Title

Limited within-host diversity and tight transmission bottlenecks limit SARS-CoV-2 evolution in acutely infected individuals

One Sentence Summary

Patterns of SARS-CoV-2 within hosts suggest efficient selection and transmission of novel variants is unlikely during typical, acute infection.

Authors

Katarina Braun[1]*, Gage Moreno[3]*, Cassia Wagner[2], Molly A. Accola[5], William M. Rehrauer[5], David Baker[3,4], Katia Koelle[6], David H. O'Connor[3,4], Trevor Bedford[2], Thomas C. Friedrich[1#], Louise H. Moncla[2#]

Affiliations

[1]Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, WI, United States of America

[2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

[3]Department of Pathology and Laboratory Medicine, University of Wisconsin-Madison, Madison, WI, United States of America

[4]Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, WI, United States of America

[5]University of Wisconsin School of Medicine and Public Health, Madison, WI, United States of America and the William S. Middleton Memorial Veterans Hospital

[6]Department of Biology, Emory University, Atlanta, GA, United States of America

*These authors contributed equally

[#]Co-corresponding

Abstract

The recent emergence of divergent SARS-CoV-2 lineages has raised concerns about the role of selection within individual hosts in propagating novel variants. Of particular concern are variants associated with immune escape and/or enhanced transmissibility. Though growing evidence suggests that novel variants can arise during prolonged infections, most infections are acute. Understanding the extent to which variants emerge and transmit among acutely infected hosts is therefore critical for predicting the pace at which variants resistant to vaccines or conferring increased transmissibility might emerge in the majority of SARS-CoV-2 infections. To characterize how within-host diversity is generated and propagated, we combine extensive laboratory and bioinformatic controls with metrics of within- and between-host diversity to 133 SARS-CoV-2 genomes from acutely infected individuals. We find that within-host diversity during acute infection is low and transmission bottlenecks are narrow, with very few viruses founding most infections. Within-host variants are rarely transmitted, even among individuals within the same household. Accordingly, we also find that within-host variants are rarely

detected along phylogenetically linked infections in the broader community. Together, these findings suggest that efficient selection and transmission of novel SARS-CoV-2 variants is unlikely during typical, acute infection.

Introduction

The recent emergence of variants of concern has spurred uncertainty about how severe acute respiratory coronavirus 2 (SARS-CoV-2) will evolve in the longer term. SARS-CoV-2 acquires a fixed consensus mutation approximately every 11 days as it replicates in a population *(1)*. Recently, however, lineages of SARS-CoV-2 have arisen harboring more variants than expected based on this clock rate, with some variants conferring enhanced transmissibility and/or antibody escape *(2, 3)*. The emergence of these lineages has raised concern that SARS-CoV-2 may rapidly evolve to evade vaccine-induced immunity, and that vaccines may need to be frequently updated. A current leading hypothesis posits that these lineages may have emerged during prolonged infections. Under this hypothesis, longer infection times, coupled with antibody selection *(4)*, may allow more time for novel mutations to be generated and selected before transmission. Studies of SARS-CoV-2 *(4–8)* and other viruses *(9, 10)* support this hypothesis. Longitudinal sequencing of SARS-CoV-2 from immunocompromised or persistently infected individuals accordingly reveals an accumulation of single-nucleotide variants (iSNVs) and short insertions and deletions (indels) during infection *(4–6, 11)*. In influenza virus and norovirus infections, variants that arose in immunocompromised patients were later detected globally, suggesting that long-term infections may mirror global evolutionary dynamics *(9, 12)*. Mutations defining novel variant lineages resulting in enhanced transmissibility and/or immune escape in SARS-CoV-2 Spike, like Δ69/70, N501Y and E484K, have already been documented arising in persistently infected and immunocompromised individuals *(4, 5)*.

While prolonged infections occur, the vast majority of SARS-CoV-2 infections are acute *(13)*. Viral evolutionary capacity is limited by the duration of infection *(14)*, and it is not yet clear whether the evolutionary patterns observed during prolonged SARS-CoV-2 infections also occur in acutely infected individuals. Replication-competent virus has rarely been recovered from individuals with mild to moderate coronavirus disease 2019 (COVID-19) beyond ~10 days following symptom onset *(15, 16)*. Multiple studies of influenza viruses show that immune escape variants are rarely detected during acute infection, even within vaccinated individuals *(17–19)*. Detailed modeling of influenza dynamics suggests that the likelihood of within-host mutation emergence depends on the interplay of immune response timing, the de-novo mutation rate, and the number of virus particles transmitted between hosts *(14)*. Understanding the speed with which SARS-CoV-2 viruses acquire novel mutations that may escape population immunity will be critical for formulating future vaccine updates. If novel immune-escape variants emerge primarily within long-term infections, then managing long-term infections in an effort to reduce any onward transmission may be critically important. Conversely, if novel variants are efficiently selected and transmitted during acute infections, then vaccine updates may need to occur frequently.

While understanding the process of variant generation and transmission is critically important, a clear consensus on how frequently variants are shared and transmitted between individuals has been elusive. Estimates of SARS-CoV-2 diversity within hosts have been highly variable, and comparing results among labs has been complicated by sensitivity to variant-calling thresholds and inconsistent laboratory controls *(20–23)*. Some data suggest that SARS-CoV-2 genetic diversity within individual hosts during acute infections is limited *(20, 24)* and shaped by genetic drift and purifying selection *(21, 25–27)*. Estimates of the size of SARS-CoV-2 transmission bottlenecks *(21, 28, 29)* have ranged considerably, and recent validation work has shown that estimates of within-host diversity and transmission bottleneck sizes are highly sensitive to

sequencing protocols and data analysis parameters, like the frequency cutoff used to define/identify within-host variants *(20, 30)*. Clarifying the extent to which within-host variants arise and transmit among acutely infected individuals, while controlling for potential error, will be critical for assessing the speed at which SARS-CoV-2 evolves and adapts.

To characterize how within-host variants are generated and propagated, we employ extensive laboratory and bioinformatic controls to characterize 133 SARS-CoV-2 samples collected from acutely-infected individuals in Wisconsin, United States. By comparing patterns of intrahost single nucleotide variants (iSNVs) to densely-sampled consensus genomes from the same geographic area, we paint a clear picture of how variants emerge and transmit within communities and households. We find that overall within-host diversity is low during acute infection, and that iSNVs detected within hosts almost never become dominant in later-sampled sequences. We find that iSNVs are infrequently transmitted, even between members of the same household, and we estimate that transmission bottlenecks between putative household pairs are narrow. This suggests that most iSNVs are transient and very rarely transmit beyond the individual in which they have originated. Our results imply that during typical, acute SARS-CoV-2 infections, the combination of limited intrahost genetic diversity and narrow transmission bottlenecks may slow the pace by which novel variants arise, are selected, and transmit onward. Finally, most individual infections likely play a minor role in SARS-CoV-2 evolution, consistent with the hypothesis that novel variants are more likely to arise in rare instances of prolonged infection.

Results

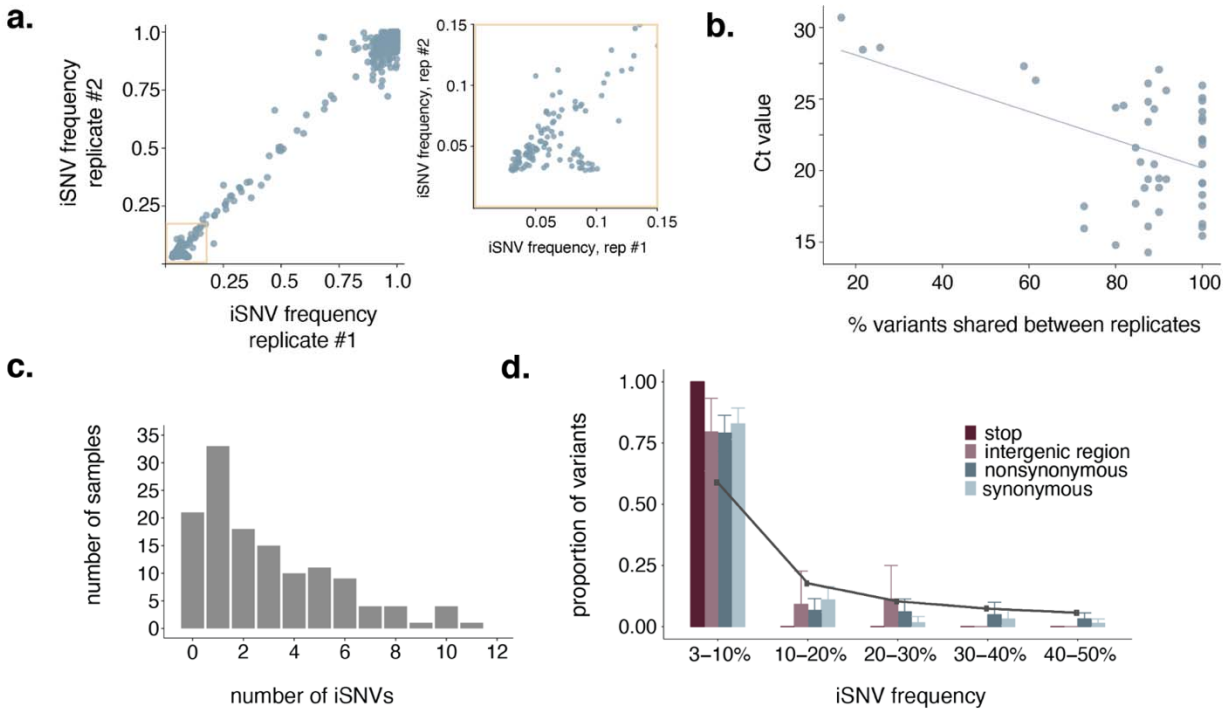
Within-host variation is limited and sensitive to iSNV-calling parameters

Viral sequence data provide rich information about how variants emerge within, and transmit beyond, individual hosts. Viral nucleotide variation generated during infection provides the raw material upon which selection can act. However, viral sequence data are sensitive to multiple sources of error *(20, 22, 23)*, which has obscured easy comparison among existing studies of SARS-CoV-2 within-host evolution. Here, we take several steps to minimize sources of error and to assess the robustness of our results against variable within-host single nucleotide variant (iSNV)-calling parameters.

First, we identified spurious iSNVs introduced by our library preparation pipeline by sequencing in duplicate a clonal, synthetic RNA transcript identical to our reference genome (MN90847.3). We considered only variants found in both technical replicates, which we refer to as "intersection iSNVs". We detected 7 intersection iSNVs at ≥1% frequency (**Supplemental Table 1**); 2 of these were previously identified by a similar experiment in Valesano et al. *(20)*. We excluded all 7 of these iSNVs from downstream analyses. To exclude laboratory contamination, we sequenced a no-template control (water) with each large sequencing batch and confirmed that these negative controls contained <10x coverage across the SARS-CoV-2 genome (**Supplemental Figure 1, Supplemental Figure 2)**. To ensure that spurious variants were not introduced by our bioinformatic pipelines, we validated our iSNV calls using a second pipeline which employs distinct trimming, mapping, and variant calling softwares. We found near-

129 equivalence between the two pipelines' iSNV calls (R^2=0.998; **Supplemental Figure 3a**),
130 providing additional independent support for our bioinformatic pipeline to accurately call iSNVs.
131
132 Viral iSNV calls are also sensitive to the variant-calling threshold (i.e., a minimum frequency at
133 which iSNVs must occur to be considered non-artefactual) applied *(22)* and the number of viral
134 input copies. Work by Grubaugh et al. *(31)* showed highly accurate iSNV calls with tiled
135 amplicon sequencing using technical replicates and a 3% frequency threshold. Consistent with
136 this observation, we observed a near-linear correlation between iSNVs called in each replicate
137 at a 3% frequency threshold (R^2=0.992) (**Figure 1a**). Unsurprisingly, we find the proportion of
138 intersection iSNVs compared to all iSNVs within a given sample increases as the frequency
139 threshold increases (**Supplemental Figure 3b**). Additionally, the majority of iSNVs detected in
140 our clonal RNA controls occur <3% frequency (**Supplemental Figure 3c**).
141
142 Consistent with previous studies, we observed a negative correlation between Ct and the
143 overlap in variants between replicates such that high-Ct (i.e., low vRNA copy number) samples
144 had fewer intersection iSNVs called in each replicate (**Figure 1b**) *(22, 31)*. Although we do not
145 have access to absolute quantification for viral input copies for our sampleset, we can use
146 results of semi-quantitative clinical assays on the sequenced specimens as a proxy for viral
147 RNA (vRNA) concentration. Using input data from two different clinical assay platforms, we find
148 no correlation between viral input copies and the number of intersection iSNVs detected
149 (**Supplemental Figure 3d** and **Supplemental Figure 3e**).
150
151 Based on these observations, we chose to use a 3% iSNV frequency cutoff for all downstream
152 analyses, and report only iSNVs that were detected in both technical replicates, at a frequency
153 ≥3%. Using these criteria, we found limited SARS-CoV-2 genetic diversity in most infected
154 individuals: 22 out of 133 samples did not harbor even a single intersection iSNV at ≥3%

155 frequency. Among the 111 samples that did harbor within-host variation, the average number of
156 iSNVs per sample was 3.5 (median=3, range=1-11) (**Figure 1c**). Most iSNVs were detected at
157 <10% frequency (**Figure 1d**). Compared to expectations under a neutral model, every type of
158 mutation we evaluated (synonymous, nonsynonymous, intergenic region, and stop) was present
159 in excess at low frequencies, consistent with purifying selection or population expansion within
160 the host (**Figure 1d**). Taken together, our results confirm that the number of iSNVs detected
161 within-host are dependent on variant-calling criteria. Once rigorous laboratory and bioinformatic
162 controls are applied, we find that most infections are characterized by very few iSNVs, and
163 primarily low-frequency variants.



