Virus Evolution, 2022, **8(2)**, 1–9

DOI: https://doi.org/10.1093/ve/veac052
Advance access publication date: 16 June 2022

Research Article

Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat

William W. Hannon,[1,2] Pavitra Roychoudhury,[3,4] Hong Xie,[4] Lasata Shrestha,[4,†] Amin Addetia,[1,4] Keith R. Jerome,[3,4,‡] Alexander L. Greninger,[3,4] and Jesse D. Bloom[2,5,6,*,§]

[1]Molecular and Cellular Biology Graduate Program, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195, USA, [2]Basic Sciences and Computational Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109, USA, [3]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [4]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, 1959 NE Pacific St, Seattle, WA 98195, USA, [5]Howard Hughes Medical Institute, 1100 Fairview Ave N, Seattle, WA 98109, USA and [6]Department of Genome Sciences, University of Washington, Seattle, WA 98109, USA

[†]https://orcid.org/0000-0002-9760-7348
[‡]https://orcid.org/0000-0002-8212-3789
[§]https://orcid.org/0000-0003-1267-3408
[*]Corresponding author: E-mail: jbloom@fredhutch.org

Abstract

The long-term evolution of viruses is ultimately due to viral mutants that arise within infected individuals and transmit to other individuals. Here, we use deep sequencing to investigate the transmission of viral genetic variation among individuals during a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) outbreak that infected the vast majority of crew members on a fishing boat. We deep-sequenced nasal swabs to characterize the within-host viral population of infected crew members, using experimental duplicates and strict computational filters to ensure accurate variant calling. We find that within-host viral diversity is low in infected crew members. The mutations that did fix in some crew members during the outbreak are not observed at detectable frequencies in any of the sampled crew members in which they are not fixed, suggesting that viral evolution involves occasional fixation of low-frequency mutations during transmission rather than persistent maintenance of within-host viral diversity. Overall, our results show that strong transmission bottlenecks dominate viral evolution even during a superspreading event with a very high attack rate.

Key words: SARS-CoV-2; transmission bottleneck; whole-genome sequencing; genomic surveillance.

Introduction

The long-term evolution of viruses is due to mutations that arise during replication within infected hosts and then transmit to new hosts. For viruses like SARS-CoV-2 or influenza that typically cause short self-limiting infections, evolution occurs over many consecutive rounds of infection, each interrupted by a transmission bottleneck. If there is a wide transmission bottleneck, then mutations can gradually increase in frequency as a virus transmits from one host to another. However, a narrow transmission bottleneck means that low-frequency mutations present in a donor host will typically be either lost or fixed in a recipient host (Zwart and Elena 2015; McCrone and Lauring 2018).

So far, efforts to understand how transmission shapes the evolution of SARS-CoV-2 have mainly focused on small household events or nosocomial pairs (Popa et al. 2020; Braun et al. 2021; Lythgoe et al. 2021; San et al. 2021; Wang et al. 2021). Such studies point to a narrow transmission bottleneck that significantly reduces viral genetic diversity at the start of each infection (Braun et al. 2021; Lythgoe et al. 2021; Martin and Koelle 2021; San et al. 2021; Wang et al. 2021). While exact estimates of the bottleneck range from one to fifteen virions, it is clear that a limited number of virions initiate most human infections. These results are broadly similar to those for influenza, another heavily studied respiratory RNA virus (McCrone et al. 2018; Xue and Bloom 2019; Valesano et al. 2020).

However, it seems possible that the transmission of viral genetic diversity could show different patterns in different settings. For example, superspreading events play a significant role in SARS-CoV-2's overall spread (Liu, Eggo, and Kucharski 2020; Lemieux et al. 2021), and such events could exhibit different patterns of evolution since they involve settings highly conducive to a viral transmission.

Here, we investigate the spread of viral genetic diversity during a SARS-CoV-2 superspreading event on a fishing boat (Addetia et al. 2020). We perform high-depth metagenomic deep sequencing on nasal swabs collected from crew members of the fishing

