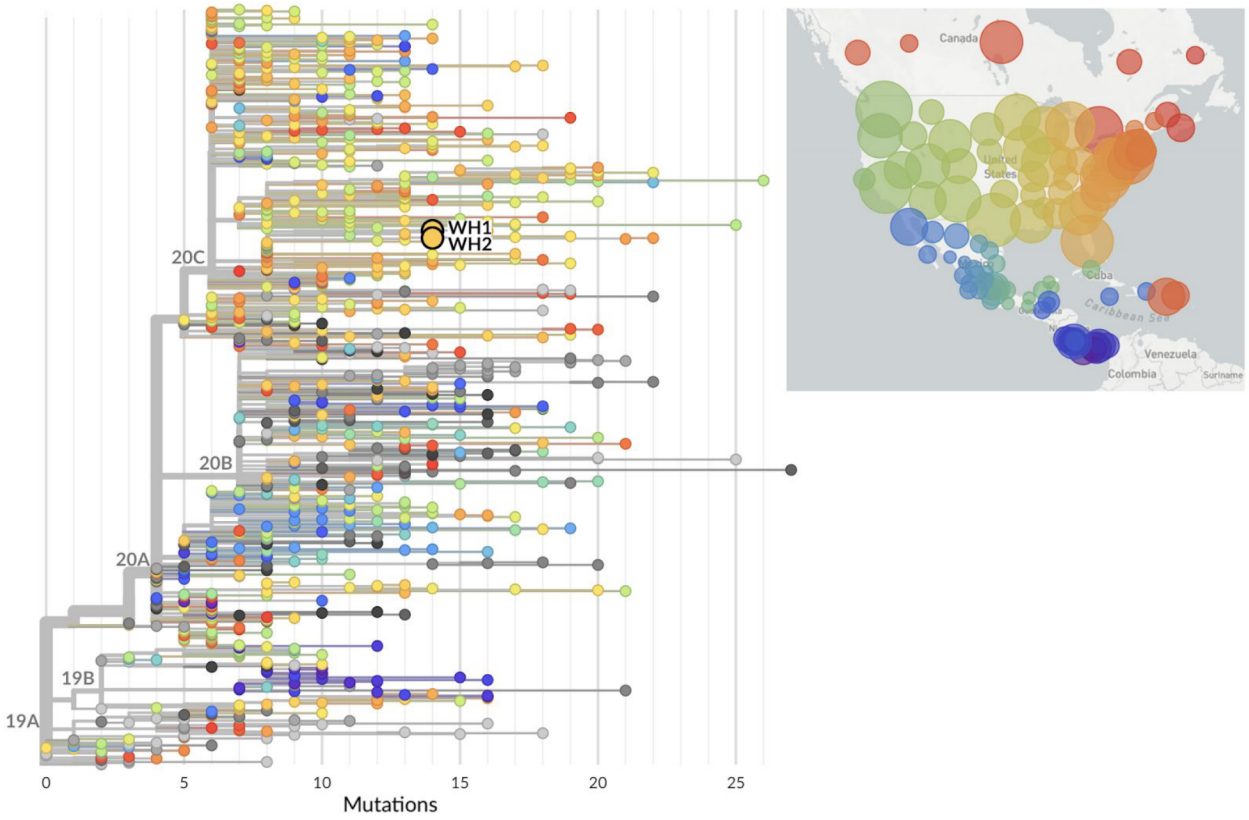

Figure 1. Phylogeny of 867 SARS-CoV-2 viruses collected from all over the world highlighting global placement of WH1 and WH2. Viruses from North America are preferentially sampled with 674 North American viruses (colored as shown in the inset map) and 193 viruses from outside North America (colored in gray). WH1 and WH2 are shown as larger circles with black outlines. An interactive version of this figure is available at nextstrain.org/community/blab/ncov-wh/background.