164

165 **Figure 1: Within host variation is limited after data quality control**

166 **a.** iSNV frequencies in replicate 1 are shown on the x-axis and frequencies in replicate 2 are shown on y-
167 axis. The yellow box highlights low-frequency iSNVs (3-15%), which is expanded out to the right. **b.** The
168 Ct value is compared to the percent of iSNVs shared between technical replicates. The blue line is a line
169 of best fit to highlight the observed negative trend. **c.** Distribution of the number of total iSNVs detected
170 per sample. Many samples harbor no iSNVs at all, and the maximum number of iSNVs in a single sample

was 11. **d.** The proportion of iSNVs that were detected at various within-host frequency bins is shown. Error bars represent the variance in the proportion of total within-host iSNVs within that frequency bin across samples in the dataset as calculated by bootstrapping. There was a single stop variant in the entire dataset, so no error bar is shown for the stop category. The solid grey line indicates the expected proportion of variants in each frequency bin under a neutral model.

Recurrent iSNVs consist of Wuhan-1 reversions and common polymorphic sites

Previous studies of SARS-CoV-2 evolution have noted the unusual observation that iSNVs are sometimes shared across multiple samples. Understanding the source and frequency of shared iSNVs is important for measuring the size of transmission bottlenecks and for identifying potential sites of selection. In our dataset, most iSNVs were unique to a single sample (**Figure 2a**). However, 41 iSNVs were detected in at least 2 samples. These “shared iSNVs” were detected across multiple sequencing runs (**Supplemental Figure 5**), and were absent in our negative controls, suggesting they are unlikely to be artefacts of method error. Most of the shared iSNVs we detect fall into two categories: iSNVs that occur within or adjacent to a homopolymer region (8/41 iSNVs, **Figure 2b**, yellow and purple bars), or iSNVs that represent “Wuhan-1 reversions” (31/41 iSNVs, **Figure 2b**, blue and purple bars). iSNVs in or near homopolymer regions were defined as those that fall within or one nucleotide outside of a span of at least 3 identical nucleotide bases. Shared iSNVs were more commonly detected in A/T hompolymer regions than in G/C homopolymer regions. We classified iSNVs as “Wuhan-1 reversions” when a sample’s consensus sequence had a near-fixed variant (50-97% frequency) relative to the Wuhan-1 reference, with the original Wuhan-1 nucleotide present as an iSNV. Overall, this suggests that shared variants in our dataset may be at least partially explained by

viral polymerase incorporation errors, potentially in A/T-rich regions, and at sites that are frequently polymorphic.

The most commonly detected iSNVs in our dataset represent Wuhan-1 reviersion at nucleotide sites 241 (detected 18 times; within/adjacent to a homopolymer region) and 3037 (detected 21 times; not in a homopolymer region). Both of these sites are polymorphic deep in the SARS-CoV-2 phylogeny near the branch point for clade 20A (Nextstrain clade nomenclature). Within-host polymorphisms at sites 241 and 3037 were also detected in recent studies in the United Kingdom and Austria *(21, 28)*. T241C and T3037C are both synonymous variants, and have emerged frequently on the global SARS-CoV-2 phylogenetic tree, suggesting that these sites may be frequently polymorphic within and between hosts across multiple geographic areas (**Figure 2c**).

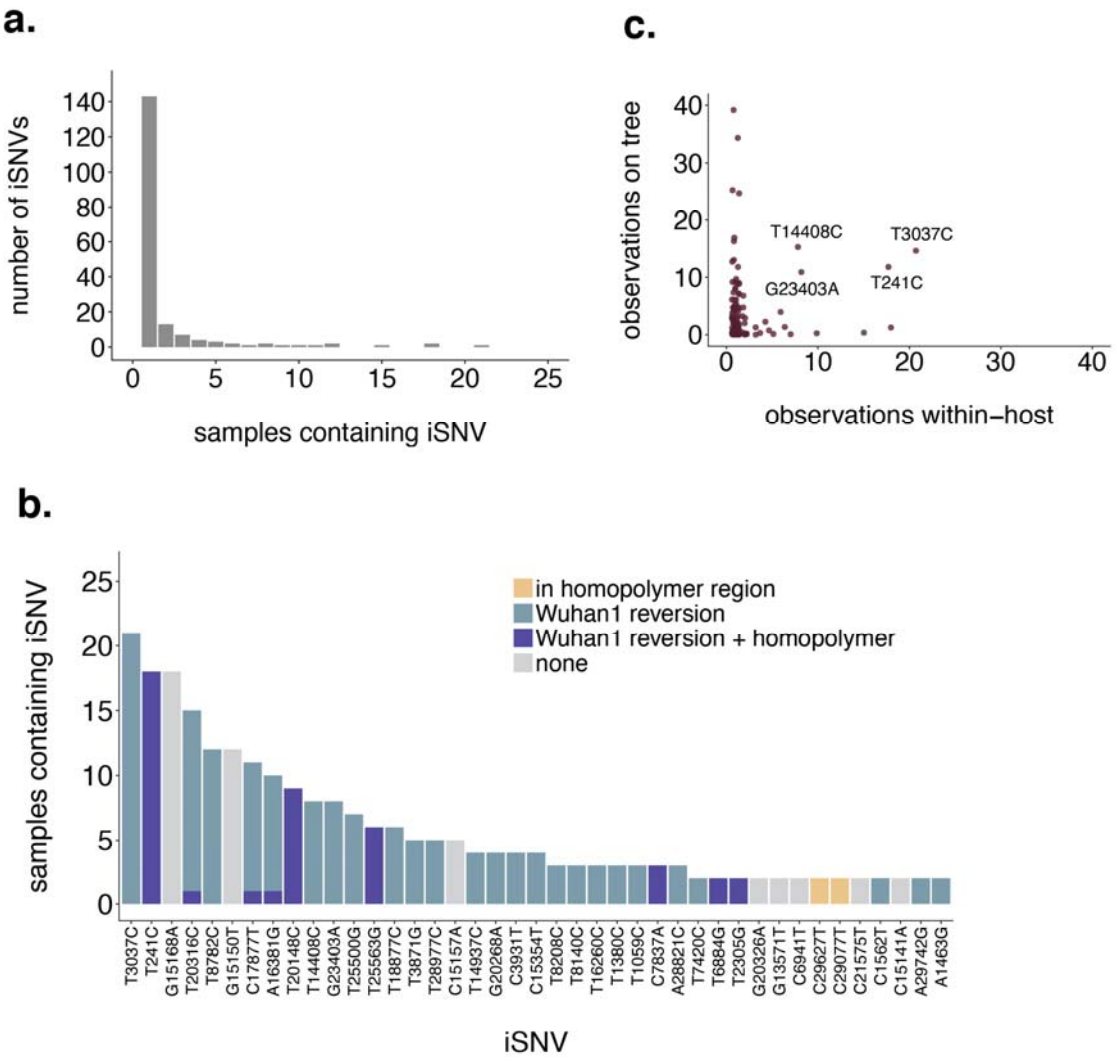


Figure 2: Shared iSNVs represent homopolymers and common polymorphic sites

a. The number of iSNVs (y-axis) present within n individuals (x-axis) is shown. The vast majority of iSNVs are found in only a single sample. 6 iSNVs are shared by at least 10 samples. **b.** Each iSNV detected in at least 2 samples is shown. Variants that occur within, or 1 nucleotide outside of, a homopolymer region (classified as a span of the same base that is at least 3 nucleotides long) are colored in yellow. Variants that represent the minor allele for variants that were nearly fixed at consensus (annotated here as "Wuhan1 reversions") are shown in blue, and variants that were both Wuhan1 reversions and occurred in homopolymer regions are colored in purple. **c.** For each unique iSNV detected within a host, the x-axis represents the number of samples in which that iSNV was detected, and the y-axis represents the

number of times it is present on the global SARS-CoV-2 phylogenetic tree. The counts on the phylogenetic tree represent the number of times the mutation arose along internal and external branches. The variants labeled with text are those that are detected at least 10 times within-host and at least 10 times on the phylogeny. Two of the most commonly detected iSNVs, T3037C and T241C (shown as the furthest to the left in panel b), are also frequently detected on the phylogenetic tree.

Most within-host variation does not contribute to consensus diversity

The emergence of divergent SARS-CoV-2 lineages has raised concerns that new variants may be selected during infection and efficiently transmitted onward. We next sought to characterize whether iSNVs arising within hosts contribute to consensus diversity sampled later in time. Using the Wisconsin-specific phylogenetic tree (**Supplemental Figure 6**), we queried whether iSNVs detected within hosts are ever found at consensus in tips sampled downstream. For each Wisconsin tip that lay on an internal node and for which we had within-host data, we traversed the tree from that tip to each subtending tip. We then enumerated each mutation that occurred along that path, and compared whether any mutations that arose on downstream branches matched iSNVs detected within-host (see **Figure 3a** for a schematic). Of the 110 Wisconsin tips harboring within-host variation, 93 occurred on internal nodes. Of those, we detect only a single instance in which an iSNV detected within a host was later detected at consensus. C1912T (a synonymous variant) was present in USA/WI-UW-214/2020 at ~4% frequency, and arose on the branch leading to USA/WI-WSLH-200068/2020 (**Figure 3b**). USA/WI-UW-214/2020 is part of a large polytomy, so this does not necessarily suggest that USA/WI-UW-214/2020 and USA/WI-WSLH-200068/2020 fall along the same transmission chain. These results indicate that despite relatively densely sampling consensus genomes from related viruses from Wisconsin, we do not find evidence that iSNVs frequently rise to consensus along phylogenetically linked infections.

If iSNVs arising during infection are adaptive and efficiently transmitted, then they should be found frequently in consensus genomes, and may be enriched on internal nodes of the phylogenetic tree. For each within-host variant detected in our dataset, we queried the number of times it occurred on the global SARS-CoV-2 phylogeny on tips and internal nodes. We then compared the ratio of detections on tips vs. internal nodes to the overall ratio of mutations on tips vs. internal nodes on the phylogeny. 42% (77/185) of iSNVs are present at least once at consensus level on the global phylogeny (**Supplemental Figure 7**). When present, iSNVs from our dataset that also occur in consensus genomes on the global tree tend to be rare, and predominantly occur on terminal nodes (**Figure 3c, Supplemental Figure 7**). Overall, iSNVs that are also found at consensus are present on internal nodes and tips at a ratio similar to that of consensus mutations overall (ratio of mutations on phylogeny nodes:tips = 4,637:17,200; ratio of iSNVs on nodes:tips = 128:411, p=0.16, Fisher's exact test). Although this is the predominant pattern, we detect one exception. C28887T is present in one sample in our dataset at a frequency of ~6%, but is found on 10 internal nodes and 15 tips (p = 0.028, Fisher's exact test) (**Figure 3c**). C28887T encodes a threonine-to-isoleucine change at position 205 in the N protein, and is a clade-defining mutation for the B.1.351 lineage. Although the functional impact of this mutation is not completely understood, N T205I may increase stability of the N protein *(32, 33)*. Despite the detection within-host and subsequent emergence of N205I globally, this iSNV was only detected in our dataset in one sample at low frequency. In general, across our dataset, the frequency with which iSNVs were detected within-host vs. on the phylogenetic tree is not correlated (**Figure 2c**). This suggests that although putative functional mutations may arise within a host, these events are rare. iSNV detection within a host, at least in typical acute infections, may therefore have limited utility for predicting future variant emergence. Together, these data suggest that with rare exception, most within-host variants are purged over time, and typically do not contribute to consensus-level diversity sampled later in time. As such, these

findings suggest that most iSNVs are not selectively beneficial and are not efficiently transmitted.