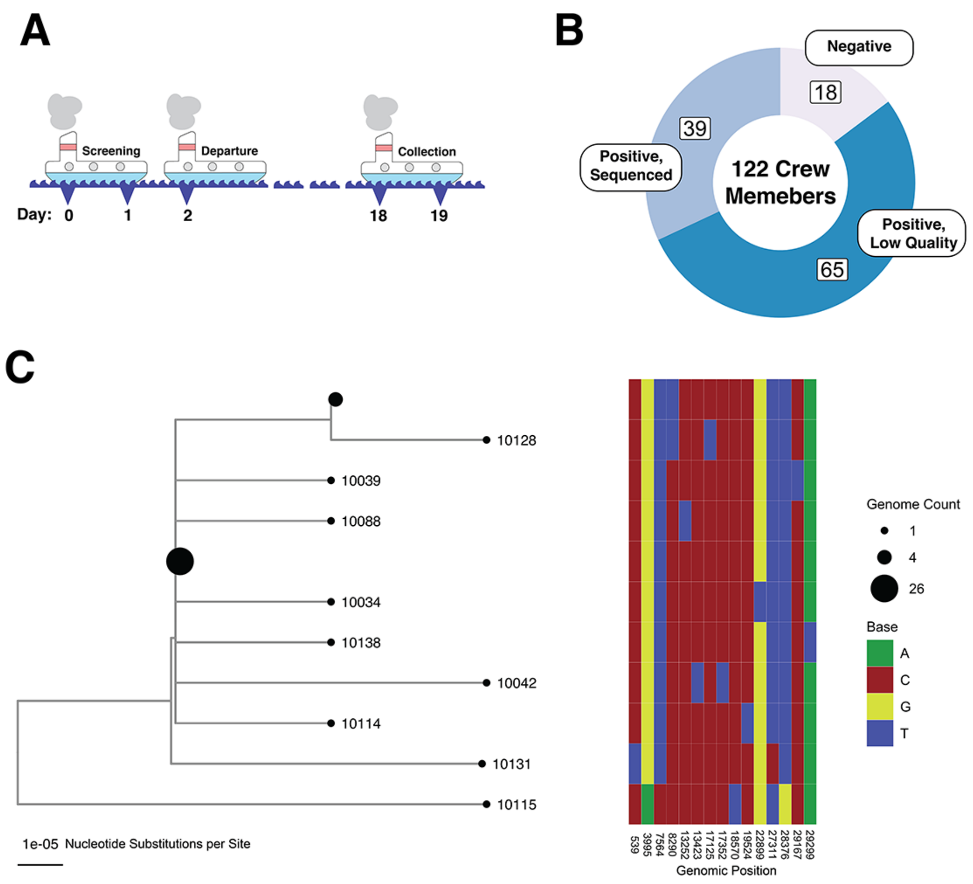

Figure 1. An outbreak of SARS-CoV-2 on an isolated fishing boat is an epidemiologically linked cluster of infections. (A) Schematic showing the timeline of the fishing vessel outbreak. All samples used in this study were taken on Day 18 as shown in the figure (relative to the start of pre-departure screening). (B) Donut plot showing the sampling breakdown for all 122 members of the crew. (C) Phylogeny of SARS-CoV-2 genome from the boat. A heatmap to the right shows the nucleotide differences between genomes on the tree. Specimen identification numbers for crew-member samples label the leaf nodes of the tree except for those nodes with more than one identical genome. Node sizes are proportional to the number of sequences: there is a node representing twenty-six identical sequences (10101, 10126, 10133, 10105, 10108, 10130, 10031, 10110, 10030, 10124, 10029, 10102, 10038, 10094, 10027, 10118, 10117, 10106, 10091, 10093, 10127, 10116, 10040, 10090, 10036, and 10089) and a node representing four identical sequences (10107, 10129, 10113, and 10028); all other nodes represent unique sequences.

boat to characterize the intra-host populations of viral variants. Our results demonstrate that epidemiologically linked individuals in a superspreading event share little to no intra-host viral diversity even at sites where mutations fix during the event, corroborating studies reporting narrow transmission bottlenecks in other settings (Braun et al. 2021; Lythgoe et al. 2021; Martin and Koelle 2021; San et al. 2021; Wang et al. 2021).

Results

A large-scale SARS-CoV-2 transmission event on a fishing boat

We analyzed samples collected from an outbreak on a fishing boat in May 2020 (Addetia et al. 2020). There were a total of 122 individuals on the boat. Two days before embarking from Seattle, 120 individuals participated in pre-departure screening for for SARS-CoV-2 , and none tested positive. Despite this, infected crew members must have boarded the boat because a large SARS-CoV-2 outbreak ensued, eventually forcing the boat to return to shore in Seattle after 16 days at sea (Fig. 1A). Over 80 per cent of crew members ultimately tested positive for SARS-CoV-2, indicating an extremely high secondary attack rate aboard the boat (Fig. 1B). Of note, only three crew members had neutralizing antibodies before the ship's departure, and none of these individuals met the case definition for infection (Addetia et al. 2020). To confirm that the secondary attack rate was high on the boat, we calculated the expected percentage of individuals infected or exposed in 16 days in a hypothetical outbreak, parameterized with a range of values for the basic reproduction number (R_0). The R_0 would need to be substantially higher ($R_0 \approx 6$–12 depending on the model used) than was usual in early 2020 ($R_0 \approx 3$) for this fraction of the boat's crew to have become infected or exposed in only 16 days (Supplementary Fig. S1A) (He, Yi, and Zhu 2020a). These results suggest that the transmission force was higher on the boat than in the typical setting of SARS-CoV-2 transmission.

Nasal swabs were collected from the crew members 2 days after the boat returned to shore. Of the samples that were positive in a SARS-CoV-2 polymerase chain reaction (PCR) test, thirty-nine had sufficiently high levels of viral RNA (Ct value less than 26) to assemble consensus viral sequences from deep-sequencing data, as previously described in Addetia et al. (Fig. 1B). These consensus viral sequences from the boat samples differed on average at fewer than two positions and were clearly diverged relative to the non-boat out-group sample (Fig. 1C). Over 75 per cent of the viral sequences from the boat were identical to at least one other sequence from the boat. When we compared the number of fixed mutations in the viral sequences from the boat to a theoretical distribution of the number of mutations expected to fix over a range of transmission intervals, the observed distribution most closely resembled that expected to accumulate in a single interval (Supplementary Fig. S1B) (Braun et al. 2021). Given the genetic similarity of viral sequences from the boat and the short time frame for infections, this cohort resembles a superspreading event where few transmission events separate all crew-member infections from the introduction of SARS-CoV-2 to the boat.

To place the superspreading event in the larger context of SARS-CoV-2's genetic diversity, we inferred a phylogeny using representative sequences from viruses circulating before the outbreak, including a subset of the most genetically similar viral sequences to those isolated from the boat. The boat clade is nearly monophyletic, although two surveillance sequences collected elsewhere in Washington state around the time of the outbreak fall in the same clade as the boat samples (Fig. 2). These sequences likely share a close common ancestor with the virus that seeded the superspreading event on the boat. We also chose one Washington state sample not from the boat for further sequencing, and as expected this sample was distinct from the boat clade on the tree. Overall, the nearly monophyletic nature of the outbreak clade and the fishing boat's isolation makes this cohort appropriate for assessing how SARS-CoV-2 genetic diversity transmits among a tightly associated group of individuals.

High-quality deep sequencing of samples with adequate viral RNA

We used deep sequencing to measure the intra-host viral genetic variation in the samples collected from infected crew members. We employed several approaches to ensure the accuracy of these measurements. First, we used a shotgun metagenomic sequencing approach to avoid potential mutational biases from specific PCR amplification of viral RNA. Of the thirty-nine nasal swabs described in the previous section, twenty-three had sufficient viral RNA (Ct value less than 20) to sequence metagenomically (Supplementary Table S1) (Charre et al. 2020). Second, we sequenced replicates starting from independent reverse transcription (RT) reactions from the same initial nasal swab. In principle, each replicate should sample from the same underlying viral population, so differences between replicates can indicate limitations due to a lack of underlying viral template molecules in the swabs due to low viral load. A lack of viral template diversity can significantly distort variant frequencies inferred from deep sequencing (Illingworth et al. 2017; Xue et al. 2018). We used a stringent cutoff for sequencing depth by only considering sequences with >80 per cent of the genome covered by 100 reads in one or more replicates in the downstream analysis (Supplementary Fig. S2B). There were no biases observed in sequencing coverage across the length of the viral genome (Supplementary Fig. S2A).