We compared WH1 and WH2 to all sequences available in the GISAID EpiCoV database [15,16] and identified viruses that are directly ancestral to WH1 and WH2 in a maximum-likelihood phylogeny (**Fig. 2A**). We find that WH1 and WH2 are descended from viruses sampled from the USA (Connecticut, Florida, New York, Texas, Washington), Canada and New Zealand in March and April 2020 with the addition of mutations A1977G, G7936T, G14250T, T18417C, C19524T and C20402T. There is a sister clade of viruses that share C20402T that were sampled in Virginia in August 2020, but these possess their own unique mutations A871T, C1549T, G2144T, G9854A, G25477T and C27879T and consequently a molecular clock analysis places the common ancestor of WH1/WH2 and this sister clade in April or early May 2020 (median estimate April 5, 95% confidence interval: March 27 to May 21) (**Fig. 2B**).

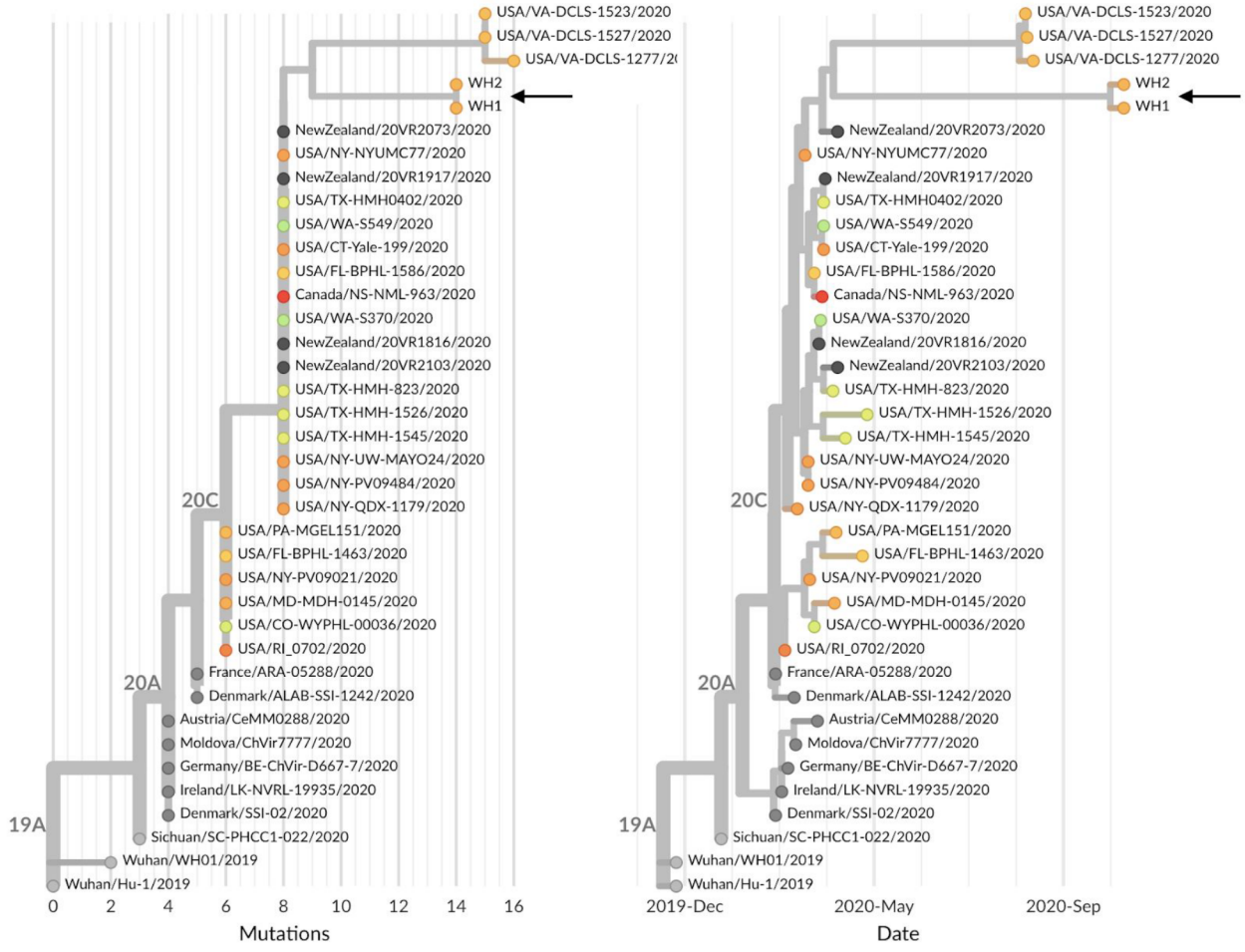

