RESEARCH ARTICLE

Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts

Andrew L. Valesano[1,2], Kalee E. Rumfelt[3], Derek E. Dimcheff[4], Christopher N. Blair[1,2], William J. Fitzsimmons[1,2], Joshua G. Petrie[3], Emily T. Martin[3], Adam S. Lauring[1,2]*

1 Division of Infectious Diseases, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, United States of America, **2** Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, United States of America, **3** Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America, **4** Division of Hospital Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, United States of America

* alauring@med.umich.edu

Abstract

Analysis of SARS-CoV-2 genetic diversity within infected hosts can provide insight into the generation and spread of new viral variants and may enable high resolution inference of transmission chains. However, little is known about temporal aspects of SARS-CoV-2 intrahost diversity and the extent to which shared diversity reflects convergent evolution as opposed to transmission linkage. Here we use high depth of coverage sequencing to identify within-host genetic variants in 325 specimens from hospitalized COVID-19 patients and infected employees at a single medical center. We validated our variant calling by sequencing defined RNA mixtures and identified viral load as a critical factor in variant identification. By leveraging clinical metadata, we found that intrahost diversity is low and does not vary by time from symptom onset. This suggests that variants will only rarely rise to appreciable frequency prior to transmission. Although there was generally little shared variation across the sequenced cohort, we identified intrahost variants shared across individuals who were unlikely to be related by transmission. These variants did not precede a rise in frequency in global consensus genomes, suggesting that intrahost variants may have limited utility for predicting future lineages. These results provide important context for sequence-based inference in SARS-CoV-2 evolution and epidemiology.

Author summary

Understanding the evolution and transmission of SARS-CoV-2 is important for designing public health interventions to prevent outbreaks. Viral genome sequencing has been widely used to reconstruct patterns of SARS-CoV-2 transmission through communities and to monitor the spread of new strains. However, because SARS-CoV-2 can transmit multiple times before a new mutation fixes, consensus sequences often cannot determine "who infected whom." Identifying individuals who share the same viral genetic variants at

OPEN ACCESS

Citation: Valesano AL, Rumfelt KE, Dimcheff DE, Blair CN, Fitzsimmons WJ, Petrie JG, et al. (2021) Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. PLoS Pathog 17(4): e1009499. https://doi.org/10.1371/journal.ppat.1009499

Editor: Andrew Pekosz, Johns Hopkins University Bloomberg School of Public Health, UNITED STATES

Received: January 21, 2021

Accepted: March 24, 2021

Published: April 7, 2021

Data Availability Statement: Raw sequence reads are available as fastq files from the Sequence Read Archive at accession number PRJNA682212, with human-mapping reads removed. Analysis code is available at https://github.com/lauringlab/SARSCov2_Intrahost. Consensus genome sequences are publicly available at the GitHub link and on GISAID.

Funding: This work was supported by a University of Michigan COVID-19 Response Innovation Grant (to ASL), National Institute of Allergy and Infectious Diseases K01 AI141579 (to JGP) and Centers for Disease Control and Prevention U01 IP000974 (to ETM). The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

low frequencies within each infection may help resolve this problem, but to do this we need to accurately identify within-host genetic variants and understand how they evolve and spread. We investigated within-host diversity of SARS-CoV-2 with samples collected in southeastern Michigan in March–May 2020. We show that there are relatively few genetic variants present in any given infection, and variants do not tend to accumulate in people over time. We also found that people who are not part of the same epidemic cluster can share the same within-host variants, due to chance or various evolutionary forces.

Introduction

Over the course of the SARS-CoV-2 pandemic, whole genome sequencing has been widely used to characterize patterns of broad geographic spread, transmission in local clusters, and the spread of specific viral variants [1–6]. Early reports demonstrated that SARS-CoV-2 exhibits genetic diversity within infected hosts, but this has been less studied than consensus-level genomic diversity [7]. Intrahost diversity is an important complement to consensus sequencing. Patterns of viral intrahost diversity throughout individual infections can suggest the relative importance of natural selection and stochastic genetic drift [8]. Shared intrahost variants between individuals can reveal loci under convergent evolution and enable measurement of the transmission bottleneck, a critical determining factor in the spread of new genetic variants [9,10]. Studies of SARS-CoV-2 intrahost diversity may shed light on selective pressures applied at the individual level, such as antivirals and antibody-based therapeutics. While a clear understanding of within-host evolution can inform how SARS-CoV-2 spreads on broader scales, there have been relatively few comprehensive studies of intrahost dynamics [9,11,12].

Sequencing of intrahost populations can also potentially be applied to genomic epidemiology [13]. A common goal in sequencing specimens from case clusters is to infer transmission linkage, which can guide future public health and infection control interventions. However, the relatively low substitution rate and genetic diversity of SARS-CoV-2 present challenges to inference of individual transmission pairs [13,14]. In the pandemic setting, there is a non-negligible chance that two individuals who are epidemiologically unrelated could be infected with nearly identical viral genomes. Viruses from a single local outbreak may have few differentiating substitutions, limiting the ability of sequencing to resolve exact transmission chains. Identification of shared intrahost variants between individuals has been explored in other pathogens to overcome this obstacle [15–19]. However, use of this approach for SARS-CoV-2 will depend on a solid understanding of the forces that shape the generation and spread of genetic variants.

There are several unresolved questions that will dictate the utility of intrahost diversity for genomic epidemiology. First, there must be sufficient intrahost diversity generated during acute infection prior to a transmission event. How much intrahost diversity is accumulated over time from infection onset is currently unknown. Second, the population bottleneck during transmission must be sufficiently wide to allow minor variants to be transmitted to recipient hosts [20,21]. Third, *de novo* generation of the same minor variants across multiple infections must be sufficiently rare. Independent generation of shared minor variants by recurrent positive selection or genetic drift in unlinked hosts could confound transmission inference [15]. Finally, measurements of intrahost diversity must be accurate and account for several potential sources of error [22,23]. Although previous studies have described within-host variation of SARS-CoV-2 [7,9,11,12,24–26], few have addressed the sources of systematic errors and batch effects in variant identification. To assess the utility of SARS-CoV-2 intrahost diversity for transmission inference, we need a clearer understanding of its temporal variation throughout infection and the extent of convergent evolution across individuals. Addressing these questions will also be valuable for understanding SARS-CoV-2 evolution.

Here, we sequenced SARS-CoV-2 genomes from 325 residual upper respiratory samples from hospitalized patients and employees at the University of Michigan. To validate our sequencing approach, we sequenced defined mixtures of two synthetic RNA controls and found that low input viral load decreases the specificity of variant calling. We find that observed intrahost diversity does not vary significantly by day since symptom onset. Intrahost variants can be shared between individuals that are unlikely to be related by transmission, suggesting that variants can arise by parallel evolution. These results inform our understanding of SARS-CoV-2 diversification in human hosts and highlight important considerations for sequence-based inference in the virus's genomic epidemiology.