We compared results between replicates for each crew member and focused our subsequent analyses on the thirteen crew members with high concordance between replicates and adequate sequencing depth (Fig. 3). Of note, the results were robust to using different methods for variant calling (Supplementary Figs S3 and S4).

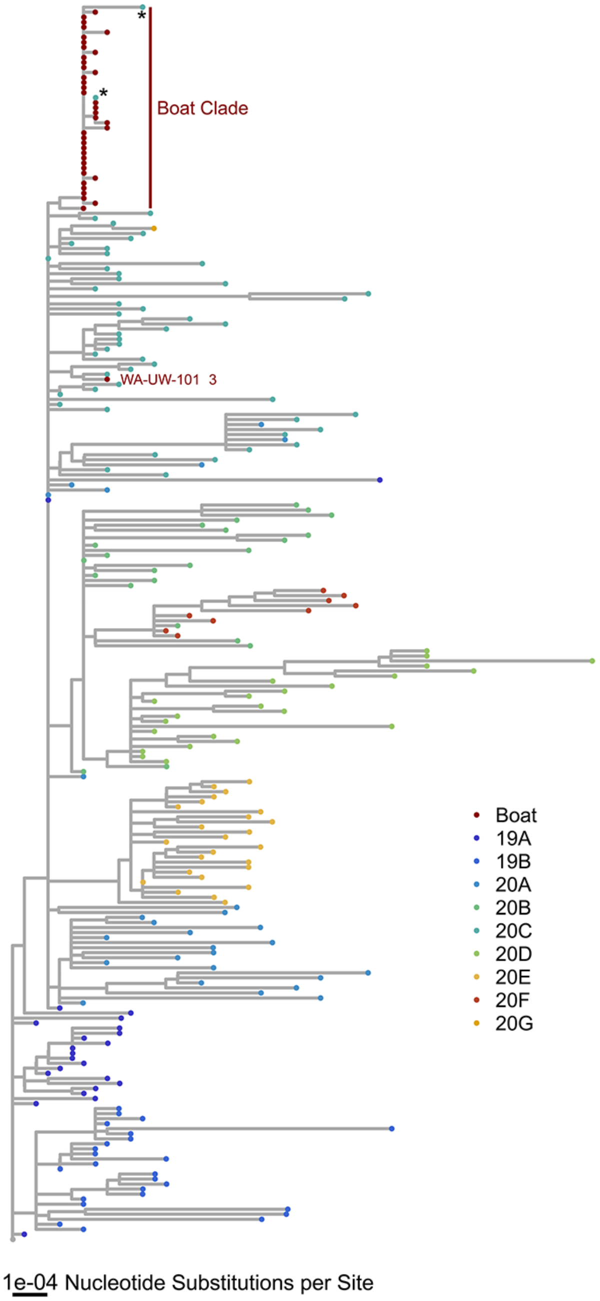

Figure 2. Sequences from the boat form a distinct clade. A phylogeny of the thirty-nine crew-member genomes and representative genomes from other circulating clades before the outbreak. Additionally, this phylogeny includes the ten closest matches to each of the thirty-nine crew-member genomes from a custom BLASTN database made with sequences collected from Washington in a 2-month interval around the time of the outbreak. We also resequenced as a control one sample not from the boat (WA-UW-10136). Most genomes isolated from the boat form a distinct clade broken only by two genomes (hCoV-19/USA/WA-UW-10510/2020 and hCoV-19/USA/WA-UW-10521/2020) annotated with an asterisk.

The intra-host virus population is relatively homogeneous

After retaining just the samples with high sequencing depth and good replicate-to-replicate correlations, we assembled a set of intra-host single nucleotide polymorphisms (SNPs) that were present in 2 per cent of at least 100 reads in both replicates. To determine the extent of within-host diversity in each patient, we converted any mutation (relative to the reference) above 50 per cent frequency to its corresponding minor allele and counted the total number of minor allele variants at >2 per cent frequency per crew member. The diversity of the virus populations within each crew member was limited, with an average of three intra-host variants per individual (range 0–5, Fig. 4A). Furthermore, most intra-host variants were at relatively low frequencies, with only a