Figure 2. Phylogeny of 38 SARS-CoV-2 viruses that are either sister lineages to WH1 and WH2 or directly ancestral in the global maximum-likelihood phylogeny. Shown are both (A) phylogeny with branch lengths scaled by number of mutations from Wuhan reference genome and (B) temporally resolved phylogeny with branch lengths estimated according to a molecular clock analysis. Both panels are colored according to state of sampling for US samples or colored gray if samples were from outside the US. An interactive version of this figure is available at nextstrain.org/community/blab/ncov-wh/lineage.

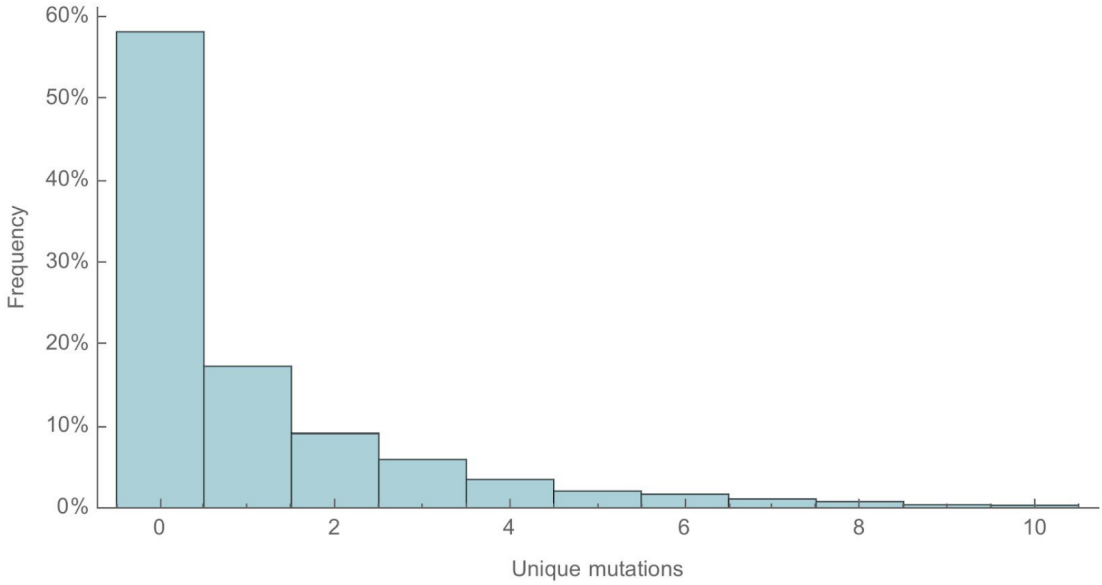

Figure 3. Distribution of number of unique mutations across sequenced SARS-CoV-2 viruses collected from the USA after August 1, 2020. This analysis uses the complete phylogeny of sequenced SARS-CoV-2 viruses and plots branch lengths for branches immediately subtending tips in this phylogeny, i.e. branches unique to a single sample.