Results

We retrieved respiratory specimens collected through diagnostic testing from March–May 2020. We sequenced samples from two groups: inpatients who were part of an observational study of COVID-19 in hospitalized individuals (n = 190), and symptomatic employees who presented to occupational health services (n = 135). All employees were diagnosed and treated in outpatient settings, except for one who was admitted as an inpatient. Basic demographic information is described in a separate work [27]. Genome copy number determined by qPCR of the nucleocapsid gene was highly variable and decreased by day from symptom onset ($p < 0.001$, linear model, Fig 1A). We obtained 212 complete genomes (Fig 1B), mostly from samples with higher viral loads (Fig 1B). Consensus genomes had a median of 7 substitutions relative to the Wuhan-Hu-1/2019 reference sequence (range 4–12). Phylogenetic analysis of whole consensus genomes identified 10 unique evolutionary lineages in our cohort (lineages determined by the PANGOLIN system, see Methods; Fig 1C). Most sequenced genomes fell in lineage B.1. We evaluated whether any employees were part of an epidemiologically linked cluster based on illness onset date, positive test status, and work location. We found that some employees were part of epidemiologically linked clusters (Fig 1C). The genomes from clusters 2, 10, 19, 20, and one pair in cluster 29 had ≤ 1 consensus difference, while the rest had 2–7 differences. Many employees in different clusters also had identical or nearly identical consensus genomes, which reflects the low genetic diversity of SARS-CoV-2 at this stage of the pandemic. We have no information on epidemiologic linkage for the remaining sequenced individuals. It is highly unlikely that there are direct transmission pairs in our dataset, but we cannot conclusively rule out coincident transmission linkage. Therefore, this population largely reflects a cross-section of infected individuals who are epidemiologically unlinked.

Identification of viral within-host variants can be prone to errors [22,23]. Therefore, we performed a mixing study to evaluate the accuracy of our pipeline for identifying intrahost single nucleotide variants (iSNV). We mixed two synthetic RNA controls that differ by seven single nucleotide substitutions at defined frequencies and input concentrations (Fig 2A). These mixtures were sequenced using the same approach as the clinical samples. We identified true iSNV at the expected frequencies at $\geq 10^3$ copies/μL (Fig 2B). There was greater variance in the observed variant frequencies at 10^2 copies/μL compared to higher input concentrations. We obtained high sensitivity for iSNV at $\geq$ 2% frequency and $\geq 10^3$ copies/μL with sufficient genome coverage. Many false positive iSNV remained at $\geq$ 2% frequency and 10^2 copies/μL despite multiple quality filters (S1 Fig). However, false positive iSNV per sample drastically decreased with input concentrations $\geq 10^3$ copies/μL. Three false positive iSNV were identified in multiple samples above 10^4 copies/μL: A3350U, G6669A, and U13248A. Mutation

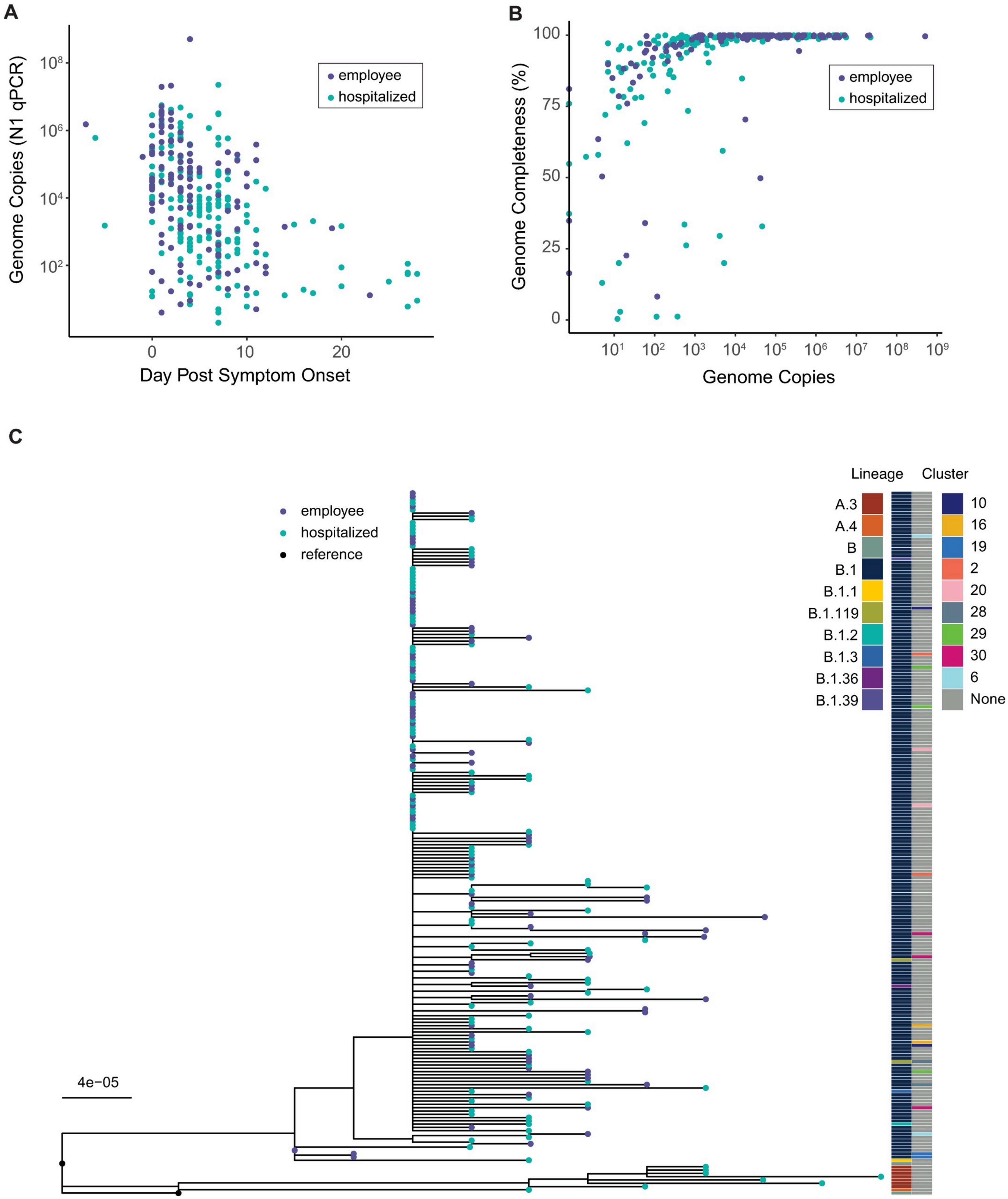

Fig 1. Viral shedding and overview of genome sequencing data. (A) Viral load by day of infection in hospitalized patients (teal) and employees (violet). Viral load, measured by qPCR of the N gene in units of genome copies per microliter of extracted RNA, is on the y-axis and day post symptom onset is on the x-axis.

(B) Genome completeness by viral load in hospitalized patients (teal) and employees (violet). Viral load as shown in (A) is on the x-axis and the fraction of the genome covered above 10x read depth is shown on the y-axis. (C) Maximum-likelihood phylogenetic tree. Tips represent complete consensus genomes from hospitalized patients (teal) and employees (violet). The axis shows divergence from the root (Wuhan-Hu-1/2019). The second genome displayed as "reference" is Wuhan/WH01/2019. Heatmaps show PANGOLIN evolutionary linage (left) and epidemiologic cluster (right).

https://doi.org/10.1371/journal.ppat.1009499.g001

U13914G recurred in multiple samples at input concentrations of 10^3 copies/μL and below. We suspect that they represent low-frequency variants present in the synthetic RNA controls, as has been observed in other studies with synthetic controls from the same manufacturer [9].