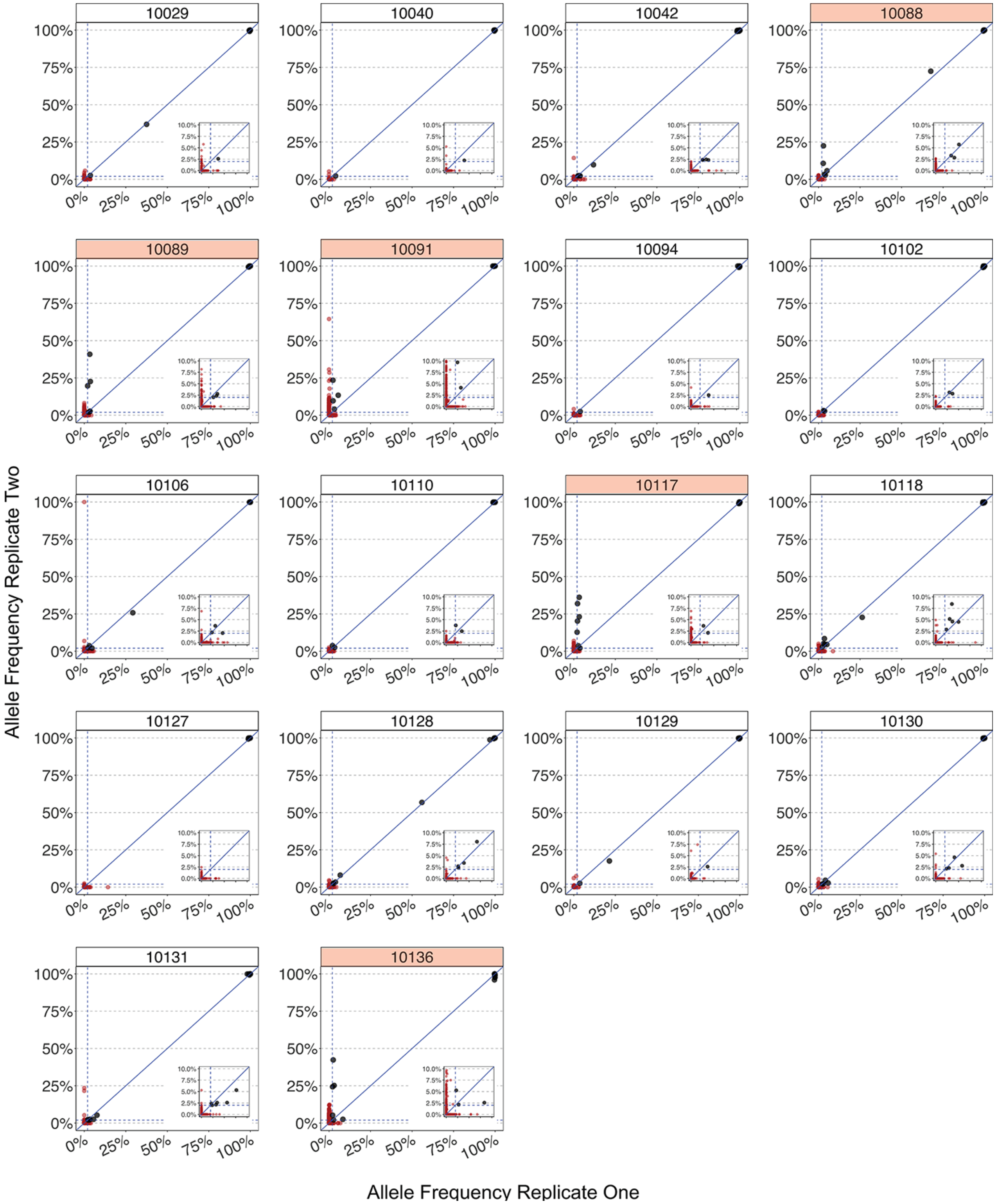

Figure 3. Robust quality control reveals false-positive variant alleles and samples of poor quality. Each plot shows the concordance between allele frequencies between replicates for every specimen that we sequenced, with both replicates having greater than 100× coverage in at least 80 per cent of the genome. Alleles that were present in less than 2 per cent of 100 reads in either replicate are colored red. The dotted line represents the 2 per cent frequency threshold. We highlighted the facet headers of 'poor' quality crew-member samples in red if there was a large discrepancy in allele frequencies between replicates. This figure also shows the non-boat sample (10136) sequenced as a control.

handful at >10 per cent (Fig. 4B). This limited within-host diversity and low-frequency-dominated allele frequency spectrum are consistent with other studies of SARS-CoV-2 intra-host diversity that have utilized robust computational and experimental controls (Fig. 4B) (Braun et al. 2021; Lythgoe et al. 2021; Martin and Koelle 2021; Valesano et al. 2021). There was no correlation

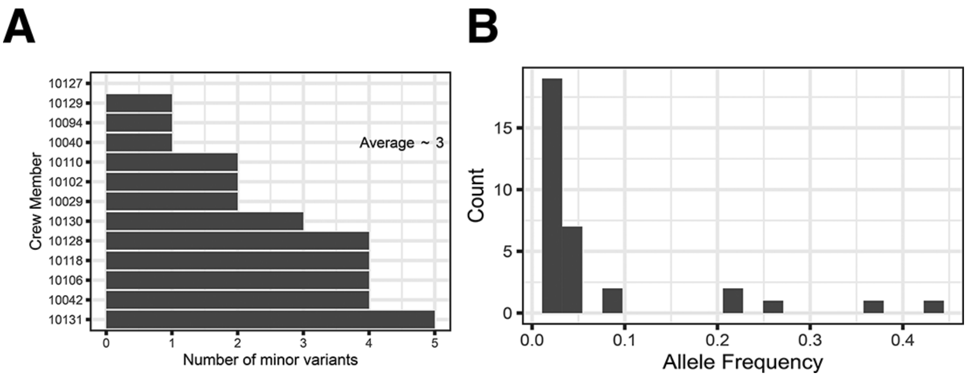

Figure 4. The intra-host spectrum of minor alleles reveals a relatively homogeneous virus population. (A) Bar graph showing the number of minor variants (<50 per cent allele frequency) identified in both replicates of each crew member. There was an average of three minor variants per infection across the ten crew members. (B) The minor allele frequency spectrum across all twelve crew-member specimens with minor variants.

between the Ct value of the nasal swab and the number of SNPs we identified (Supplementary Fig. S4). Additionally, there was no discernable pattern in the location of SNPs in the genome (Supplementary Figs S5 and S6).

Mutations that fix on the boat are not observed at intermediate frequencies

We next considered two possible conceptual models for how mutations could spread and fix on the boat. The first model assumes that the transmission bottleneck is narrow, and variants will be either lost during transmission or, less frequently, fixed during a single transmission event. The second model assumes that the transmission bottleneck is wide, and variants will transmit between multiple infections and gradually rise in frequency until they fix (Fig. 5A).

To determine which conceptual model best describes viral transmission on the boat, we plotted the frequency of every variant allele for each crew member and sorted the crew members by allele frequency. We identified variants relative to the inferred ancestral sequence for the root of the boat clade (which is also the consensus and most common sequence on the boat, see Fig. 1C). If the transmission bottleneck is narrow, most non-fixed variants would be private to single individuals, and at sites with fixed variants, the mutations will generally be present at ∼0 per cent or ∼100 per cent frequency. If transmission bottlenecks were wide on the boat, variants would be observed in multiple individuals at intermediate frequencies. We observed that most low-frequency variants were private to single individuals, and fixed variants were also never observed at intermediate frequencies (Fig. 5B). The lack of a gradient in the frequency for fixed variants on the boat suggests that viral evolution on the boat is dominated by a narrow transmission bottleneck.