These observations suggest a transmission chain leading to WH1 and WH2 from viruses circulating in the USA in March and April that collects an additional 5 mutations over these 6 months of circulation, consistent with the overall observed rate of molecular evolution of SARS-CoV-2 and natural Poisson variation [9]. Due to this 6-month gap in genomic surveillance of the WH lineage, we can say very little about the more recent geographic ancestry of WH1 and WH2. However, given its membership in lineage B.1.26 and its immediate common ancestors, a US origin is most parsimonious, as 306 out 324 (94%) of B.1.26 lineage viruses collected after July 1, 2020, are from the USA, while 11% of overall sequenced viruses collected after July 1, 2020, are from the USA.

Remarkably, academic, medical and public health institutions have sequenced and shared over 35,000 SARS-CoV-2 virus genomes from the USA epidemic [15,16]. However, this is a small fraction of the >9M confirmed cases of COVID-19 (and an even smaller fraction of SARS-CoV-2 infections) and so it may not be uncommon to have individual viruses that are divergent from sequenced viral diversity. We conducted a rough analysis of how uncommon it is to observe a sequenced sample that is 5 mutations divergent from its common ancestor by looking at the distribution of the number of unique mutations possessed by sequenced SARS-CoV-2 viruses collected from the USA after August 1, 2020 (**Fig. 3**). We observe 102 viruses with 5 or more unique mutations out of 2112, representing 4.8% of the total. If we instead look after September 1, 2002, we observe 35 viruses with 5 or more unique mutations out of 759, representing 4.6% of the total. Thus, the finding that WH1 possesses 5 unique mutations is rare, but not entirely unexpected. Note that this calculation does not take into account state or regional discrepancies in sequencing of SARS-CoV-2 genomes, which may skew the distribution further.

Discussion

Viral genome sequencing represents a powerful new tool for epidemiological investigation. However, it generally requires a decent fraction of infections to be sequenced to provide critical comparisons for the sequence or group of sequences in question. Nonetheless, there are situations where prompt genome sequencing of even a handful of viral genomes can be highly informative. For example, in late February 2020, genome sequencing of early SARS-CoV-2 infections in the Washington State area by us and colleagues strongly suggested cryptic spread of COVID-19 during January and February 2020, before active community surveillance was implemented [9].

The USA and the world have sequenced and publicly shared SARS-CoV-2 genomes more rapidly and at a far larger scale than any previous outbreak, epidemic or pandemic. To date, over 35,000 SARS-CoV-2 virus genomes have been sequenced and publicly shared from the USA alone [15,16]. However, because SARS-CoV-2 is so prevalent, there are many chains of transmission represented by viral lineages that have not been sequenced. The White House COVID-19 outbreak appears to descend from one such lineage that escaped detection for the previous ~6 months (**Fig. 1, Fig. 2**). As it stands, the WH1 and WH2 sequences are analogous to having two puzzle pieces that connect to one another, but with nothing else to directly connect to. This scenario is relatively rare but not too surprising overall (**Fig. 3**).

Given their genetic similarity, especially with sharing rare mutation 7936T, we believe that WH1 and WH2 are closely related epidemiologically. Given that the individuals in question did not have any direct contact with one another and both attested exposure at events associated with the White House COVID-19 outbreak, we believe that a shared epidemiological connection through the White House COVID-19 outbreak is the most parsimonious explanation for their infections' genetic similarity. This would imply that the WH lineage identified here was responsible for other infections in the White House cohort as well. However, we cannot completely rule out that the WH lineage identified here was circulating more broadly in the DC area and both individuals independently acquired infections of this lineage outside of White House associated exposure. Further sequencing of DC area infections from this time period would help to definitively rule out this possibility. Sequencing of additional infections in the White House cohort would be helpful in this regard as well.

