SARS-CoV-2 within-host diversity and transmission

Oxford Virus Sequencing Analysis Group (OVSG); The COVID-19 Genomics UK (COG-UK) Consortium; Beggs, Andrew

DOI:
10.1126/science.abg0821

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

SARS-CoV-2 within-host diversity and transmission

Katrina A. Lythgoe†*, Matthew Hall†*, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cocket, Amy Trebes, Monique Andersson, Newton Otecko, Emma L. Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M. Nicholls, M. Azim Ansari, Lucie Abeler-Dörner, Catrin E. Moore, Timothy E. A. Peto, David W. Eyre, Robert Shaw, Peter Simmonds, David Buck, John A. Todd on behalf of the Oxford Virus Sequencing Analysis Group (OVSG)‡, Thomas R. Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C. Thomson, The COVID-19 Genomics UK (COG-UK) Consortium§, David Bonsall, Christophe Fraser, Tanya Golubchik*

INTRODUCTION: Genome sequencing at an unprecedented scale during the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic is helping to track spread of the virus and to identify new variants. Most of this work considers a single consensus sequence for each infected person. Here, we looked beneath the consensus to analyze genetic variation within viral populations making up an infection and studied the fate of within-host mutations when an infection is transmitted to a new individual. Within-host diversity offers the means to help confirm direct transmission and identify new variants of concern.

RATIONALE: We sequenced 1313 SARS-CoV-2 samples from the first wave of infection in the United Kingdom. We characterized within-host diversity and dynamics in the context of transmission and ongoing viral evolution.

RESULTS: Within-host diversity can be described by the number of intrahost single nucleotide variants (iSNVs) occurring above a given minor allele frequency (MAF) threshold. We found that in lower-viral-load samples, stochastic sampling effects resulted in a higher variance in MAFs, leading to more iSNVs being detected at any threshold. Based on a subset of 27 pairs of high-viral-load replicate RNA samples (>50,000 uniquely mapped veSEQ reads, corresponding to a cycle threshold of ~22), iSNVs with a minimum 3% MAF were highly reproducible. Comparing samples from two time points from 41 individuals, taken on average 6 days apart (interquartile ratio 2 to 10), we observed a dynamic process of iSNV generation and loss. Comparing iSNVs among 14 household contact pairs, we estimated transmission bottleneck sizes of one to eight viruses. Consensus differences between individuals in the same household, where sample depth allowed iSNV detection, were explained by the presence of an iSNV at the same site in the paired individual, consistent with direct transmission leading to fixation. We next focused on a set of 563 high-confidence iSNV sites that were variant in at least one high-viral-load sample (>50,000 uniquely mapped); low-confidence iSNVs unlikely to represent genomic diversity were excluded. Within-host diversity was limited in high-viral-load samples (mean 1.4 iSNVs per sample). Two exceptions, each with >14 iSNVs, showed variant frequencies consistent with coinfection or contamination. Overall, we estimated that 1 to 2% of samples in our dataset were coinfected and/or contaminated. Additionally, one sample was coinfected with another coronavirus (OC43), with no detectable impact on diversity. The ratio of nonsynonymous to synonymous (*dN/dS*) iSNVs was consistent with within-host purifying selection when estimated across the whole genome [*dN/dS* = 0.55, 95% confidence interval (95% CI) = 0.49 to 0.61] and for the Spike gene (*dN/dS* = 0.60, 95% CI = 0.45 to 0.82). Nevertheless, we observed Spike variants in multiple samples that have been shown to increase viral infectivity (L5F) or resistance to antibodies (G446V and A879V). We observed a strong association between high-confidence iSNVs and a consensus change on the phylogeny (153 cases), consistent with fixation after transmission or de novo mutations reaching consensus. Shared variants that never reached consensus (261 cases) were not phylogenetically associated.

CONCLUSION: Using robust methods to call within-host variants, we uncovered a consistent pattern of low within-host diversity, purifying selection, and narrow transmission bottlenecks. Within-host emergence of vaccine and therapeutic escape mutations is likely to be relatively rare, at least during early infection, when viral loads are high, but the observation of immune-escape variants in high-viral-load samples underlines the need for continued vigilance. ■