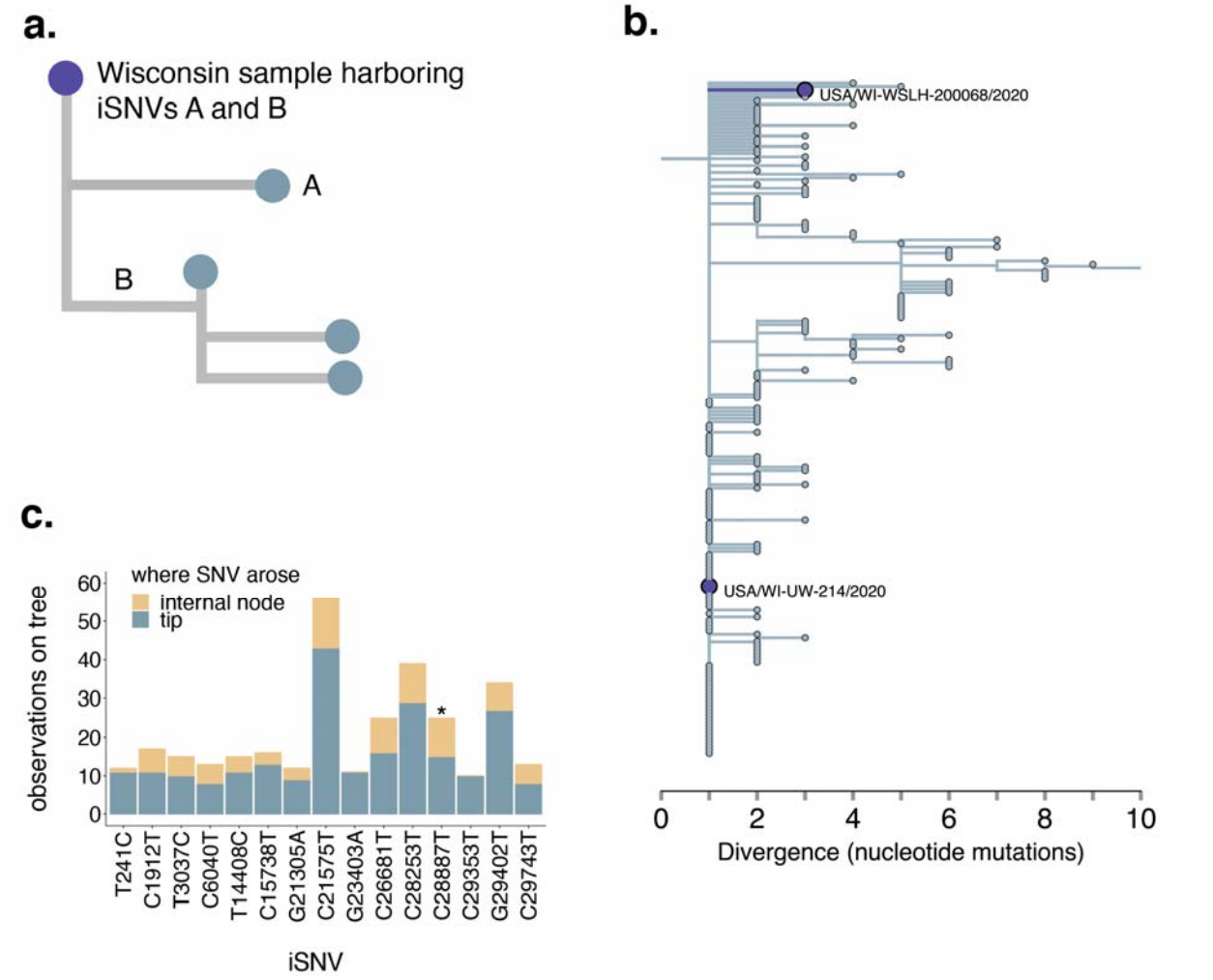


Figure 3: Variants are not common in consensus sequences or in downstream branches

a. We traversed the Wisconsin-focused full-genome SARS-CoV-2 phylogeny from root to tip. For each Wisconsin tip for which we had within-host data, we queried whether any of the iSNVs detected in that sample were ever detected in downstream branches at consensus. In this example, the purple tip represents a Wisconsin sample for which we have within-host data. This sample harbors 2 iSNVs, A and B. iSNV A arises on a tip that falls downstream from the starting, purple tip. iSNV B is present on a downstream branch leading to an internal node. Both A and B would be counted as instances in which an iSNV was detected at consensus in a downstream branch. **b.** In the Wisconsin-specific phylogenetic tree, we applied the metric described in **a.** Among 110 Wisconsin samples that harbored within-host variation,

93 occurred on internal nodes. Of those, we detect one instance in which a mutation detected as an iSNV in one sequence was detected in a downstream consensus sequence. (C1912T, an iSNV in USA/WI-UW-214/2020, was detected downstream in USA/WI-WSLH-200068/2020.) **c.** For each iSNV identified in the study (in at least 1 sample), we enumerated the number of times that variant occurred on the global SARS-CoV-2 phylogeny on an internal node (yellow) or on a tip (blue). The results for every variant are shown in **Supplemental Figure 6**. Here, we show only the variants that were detected at least 10 times on the global phylogeny. Each such iSNV is found at internal nodes and tips at a ratio comparable to overall mutations on the tree, except for C28887T, which is enriched on internal nodes (p=0.028, Fishers' exact test). * indicates p-value < 0.05.

Variation is shared among some household samples, but is likely insufficient for transmission resolution

Household studies provide the opportunity to investigate transmission dynamics in a setting of known epidemiologic linkage. We analyzed 44 samples collected from 19 households from which multiple individuals were infected with SARS-CoV-2. To define putative transmission pairs from our household dataset, we modeled the expected number of mutations that should differ between consensus genomes given one serial interval as previously described *(34)*(see Methods for details and rationale). We estimate that members of a transmission pair should generally differ by 0 to 2 consensus mutations (**Figure 4a**), and classify all such pairs within a household as putative transmission pairs. While most samples derived from a single household had near-identical consensus genomes, we observed a few instances in which consensus genomes differed substantially. In particular, USA/WI-UW-476/2020 differed from both other genomes from the same household by 11 mutations, strongly suggesting that this individual was independently infected.

To determine whether putative household transmission pairs shared more within-host variation than randomly sampled pairs of individuals, we performed a permutation test. We randomly sampled individuals with replacement and computed the proportion of iSNVs shared among random pairs to generate a null distribution (**Figure 4b**, grey bars). We then computed the proportion of variants shared among each putative household transmission pair. Finally, we compared the distribution of shared variants among household pairs and random pairs **(Figure 4b**). 90% of random pairs do not share any iSNVs. Although household pairs share more iSNVs than random pairs on average, half (14/28) of all household pairs share no iSNVs at all. Only 7 out of 28 of household pairs share more iSNVs than expected by chance ($p < 0.05$).

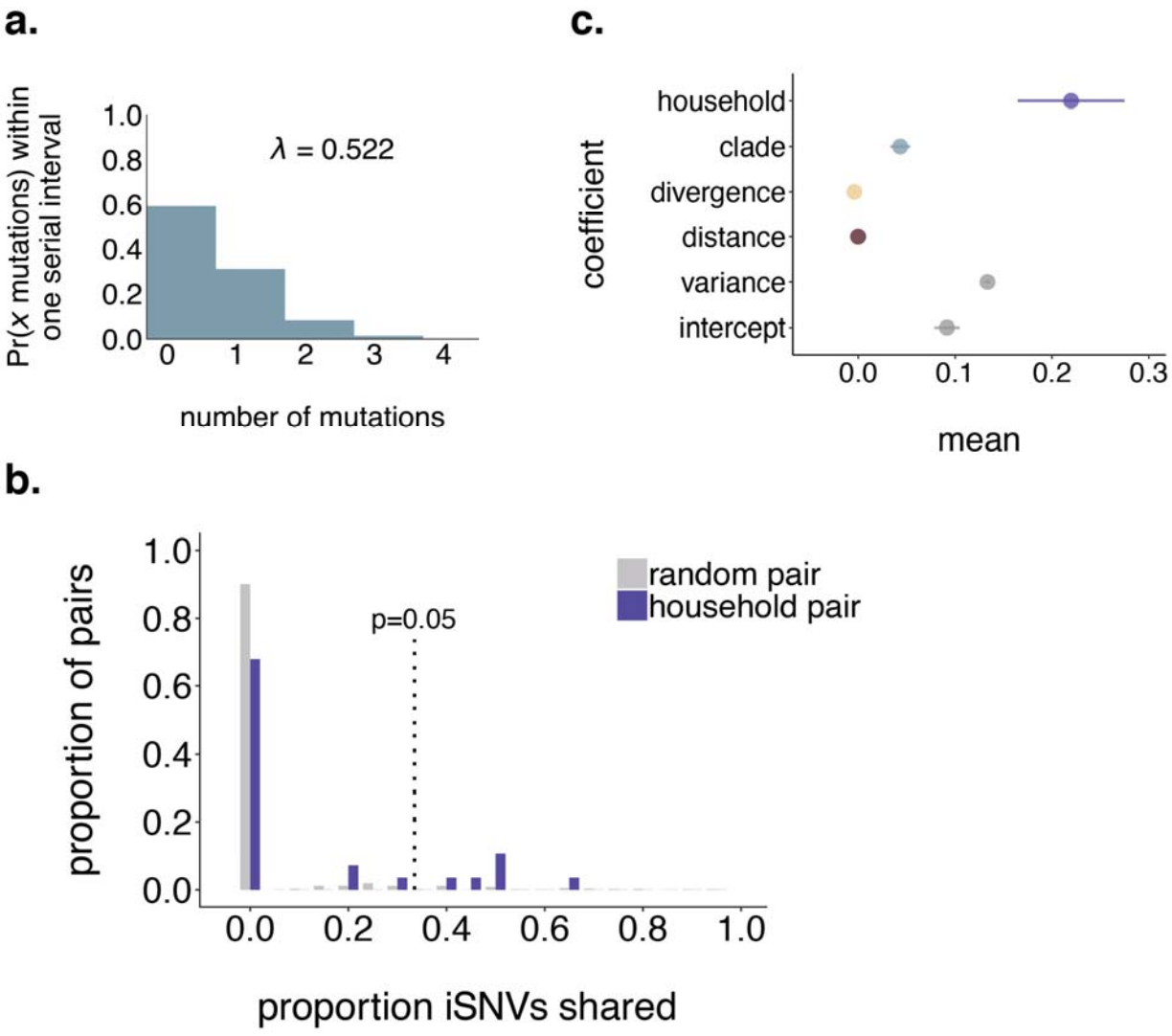


Figure 4: Household pairs share a modest degree of within-host variation

a. We modeled the probability that 2 consensus genomes will share x mutations as Poisson-distributed with lambda equal to the number of mutations expected to accumulate in the SARS-CoV-2 genome over 5.8 days *(35)* given a substitution rate of 1.10×10^{-3} substitutions per site per year *(1)*. Exploration of how these probabilities change using a range of plausible serial intervals and substitution rates is shown in **Supplemental Figure 8**. The vast majority of genomes that are separated by one serial interval are expected to differ by ≤2 consensus mutations. **b.** The proportion of random pairs (grey) and putative household transmission pairs (purple) is shown on the y-axis vs. the proportion of iSNVs shared. The dotted line indicates the 95th percentile among the random pairs. Household pairs that share a greater proportion of iSNVs than 95% of random pairs (i.e., are plotted to the right of the dotted line) are considered statistically significant at p=0.05. iSNVs had to be present at a frequency of ≥3% to be considered in this analysis. **c.** We assessed the impact of household membership, clade membership, phylogenetic divergence, and geographic distance on the proportion of iSNVs shared between each pair of samples in our dataset. The mean of each estimated coefficient in the combined linear regression model including all predictors is shown on the x-axis, with lines of spread indicating the range of the estimated 95% highest posterior density interval (HPDI).