There are currently no other viruses in the GISAID EpiCoV database collected from Washington DC after August 1, 2020. Retrospective sequencing of samples from Washington DC in this time period could yield other viruses that are part of the same lineage, but this is far from certain, as the WH lineage may have been introduced from elsewhere in the US. Generally, there continues to be substantial backfill of publicly available sequences and it is possible that viruses closer to WH1 and WH2 will be sequenced and shared in the coming months. This has the potential to shed light on the geographic origins of the WH lineage.

We believe that the WH lineage is likely to be at least relatively rare given its lack of sampling in publicly available sequences. Looking forward, this relative rarity should make it possible to identify infections that likely descend from the White House outbreak. If viruses from November onwards are discovered that bear that same constellation of mutations as WH1 and WH2, then the inference would be that these infections are the downstream repercussions of the sizable October White House outbreak. This has precedent in the downstream impact of a superspreader event at a business conference in Boston in February, where conference-associated mutations were later seen at high frequency in the downstream Massachusetts epidemic [17].

The COVID-19 pandemic has damaged health and health systems and disrupted society globally to a degree and with a speed unprecedented since the 1918 influenza pandemic. Science has made a great many discoveries and innovations since then, with genome sequencing being a fairly recent addition to the toolkit to combat infectious disease. We, as a society, have the tools to control COVID-19, they just have to be employed.

Methods

Sample collection

Individuals were enrolled as part of the HAARVI study. All participants completed informed consent. Previously collected samples, as well as prospectively collected samples, were used for this analysis. This study was approved by the University of Washington IRB (protocol #STUDY00000959).

Specimens were shipped to the Brotman Baty Institute for Precision Medicine via commercial couriers or the US Postal Service at ambient temperatures and opened in a class II biological safety cabinet in a biosafety level-2 laboratory. Dry swabs were rehydrated with 1 mL low TE, agitated for a minute, and allowed to incubate for 10 minutes at room temperature. Two 400 µL aliquots of low TE were collected from each specimen and stored at 4˚C until the time of nucleic acid extraction, performed with the MagnaPure 96 small volume total nucleic acids kit (Roche) or MagMAX Viral Pathogen II Nucleic Acid Isolation Kit (ThermoFisher). SARS-CoV-2 detection was performed using real-time RT-PCR with a probe sets targeting Orf1b and S with a FAM fluor (Life Technologies 4332079 assays # APGZJKF and APXGVC4APX) multiplexed with an RNaseP probe set with a HEX fluor (Life Technologies A30064 or IDT custom) each in duplicate on a QuantStudio 6 instrument (Applied Biosystems).

Sequencing

SARS-CoV-2 genome sequencing was conducted using both hybrid-capture and metagenomic approaches. In both cases, RNA was fragmented and converted to cDNA using random hexamers and reverse transcriptase (Superscript IV, Thermo) and a sequencing library was constructed using the Illumina TruSeq RNA Library Prep for Enrichment kit. For hybrid capture, we used the Ct value as a proxy for viral load to balance the samples and pooled 24-plex with Seattle-based samples for the hybrid capture reaction. Capture pools were incubated overnight with probes targeting the Wuhan-Hu-1 isolate, synthesized by Twist Biosciences. The

manufacturer's protocol was followed for the hybrid capture reaction and target enrichment washes. The hybrid capture pools were sequenced on the Illumina MiSeq instrument using 2x150bp reads and the metagenomic sample was sequenced on an Illumina MiSeq instrument using 2x100bp reads. The resulting reads were assembled against the SARS-CoV-2 reference genome Wuhan-Hu-1/2019 (Genbank accession MN908947) using the bioinformatics pipeline github.com/seattleflu/assembly. The consensus genome sequence of WH1 was deposited to the GISAID EpiCoV database as strain USA/DC-BBI1/2020 with accession EPI_ISL_603248. GISAID does not allow partial genome sequences with >50% Ns and so WH2 has not been deposited. Consensus genome sequences for both WH1 and WH2 are also available from github.com/blab/ncov-wh.