Although most variants were either fixed or private to single crew members, four low-frequency alleles were present in multiple individuals on the boat (A4229C, C9502T, G14335T, and T18402A in Fig. 5B). However, none of these variants ever reached more than 5 per cent frequency. Furthermore, several characteristics of these shared low-frequency variants suggest that they are sequencing artifacts rather than true mutations. First, these same variants are also observed in our deep sequencing of a control sample not collected from the boat but sequenced in the same run as the boat samples (Supplementary Fig. S7). Furthermore, one variant, C9502T, is present in a homopolymeric stretch of thymines, a known correlation with spurious variant calls in SARS-CoV-2 sequencing data (Pfeiffer et al. 2018; Braun et al. 2021). Additionally, G14335T and A4229C exhibit significant positional bias in the aligned reads, with most observations at the beginning of the read. Read position correlates with false-positive variant calls in experimental studies of viral deep-sequencing data (McCrone and Lauring 2016). Finally, T18402A demonstrates significant divergence in its frequency between replicates. These four shared variant alleles are therefore likely technical artifacts that survived our quality checks.

Discussion

This study examined the spread of SARS-CoV-2 genetic diversity during a superspreading event on a boat. We found low rates of intra-host viral diversity among infected individuals, and mutations that did fix appeared to do so during single transmission events. Our results demonstrate that the transmission of intra-host viral diversity is limited even during superspreading events that are highly conducive to transmission. These findings are consistent with studies of SARS-CoV-2 transmission in other settings such as households or hospitals (Braun et al. 2021; Lythgoe et al. 2021; Martin and Koelle 2021; San et al. 2021; Wang et al. 2021), suggesting that narrow transmission bottlenecks are a common feature of the virus's transmission. Similar narrow transmission bottlenecks also dominate the evolution of influenza virus (McCrone et al. 2018; Xue and Bloom 2019; Valesano et al. 2020).

A key aspect of our study was sequencing duplicates and rigorous variant calling. False-positive variants shared between multiple samples significantly biased the results of Popa et al., leading to an estimate of the bottleneck nearly 10-fold higher than other studies (Popa et al. 2020; Martin and Koelle 2021). Martin and Koelle reanalyzed these data with a more stringent allele frequency filter, and the bottleneck estimate dropped from greater than 1,000 founding viruses to between one and three founding viruses (Martin and Koelle 2021). Despite our attempts to remove low-frequency false-positive variants, some survived our quality controls. Further research to determine the cause of shared false-positive variants in clinical SARS-CoV-2 deep sequencing could further improve the accuracy of these studies.

Our study has several limitations. First, we were able to obtain high-quality sequencing for only some of the boat's crew members. After accounting for samples that passed our quality controls, only 13 of the 122 crew members were available for analysis. Therefore, we might be missing instances where a variant rises to fixation over multiple transmission events. Another limitation of this study is that we cannot quantitatively estimate the

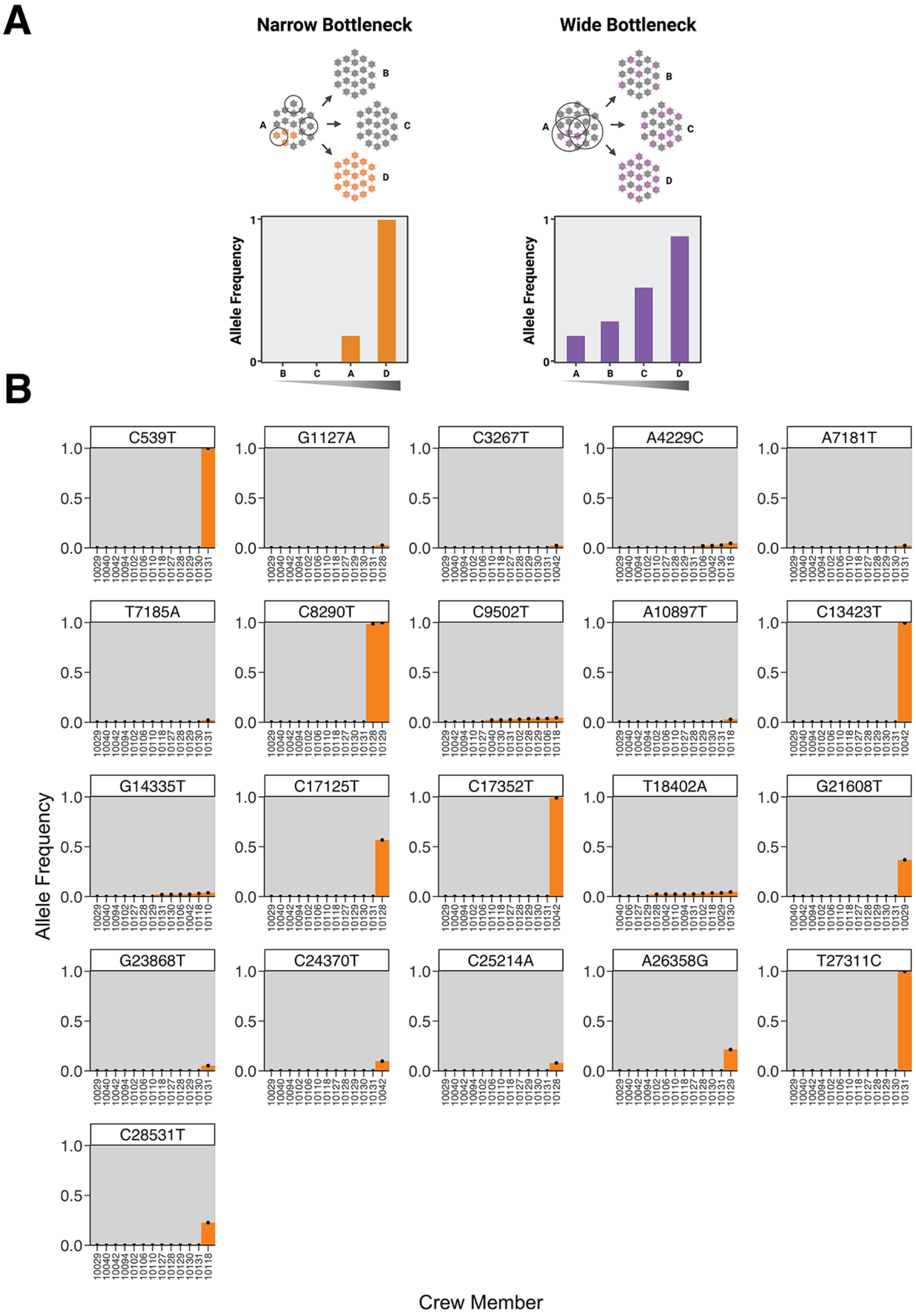

Figure 5. The spectrum of shared minor variation suggests that the transmission bottleneck is narrow. (A) A schematic showing the expected pattern of observed allele frequencies for shared variants in either a narrow or wide bottleneck scenario. (B) Each plot represents the frequency of an SNP across crew members. Variants are called relative to the ancestral sequence of the virus introduced to the boat as inferred from the phylogeny of crew-member genomes. The x-axis is ordered by variant frequency.