While we hypothesized that putative transmission linkage would be the best predictor of sharing iSNVs, other processes could also result in shared iSNVs. For example, if transmission bottlenecks are wide and iSNVs are efficiently transmitted along transmission chains, then iSNVs may be propagated during community transmission. If so, then iSNVs should be shared among samples that are phylogenetically close together. If transmission chains circulate within local geographic areas, then iSNVs may be commonly shared by samples from the same geographic location. Finally, if iSNVs are strongly constrained by genetic backbone, then variants may be more likely to be shared across samples from the same clade.

To measure the contribution of these factors, we computed the proportion of iSNVs shared by each pair of samples in our dataset (including household and non-household samples), and model the proportion of shared iSNVs as the combined effect of phylogenetic divergence between the tips (i.e., the branch length in mutations between tips), clade membership, geographic distance between sampling locations, and household membership. Phylogenetic divergence and geographic distance between sampling locations have minimal predicted impact on iSNV sharing (**Figure 4c and Supplemental Figure 9**). The strongest predictor of sharing iSNVs is being sampled from the same household, which increased the predicted proportion of shared iSNVs by 0.22 (0.16 - 0.27, 95% HPDI). Belonging to the same clade increases the predicted proportion of shared iSNVs by 0.043 (0.033 - 0.053, 95% HPDI), likely because sharing a within-host variant is contingent on sharing the same consensus base. Taken together, being sampled from the same household is the strongest predictor of sharing iSNVs, and some household pairs share more variation than expected by chance. However, these effects are modest. Given the low overall diversity within hosts and presence of shared iSNVs, the degree of sharing we observe is unlikely sufficient for inferring transmission linkage independent of epidemiologic investigation.

Transmission bottlenecks are likely narrow, and sensitive to variant calling threshold

The number of viral particles that found infection is a crucial determinant of the pace at which novel, beneficial variants can emerge. Narrow transmission bottlenecks can induce a founder effect that purges low-frequency iSNVs, regardless of their fitness. Conversely, wide transmission bottlenecks result in many viral particles founding infection, reducing the chance that beneficial variants are lost. Understanding the size of the transmission bottleneck is therefore important for evaluating the probability that novel SARS-CoV-2 variants arising during

acute infection will be transmitted onward. To infer transmission bottleneck sizes, we applied the beta-binomial inference method *(36)*. We inferred transmission directionality using the date of symptom onset or date of sample collection (see methods for details). If this information was not informative, we calculated a bottleneck size bi-directionally evaluating each individual as the possible donor. In total, we performed 40 transmission bottleneck size estimates in 28 putative household pairs.

iSNV frequencies in donor and recipient pairs are plotted in **Figure 5a**. Most iSNVs detected in the donor are either lost or fixed following transmission in the recipient. However, there are a few low-frequency and near-fixed iSNVs which are shared in donor-recipient pairs. The combined maximum likelihood estimate for mean transmission bottleneck size at our defined 3% frequency threshold is 15 (95% CI: 11-21), although results vary across pairs (**Figure 5b**)**.** Prior transmission bottleneck estimates have changed based on the variant-calling threshold employed *(28, 30)*. To determine whether our estimates were sensitive to our choice of a 3% variant threshold, we evaluated bottleneck sizes using variant thresholds ranging from 1% to 20%. We estimate the highest mean transmission bottleneck size when we employ a 1% frequency threshold (38, 95% CI: 33-43), and lowest when we use a ≥7% frequency threshold (2, 95% CI: 1-4) (**Figure 5c; Supplemental Figure 10**). The finding of larger bottleneck sizes at a 1% threshold may be due to increased false-positive iSNVs at lower thresholds, in agreement with our findings that a majority of iSNVs detected in the clonal RNA control occurred at frequencies <3%. Importantly though, while variant threshold clearly impacts estimated bottleneck size, our estimates are quite consistent. Even across a wide range of thresholds, our transmission bottleneck size estimates range from 2-43, and never exceed 50.

The beta-binomial inference method assumes that shared variation in donor-recipient pairs is due to transmission. However, it is possible that shared low-frequency iSNVs are recurring

mutations (i.e. homoplasies) that should be excluded from the beta-binomial analysis. One site in particular, a synonymous change at nucleotide 15,168 in ORF1ab, was commonly found at low frequencies in donor-recipient pairs. To account for the possibility that this variant is a homoplasy rather than shared via transmission, we dropped this site from our dataset and re-calculated bottleneck sizes. While bottleneck size estimates decrease in individual pairs where this variant is found (**Supplemental Figure 10c**), the average bottleneck size across all transmission pairs remains low (mean = 9, 95% CI: 6-14).

It is possible that some of the pairs evaluated were not direct transmission pairs. Instead individuals may be part of the same transmission chain or share a common source of infection. We reasoned if two individuals were infected from a common source, then they may have developed symptoms around the same time. In contrast, if one individual infected the other, then their symptom onset dates should be staggered. To assess this, we compared bottleneck sizes to the time between symptom onset in donor-recipient pairs for which symptom onset dates were available (n=17) (**Supplemental Figure 11**). We observed no clear trend between bottleneck size and symptom onset intervals. Finally, all bottleneck estimates are inherently limited by access to a single time point from each donor and recipient. Because it is impossible to know the exact date of infection and transmission, the donor iSNV frequencies may not reflect the true diversity present at the time of transmission. Taken together, we find that even among household pairs, the number of transmitted viruses is likely small. Although bottleneck size estimates vary by variant calling threshold, we find consistent support for fewer than 50 viruses founding infection and suspect that the majority of transmission events are founded by very few viruses (<10). Our data suggest that iSNVs generated within-host are generally lost during the transmission event, and are not efficiently propagated among epidemiologically linked individuals.

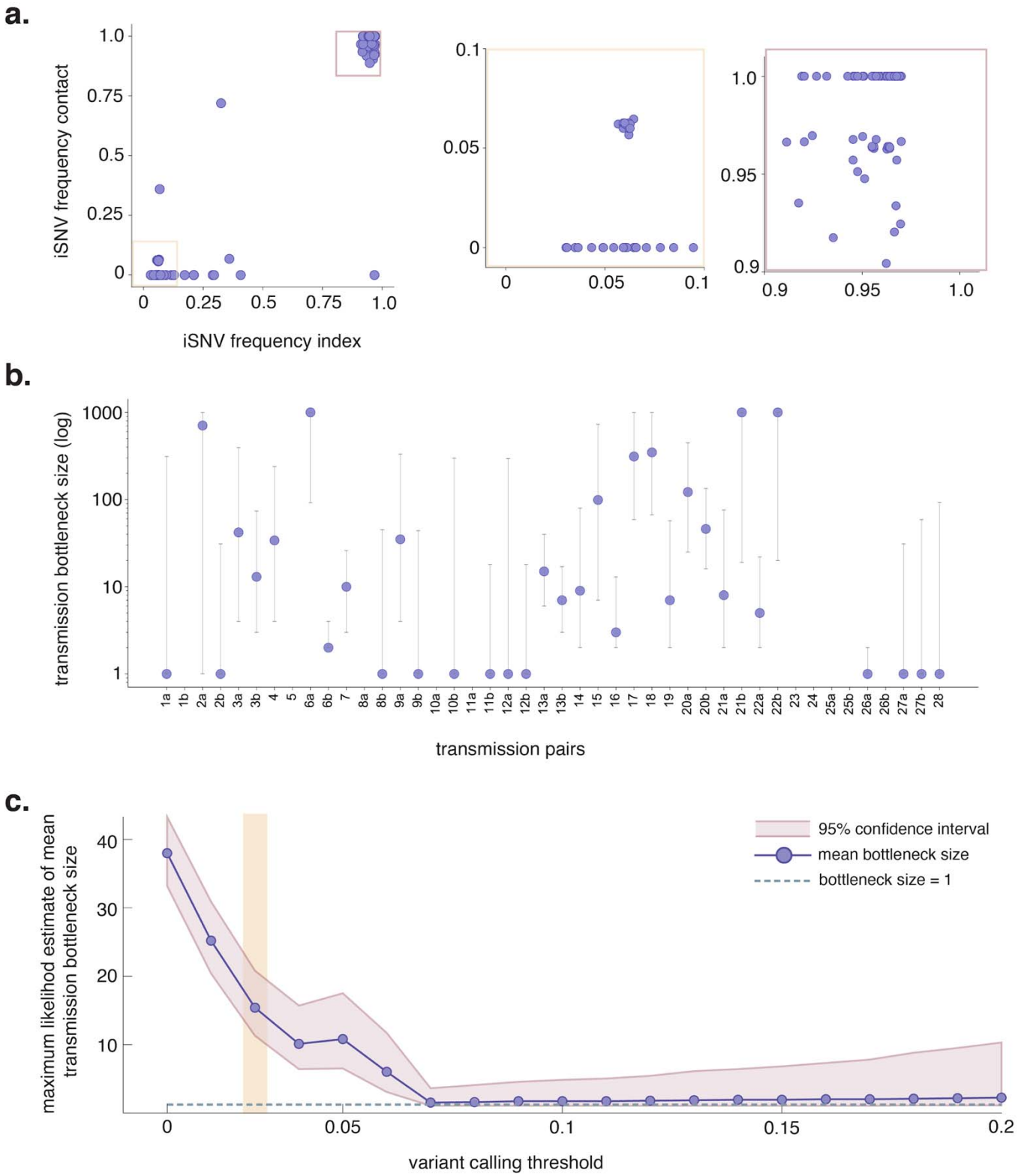


Figure 5: SARS-CoV-2 transmission bottlenecks in household transmission pairs

a. "TV plots" showing intersection iSNV frequencies in all 44 donor-recipient pairs using a 3% frequency threshold. The yellow box highlights low-frequency iSNVs (3-10%) and the mauve box highlights high-frequency iSNVs (90-100%). **b.** Maximum likelihood estimates for mean transmission bottleneck size in individual donor-recipient pairs. Bottleneck sizes could not be estimated for a few pairs (e.g. pairs 5, 10a, 11a, etc) because there were no polymorphic sites detected in the donor. **c.** Bidirectional comparisons are denoted with an "a" and "b" following the pair number. Combined maximum likelihood estimates across all 44 donor-recipient pairs plotted against variant calling thresholds ranging from 1-20%. The purple line shows combined estimates at each variant calling threshold shown and the mauve band displays the 95% confidence interval for this estimate. The dashed grey line indicates a bottleneck size equal to 1. The vertical yellow band highlights the combined transmission bottleneck size using a 3% variant calling threshold.

Discussion

The emergence of divergent SARS-CoV-2 lineages has called into question the role of within-host selection in propagating novel variants. Our results suggest that very limited variation is generated and transmitted during acute SARS-CoV-2 infection. Most infections in our dataset are characterized by fewer than 5 total intersection iSNVs, the majority of which are low-frequency. Most iSNVs are not detected in global consensus genomes, and are rarely detected in downstream branches on the phylogenetic tree. We show that even among putative household transmission pairs, iSNVs are shared infrequently, and we estimate that a small number of viruses found infection after most transmission events. The combination of low overall within-host diversity, tight transmission bottlenecks, and infrequent propagation along transmission chains may slow the rate of novel variant emergence among acutely infected individuals. Importantly, our results imply that the accumulation of multiple iSNVs is unlikely during typical, acute infection. Together, our findings are consistent with a regime in which typical acute infections play a limited role in the generation and spread of new SARS-CoV-2

variants, and argue for the need to better understand the role of prolonged infections as a source of consequential new variants. Targeted interventions to prevent the number of long-term infections and to prevent transmission from persistently infected individuals may be particularly fruitful for slowing the rate of emergence of novel variants of concern.