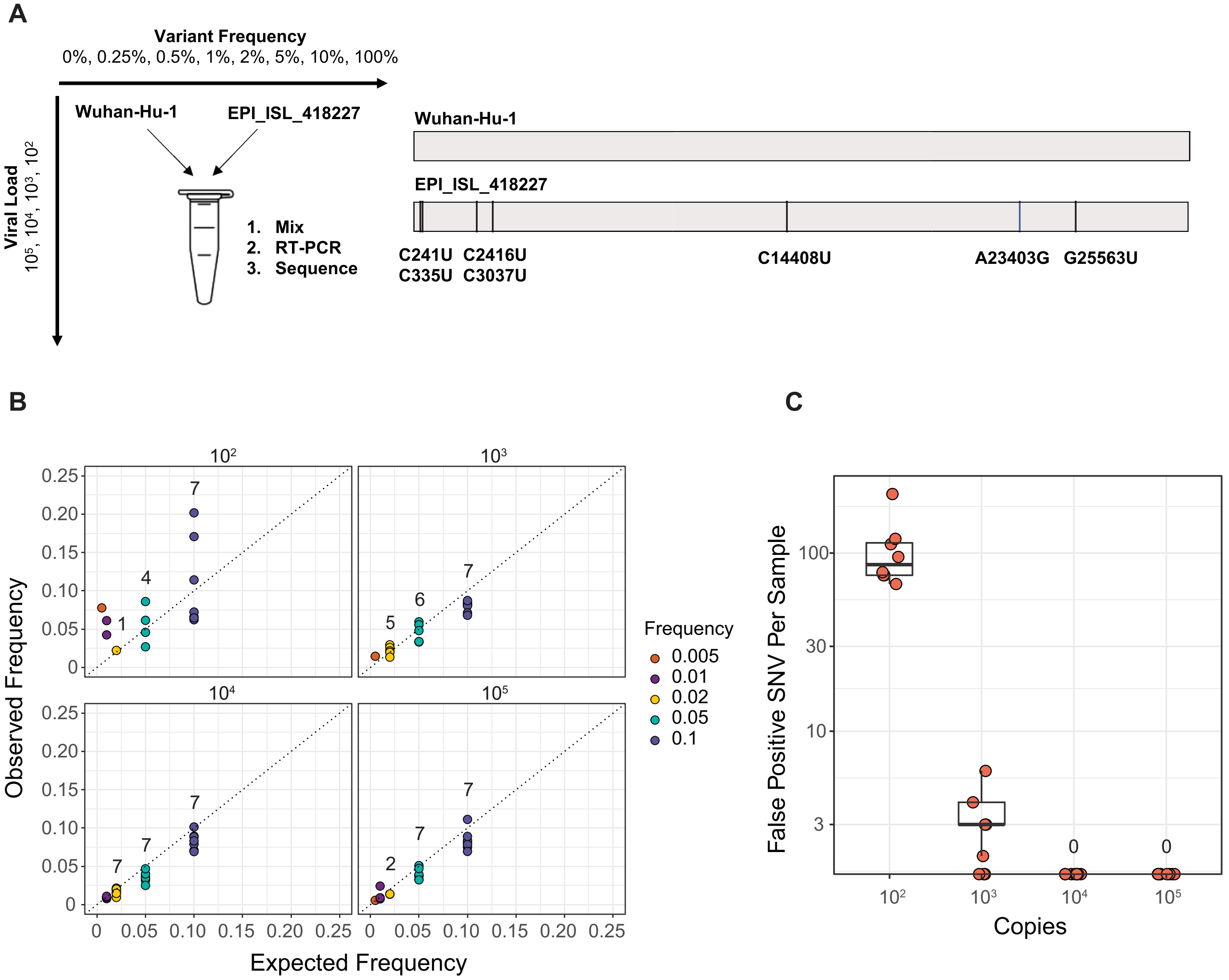

Fig 2. Assessing accuracy of intrahost variant detection by sequencing defined viral mixtures. (A) Schematic of the experiment. Wuhan-Hu-1 (reference) and EPI_ISL_418227 (variant) RNA were mixed at the given frequencies and viral loads (units of genome copies per microliter, representing the resulting mixture). Mixtures of RNA were amplified and sequenced in the same fashion as the clinical specimens. Reference and variant genomes differ by seven single nucleotide substitutions. (B) Observed frequency by expected frequency. Observed frequency of the true positive intrahost single nucleotide variants (iSNV) is on the y-axis and expected iSNV frequency is on the x-axis. Synthetic RNA copy number in units of genome copies per microliter of RNA is shown above each facet. Values above the points indicate the number of variants detected in that group (maximum of seven per group). (C) False positive iSNV. Number of false positive iSNV per sample is shown on the y-axis (base 10 log scale) and viral load as shown in (B) is on the x-axis, excluding iSNV at positions 3350, 6669, 13248, and 13419. Each point represents a unique sample and the boxplots represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range.

https://doi.org/10.1371/journal.ppat.1009499.g002

Excluding these recurrent sites, there were few false positive iSNV per sample with input concentrations above 10^3 copies/μL (Fig 2C). Together, these data indicate that sufficient input viral load is a critical factor for accurate identification of iSNV.

Based on our benchmarking experiment, we identified iSNV in 178 specimens with viral loads $\geq 10^3$ copies/μL (Fig 3A). We excluded position 11083, which is near a natural poly-U site and prone to sequencing errors [28], as well as the four sites with recurrent false positives (nucleotide positions 3350, 6669, 13248, and 13914). Most specimens exhibited fewer than ten minor iSNV (median 1, IQR 0–2, Fig 3B). There were four outlier specimens with greater than 15 iSNV. In these samples, iSNV were dispersed throughout the genome at various frequencies, so it is difficult to determine whether they represent mixed infections [11]. The locations of these samples on sequencing plates were not suggestive of cross-contamination. There was no difference in minor iSNV richness between hospitalized patients and employees treated as outpatients (p = 0.25, Mann-Whitney U test, S2 Fig). We identified more minor iSNV encoding non-synonymous changes than synonymous ones across most open reading frames (Fig 3C) and identified more iSNV at lower frequencies (Fig 3D), which together is suggestive of mild within-host purifying selection. Sample iSNV richness decreased with higher viral loads by about 1 iSNV per 10-fold increase in viral load (p = 0.01, multiple linear model, S3 Fig). Sample iSNV richness did not correlate with day from symptom onset (p = 0.79, multiple linear model, Fig 3E). This result was robust to exclusion of the four outlier samples and exclusion of viral load from the model. These results show that within-host diversity is low and remains that way over the duration of most SARS-CoV-2 infections.

Next, we investigated patterns of shared intrahost diversity between individuals. Most iSNV were unique to a single individual. However, 18 iSNV were present in multiple specimens (Fig 4A). None of these mutations were located at sites known to commonly produce errors or homoplasies [28,29]. Two iSNV were present in three individuals (G12331A and A11782G, both synonymous changes in ORF1a). There was no clear phylogenetic clustering of genomes exhibiting these shared iSNV (S4 Fig), and G12331A was shared between samples from different viral lineages (13 substitutions). These two mutations were first detected in our samples in late March 2020 (Fig 4B). None reached $>$ 1% frequency per week in consensus sequences submitted to GISAID through mid-November 2020. These results suggest that iSNV that arise convergently across viral lineages are not necessarily predictive of subsequent global spread of those mutations.