Analysis

Consensus genome sequence of WH1 was combined with SARS-CoV-2 genomes downloaded from GISAID [15,16] and processed using the Nextstrain [13] bioinformatics pipeline Augur to align genomes via MAFFT v7.4 [18], build maximum likelihood phylogeny via IQ-TREE v1.6 [19] and reconstruct nucleotide and amino acid changes on the ML tree. Branch lengths were temporally resolved using TreeTime v0.7.4 [20]. The resulting tree was visualized in the Nextstrain web application Auspice to view resulting inferences. Workflows to reproduce phylogenetic trees shown in **Figure 1** and **Figure 2** are available from github.com/blab/ncov-wh.

The global phylogeny of SARS-CoV-2 was downloaded from github.com/roblanf/sarscov2phylo [21] and tip labels along with immediately subtending branch lengths extracted. The distribution of these tip lengths is shown in **Figure 3**.

Acknowledgements

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu Database on which this research is based. A full Acknowledgments table is available as supplementary materials. This work was supported by funding from the Brotman Baty Institute for Precision Medicine.

Competing interests

HYC is a consultant for Merck and GlaxoSmithKline. JS is a consultant with Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Nanostring, Phase Genomics, Adaptive Biotechnologies, and Stratos Genomics, and he has a research collaboration with Illumina. All other authors declare no competing interests.

References

1. World Health Organization. COVID-19 Weekly Epidemiological Update - 27 October 2020. 2020, October 25. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

2. Centers for Disease Control and Prevention. CDC COVID Data Tracker. 2020, Oct 30.

Available: https://covid.cdc.gov/covid-data-tracker/

3. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine. 2020. doi:10.1038/s41591-020-1092-0

4. Furuse Y, Sando E, Tsuchiya N, Miyahara R, Yasuda I, Ko YK, et al. Clusters of Coronavirus Disease in Communities, Japan, January-April 2020. Emerg Infect Dis. 2020;26. doi:10.3201/eid2609.202272

5. Washington State Department of Health. Statewide COVID-19 Outbreak Report. 2020, October 29. Available: https://www.doh.wa.gov/Portals/1/Documents/1600/coronavirus/StatewideCOVID-19OutbreakReport.pdf

6. Buchanan L, Gamio L, Leatherby L, Keefe J, Koettl C, Schoenfeld Walker A. Tracking the White House Coronavirus Outbreak. 2020, Oct 14. Available: https://www.nytimes.com/interactive/2020/10/02/us/politics/trump-contact-tracing-covid.html

7. White House COVID-19 outbreak. In: Wikipedia [Internet]. 2020, Oct 30. Available: https://en.wikipedia.org/wiki/White_House_COVID-19_outbreak

8. Sun LH, Abutaleb Y, Dawsey J. A week into a White House outbreak, CDC still playing only a limited role. In: The Washington Post [Internet]. 2020, October 8. Available: https://www.washingtonpost.com/health/2020/10/08/covid-white-house-contact-tracing/

9. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L, et al. Cryptic transmission of SARS-CoV-2 in Washington State. Science. 2020; eabc0523.

10. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria NR, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 2020;369: 582–587.

11. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science. 2020. p. eabc1917. doi:10.1126/science.abc1917

12. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. 2020; veaa061.

13. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34: 4121–4123.

14. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5: 1403–1407.

15. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017;1: 33–46.

16. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22. doi:10.2807/1560-7917.ES.2017.22.13.30494

17. Lemieux J, Siddle KJ, Shaw BM, Loreth C, Schaffner S, Gladden-Young A, et al. Phylogenetic analysis of SARS-CoV-2 in the Boston area highlights the role of recurrent importation and superspreading events. medRxiv. 2020; 2020.08.23.20178236.

18. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.

19. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.

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

21. Lanfear R. A global phylogeny of SARS-CoV-2 sequences from GISAID. 2020. doi:10.5281/zenodo.3958883