Relatively few studies have reported on SARS-CoV-2 within-host diversity, and their results have varied. SARS-CoV-2 within-host sequence data appear to be particularly vulnerable to method error, including sensitivity to cycle threshold *(20, 21)*, putative false positive iSNV calls in control runs *(20)*, an uncertain degree of recurrent mutations shared across unrelated samples *(21, 28, 29, 37)*, and variation between technical replicates. Complicating matters, each lab employs its own sample preparation and variant calling pipelines, making comparison across datasets challenging, and concern has been raised regarding recurrent errors that are platform- and lab-specific *(38)*. iSNVs that recur in nature pose a challenge because they result in the same data pattern that would be expected from recurrent pipeline errors. We have attempted to employ multiple, overlapping controls to mitigate errors that could arise from sample preparation, bioinformatic processing, and improper variant thresholds. In particular, our results emphasize the importance of duplicate sequencing for any studies relying on low-frequency iSNVs to infer biological processes. Like Valesano et al. *(20)*, we observe that SARS-CoV-2 variant calls are sensitive to Ct and variant-calling criteria. We echo their expressed caution in interpreting SARS-CoV-2 within-host data in the absence of pipeline-specific controls.

Similar to work reported by others *(20, 21, 37)*, we find that most samples harbor very few iSNVs, and that most variants are low-frequency. Although we employ distinct methods, we corroborate findings by Lythgoe & Hall et al. *(21)* that iSNVs do not cluster geographically or phylogenetically, suggesting that they are not transmitted efficiently within communities. One difference is that we detect a higher number of shared/recurrent iSNVs in our dataset than

reported by Lythgoe & Hall et al. *(21)*, Valesano et al. *(20)*, and Shen et al. *(37)*, but fewer than Popa & Genger et al. *(28)* and James et al. *(29)*. While some degree of shared iSNVs is reported across most SARS-CoV-2 datasets *(20, 21, 28, 29, 37)* the exact frequency of shared sites is highly variable. The higher number of shared iSNVs in our results may be partially accounted for by our method of variant reporting. While most studies mapped reads to the Wuhan-1 reference and report variants present at <50% frequency *(20, 21, 28, 37)*, we converted consensus-level variants to their low-frequency counterparts, and counted the minor allele for near-fixed variants. The higher level of shared iSNVs we observe could also be explained by sampling many closely related, cohabiting individuals. Though relatively few, some household transmission pairs do share iSNVs, likely accounting for some of the shared variation we observe. Future work will be necessary to determine the precise degree to which iSNVs recur across unrelated individuals and the extent to which factors like viral copy number, time of infection, host factors including pre-existing immunity, and sequencing pipeline influence these estimates.

Four other groups have previously estimated the size of the SARS-CoV-2 transmission bottleneck, although the total number of transmission events evaluated to date across studies remains small (~66). Lythgoe & Hall et al. (n=14 pairs) *(39)*, James & Ngcapu et al. (n=11 pairs) *(29)*, and Wang et al. (n=2 pairs) *(40)* report narrow bottlenecks, in which infection is founded by fewer than 10 viruses. Popa & Genger et al. (n=39 pairs) *(28)* report bottleneck sizes ranging from 10 to 5000, and an average size of 1000. Reanalysis of the Popa & Genger data using a more conservative variant dataset resulted in an average bottleneck size of 1-3 *(30)*. Similarly, we find a combined average bottleneck size of 15 using a 3% frequency threshold, and 2 using a 7% frequency threshold. Thus, current evidence is converging to support narrow transmission bottlenecks for SARS-CoV-2, similar to influenza virus *(18, 41, 42)*. Still, these estimates rely on a small number of putative transmission events, including the pairs analyzed here. Genuine

differences in the SARS-CoV-2 transmission bottleneck size, depending on route of transmission *(43)* and host factors may exist.

When transmission bottlenecks are narrow, even beneficial variants present at low frequencies in the transmitting host are likely to be lost. However, the recent emergence of multiple divergent lineages, some of which increase infectiousness, underscore that transmission of such variants clearly can occur *(44)*. This raises the question: how did these variants make their way out of individual hosts? Narrow transmission bottlenecks generally purge within-host diversity through a founder effect. Although rare, a low-frequency variant that successfully passes through a transmission bottleneck could quickly become the dominant variant in the next host. Such events would become increasingly common as the total number of infected individuals and transmission events occurring in the population climbs, making it possible to observe these rare events.

The model outlined above aligns with the hypothesis that prolonged SARS-CoV-2 infection leads to accumulation of intrahost mutations *(4–8)*. Prolonged infections may permit additional cycles of viral replication, allowing for more variants to be generated and more time for selection to increase the frequency of beneficial variants. Even a modest increase in frequency within a donor enhances the likelihood of a beneficial variant becoming fixed following transmission in the setting of a narrow transmission bottleneck. Alternatively, it is possible for selection to act during transmission such that some viruses harboring a particular mutation or group of mutations are preferentially transmitted *(45)*. In a previous study evaluating SARS-CoV-2 genetic diversity within and between domestic cats, we documented modest evidence supporting preferential transmission of a particular nonsynonymous variant in Spike *(25)*. However, we saw no evidence for selective bottlenecks in this study. Additional studies

evaluating the SARS-CoV-2 transmission bottleneck are needed, in particular in the setting of long-term infections and immunocompromised hosts.

Our findings that within-host variation is limited and infrequently transmitted are important. Our data, combined with findings from others, suggest that rapid accumulation of novel mutations within-host is not the norm during acute infection. Like influenza viruses, a significant portion of variation generated within one infected host is likely lost during transmission. The combination of within-host limited diversity and tight transmission bottlenecks should slow the pace at which novel, beneficial variants could emerge during transmission among acutely infected individuals. Future studies that compare within-host diversity in individuals with and without SARS-CoV-2 antibodies will be necessary to evaluate whether immunity imposes signatures of within-host selection. Finally, given the increasing appreciation for the potential role of long infections to promote variant emergence, within-host data may provide its maximum benefit for dissecting the process of variant evolution during prolonged infections.

Materials and Methods

Study design

The goal of this study was to characterize the underlying evolutionary processes acting on SARS-CoV-2 within and between hosts during acute infection, and to understand the processes that drive iSNVs to consensus level. For this purpose, isolated viral RNA from 3,351 samples (March 2020 to March 2021) was processed for broad surveillance sequencing in Wisconsin, USA. Additional analyses on a subset of samples (n=133) consisted of calling iSNVs across the genome, enumerating iSNVs along the phylogeny, and estimating the transmission bottleneck size in household transmission pairs. Samples were selected for geographic representation

across two Wisconsin counties (Dane or Milwaukee county) and to ensure all dominant phylogenetic clades in spring-summer of 2020 were represented (Nextstrain clades 19A, 19B, and 20A). In addition, we prioritized samples if more than one sample was available per household residence within a two week period.

Sample approvals and sample selection criteria

Sequences that were selected for deep sequencing and iSNV characterization were derived from 150 nasopharyngeal (NP) swab samples collected from March 2020 though July 2020. Samples originated from the University of Wisconsin Hospital and Clinics and the Milwaukee Health Department Laboratories. Submitting institutions provided a cycle threshold (Ct) or relative light unit (RLU) for all samples. Sample metadata, including GISAID and SRA accession identifiers, are available in **Supplemental Table 2**.

We obtained a waiver of HIPAA Authorization and were approved to obtain the clinical samples along with a Limited Data Set by the Western Institutional Review Board (WIRB #1-1290953-1) and the FUE IRB 2016-0605. This limited dataset contains sample collection data and county of collection. Additional sample metadata, e.g. race/ethnicity, were not shared.

Diagnostic assays for the samples included in this study were performed at the University of Wisconsin Hospital and Clinical diagnostic laboratory using CDC's diagnostic RT-PCR *(46)*, the Hologic Panther SARS-CoV-2 assay *(47)*, or the Aptima SARS-CoV-2 assay *(48)*.

Nucleic acid extraction

Viral RNA (vRNA) was extracted from 100 µl of VTM using the Viral Total Nucleic Acid Purification kit (Promega, Madison, WI, USA) on a Maxwell RSC 48 instrument and eluted in 50 µL of nuclease-free H2O.

Complementary DNA (cDNA) generation

Complementary DNA (cDNA) was synthesized according to a modified ARTIC Network approach *(49, 50)*. RNA was reverse transcribed with SuperScript IV VILO (Invitrogen, Carlsbad, CA, USA) according to manufacturer guidelines. Samples were incubated at room temperature (25°C) for 10 minutes, heated to 55°C for 10 minutes, heated to 85°C for 5 minutes, and then cooled to 4°C for 1 minute *(49, 50)*.

Multiplex PCR for SARS-CoV-2 genomes

A SARS-CoV-2-specific multiplex PCR for Nanopore sequencing was performed using the ARTIC v3 primers. Primers used in this manuscript were designed by ARTIC Network and are shown in **Supplemental Table 3**. Specifically, cDNA (2.5 µL) was amplified in two multiplexed PCR reactions using Q5 Hot-Start DNA High-fidelity Polymerase (New England Biolabs, Ipswich, MA, USA) using the following cycling conditions; 98ºC for 30 seconds, followed by 25 cycles of 98ºC for 15 seconds and 65ºC for 5 minutes, followed by an indefinite hold at 4ºC (*(49, 50)*. Following amplification, samples were pooled prior to beginning library preparations.

TruSeq Illumina library prep and sequencing for minor variants

All Wisconsin surveillance samples were prepped and sequenced by Oxford Nanopore Technologies (details below) and a subset described in this paper were additionally prepped for

584 sequencing on an Illumina MiSeq. These SARS-CoV-2 samples (n=150) consisted of household
585 pairs as well as a random sampling of the surveillance cohort selective for enhanced iSNV
586 characterization. Amplified cDNA was purified using a 1:1 concentration of AMPure XP beads
587 (Beckman Coulter, Brea, CA, USA) and eluted in 30 µL of water. PCR products were quantified
588 using Qubit dsDNA high-sensitivity kit (Invitrogen, USA) and were diluted to a final concentration
589 of 2.5 ng/µl (150 ng in 50 µl volume). Each sample was then made compatible for deep
590 sequencing using the Nextera TruSeq sample preparation kit (Illumina, USA). Specifically, each
591 sample was enzymatically end repaired. Samples were then purified using two consecutive
592 AMPure bead cleanups (0.6x and 0.8x) and were quantified once more using Qubit dsDNA
593 high-sensitivity kit (Invitrogen, USA). A non-templated nucleotide was attached to the 3′ ends of
594 each sample, followed by adaptor ligation. Samples were again purified using an AMPure bead
595 cleanup (1x) and eluted in 25 µL of resuspension buffer. Lastly, samples were indexed using 8
596 PCR cycles, cleaned with a 1:1 bead clean-up, and eluted in 30 µL of resuspension buffer. The
597 average sample fragment length and purity was determined using the Agilent High Sensitivity
598 DNA kit and the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA). After passing quality
599 control measures, samples were pooled into equimolar concentrations to a final concentration of
600 4 nM. 5 µl of each 4 nM pool was denatured in 5 µl of 0.2 N NaOH for 5 min. Sequencing pools
601 were denatured to a final concentration of 10 pM with a PhiX-derived control library accounting
602 for 1% of total DNA and were loaded onto a 500-cycle v2 flow cell. Average quality metrics were
603 recorded, reads were demultiplexed, and FASTQ files were generated on Illumina's BaseSpace
604 platform. The samples included in this study were sequenced across seven distinct MiSeq runs.
605 Each sample was library prepped and sequenced in technical replicate. Replicates were true
606 replicates in that we started from two aliquots taken from the original samples.

607 Oxford nanopore library preparation and sequencing for
608 consensus sequences

All consensus-level surveillance sequencing of SARS-CoV-2 was performed using Oxford Nanopore sequencing (n=3,351). Amplified PCR product was purified using a 1:1 concentration of AMPure XP beads (Beckman Coulter, Brea, CA, USA) and eluted in 30 µL of water. PCR products were quantified using Qubit dsDNA high-sensitivity kit (Invitrogen, USA) and were diluted to a final concentration of 1 ng/µl. A total of 5ng for each sample was then made compatible for deep sequencing using the one-pot native ligation protocol with Oxford Nanopore kit SQK-LSK109 and its Native Barcodes (EXP-NBD104 and EXP-NBD114) *(50)*. Samples were then tagged with ONT sequencing adaptors according to the modified one-pot ligation protocol *(50)*. Up to 24 samples were pooled prior to being run on the appropriate flow cell (FLO-MIN106) using the 24hr run script.

Processing raw ONT data

Sequencing data was processed using the ARTIC bioinformatics pipeline scaled up using on campus computing cores (https://github.com/artic-network/artic-ncov2019). The entire ONT analysis pipeline is available at https://github.com/gagekmoreno/SARS-CoV-2-in-Southern-Wisconsin.