Transmission inference based on shared iSNV integrates information such as consensus genome sequences, sample dates, and shared iSNV [15]. Therefore, we compared shared iSNV across all unique pairs of specimens used for variant calling (n = 15753, Fig 5). Because most iSNV were unique to an individual, most pairs did not share iSNV and only 0.14% of pairs shared one iSNV. Many pairs with shared iSNV were sequenced in separate batches, which reduces the likelihood that shared iSNV are due to cross-contamination. No employee pairs in the same epidemiologic cluster shared iSNV, which are the only pairs in our dataset who are likely part of the same transmission network. The rest of the pairs of individuals are most likely not directly linked by transmission and probably share iSNV by chance. We identified nine unique pairs with shared iSNV between genomes that were near-identical (0–1 consensus differences), three of which were collected within one week of each other. We also identified shared iSNV between 13 pairs separated by $\geq$ 2 consensus substitutions (Fig 5A and 5B) and 15 pairs with collection dates 7–28 days apart (Fig 5B). Due to differences in viral lineage and time of collection, these are very unlikely to be transmission pairs. Together, these data indicate that iSNV can arise convergently between individuals who are unlikely to be related by transmission.

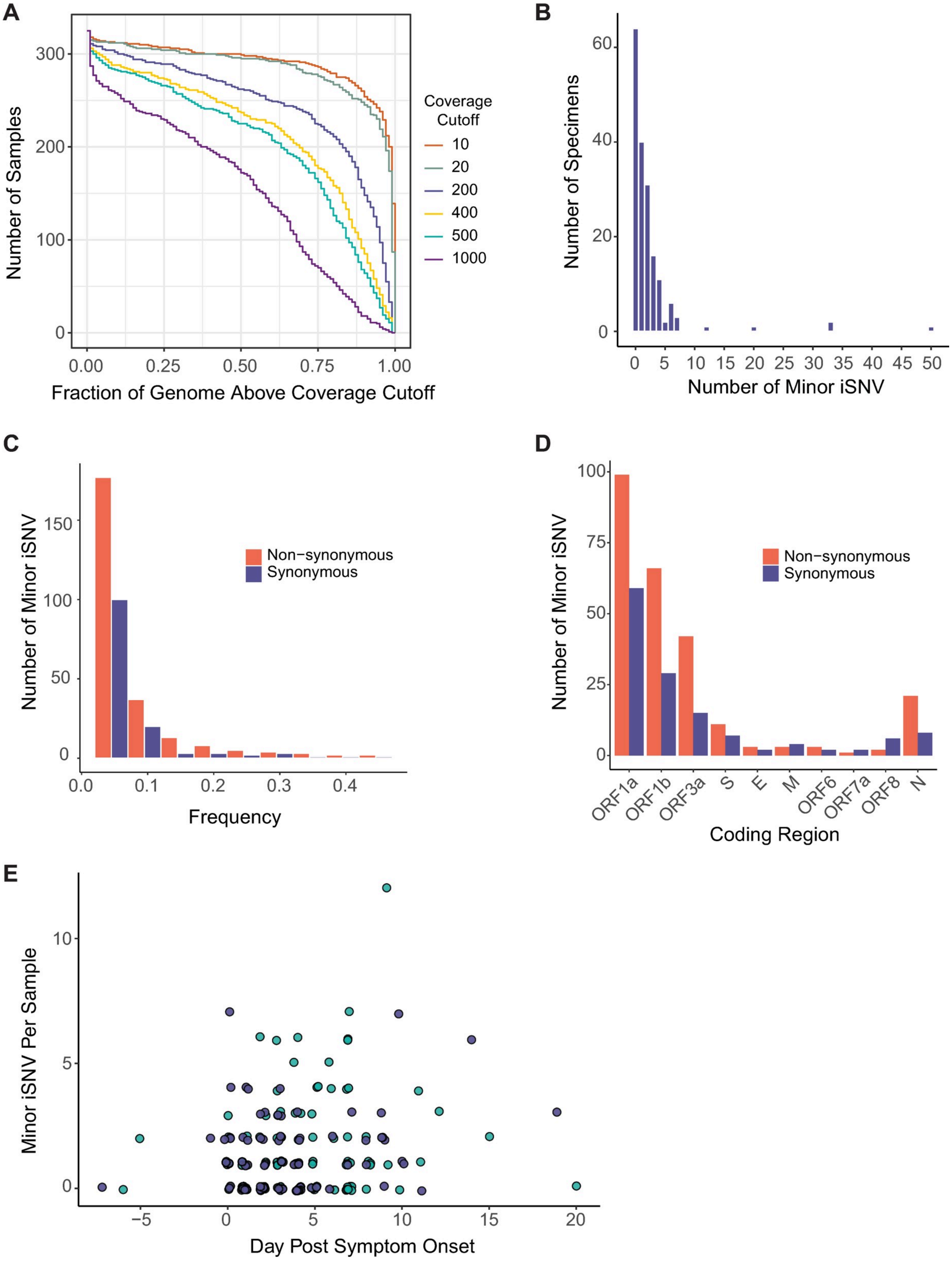

Fig 3. SARS-CoV-2 intrahost single nucleotide variant (iSNV) diversity. (A) Sequencing coverage for clinical samples. The number of clinical samples (y-axis) is shown by the fraction of the genome above a given read depth threshold (x-axis). The different lines show the data evaluated with six read depth thresholds. (B) Histogram of the number of specimens (y-axis) by the number of minor iSNV per sample (x-axis), n = 178. (C) Number of minor iSNV by frequency with a bin width of 0.05. Non-synonymous iSNV are shown in orange and synonymous iSNV are shown in violet. (D) Number of minor iSNV by coding region. Non-synonymous iSNV are shown in orange

and synonymous iSNV are shown in violet. (E) Scatterplot of the number of minor iSNV per sample (y-axis) by the day post symptom onset (x-axis). Hospitalized patients are shown in teal and employees shown in violet. The four samples with > 15 iSNV shown in (B) are excluded from the plot for visualization.

https://doi.org/10.1371/journal.ppat.1009499.g003

Discussion

Accurate characterization of SARS-CoV-2 intrahost diversity is important for understanding the spread of new genetic variants and its potential use in transmission inference. In this study, we sequenced upper respiratory specimens from a cohort of hospitalized COVID-19 patients and infected employees. We found that intrahost diversity is low and its distribution does not vary by time since symptom onset. We identified iSNV shared across viral genomes separated by time and disparate evolutionary lineages, indicating that iSNV can arise convergently. Because variants may be shared through parallel mutation rather than transmission, caution is warranted in the use of shared iSNV alone for inferring transmission chains. Intrahost variants shared across multiple individuals did not precede an increase in frequency in global consensus genomes, which suggests that identifying convergent iSNV may have limited utility in tracking broader SARS-CoV-2 evolution.