transmission bottleneck because we do not know which passengers infected one another. Finally, we must also consider that the lack of initial viral diversity in acute SARS-CoV-2 infections and the possibility of within-host bottlenecks between infection and sampling limit our statistical power to make claims about the size of the transmission bottleneck. However, the absence of shared high-frequency alleles, which are highly likely to survive within-host founder effects and transmit between crew members if the bottleneck is wide, suggests a generally narrow transmission bottleneck.

Overall, our study corroborates the finding of limited shared intra-host viral diversity that has been observed in studies of acute infections with SARS-CoV-2 in other settings. Therefore, even superspreading events in poorly ventilated, close-quarters environments appear insufficient to alter the dominant role of transmission bottlenecks in shaping the evolution of SARS-CoV-2.

Methods

Ethics statement

The use of residual clinical specimens was approved by the University of Washington IRB (protocol STUDY00000408) with a waiver of informed consent.

Sample collection and preparation

RNA was extracted from positive SARS-CoV-2 nasal swabs from crew members using the Roche MagNa Pure 96 (Nalla et al., n.d.). The initial sequencing libraries were constructed as previously described and sequenced on a 1 × 75 bp Illumina NextSeq run (Addetia et al. 2020). RNA was DNase treated using the Turbo DNA-Free kit (Thermo Fisher). First-strand complementary DNA (cDNA) synthesis was performed using Superscript IV (Thermo Fisher) and 2.5 μM random hexamers (Integrated DNA Technologies), and second-strand synthesis with Sequenase version 2.0 DNA polymerase (Thermo Fisher). Double-stranded cDNA was purified using AMPure XP beads (Beckman Coulter), and libraries were constructed using a Nextera Flex DNA pre-enrichment kit with twelve cycles of PCR amplification (Illumina). We resequenced samples from these original libraries to increase their depth if they had an RT-quantitative PCR (RT-qPCR) Ct value less than 20 from an RT-qPCR as measured in a previous paper (Addetia et al. 2020). Samples with a Ct value less than 20 were deemed to have enough RNA to be sequenced without specific amplification of viral RNA by PCR with targeted primers.

Additionally, we made duplicate libraries starting from the same nasal swabs as the initial library using independent RT reactions and identical library preparation methodology. In principle, each replicate should sample from the same underlying virus population, so differences between replicates can indicate limitations due to a lack of underlying template molecules in the swabs (Xue et al. 2018; Xue and Bloom 2019). Of note, one specimen from the original paper, 10136, was subsequently determined not to have been isolated from the boat but from general viral surveillance in Seattle. We kept this sample and resequenced it as non-boat control. We obtained an average of 1,113,690 mapped reads per library.

Sequencing data processing

All data processing from the raw unaligned sequencing files onward was handled by our Snakemake pipeline available on Github—https://github.com/jbloomlab/SARS-CoV-2_bottleneck (Köster and Rahmann 2012). Sequencing reads from the raw FASTQ files from each sequencing run were trimmed for adapter sequences and long (>10) homopolymer sequences at the ends of reads with fastp (Chen et al. 2018). Fastp was also used to filter reads from the FASTQ file if they contained more than 50 per cent unqualified bases (Phred < 15) or were less than 50 bp in length. Following quality filtering, SARS-CoV-2 specific reads were selected from the FASTQ files by matching thirty-one base long *k*-mers to the Wuhan-1/2019 reference genome (NC_045512.2) using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/).

After quality filtering and selection of reads containing SARS-CoV-2 sequences, the FASTQ files were aligned to the Wuhan-1/2019 reference (NC_045512.2) with BWA mem (Li 2013). Libraries that were resequenced for greater depth were joined together after alignment with Samtools merge (Li 2013). The aligned BAM files were checked for quality using Samtools to obtain average coverage, base quality, and completeness.

Phylogenetic analysis

We used aligned BAM files to make consensus sequences for each crew member. Individual consensus sequences were created for each replicate by taking the most represented base at every position if that position had more than 100 reads with a base quality score of greater than 25; otherwise, an *N* was added to the sequence. Then, we combined the consensus sequences from each replicate and filled in *N*s where possible. If the consensus from each replicate disagreed at a position, an *N* was inserted. In addition to the consensus genomes from twenty-three crew-member samples, we deep-sequenced in duplicate, we included sixteen consensus genomes from the boat assembled in the previous study downloaded from GISAID (Addetia et al. 2020). Following the assembly of consensus genomes for each crew member, we aligned the genomes with MAFFT (Katoh and Standley 2013). We masked the non-coding 3′ and 5′ portions of the genome. Using these aligned genomes, we built a phylogenetic tree with IQtree using 1,000 bootstrap iterations with an invariable site plus a discrete gamma model and ancestral state reconstruction (Hoang et al. 2018; Minh et al. 2020). The ancestral state reconstruction was used to infer the ancestral sequence of the genomes from the boat. The tree was rooted using midpoint rooting as implemented by the R package phytools and plotted with ggtree (Revell 2012; Yu et al. 2017) (Fig. 1C). We collapsed weakly supported branches into polytomies if the branch was not supported in more than 60 per cent of the bootstraps.