Processing raw Illumina data

Raw FASTQ files were analyzed using a workflow called “SARSquencer”. The complete “SARSquencer” pipeline is available in the following GitHub repository – https://github.com/gagekmoreno/SARS_CoV-2_Zequencer. Reads were paired and merged using BBMerge (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmerge-guide/) and mapped to the Wuhan-Hu-1/2019 reference (Genbank accession MN908947.3) using BBMap (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/). Mapped reads were imported into Geneious (https://www.geneious.com/) for visual inspection. Variants

were called using callvariants.sh (contained within BBMap) and annotated using SnpEff (https://pcingola.github.io/SnpEff/). Variants were called at ≥0.01% in high-quality reads (phred score >30) that were ≥100 base pairs in length and supported by a minimum of 10 reads. The total minimum read support was set to 10 to generate initial VCF files with complete consensus genomes for the few samples where coverage fell below 100 reads in a few areas. Substantial downstream variant cleaning was performed as outlined below.

iSNV quality control

BBMap's output VCF files were cleaned using custom Python scripts, which can be found in the GitHub accompanying this manuscript (https://github.com/lmoncla/ncov-WI-within-host). First, any samples without technical replicates were excluded. This occurred due to limited sample volume, degraded RNA, or limited deep sequence reads in one or both replicates (n=5; tube/filename identifiers = 19, 188, 1049, 1064, and 1144). Next, we discarded all iSNVs which occurred at primer-binding sites (**Supplemental Table 3**). These "recoded" VCFs can be found in the GitHub repository in "data/vcfs-recode". We then filtered these recoded VCF files and for variants with (1) 100x coverage; (2) found at ≥3% frequency (more in "Within-host variation is limited once sources of sequencing error are properly accounted for"); (3) and found between nucleotides 54 and 29,837 (based on the first and last ARTIC v3 amplicon). We excluded all indels from our analysis, including those that occur in intergenic regions.

We next inspected our filtered iSNV datasets across replicate pairs. We visually inspected each replicate pair VCF and plotted replicate frequencies against each other (available in the GitHub repository). This identified a few samples which were outliers for having very limited overlap in their iSNV populations. This could be traced to low coverage or amplicon drop-out in each sample. FASTQs for these samples are available in GenBank, but we have excluded them from downstream analyses presented here (n=11; tube/filename identifier 65, 124, 125, 303, 316,

1061, 1388, 1103, 1104, 1147, and 1282) (iSNVs in technical replicates are shown for sample 1104 in **Supplemental Figure 4b**).

We generated one cleaned VCF file by averaging the frequencies found for overlapping iSNVs and discarding all iSNVs which were only found in one replicate. In addition to the SARS-CoV-2 diagnostic swabs, we sequenced a SARS-CoV-2 synthetic RNA control (Twist Bioscience, San Francisco, CA) representing the Wuhan-Hu-1 sequence (Genbank: MN908947.3) in order to identify variants which are likely to arise during library prep and sequencing. We amplified and sequenced technical replicates of this vRNA synthetic control as described above, using $1x10^6$ template copies per reaction. We then excluded variants detected in the synthetic RNA control (**Supplemental Table 4**) from all downstream analyses. Notably, this filter removed a single variant at nucleotide position 6,669 from our analysis *(20)*. Finally, within-host variants called at ≥50% and <97% frequency comprise consensus-level mutations relative to the Wuhan-Hu-1/2019 reference sequence. To ensure that the corresponding minor variant was reported we report the opposite minor allele at a frequency of 1 - the consensus variant frequency. For example, a C to T variant detected at 75% frequency relative to the Wuhan-1 reference was converted to a T to C variant at 25% frequency.

Processing of the raw sequence data, mapping, and variant calling with the Washington pipeline

To assess the sensitivity of our iSNV calls to bioinformatic pipelines, we generated VCF files using an independent bioinformatic pipeline. Raw reads were assembled against the SARS-CoV-2 reference genome Wuhan-Hu-1/2019 (Genbank accession MN908947.3; the same reference used for the alternative basecalling method) to generate pileup files using the bioinformatics pipeline available at https://github.com/seattleflu/assembly. Briefly, reads were trimmed with Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) *(51)* in paired

end mode, in sliding window of 5 base pairs, discarding all reads that were trimmed to <50 base pairs. Trimmed reads were mapped using Bowtie 2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) *(52)*, and pileups were generated using samtools mpileup (http://www.htslib.org/doc/samtools-mpileup.html). Variants were then called from pileups using varscan mpileup2cns v2.4.4 (http://varscan.sourceforge.net/using-varscan.html#v2.3_mpileup2cns). Variants were called at ≥1% frequency, with a minimum coverage of 100, and were supported by a minimum of 2 reads.

Phylogenetic analysis

All available full-length sequences from Wisconsin through February 16, 2021 were used for phylogenetic analysis using the tools implemented in Nextstrain custom builds (https://github.com/nextstrain/ncov) *(53, 54)*. Time-resolved and divergence phylogenetic trees were built using the standard Nextstrain tools and scripts *(53, 54)*. We used custom python scripts to filter and clean metadata. A custom "Wisconsin" profile was made to create a Wisconsin-centric subsampled build to include representative sequences. The scripts and output are available at https://github.com/gagekmoreno/Wisconsin-SARS-CoV-2.

Household pairs permutation test

For household groups, we performed all pairwise comparisons between members of the household, excluding pairs for which the consensus genomes differed by >2 nucleotide changes. We determined this cutoff by modeling the probability that 2 consensus genomes separated by one serial interval differ by n mutations. We model this process as Poisson-distributed with lambda equal to the expected number of substitutions per serial interval, as described previously *(34)*. We chose to model this expectation using the serial interval rather than the generation interval for the following reason. The sequence data we have represent

cases that were sampled via passive surveillance, usually from individuals seeking testing after developing symptoms. Differences in the genome sequences from two individuals therefore represent the evolution that occurred between the sampling times of those two cases. Although neither the serial interval nor the generation interval perfectly matches this sampling process, we reasoned that the serial interval, or the time between the symptom onsets of successive cases, may more accurately capture how the data were sampled. We evaluated probabilities across a range of serial interval and clock rates. For serial interval, we use the values inferred by He et al, of a mean of 5.8 days with a 95% confidence interval of 4.8-6.8 days *(35)*. For substitution rate, we employ estimates from Duchene et al, who estimate a mean substitution rate of 1.10×10^{-3} substitutions per site per year, with a 95% credible interval of 7.03×10^{-4} and 1.15×10^{-3} *(1)*. To model the expectation across this range of values, we evaluate the probabilities for serial intervals at the mean (5.8), as well as for 4, 5, 6, 7, and 8 days, and substitution rates at the mean (1.10×10^{-3}) and at the bounds of the 95% credible interval. For each combination of serial interval and substitution rate, we calculate the expected substitutions in one serial interval as: (substitution rate per site per year * genome length/365 days) *serial interval. The results using the mean serial interval (5.8 days) and substitution rate (1.10×10^{-3}) are shown in the main text, while the full set of combinations is shown in the supplement. Under this model, the vast majority of consensus genomes derived from cases separated by a single serial interval are expected to differ by ≤2 mutations. The probability that two genomes that are separated by one serial interval differ by 3 mutations ranges from 0.0016-0.059. Only in the case of an 8 day serial interval with the highest bound of the substitution rate do we infer a probability of 3 mutations that is greater than 0.05. We therefore classified all pairs of individuals from each household that differed by ≤2 consensus mutations and who were tested within 14 days of each other as putative transmission pairs.

To determine whether putative household transmission pairs shared more variants than individuals without an epidemiologic link, we performed a permutation test. At each iteration, we randomly selected a pair of samples (with replacement) and computed the proportion of variants they share as: (2 x total number of shared variants) / (the total number of variants detected among the two samples). For example, if sample A contained 5 iSNVs relative to the reference (Wuhan-1, Genbank accession MN908947.3), sample B harbored 4 iSNVs, and 1 iSNV was shared, then the proportion of sample A and B's variants that are shared would be 2/9 = 0.22. We performed 10,000 iterations in which pairs were sampled randomly to generate a null distribution. We then compared the proportion of variants shared by each putative household transmission pair to this null distribution. The proportion of variants shared by a household pair was determined to be statistically significant if it was greater than 95% of random pairs.

Transmission bottleneck calculation

The beta-binomial method, explained in detail in *(36)*, was used to infer the transmission bottleneck size N_b. N_b quantifies the number of virions donated from the index individual to the contact (recipient) individual that successfully establish lineages in the recipient that are present at the sampling time point. The method statistically incorporates sampling noise arising from a finite number of reads and accounts for the possibility of false-negative variants that are not called in the recipient host due to conservative variant calling thresholds (≥3% in both technical replicates). The beta-binomial method adopts several important assumptions. It assumes viral genetic diversity is neutral and variant frequencies are not impacted by selection; it also assumes variant sites are independent, which may not be true given that SARS-CoV-2 contains a continuous genome thought to undergo limited recombination *(55)*. In addition, the beta-binomial method assumes that identical variants found in the index and contact are shared as a result of transmission, though it is possible that identical variants occurring in a donor and a

752 recipient individual occurred independently of one another and are not linked through
753 transmission. We consider this possibility at one site in particular which commonly appears at
754 low frequencies in donor-recipient pairs. Code for estimating transmission bottleneck sizes
755 using the beta-binomial approach has been adapted from the original scripts
756 (https://github.com/koellelab/betabinomial_bottleneck) and is included in the GitHub
757 accompanying this manuscript (https://github.com/lmoncla/ncov-WI-within-host).
758
759 We calculated individual transmission bottleneck size estimates for each household
760 transmission pair as were identified in the household permutation test (n=28). We used the date
761 of symptom onset and/or date of sample collection to assign donor and recipient within each
762 pair. Within each pair, if the date of symptom onset differed by ≥3 days, we assigned the
763 individual with the earlier date as the donor. If this information was unavailable or uninformative
764 (<3 days) for both individuals in a pair, we looked at the date of sample collection and if these
765 dates differed by ≥3 days, we assigned the individual with the earlier date as the donor. If this
766 information was also not available or was not informative (<3 days), we calculated the
767 bottleneck size with each individual as a donor. These bidirectional comparisons are denoted
768 with an “a” or “b” appended to the filename (n=16 pairs were analyzed bidirectionally). In total,
769 we analyzed 44 pairs (including bidirectional comparisons). Metadata and GISAID accession
770 numbers for each pair are described in **Supplemental Table 4.**
771
772 Combined transmission bottleneck size estimates (as seen in **Figure 6c**) were estimated as
773 described in the supplemental methods in Martin & Koelle *(30)*. Briefly, overall transmission
774 bottleneck sizes were estimated based on the assumption that transmission bottleneck sizes
775 are distributed according to a zero-truncated Poisson-distribution and bidirectional bottleneck
776 estimates were each assigned 50% of the weight in this calculation compared to the

unidirectional pairs. Matlab code to replicate the combined bottleneck estimates can be found in the GitHub accompanying this paper (https://github.com/lmoncla/ncov-WI-within-host).

Enumerating mutations along the phylogeny

We used the global Nextstrain *(53)* phylogenetic tree (nextstrain.org/ncov/global) accessed on February 24, 2021 to query whether mutations detected within-host are detected on the global tree. We accessed the tree in JSON format and traverse the tree using baltic *(56)*. To determine the fraction of within-host variants detected on the phylogenetic tree, we traversed the tree from root to tip, gathering each mutation that arose on the tree in the process. For each mutation, we counted the number of times it arose on an internal and a terminal node. We then compared the fraction of times each iSNV identified within-host was detected on an internal node vs. a terminal node. To determine whether particular iSNVs were enriched at internal nodes, we compared the frequency of that iSNV's detection against the overall ratio of mutations arising on internal vs. terminal nodes in the phylogeny with a Fisher's exact test.

To query whether iSNVs ever became dominant in tips sampled downstream, we used a transmission metric developed previously *(57)*. Using the tree JSON output from the Nextstrain pipeline *(53)*, we traversed the tree from root to tip. We collapsed very small branches (those with branch lengths less than 1×10^{-16}) to obtain polytomies. For each tip for which we had within-host data that lay on an internal node, i.e., had a branch length of nearly 0 ($< 1 \times 10^{-16}$), we then determined whether any subsequent tips occurred in the downstream portion of the tree, i.e., tips that fall along the same lineage but to the right of the parent tip. We then traversed the tree and enumerated every mutation that arose from the parent tip to each downstream tip. If any mutations along the path from the parent to downstream tip matched a mutation found within-host in the parent, this was classified as a potential instance of variant transmission. A diagram of how "downstream tips" and mutations were classified is shown in **Figure 4a**.