Specimen viral load is important when measuring intrahost diversity. We and others have shown that samples with low viral loads are prone to false positive iSNV and lower sensitivity [22,23,30]. A strength of our study is that we experimentally validated the accuracy of our variant calling by sequencing defined populations. Based on these results, we excluded samples with low viral load from subsequent analyses. Future studies of SARS-CoV-2 intrahost diversity should report and account for specimen viral loads to avoid this common source of error. We did not benchmark our sequencing approach for detecting insertions and deletions (indels) and therefore did not report these for the clinical specimens. Intrahost indels could conceivably provide useful information about within-host evolution, but accurate detection is also subject to similar issues of sample quality and viral load.

The low level of intrahost diversity that we found here is consistent with a recent preprint by Lythgoe et al. [9]. The fact that our work and the study by Lythgoe et al. were performed

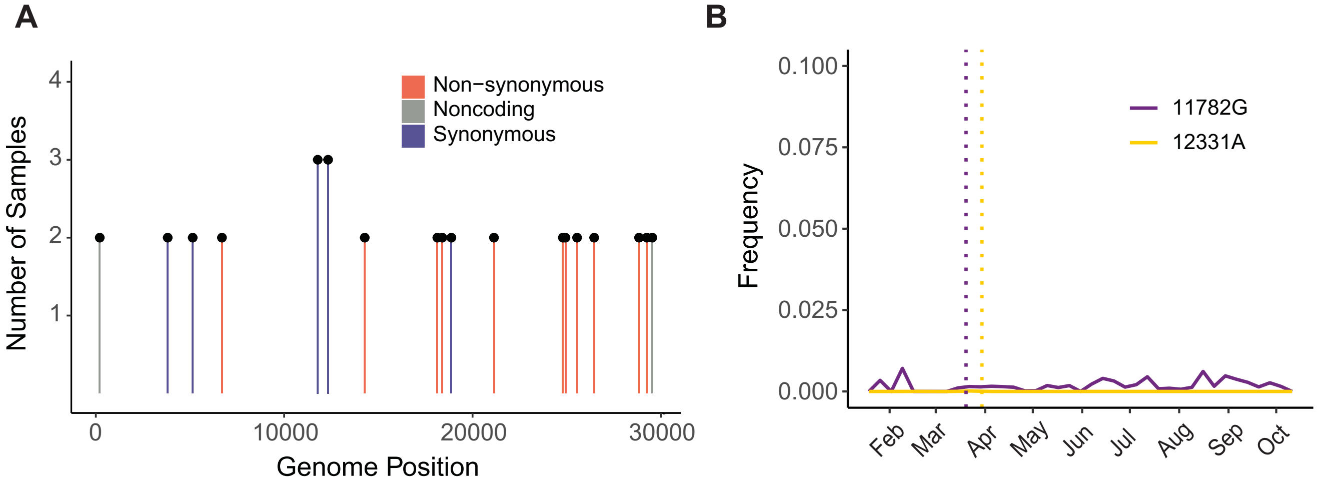

Fig 4. Shared iSNV across samples and their frequency in global consensus genomes. (A) Shared iSNV across samples, with the number of samples sharing the iSNV (y-axis) by the genome position (x-axis). Colors indicate the iSNV coding change relative to the reference. (B) The frequency (y-axis) of three iSNV shared by three or more samples over time (x-axis). The consensus genomes are from GISAID, as available on 2020-11-11. The vertical dotted lines represent the earliest time we detected each iSNV in our samples.

https://doi.org/10.1371/journal.ppat.1009499.g004

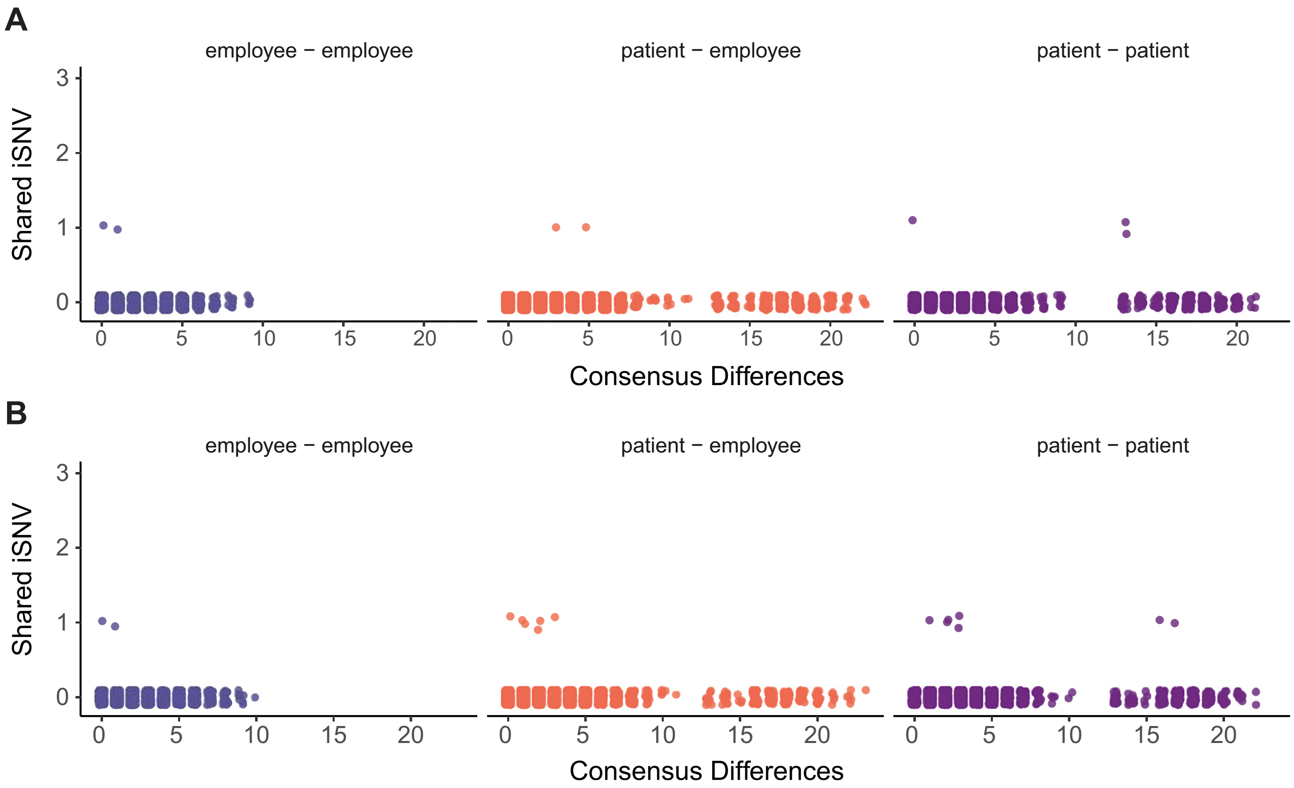

Fig 5. Pairwise comparisons of shared iSNV. Each unique pair is shown as a single point, with employee-employee pairs in violet (left), patient-employee pairs in orange (middle), and patient-patient pairs in purple (right). The number of iSNV shared by each pair is shown on the y-axis with the number of consensus differences between the pair of genomes on the x-axis. Pairs of samples collected within seven days of each other are displayed in (A), and pairs of samples collected greater than seven days apart are shown in (B).

https://doi.org/10.1371/journal.ppat.1009499.g005

with different geographical areas, sequencing approaches (ARTIC Network amplicons vs. veSEQ metagenomic sequencing), and analysis methods lends credence to the results. Lythgoe et al. reported more shared variation than seen here, but this is most likely due to sequencing a greater number of samples among individuals within known epidemiologic clusters. We and Lythgoe et al. measure a lower level of intrahost diversity at the 2% frequency threshold compared to a recent study in Austria [12]. The reasons for this are not clear, but it is likely due to differences in sample viral loads and variant calling methods. We did not find a difference in intrahost diversity between hospitalized COVID-19 patients and those treated as outpatients, which suggests that viral diversity may not be a reliable marker for disease severity.