To determine where all of the available boat sequences fit in the coincident global phylogeny, we downloaded at most twenty-five genomes from GISAID that met our quality criteria (<5 per cent Ns, high coverage, complete coverage, and human host) from each of the circulating Nextrain clades at and before the time of the outbreak on 5 May 2020 (19A, 19B, 20A, 20B, 20C, 20D, 20E, and 20 F) (Hadfield et al. 2018). Additionally, to include genomes that were similar to those on the boat, we built a BLASTN database from all sequences collected in Washington state at and before the time of the outbreak (5 May 2020) that met the same quality standards described above. We took the ten closest matches to each of the twenty-four consensus genomes to include in the phylogeny. We aligned these sequences using MAFFT; however, we also aligned to the Wuhan-1/2019 (NC_045512.2) reference and standardized the length of each sequence. Following alignment, we masked the sequence before the start of ORF1ab and after position 29675 to control for sequencing errors at the start and end of the genome. We built a phylogeny with IQtree using the same parameters as above. The tree was rooted using out-group rooting with the Wuhan-1/2019 reference (NC_045512.2) as the out-group as implemented by the R package ape and plotted with ggtree with weakly supported branches also collapsed into polytomies (Paradis and Schliep 2019; Yu et al. 2017) (Fig. 2).

The code to run all of the phylogenetic analyses is provided on Github at https://github.com/jbloomlab/SARS-CoV-2_bottleneck/blob/master/workflow/notebooks/Phylogenetic-Analysis.ipynb.

The GISAID IDs for sequences used to conduct this analysis are listed in the supplement along with their submitting lab (Supplementary Table S2).

Variant calling and filtering

Variants were identified using a custom Python script (https://github.com/jbloomlab/SARS-CoV-2_bottleneck/blob/master/workflow/scripts/process_pysam_pileup.py). Briefly, we counted the coverage of each base at every position in the reference genome using the Python/Samtools interface Pysam (https://github.com/pysam-developers/pysam). Bases were only included if they surpassed a Phred quality score of 25. After identifying SNPs, our program goes back through the BAM file and identifies reads that overlap these polymorphic sites. We record the total number of occurrences of the SNP, the average position in each read, and the strand ratio. SNPs present after position 29860 in the genome were excluded from the output to avoid sequencing artifacts. The final SNPs were annotated for coding effect and position in the genome using another custom script (https://github.com/jbloomlab/SARS-CoV-2_bottleneck/blob/master/workflow/scripts/annotate_coding_changes.py).

In addition to our custom approach, we also called variants using three different variant calling programs, ivar, varscan2, and lofreq (Koboldt et al. 2012; Wilm et al. 2012; Grubaugh et al. 2019). Where applicable, the same heuristic filters were used in each program. The minimum base quality score was 25, the minimum coverage was 100×, at least ten reads were needed to contain a given SNP, and the minimum allele frequency was 0.5 per cent. Filters that could not be applied in a given program were standardized *post hoc* in R. Variants from each program were standardized into a similar format and added to a single table. Insertions and deletions were removed as we did not benchmark our pipeline to detect these. We annotated the coding effect of each SNP using SnpEff (Cingolani et al. 2012). These extra sets of shared variants were used to cross-check the results of our approach with that of others.

Finally, to identify variants that were shared between individuals on the boat and determine how variants came to be fixed (Fig. 5), we considered all variants relative to the ancestral boat sequence inferred by IQtree using a phylogeny of the boat sequences. Therefore, the included fixed mutations arose after the introduction of the virus to the boat.

Outbreak modeling

To support the claim that the outbreak occurred in a high-transmissibility environment where the total number of secondary cases was larger than the number of primary cases, we calculated the expected percentage of individuals infected or exposed over 16 days in a hypothetical outbreak, parameterized with a range of values for the basic reproduction (R_0) number between 1 and 15. We used two standard epidemiological models of infection: one that calculates the percentage of a population that is susceptible, infected, or removed and one that additionally accounts for latency between exposure and infectiousness. We defined an outbreak with 122 individuals and a single introduction. We used a latency period of 5.08 days and 8 days of infectiousness until recovery (He, Yi, and Zhu 2020a; van Kampen et al. 2021). We used these models to calculate the point at which more than 85 per cent of the crew would have been either infected or exposed to SARS-CoV-2.

Substitutions in a serial interval

We implemented a simple Poisson model of mutation accumulation from Braun et al. to get a theoretical distribution of the number of fixed mutations expected to accumulate in a transmission event (Braun et al. 2021). This model defines a transmission event as a single serial interval, i.e. the length of time between symptom onset in a primary and secondary case. The lambda parameter of the Poisson distribution was derived from the number of substitutions per site in the genome per year (0.0011 substitutions/site/year) and the average length of a serial interval (5.8 days) (Duchene et al. 2020; He et al. 2020b). The outbreak took place over 16 days; therefore, at most, three intervals could separate the index case from the final infection. We compared the distribution of consensus differences that separated the clade encompassing every sample that qualified for deep sequencing from its inferred root to the theoretical distribution of mutations expected to fix in 1, 2, and 3 serial intervals.

Data availability

All sequencing data are available on the NCBI SRA at the project accession PRJNA803551. All codes used to run the analyses described in this paper are archived on Github (https://github.com/jbloomlab/SARS-CoV-2_bottleneck) and Zenodo at DOI: 10.5281/zenodo.6456186.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

We thank collaborators in the Jerome lab and Greninger lab. J.D.B. is an Investigator of the Howard Hughes Medical Institute. P.R. is a University of Washington / Fred Hutch Center for AIDS Research (CFAR) New Investigator award recipient. We acknowledge all authors from originating and submitting laboratories of the sequences from GISAID's Epi-CoV database. An acknowledgments table is available in the Supplementary material.

Funding

National Institutes of Health/National Institute of Allergy and Infectious Diseases (HHS75N93021C00015); R01AI141707 (to J.D.B.) and Washington / Fred Hutch Center for AIDS Research (CFAR) New Investigator award AQ15NIH AI027757 (to P.R.). J.D.B. is an Investigator of the Howard Hughes Medical Institute.

Competing interests: J.D.B. consults for Moderna, Flagship Labs 77, and Oncorus. J.D.B. is an inventor on Fred Hutch licensed patents related to viral deep mutational scanning. A.L.G. reports central laboratory testing contracts from Abbott Molecular, grants from Merck, and grants from Gilead. W.W.H., P.R., H.X., L.S., A.A., and K.R.J. have nothing to disclose.