Linear regression model

To determine the relative contributions of phylogenetic divergence, geographic distance, clade membership, and household membership to the probability of sharing within-host variants, we fit linear regression models to the data in R. As our outcome variable, we performed pairwise comparisons for each pair of samples in the dataset (including household and non-household pairs) and compute the proportion of variants shared for each pair. We then model the proportion of shared variants as the combined function of 4 predictor variables as follows: Proportion of variants shared ~ $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, where x_1 represents a 0 or 1 value for household, where a 1 indicates the same household and a 0 indicates no household relationship. X_2 denotes the divergence, i.e., the branch length in mutations between tip A and tip B as a continuous variable, x_3 indicates the great circle distance in kilometers between the location of sample collection as a continuous variable, and x_4 denotes a 0 or 1 for whether the two tips belong to the same clade (same clade coded as a 1, different clade coded as a 0). We fit a univariate model for each variable independently, a model with an intercept alone, and a combined model using the Rethinking package in R (https://www.rdocumentation.org/packages/rethinking/versions/1.59). We perform model comparison with the WAIC metric and select the combined model as the one with the best fit. We compute mean coefficient estimates and 95% highest posterior density intervals (HPDI) by sampling and summarizing 10,000 values from the posterior distribution.

Data and code availability

Consensus genomes have been deposited in GISAID with accession numbers available in **Supplemental Table 1**. Raw Illumina reads are available in the Short Read Archive under bioproject PRJNA718341. All raw Nanopore reads are available in the Short Read Archive

under bioproject PRJNA614504. All code used to analyze the data and generate the figures shown in this manuscript are available at https://github.com/lmoncla/ncov-WI-within-host.

Statistical analysis

Throughout the manuscript, we have opted to show individual data points rather than summary statistics whenever possible, and to include measures of spread for estimated variables. For the test comparing the frequency of iSNVs on internal nodes and tips on the phylogeny, we evaluate these ratios with Fisher's exact tests. To test whether putative household transmission pairs share more variants than expected by chance, we devise our own permutation test. We construct a null distribution by computing the proportion of shared iSNVs between randomly selected pairs of individuals 10,000 times, and report true pairs as sharing a statistically significant proportion of variants at an alpha of 0.05 if they fall in the upper 5% of random pairs in the null distribution. We present both the null distribution and distribution to true values, along with a line indicating the 95th percentile for completeness. For the regression analysis, we use a Bayesian implementation of multiple linear regression in R. Each predictor variable was evaluated in a univariate model as well as in the combined, multivariate mode, and models were compared using an information criterion (WAIC) that penalizes additional parameters. Estimated coefficient values, along with the estimated variance and intercept, for the multivariate model are shown as the computed mean with the 95% highest posterior density interval (HPDI) to express the spread of the results.

References and Notes

1. S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, G. Baele, Temporal signal and the phylodynamic threshold of SARS-CoV-2, *Virus Evol* **6**, veaa061 (2020).

2. Public Health England, Investigation of novel SARS-CoV-2 variant: Variant of Concern 202012/01 (2020) (available at https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201).

3. arambaut, garmstrong, isabel, Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations (2020) (available at https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563).

4. S. A. Kemp, D. A. Collier, R. P. Datir, I. A. T. M. Ferreira, S. Gayed, A. Jahun, M. Hosmillo, C. Rees-Spear, P. Mlcochova, I. U. Lumb, D. J. Roberts, A. Chandra, N. Temperton, CITIID-NIHR BioResource COVID-19 Collaboration, COVID-19 Genomics UK (COG-UK) Consortium, K. Sharrocks, E. Blane, Y. Modis, K. E. Leigh, J. A. G. Briggs, M. J. van Gils, K. G. C. Smith, J. R. Bradley, C. Smith, R. Doffinger, L. Ceron-Gutierrez, G. Barcenas-Morales, D. D. Pollock, R. A. Goldstein, A. Smielewska, J. P. Skittrall, T. Gouliouris, I. G. Goodfellow, E. Gkrania-Klotsas, C. J. R. Illingworth, L. E. McCoy, R. K. Gupta, SARS-CoV-2 evolution during treatment of chronic infection, *Nature* (2021), doi:10.1038/s41586-021-03291-y.

5. B. Choi, M. C. Choudhary, J. Regan, J. A. Sparks, R. F. Padera, X. Qiu, I. H. Solomon, H.-H. Kuo, J. Boucau, K. Bowman, U. D. Adhikari, M. L. Winkler, A. A. Mueller, T. Y.-T. Hsu, M. Desjardins, L. R. Baden, B. T. Chan, B. D. Walker, M. Lichterfeld, M. Brigl, D. S. Kwon, S. Kanjilal, E. T. Richardson, A. H. Jonsson, G. Alter, A. K. Barczak, W. P. Hanage, X. G. Yu, G. D. Gaiha, M. S. Seaman, M. Cernadas, J. Z. Li, Persistence and Evolution of SARS-CoV-2 in an

868 Immunocompromised Host, *N. Engl. J. Med.* **383**, 2291–2293 (2020).

869 6. J. H. Baang, C. Smith, C. Mirabelli, A. L. Valesano, D. M. Manthei, M. A. Bachman, C. E.
870 Wobus, M. Adams, L. Washer, E. T. Martin, A. S. Lauring, Prolonged Severe Acute Respiratory
871 Syndrome Coronavirus 2 Replication in an Immunocompromised Patient*The Journal of*
872 *Infectious Diseases* **223**, 23–27 (2021).

873 7. T. T. Truong, A. Ryutov, U. Pandey, R. Yee, L. Goldberg, D. Bhojwani, P. Aguayo-Hiraldo, B.
874 A. Pinsky, A. Pekosz, L. Shen, S. D. Boyd, O. F. Wirz, K. Röltgen, M. Bootwalla, D. T. Maglinte,
875 D. Ostrow, D. Ruble, J. H. Han, J. A. Biegel, M. L. ScM, C. Huang, M. K. Sahoo, P. S. Pannaraj,
876 M. O'Gorman, A. R. Judkins, X. Gai, J. D. Bard, Persistent SARS-CoV-2 infection and
877 increasing viral variants in children and young adults with impaired humoral immunity, *medRxiv*
878 (2021), doi:10.1101/2021.02.27.21252099.

879 8. M. C. Choudhary, C. R. Crain, X. Qiu, W. Hanage, J. Z. Li, SARS-CoV-2 sequence
880 characteristics of COVID-19 persistence and reinfection*bioRxiv* (2021),
881 doi:10.1101/2021.03.02.21252750.

882 9. K. S. Xue, T. Stevens-Ayers, A. P. Campbell, J. A. Englund, S. A. Pergam, M. Boeckh, J. D.
883 Bloom, Parallel evolution of influenza across multiple spatiotemporal scales, *Elife* **6** (2017),
884 doi:10.7554/eLife.26875.

885 10. J. van Beek, A. A. van der Eijk, P. L. A. Fraaij, K. Caliskan, K. Cransberg, M. Dalinghaus, R.
886 A. S. Hoek, H. J. Metselaar, J. Roodnat, H. Vennema, M. P. G. Koopmans, Chronic norovirus
887 infection among solid organ recipients in a tertiary care hospital, the Netherlands, 2006-2014,
888 *Clin. Microbiol. Infect.* **23**, 265.e9–265.e13 (2017).

889 11. V. A. Avanzato, M. J. Matson, S. N. Seifert, R. Pryce, B. N. Williamson, S. L. Anzick, K.
890 Barbian, S. D. Judson, E. R. Fischer, C. Martens, T. A. Bowden, E. de Wit, F. X. Riedo, V. J.

Munster, Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer, *Cell* **183**, 1901–1912.e9 (2020).

12. K. Debbink, L. C. Lindesmith, M. T. Ferris, J. Swanstrom, M. Beltramello, D. Corti, A. Lanzavecchia, R. S. Baric, Within-Host Evolution Results in Antigenically Distinct GII.4 Noroviruses*Journal of Virology* **88**, 7244–7255 (2014).

13. C. Rhee, S. Kanjilal, M. Baker, Duration of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infectivity: when is it safe to discontinue isolation?, *Clin. Infect. Dis.* (2020) (available at https://academic.oup.com/cid/advance-article-abstract/doi/10.1093/cid/ciaa1249/5896916).

14. D. H. Morris, V. N. Petrova, F. W. Rossine, E. Parker, B. T. Grenfell, R. A. Neher, S. A. Levin, C. A. Russell, Asynchrony between virus diversity and antibody selection limits influenza virus evolution, *Elife* **9** (2020), doi:10.7554/eLife.62105.

15. J. Bullard, K. Dust, D. Funk, J. E. Strong, D. Alexander, L. Garnett, C. Boodman, A. Bello, A. Hedley, Z. Schiffman, K. Doan, N. Bastien, Y. Li, P. G. Van Caeseele, G. Poliquin, Predicting Infectious Severe Acute Respiratory Syndrome Coronavirus 2 From Diagnostic Samples, *Clin. Infect. Dis.* **71**, 2663–2666 (2020).

16. R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, C. Wendtner, Virological assessment of hospitalized patients with COVID-2019, *Nature* **581**, 465–469 (2020).

17. K. Debbink, J. T. McCrone, J. G. Petrie, R. Truscon, E. Johnson, E. K. Mantlo, A. S. Monto, A. S. Lauring, D. R. Perez, Ed. Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses, *PLoS Pathog.* **13**, e1006194 (2017).

914 18. J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, A. S. Lauring,
915 Stochastic processes constrain the within and between host evolution of influenza virus, *Elife* **7**
916 (2018), doi:10.7554/eLife.35962.

917 19. J. M. Dinis, N. W. Florek, O. O. Fatola, L. H. Moncla, J. P. Mutschler, O. K. Charlier, J. K.
918 Meece, E. A. Belongia, T. C. Friedrich, S. Schultz-Cherry, Ed. Deep Sequencing Reveals
919 Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans, *J. Virol.*
920 **90**, 3355–3365 (2016).

921 20. A. L. Valesano, K. E. Rumfelt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E.
922 T. Martin, A. S. Lauring, Temporal dynamics of SARS-CoV-2 mutation accumulation within and
923 across infected hosts, *bioRxiv* (2021), doi:10.1101/2021.01.19.427330.

924 21. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M.
925 Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, R. Williams,
926 G. Vernet, A. Justice, A. Green, S. M. Nicholls, M. A. Ansari, L. Abeler-Dörner, C. E. Moore, T.
927 E. A. Peto, D. W. Eyre, R. Shaw, P. Simmonds, D. Buck, J. A. Todd, Oxford Virus Sequencing
928 Analysis Group (OVSG), T. R. Connor, S. Ashraf, A. da Silva Filipe, J. Shepherd, E. C.
929 Thomson, COVID-19 Genomics UK (COG-UK) Consortium, D. Bonsall, C. Fraser, T. Golubchik,
930 SARS-CoV-2 within-host diversity and transmission, *Science* (2021),
931 doi:10.1126/science.abg0821.

932 22. J. T. McCrone, A. S. Lauring, Measurements of Intrahost Viral Diversity Are Extremely
933 Sensitive to Systematic Errors in Variant Calling, *J. Virol.* **90**, 6884–6895 (2016).

934 23. K. S. Xue, J. D. Bloom, Reconciling disparate estimates of viral genetic diversity during
935 human influenza infections*Nat. Genet.* **51**, 1298–1301 (2019).

936 24. J. W. Rausch, A. A. Capoferri, M. G. Katusiime, S. C. Patro, M. F. Kearney, Low genetic

diversity may be an Achilles heel of SARS-CoV-2*Proc. Natl. Acad. Sci. U. S. A.* **117**, 24614–24616 (2020).

25. K. M. Braun, G. K. Moreno, P. J. Halfmann, D. A. Baker, E. C. Boehm, A. M. Weiler, A. K. Haj, M. Hatta, S. Chiba, T. Maemura, Y. Kawaoka, K. Koelle, D. H. O'Connor, T. C. Friedrich, Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck, *bioRxiv* (2020), doi:10.1101/2020.11.16.384917.

26. A. Graudenzi, D. Maspero, F. Angaroni, R. Piazza, D. Ramazzotti, Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity, *iScience* , 102116 (2021).