Measuring viral diversity over the course of infection is relevant for understanding how variants are transmitted to new hosts. Only genetic variants present at the time of a transmission event will have the opportunity to spread. Because SARS-CoV-2 usually transmits just before or several days after symptom onset [31,32], it is important to define viral diversity in this window. Our cross-sectional analysis of diversity by time since symptom onset indicates that diversity does not significantly increase over the course of infection. A significant fraction of samples may not exhibit any iSNV at the time of transmission, which could limit the utility of iSNV for linking transmission pairs. Only a large bottleneck would lead to onward spread of most iSNV present during early infection. However, it is important to recognize that although the absolute level of diversity may not change over time, different variants may arise or go extinct during a given infection. This phenomenon was observed in a recent study by Tonkin-Hill et al. [11]. Serial samples from individuals could address this issue with higher resolution. Low diversity within hosts also shapes our expectations for emergence of resistance to drugs and monoclonal antibodies. With such limited substrate for selection to act upon, the short window of time between treatment and transmission could limit the spread of a variant selected within a host. Even during prolonged infections in immunocompromised hosts, there is only limited evidence of resistance to various COVID-19 therapeutics [33–35].

Parallel evolution is a critical factor to consider in the interpretation of shared intrahost variation [15]. Even if iSNV identification were perfectly specific, iSNV can arise in parallel due to biological processes such as natural selection and genetic drift. A key finding of this work is that iSNV can arise in genomes that are unrelated by local transmission, specifically those across large time intervals and lineages. Shared iSNV between individuals with identical genomes collected the same week may also have arisen in parallel. These pairs are most likely not epidemiologically linked, but we are unable to rule out coincident local transmission in the community. Because iSNV can arise in parallel in genomes that are not linked by transmission, caution is needed when relying entirely on shared iSNV for transmission inference [11,13].

We also found that identifying iSNV across multiple individuals did not precede an increase of those mutations in frequency in global consensus genomes. It is unclear whether these mutations arose due to positive selection, chance, or mutational "hotspots" [11]. It is possible that these mutations were lost due to purifying selection within hosts or during transmission [8,36]. These results suggest that iSNV may have lower utility for tracking broader SARS-CoV-2 evolution, but larger sample sizes in more geographic areas are necessary to evaluate this.

One of the most important variables for transmission inferences is the size of the transmission bottleneck [15]. If parallel evolution of iSNV occurs regularly and the transmission bottleneck is very small, that would increase the likelihood that shared iSNV are due to convergence rather than transmission. However, if the bottleneck is large, then iSNV may become more valuable for detecting transmission networks when consensus genomes are limited. There are currently conflicting results on the SARS-CoV-2 bottleneck size. Popa et al. estimated a bottleneck size of greater than 1000 unique genomes [12]. In contrast, Lythgoe et al. estimated a bottleneck size range from 1–8 unique genomes based on 14 household pairs [9]. Lythgoe et al. in particular used extensive controls and validation for preventing contamination and identifying sequencing errors. Other studies both in humans and in domestic cats have estimated small bottlenecks [37,38]. It is difficult to interpret these contrasting results because each study used different sequencing and analysis methodologies. In recent work on influenza A virus, a study of methodological differences was key for resolving different conclusions about the bottleneck size [39]. One factor that has not yet been clearly defined is how the time interval between donor-recipient pairs affects SARS-CoV-2 bottleneck estimates. We expect that further work will clarify the reasons behind these conflicting estimates.

Because of the high incidence and low mutation rate of SARS-CoV-2, genomic epidemiology is necessarily constrained in its ability to determine exact transmission chains in an outbreak. Using minor genetic variation to increase the resolution of genomic epidemiology requires attention to the underlying processes of within-host viral evolution and awareness of possible confounders. Unified statistical frameworks that incorporate sequences, metadata, and epidemiological models are likely the most robust approaches for integrating intrahost variants, but these models also must account for parallel evolution [15–17]. As others have recently suggested [11], we caution against assigning transmission pairs solely by virtue of shared iSNV in the absence of clear epidemiologic information.

Materials and methods

Ethics statement

We collected clinical metadata and residual diagnostic specimens positive for SARS-CoV-2 from hospitalized patients enrolled in the CDC HAIVEN (Hospitalized Adult Influenza Vaccine Effectiveness Network) study and infected employees enrolled in the HARVI (hospital associated respiratory virus infection) study. These studies were reviewed by the University of Michigan Institutional Review Board (HUM 150524 and HUM 105491) and operated under waivers of informed consent for collection of limited datasets and use of residual clinical specimens.

Date of illness onset for hospitalized patients was collected individually via medical chart abstraction from physician notes. Michigan Medicine employees with any suspected COVID-19 symptoms were asked to call a COVID-19 healthcare worker hotline before reporting to work. Date of symptom onset, a list of symptoms, close contacts, travel history, and work location and description were recorded. After testing, employee clusters were determined by illness onset date, positive test status, and work location.

Genome amplification and sequencing

Residual samples from nasopharyngeal swabs and sputum specimens were centrifuged at 1200 x g. and 200 microliters were aliquoted. RNA was extracted with the Invitrogen PureLink Pro 96 Viral RNA/DNA Purification Kit and eluted in volumes of 100 microliters. Complementary DNA was reverse transcribed with SuperScript IV (ThermoFisher). The SARS-CoV-2 genome was amplified in two multiplex PCR reactions using the ARTIC Network V3 primer sets. Sequencing libraries were prepared with the NEBNext Ultra II kit and pooled in equal volumes after barcoding. The pooled sequencing library was gel extracted to remove adapter dimers. Libraries were sequenced on an Illumina MiSeq at the University of Michigan Microbiome Core facility (v2 chemistry, 2x250 cycles). To validate this approach, we used two synthetic RNA controls that differ by seven single nucleotide mutations, Wuhan-Hu-1 and EPI_ISL_418227 (Twist Bioscience, San Francisco, CA). We mixed the two RNAs at various copy numbers (10^5, 10^4, 10^3, 10^2 genome copies/μL) and frequencies (0%, 0.25%, 0.5%, 1%, 2%, 5%, 10%, and 100%). We amplified and sequenced each RNA mixture as described above. We defined true positive iSNV as mutations at the seven sites that differ between the two synthetic RNA controls (C241U, C335U, C2416U, C3037U, C14408U, A23403G, G25563U). We defined false positives as any iSNV other than the seven true-positive mutations.