References

Addetia, A. et al. (2020) 'Neutralizing Antibodies Correlate with Protection from SARS-CoV-2 in Humans during a Fishery Vessel Outbreak with a High Attack Rate', *Journal of Clinical Microbiology*, 58: e02107-20.

Braun, K. M. et al. (2021) 'Acute SARS-CoV-2 Infections Harbor Limited Within-Host Diversity and Transmit via Tight Transmission Bottlenecks', *PLOS Pathogens*, 17: e1009849.

Charre, C. et al. (2020) 'Evaluation of NGS-based Approaches for SARS-CoV-2 Whole Genome Characterisation', *Virus Evolution*, 6: veaa075.

Chen, S. et al. (2018) 'Fastp: An Ultra-fast All-in-One FASTQ Preprocessor', *Bioinformatics*, 34: i884–90.

Cingolani, P. et al. (2012) 'A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3', *Fly*, 6: 80–92.

Duchene, S. et al. (2020) 'Temporal Signal and the Phylodynamic Threshold of SARS-CoV-2', *Virus Evolution*, 6: veaa061.

Grubaugh, N. D. et al. (2019) 'An Amplicon-based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and iVar', *Genome Biology*, 20: 8.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics (Oxford, England)*, 34: 4121–3.

He, W., Yi, G. Y., and Zhu, Y. (2020a) 'Estimation of the Basic Reproduction Number, Average Incubation Time, Asymptomatic Infection Rate, and Case Fatality Rate for COVID-19: Meta-analysis and Sensitivity Analysis', *Journal of Medical Virology*, 92: 2543–50.

He, X. et al. (2020b) 'Temporal Dynamics in Viral Shedding and Transmissibility of COVID-19', *Nature Medicine*, 26: 672–5.

Hoang, D. T. et al. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.

Illingworth, C. J. R. et al. (2017) 'On the Effective Depth of Viral Sequence Data', *Virus Evolution*, 3: 2.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Koboldt, D. C. et al. (2012) 'VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing', *Genome Research*, 22: 568–76.

Köster, J., and Rahmann, S. (2012) 'Snakemake—A Scalable Bioinformatics Workflow Engine', *Bioinformatics (Oxford, England)*, 28: 2520–2.

Lemieux, J. E. et al. (2021) 'Phylogenetic Analysis of SARS-CoV-2 in Boston Highlights the Impact of Superspreading Events', *Science*, 371: eabe3261.

Li, H. (2013) *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. arXiv:1303.3997 [q-Bio]. <http://arxiv.org/abs/1303.3997> last accessed 9 Sept 2021.

Liu, Y., Eggo, R. M., and Kucharski, A. J. (2020) 'Secondary Attack Rate and Superspreading Events for SARS-CoV-2', *The Lancet*, 395: e47.

Lythgoe, K. A. et al. (2021) 'SARS-CoV-2 Within-Host Diversity and Transmission', *Science*, 372: eabg0821.

Martin, M. A., and Koelle, K. (2021) 'Comment on "Genomic Epidemiology of Superspreading Events in Austria Reveals Mutational Dynamics and Transmission Properties of SARS-CoV-2"', *Science Translational Medicine*, 13: eabh1803.

McCrone, J. T., and Lauring, A. S. (2016) 'Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling', *Journal of Virology*, 90: 6884–95.

——— (2018) 'Genetic Bottlenecks in Intraspecies Virus Transmission', *Current Opinion in Virology*, 28: 20–5.

McCrone, J. T. et al. (2018) 'Stochastic Processes Constrain the within and between Host Evolution of Influenza Virus', *ELife*, 7: e35962.

Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–4.

Nalla, A. K. et al. (n.d.) 'Comparative Performance of SARS-CoV-2 Detection Assays Using Seven Different Primer-Probe Sets and One Assay Kit', *Journal of Clinical Microbiology*, 58: e00557–20.

Paradis, E., and Schliep, K. (2019) 'Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R', *Bioinformatics (Oxford, England)*, 35: 526–8.

Pfeiffer, F. et al. (2018) 'Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing', *Scientific Reports*, 8: 10950.

Popa, A. et al. (2020) 'Genomic Epidemiology of Superspreading Events in Austria Reveals Mutational Dynamics and Transmission Properties of SARS-CoV-2', *Science Translational Medicine*, 12: eabe2555.

Revell, L. J. (2012) 'Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things)', *Methods in Ecology and Evolution*, 3: 217–23.

San, J. E. et al. (2021) 'Transmission Dynamics of SARS-CoV-2 Within-Host Diversity in Two Major Hospital Outbreaks in South Africa', *Virus Evolution*, 7: 1.

Valesano, A. L. et al. (2020) 'Influenza B Viruses Exhibit Lower Within-Host Diversity Than Influenza A Viruses in Human Hosts', *Journal of Virology*, 94: 5.

——— et al. (2021) 'Temporal Dynamics of SARS-CoV-2 Mutation Accumulation within and across Infected Hosts', *PLOS Pathogens*, 17: e1009499.

van Kampen, J. J. A. et al. (2021) 'Duration and Key Determinants of Infectious Virus Shedding in Hospitalized Patients with Coronavirus Disease-2019 (COVID-19)', *Nature Communications*, 12: 267.

Wang, D. et al. (2021) 'Population Bottlenecks and Intra-host Evolution during Human-to-Human Transmission of SARS-CoV-2', *Frontiers in Medicine*, 8:

Wilm, A. et al. (2012) 'LoFreq: A Sequence-Quality Aware, Ultra-sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets', *Nucleic Acids Research*, 40: 11189–201.

Xue, K. S., and Bloom, J. D. (2019) 'Reconciling Disparate Estimates of Viral Genetic Diversity during Human Influenza Infections', *Nature Genetics*, 51: 1298–301.

Xue, K. S. et al. (2018) 'Within-Host Evolution of Human Influenza Virus', *Trends in Microbiology*, 26: 781–93.

Yu, G. et al. (2017) 'Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data', *Methods in Ecology and Evolution*, 8: 28–36.

Zwart, M. P., and Elena, S. F. (2015) 'Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution', *Annual Review of Virology*, 2: 161–79.