27. G. Tonkin-Hill, I. Martincorena, R. Amato, A. R. J. Lawson, M. Gerstung, I. Johnston, D. K. Jackson, N. R. Park, S. V. Lensing, M. A. Quail, S. Gonçalves, C. Ariani, M. S. Chapman, W. L. Hamilton, L. W. Meredith, G. Hall, A. S. Jahun, Y. Chaudhry, M. Hosmillo, M. L. Pinckert, I. Georgana, A. Yakovleva, L. G. Caller, S. L. Caddy, T. Feltwell, F. A. Khokhar, C. J. Houldcroft, M. D. Curran, S. Parmar, The COVID-19 Genomics UK (COG-UK) Consortium, A. Alderton, R. Nelson, E. Harrison, J. Sillitoe, S. D. Bentley, J. C. Barrett, M. Estee Torok, I. G. Goodfellow, C. Langford, D. Kwiatkowski, Wellcome Sanger Institute COVID-19 Surveillance Team, Patterns of within-host genetic diversity in SARS-CoV-2*bioRxiv* , 2020.12.23.424229 (2020).

28. A. Popa, J.-W. Genger, M. D. Nicholson, T. Penz, D. Schmid, S. W. Aberle, B. Agerer, A. Lercher, L. Endler, H. Colaço, M. Smyth, M. Schuster, M. L. Grau, F. Martínez-Jiménez, O. Pich, W. Borena, E. Pawelka, Z. Keszei, M. Senekowitsch, J. Laine, J. H. Aberle, M. Redlberger-Fritz, M. Karolyi, A. Zoufaly, S. Maritschnik, M. Borkovec, P. Hufnagl, M. Nairz, G. Weiss, M. T. Wolfinger, D. von Laer, G. Superti-Furga, N. Lopez-Bigas, E. Puchhammer-Stöckl, F. Allerberger, F. Michor, C. Bock, A. Bergthaler, Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2, *Sci.*

961 *Transl. Med.* **12** (2020), doi:10.1126/scitranslmed.abe2555.

962 29. S. E. James, S. Ngcapu, A. M. Kanzi, H. Tegally, V. Fonseca, J. Giandhari, E. Wilkinson, B.
963 Chimukangara, S. Pillay, L. Singh, M. Fish, I. Gazy, K. Khanyile, R. Lessells, T. de Oliveira,
964 High Resolution analysis of Transmission Dynamics of Sars-Cov-2 in Two Major Hospital
965 Outbreaks in South Africa Leveraging Intrahost Diversity, *medRxiv* (2020),
966 doi:10.1101/2020.11.15.20231993.

967 30. M. A. Martin, K. Koelle, Reanalysis of deep-sequencing data from Austria points towards a
968 small SARS-COV-2 transmission bottleneck on the order of one to three virions, *bioRxiv* (2021)
969 (available at https://www.biorxiv.org/content/10.1101/2021.02.22.432096v1.abstract).

970 31. N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L.
971 Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F.
972 Michael, L. L. Coffey, N. J. Loman, K. G. Andersen, An amplicon-based sequencing framework
973 for accurately measuring intrahost virus diversity using PrimalSeq and iVar, *Genome Biol.* **20**, 8
974 (2019).

975 32. J. K. Das, A. Sengupta, P. P. Choudhury, S. Roy, Characterizing genomic variants and
976 mutations in SARS-CoV-2 proteins from Indian isolates, *Gene Rep* , 101044 (2021).

977 33. J. Singh, J. Samal, V. Kumar, J. Sharma, U. Agrawal, N. Z. Ehtesham, D. Sundar, S. A.
978 Rahman, S. Hira, S. E. Hasnain, Structure-Function Analyses of New SARS-CoV-2 Variants
979 B.1.1.7, B.1.351 and B.1.1.28.1: Clinical, Diagnostic, Therapeutic and Public Health
980 Implications, *Viruses* **13** (2021), doi:10.3390/v13030439.

981 34. E. Kinganda-Lusamaki, A. Black, D. Mukadi, J. Hadfield, P. Mbala-Kingebeni, C. B. Pratt, A.
982 Aziza, M. M. Diagne, B. White, N. Bisento, B. Nsunda, M. Akonga, M. Faye, O. Faye, F. Edidi-
983 Atani, M. Matondo, F. Mambu, J. Bulabula, N. D. Paola, G. Palacios, E. Delaporte, A. A. Sall, M.

984 Peeters, M. R. Wiley, S. Ahuka-Mundeke, T. Bedford, J.-J. Muyembe Tamfum, Operationalizing
985 genomic epidemiology during the Nord-Kivu Ebola outbreak, Democratic Republic of the
986 Congo*bioRxiv* (2020), doi:10.1101/2020.06.08.20125567.

987 35. X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X.
988 Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang,
989 B. J. Cowling, F. Li, G. M. Leung, Temporal dynamics in viral shedding and transmissibility of
990 COVID-19, *Nat. Med.* **26**, 672–675 (2020).

991 36. A. Sobel Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Transmission
992 Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to
993 Human Influenza A Virus, *J. Virol.* **91** (2017), doi:10.1128/JVI.00171-17.

994 37. Z. Shen, Y. Xiao, L. Kang, W. Ma, L. Shi, L. Zhang, Z. Zhou, J. Yang, J. Zhong, D. Yang, L.
995 Guo, G. Zhang, H. Li, Y. Xu, M. Chen, Z. Gao, J. Wang, L. Ren, M. Li, Genomic Diversity of
996 Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease
997 2019*Clin. Infect. Dis.* **71**, 713–720 (2020).

998 38. De Maio, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, Nick Goldman,
999 Nicola, Issues with SARS-CoV-2 sequencing data (2020) (available at
1000 https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473).

1001 39. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M.
1002 Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, R. Williams,
1003 G. Vernet, A. Justice, A. Green, S. M. Nicholls, M. Azim Ansari, L. Abeler-Dörner, C. E. Moore,
1004 T. E. A. Peto, D. W. Eyre, R. Shaw, P. Simmonds, D. Buck, J. A. Todd, on behalf of OVSG
1005 Analysis Group, T. R. Connor, A. da Silva Filipe, J. Shepherd, E. C. Thomson, The COVID-19
1006 Genomics UK (COG-UK) consortium, D. Bonsall, C. Fraser, T. Golubchik, Within-host genomics

of SARS-CoV-2*Cold Spring Harbor Laboratory* , 2020.05.28.118992 (2020).

40. D. Wang, Y. Wang, W. Sun, L. Zhang, J. Ji, Z. Zhang, X. Cheng, Y. Li, F. Xiao, A. Zhu, B. Zhong, S. Ruan, J. Li, P. Ren, Z. Ou, M. Xiao, M. Li, Z. Deng, H. Zhong, F. Li, W.-J. Wang, Y. Zhang, W. Chen, S. Zhu, X. Xu, X. Jin, J. Zhao, N. Zhong, W. Zhang, J. Zhao, J. Li, Y. Xu, Population Bottlenecks and Intra-host Evolution during Human-to-Human Transmission of SARS-CoV-2, doi:10.1101/2020.06.26.173203.

41. A. L. Valesano, W. J. Fitzsimmons, J. T. McCrone, J. G. Petrie, A. S. Monto, E. T. Martin, A. S. Lauring, Influenza B Viruses Exhibit Lower Within-Host Diversity than Influenza A Viruses in Human Hosts, *J. Virol.* **94** (2020), doi:10.1128/JVI.01710-19.

42. H. Zaraket, T. Baranovich, B. S. Kaplan, R. Carter, M.-S. Song, J. C. Paulson, J. E. Rehg, J. Bahl, J. C. Crumpton, J. Seiler, M. Edmonson, G. Wu, E. Karlsson, T. Fabrizio, H. Zhu, Y. Guan, M. Husain, S. Schultz-Cherry, S. Krauss, R. McBride, R. G. Webster, E. A. Govorkova, J. Zhang, C. J. Russell, R. J. Webby, Mammalian adaptation of influenza A(H7N9) virus is limited by a narrow genetic bottleneck*Nature Communications* **6** (2015), doi:10.1038/ncomms7553.

43. A. Varble, R. A. Albrecht, S. Backes, M. Crumiller, N. M. Bouvier, D. Sachs, A. García-Sastre, B. R. tenOever, Influenza A virus transmission bottlenecks are defined by infection route and recipient host, *Cell Host Microbe* **16**, 691–700 (2014).

44. CDC, Science brief: Emerging SARS-CoV-2 variants (2021) (available at https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/scientific-brief-emerging-variants.html).

45. L. H. Moncla, G. Zhong, C. W. Nelson, J. M. Dinis, J. Mutschler, A. L. Hughes, T. Watanabe, Y. Kawaoka, T. C. Friedrich, Selective Bottlenecks Shape Evolutionary Pathways Taken during Mammalian Adaptation of a 1918-like Avian Influenza Virus, *Cell Host Microbe* **19**, 169–180

(2016).

46. CDC, CDC Diagnostic Tests for COVID-19 (2020) (available at https://www.cdc.gov/coronavirus/2019-ncov/lab/testing.html).

47. Panther Fusion® SARS-CoV-2 Assay (available at https://www.hologic.com/package-inserts/diagnostic-products/panther-fusionr-sars-cov-2-assay).

48. Hologic (available at https://www.hologic.com/package-inserts/diagnostic-products/aptimar-sars-cov-2-assay-pantherr-system).

49. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara Jr, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, N. J. Loman, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples, *Nat. Protoc.* **12**, 1261–1276 (2017).

50. J. Quick, nCoV-2019 sequencing protocol v1 (protocols.io.bbmuik6w)*protocols.io* (2020), doi:10.17504/protocols.io.bbmuik6w.

51. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* **30**, 2114–2120 (2014).

52. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2, *Nat. Methods* **9**, 357–359 (2012).

53. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, J. Kelso, Ed. Nextstrain: real-time tracking of pathogen evolution, *Bioinformatics* **34**, 4121–4123 (2018).

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

55. D. Richard, C. J. Owen, L. van Dorp, F. Balloux, No detectable signal for ongoing genetic recombination in SARS-CoV-2, *bioRxiv* (2020) (available at https://www.biorxiv.org/content/10.1101/2020.12.15.422866v1.abstract).

56. G. Dudas, *baltic* (Github; https://github.com/evogytis/baltic).

57. L. H. Moncla, A. Black, C. DeBolt, M. Lang, N. R. Graff, A. C. Pérez-Osorio, N. F. Müller, D. Haselow, S. Lindquist, T. Bedford, Repeated introductions and intensive community transmission fueled a mumps virus outbreak in Washington State*bioRxiv* (2020), doi:10.1101/2020.10.19.20215442.

Acknowledgments

LHM is supported by NIAID grant number K99 AI147029-01. GKM is supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359).

Supplementary Materials

1. **Supplemental Figure 1:** Read depth for MiSeq runs 627, 628, 643, and 644.
2. **Supplemental Figure 2:** Read depth for MiSeq runs 645, 667, and 671.
3. **Supplemental Figure 3:** Additional iSNV quality control information.
4. **Supplemental Figure 4:** iSNVs in technical replicates across all samples.
5. **Supplemental Figure 5:** iSNVs do not cluster by sequencing run.
6. **Supplemental Figure 6:** Wisconsin divergence phylogeny.

7. **Supplemental Figure 7:** Most iSNVs are not detected on the phylogeny. Please note, you may need to zoom into this figure in order to read each iSNV along the x-axis.
8. **Supplemental Figure 8:** Modeling the expected number of mutations distinguishing genomes separated by one serial interval.
9. **Supplemental Figure 9:** Posterior density estimates for regression coefficients.
10. **Supplemental Figure 10:** Sensitivity testing of transmission bottleneck estimates.
11. **Supplemental Figure 11:** Variance in transmission bottleneck size cannot be explained by time between symptom onset in donor:recipient pairs.

12. **Supplemental Table 1.** iSNVs detected in replicate sequencing of the synthetic RNA control (Twist-Biosciences).
13. **Supplemental Table 2**. Sample identifiers and accession numbers. This table includes strain name, tube/filename, state of collection, county of collection, collection date, GISAID accession number, Genbank accession number, as well as Ct values and RLU values where available for each sample included in this study.
14. **Supplemental Table 3**. ARTIC v3 primer sequences used to amplify cDNA for library preparation.
15. **Supplemental Table 4**. Household transmission pair metadata including accession numbers, difference in days between symptom onset, difference in days between collection dates, and pair identifier.