Viral load measurements

We measured SARS-CoV-2 genome copy concentration for each sample by qPCR using conditions outlined in the CDC 2019-Novel Coronavirus EUA protocol (https://www.fda.gov/media/134922/download). The nucleocapsid gene was amplified using the CDC N1 primer and probe set as follows: 2019-nCoV_N1 Forward Primer GACCCCAAAATCAGCGAAAT; 2019-nCoV_N1 Reverse Primer TCTGGTTACTGCCAGTTGAATCTG; 2019-nCoV_N1 Probe ACCCCGCATTACGTTTGGTGGACC. Probe sequences were FAM labeled with Iowa Black quencher (Integrated DNA Technologies, Coralville, IA). Reactions were performed using TaqPath 1-step RT-qPCR master mix (Thermofisher, Waltham, MA) with 500 nM of each primer and 250 nM of each probe in a total reaction volume of 20 μl. Cycling conditions were as follows: 2 min at 25˚C, 15 min at 50˚C, 2 min at 95˚C, and 45 cycles of 3 seconds at 95˚C, 30 seconds at 55˚C. Samples were run on an Applied Biosystems 7500 FAST real-time PCR system. Cycle threshold (Ct) was designated uniformly across PCR runs.

Standard curves based on serial dilutions of a plasmid containing the nucleocapsid sequence were used to determine copy number for each plate of samples. Copy number is expressed in genome copies per microliter of extracted viral RNA.

Analysis of sequence reads

We aligned reads to the MN908947.3 reference genome with BWA-MEM version 0.7.15 [40]. We removed sequencing adaptors and trimmed ARTIC primer sequences with iVar 1.2.1 [23]. We determined the consensus sequences with iVar 1.2.1, taking the most common base as the consensus (>50% frequency). We placed an N at positions along the MN908947.3 reference with fewer than 10 reads. We manually inspected insertions and deletions by visualizing alignments with IGV (version 2.8.0) [41]. We identified intrahost single nucleotide variants relative to the MN908947.3 reference genome with iVar 1.2.1 using the following parameters: sample with viral load $\geq 10^3$ copies/μL; sample with consensus genome length of ≥ 29000; sample with $\geq 80\%$ of genome sites above 200x coverage; iSNV frequency of 2–50%; read depth of ≥ 100 at iSNV sites; ≥ 10 reads with average Phred score of > 35 supporting a given iSNV; iVar p-value of < 0.0001. All samples on which we called variants had $> 50{,}000$ mapped reads. We accounted for strand bias by performing a two-sided Fisher's exact test for hypothesis that the forward/reverse strand counts supporting the variant base are derived from the same distribution as the consensus base. We then applied a Bonferroni multiple test correction and excluded variants with an adjusted p-value < 0.05. We used a multiple linear model to evaluate the correlation of sample iSNV richness to day post symptom onset and viral load (base 10 log). To generate a phylogenetic tree, we aligned consensus genomes with MUSCLE 3.8.31 and masked positions that are known to commonly exhibit homoplasies or sequencing errors [42]. We generated a maximum likelihood phylogeny with IQ-TREE, using a GTR model and 1000 ultrafast bootstrap replicates [43,44]. Evolutionary lineages (Pango lineages) were assigned with PANGOLIN [45].

Supporting information

S1 Fig. True and false positive iSNV in RNA mixture validation experiment. Each iSNV is shown as a point, with the frequency on the y-axis and genome position on the x-axis. True positive iSNV are shown in violet and false positive iSNV are shown in orange. All iSNV displayed have a frequency of 2% or greater. Viral loads are shown above each facet, in units of genome copies per microliter of RNA.
(EPS)

S2 Fig. Number of minor iSNV per sample (y-axis) across groups, with hospitalized patients shown by teal points and employees shown by violet points. Boxplots for each group represent the median and 25th and 75th percentiles, with whiskers extending to the most extreme point within the range of the median ± 1.5 times the interquartile range.
(EPS)

S3 Fig. Number of minor iSNV per sample (y-axis) by genome copies per microliter of RNA (x-axis). Hospitalized patients are shown by teal points and employees shown by violet points.
(EPS)

S4 Fig. Maximum likelihood phylogenetic tree as shown in Fig 1C. Tips represent complete consensus genomes from hospitalized patients (teal) and employees (violet). The x-axis shows divergence from the root (Wuhan-Hu-1/2019). Heatmaps show samples that contain each mutation as an iSNV.
(EPS)

Acknowledgments

We thank the University of Michigan Clinical Microbiology Laboratory and the University of Michigan Central Biorepository for their assistance in providing samples. We thank Christina Cartaciano and the University of Michigan Microbiome Core for their assistance in sequencing. We thank Emily Stoneman from Michigan Medicine Occupational Health Services for assistance with employee data.

Author Contributions

Conceptualization: Joshua G. Petrie, Emily T. Martin, Adam S. Lauring.

Data curation: Andrew L. Valesano, Kalee E. Rumfelt, Derek E. Dimcheff, Joshua G. Petrie.

Formal analysis: Andrew L. Valesano, Kalee E. Rumfelt, Derek E. Dimcheff.

Funding acquisition: Joshua G. Petrie, Emily T. Martin, Adam S. Lauring.

Investigation: Andrew L. Valesano, Adam S. Lauring.

Methodology: Andrew L. Valesano, Christopher N. Blair, William J. Fitzsimmons.

Project administration: Adam S. Lauring.

Resources: Kalee E. Rumfelt, Joshua G. Petrie, Emily T. Martin.

Supervision: Joshua G. Petrie, Emily T. Martin, Adam S. Lauring.

Writing – original draft: Andrew L. Valesano, Adam S. Lauring.

Writing – review & editing: Andrew L. Valesano, Adam S. Lauring.

References

1. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, et al. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell. 2020 May 28; 181(5):990–996.e5. https://doi.org/10.1016/j.cell.2020.04.021 PMID: 32386545
2. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis [Internet]. 2020 Jul 14 [cited 2020 Aug 5];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30562-4/abstract PMID: 32679081
3. Munnink BBO, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med. 2020 Jul 16;1–6. https://doi.org/10.1038/s41591-019-0740-8 PMID: 31932805
4. Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, et al. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. Proc Natl Acad Sci [Internet]. 2020 Jul 28 [cited 2020 Jul 31]; Available from: http://www.pnas.org/content/early/2020/07/27/2006824117 https://doi.org/10.1073/pnas.2006824117 PMID: 32723824
5. Geoghegan JL, Ren X, Storey M, Hadfield J, Jelley L, Jefferies S, et al. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. Nat Commun. 2020 Dec 11; 11(1):6351. https://doi.org/10.1038/s41467-020-20235-8 PMID: 33311501
6. Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nat Commun. 2020 Nov 2; 11(1):5518. https://doi.org/10.1038/s41467-020-19248-0 PMID: 33139704
7. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. Clin Infect Dis. 2020 Jul 28; 71 (15):713–20.
8. Lauring AS. Within-Host Viral Diversity: A Window into Viral Evolution. Annu Rev Virol [Internet]. 2020 Jun 8 [cited 2020 Jun 9]; Available from: https://www.annualreviews.org/doi/10.1146/annurev-virology-010320-061642 PMID: 32511081

9. Lythgoe KA, Hall M, Ferretti L, Cesare M de, MacIntyre-Cockett G, Trebes A, et al. Within-host genomics of SARS-CoV-2. bioRxiv. 2020 Dec 10;2020.05.28.118992.

10. Gutierrez B, Escalera-Zamudio M, Pybus OG. Parallel molecular evolution and adaptation in viruses. Curr Opin Virol. 2019 Feb 1; 34:90–6. https://doi.org/10.1016/j.coviro.2018.12.006 PMID: 30703578

11. Tonkin-Hill G, Martincorena I, Amato R, Lawson ARJ, Gerstung M, Johnston I, et al. Patterns of within-host genetic diversity in SARS-CoV-2. bioRxiv. 2020 Dec 25;2020.12.23.424229.

12. Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med [Internet]. 2020 Dec 9 [cited 2020 Dec 22]; 12(573). Available from: http://stm.sciencemag.org/content/12/573/eabe2555 https://doi.org/10.1126/scitranslmed.abe2555 PMID: 33229462

13. Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. Nat Microbiol. 2020 Jul; 5(7):876–7. https://doi.org/10.1038/s41564-020-0738-5 PMID: 32427978

14. Sikkema RS, Pas SD, Nieuwenhuijse DF, O'Toole Á, Verweij J, Linden A van der, et al. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. Lancet Infect Dis [Internet]. 2020 Jul 2 [cited 2020 Jul 4];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30527-2/abstract PMID: 32622380

15. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol. 2017 Nov 15; 186(10):1209–16. https://doi.org/10.1093/aje/kwx182 PMID: 29149252

16. Maio ND, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. PLOS Comput Biol. 2018 Apr 18; 14(4):e1006117. https://doi.org/10.1371/journal.pcbi.1006117 PMID: 29668677

17. Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics. 2018 Jan 1; 34(1):163–70. https://doi.org/10.1093/bioinformatics/btx402 PMID: 29304222

18. Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. PLOS Comput Biol. 2014 Mar 27; 10(3): e1003549. https://doi.org/10.1371/journal.pcbi.1003549 PMID: 24675511

19. Martin MA, Lee RS, Cowley LA, Gardy JL, Hanage WP. Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. Microb Genomics. 2018; 4(10):e000217. https://doi.org/10.1099/mgen.0.000217 PMID: 30303479

20. McCrone JT, Lauring AS. Genetic bottlenecks in intraspecies virus transmission. Curr Opin Virol. 2018 Feb 1; 28:20–5. https://doi.org/10.1016/j.coviro.2017.10.008 PMID: 29107838

21. Zwart MP, Elena SF. Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. Annu Rev Virol. 2015 Nov 6; 2(1):161–79. https://doi.org/10.1146/annurev-virology-100114-055135 PMID: 26958911

22. McCrone JT, Lauring AS. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. J Virol. 2016 Aug 1; 90(15):6884–95. https://doi.org/10.1128/JVI.00667-16 PMID: 27194763

23. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019 Jan 8; 20(1):8. https://doi.org/10.1186/s13059-018-1618-7 PMID: 30621750

24. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host Variation and Evolutionary Dynamics of SARS-CoV-2 Population in COVID-19 Patients. bioRxiv. 2020 May 20;2020.05.20.103549.

25. Moreno GK, Braun KM, Halfmann PJ, Prall TM, Riemersma KK, Haj AK, et al. Limited SARS-CoV-2 diversity within hosts and following passage in cell culture. bioRxiv. 2020 Apr 20;2020.04.20.051011.

26. James SE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, et al. High Resolution analysis of Transmission Dynamics of Sars-Cov-2 in Two Major Hospital Outbreaks in South Africa Leveraging Intrahost Diversity. medRxiv. 2020 Nov 16;2020.11.15.20231993. https://doi.org/10.1101/2020.11.15.20231993 PMID: 33236025

27. Dimcheff DE, Valesano AL, Rumfelt KE, Fitzsimmons WJ, Blair C, Mirabelli C, et al. SARS-CoV-2 Total and Subgenomic RNA Viral Load in Hospitalized Patients. medRxiv. 2021 Mar 1;2021.02.25.21252493. https://doi.org/10.1101/2021.02.25.21252493 PMID: 33688671

28. Issues with SARS-CoV-2 sequencing data [Internet]. Virological. 2020 [cited 2020 Dec 23]. Available from: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473

29. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020 Sep 1; 83:104351. https://doi.org/10.1016/j.meegid.2020.104351 PMID: 32387564

30. Valesano AL, Taniuchi M, Fitzsimmons WJ, Islam MO, Ahmed T, Zaman K, et al. The Early Evolution of Oral Poliovirus Vaccine Is Shaped by Strong Positive Selection and Tight Transmission Bottlenecks. Cell Host Microbe [Internet]. 2020 Nov 18 [cited 2020 Dec 23];0(0). Available from: https://www.cell.com/cell-host-microbe/abstract/S1931-3128(20)30574-6 PMID: 33212020

31. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. 2020 Apr 15;1–4. https://doi.org/10.1038/s41591-019-0740-8 PMID: 31932805

32. Rhee C, Kanjilal S, Baker M, Klompas M. Duration of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infectivity: When Is It Safe to Discontinue Isolation? Clin Infect Dis [Internet]. [cited 2020 Dec 28]; Available from: http://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa1249/5896916 PMID: 33029620

33. Baang JH, Smith C, Mirabelli C, Valesano AL, Manthei DM, Bachman MA, et al. Prolonged Severe Acute Respiratory Syndrome Coronavirus 2 Replication in an Immunocompromised Patient. J Infect Dis. 2021 Jan 4; 223(1):23–7. https://doi.org/10.1093/infdis/jiaa666 PMID: 33089317

34. Buckland MS, Galloway JB, Fhogartaigh CN, Meredith L, Provine NM, Bloor S, et al. Treatment of COVID-19 with remdesivir in the absence of humoral immunity: a case report. Nat Commun. 2020 Dec 14; 11(1):6385. https://doi.org/10.1038/s41467-020-19761-2 PMID: 33318491

35. Kemp S, Harvey W, Datir R, Collier D, Ferreira I, Carabelii A, et al. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. bioRxiv. 2020 Dec 21;2020.12.14.422555.

36. Xue KS, Moncla LH, Bedford T, Bloom JD. Within-Host Evolution of Human Influenza Virus. Trends Microbiol. 2018 Sep 1; 26(9):781–93. https://doi.org/10.1016/j.tim.2018.02.007 PMID: 29534854

37. Wang D, Wang Y, Sun W, Zhang L, Ji J, Zhang Z, et al. Population Bottlenecks and Intra-host Evolution during Human-to-Human Transmission of SARS-CoV-2. bioRxiv. 2020 Jun 26;2020.06.26.173203.

38. Braun KM, Moreno GK, Halfmann PJ, Baker DA, Boehm EC, Weiler AM, et al. Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. bioRxiv. 2020 Nov 17;2020.11.16.384917. https://doi.org/10.1101/2020.11.16.384917 PMID: 33236011

39. Xue KS, Bloom JD. Reconciling disparate estimates of viral genetic diversity during human influenza infections. Nat Genet. 2019 Feb 25; 1. https://doi.org/10.1038/s41588-019-0349-3 PMID: 30804564

40. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. 2013 May 26 [cited 2020 Sep 3]; Available from: http://arxiv.org/abs/1303.3997

41. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan; 29(1):24–6. https://doi.org/10.1038/nbt.1754 PMID: 21221095

42. Masking strategies for SARS-CoV-2 alignments [Internet]. Virological. 2020 [cited 2020 Dec 23]. Available from: https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480

43. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 1; 32(5):1792–7. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

44. 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 Jan 1; 32(1):268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

45. 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 Jul 15;1–5. https://doi.org/10.1038/s41564-019-0652-x PMID: 31857732