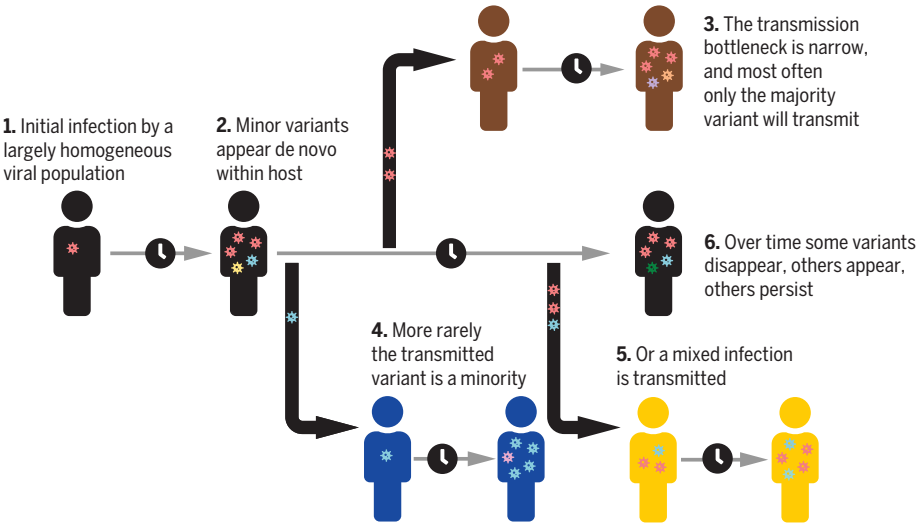

Diagram showing low SARS-CoV-2 within-host genetic diversity and narrow transmission bottleneck. Individuals with high viral load typically have few, if any, within-host variants. Narrow transmission bottlenecks mean that the major variant in the source individual was typically transmitted and the minor variants lost. Occasionally, the minor variant was transmitted, leading to a consensus change, or multiple variants were transmitted, resulting in a mixed infection. Credit: FontAwesome, licensed under CC BY 4.0.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: Tanya.Golubchik@bdi.ox.ac.uk (T.G.); Katrina.Lythgoe@bdi.ox.ac.uk (K.A.L.); Matthew.Hall@bdi.ox.ac.uk (M.H.)
†These authors contributed equally to this work.
‡The full list of the OVSG members is provided in the supplementary materials.
§The full list of names and affiliations of COG-UK members is provided in the supplementary materials.

Cite this article as K. A. Lythgoe *et al.*, *Science* 372, eabg0821 (2021). DOI: 10.1126/science.abg0821

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abg0821

RESEARCH ARTICLE

CORONAVIRUS

SARS-CoV-2 within-host diversity and transmission

Katrina A. Lythgoe[1,2]†*, Matthew Hall[1]†*, Luca Ferretti[1], Mariateresa de Cesare[1,3], George MacIntyre-Cockett[1,3], Amy Trebes[3], Monique Andersson[4,5], Newton Otecko[1], Emma L. Wise[6,7], Nathan Moore[6], Jessica Lynch[6], Stephen Kidd[6], Nicholas Cortes[6,8], Matilde Mori[9], Rebecca Williams[6], Gabrielle Vernet[6], Anita Justice[4], Angie Green[3], Samuel M. Nicholls[10], M. Azim Ansari[11], Lucie Abeler-Dörner[1], Catrin E. Moore[1], Timothy E. A. Peto[4,12], David W. Eyre[4,13], Robert Shaw[4], Peter Simmonds[11], David Buck[3], John A. Todd[3] on behalf of the Oxford Virus Sequencing Analysis Group (OVSG)‡, Thomas R. Connor[14,15], Shirin Ashraf[16], Ana da Silva Filipe[16], James Shepherd[16], Emma C. Thomson[16], The COVID-19 Genomics UK (COG-UK) Consortium§, David Bonsall[1,3,4], Christophe Fraser[1,3,17], Tanya Golubchik[1,2]*

Extensive global sampling and sequencing of the pandemic virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have enabled researchers to monitor its spread and to identify concerning new variants. Two important determinants of variant spread are how frequently they arise within individuals and how likely they are to be transmitted. To characterize within-host diversity and transmission, we deep-sequenced 1313 clinical samples from the United Kingdom. SARS-CoV-2 infections are characterized by low levels of within-host diversity when viral loads are high and by a narrow bottleneck at transmission. Most variants are either lost or occasionally fixed at the point of transmission, with minimal persistence of shared diversity, patterns that are readily observable on the phylogenetic tree. Our results suggest that transmission-enhancing and/or immune-escape SARS-CoV-2 variants are likely to arise infrequently but could spread rapidly if successfully transmitted.

The ongoing evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been the topic of considerable interest as the pandemic has unfolded. Clear lineage-defining single nucleotide polymorphisms (SNPs) have emerged (*1*), enabling tracking of viral spread (*2*, *3*) but also raising concerns that new mutations, or combinations of mutations, may confer selective advantages on the virus, hampering efforts at control. There is compelling evidence that the D614G mutation in the Spike protein (S), which spread globally during the first year of the pandemic, increases viral transmissibility (*4*–*6*). Current variants of concern include the B.1.1.7. lineage (*7*, *8*), with an estimated transmission advantage of ~50% (*9*), and the B.1.351 and P.1 lineages (*10*, *11*), which may have decreased sensitivity to natural and/or vaccine-acquired immunity (*12*–*14*). Lineage codes given here are as designated by Pangolin software (*1*).

Most analyses have been focused on mutations observed in viral consensus genomes, which represent the dominant variants within infected individuals. Ultimately though, new mutations emerge within individuals, so knowledge of the full underlying within-host diversity of the virus at the population level and how frequently this is transmitted is important for understanding adaptation and patterns of spread.

The United Kingdom experienced one of the most severe first waves of infection, with >1000 independent importation events contributing to substantial viral diversity during this period (*15*). In this study, we analyzed 1390 SARS-CoV-2 genomes from 1313 nasopharyngeal swabs sampled predominantly from symptomatic individuals on admission to the hospital and from health care workers during the first wave of infection (March to June 2020; table S1). The dataset comprised samples from 1173 unique individuals, including 41 with samples at two to four time points, plus 93 anonymous samples, with multiple RNA aliquots from 76/1313 samples resequenced to test for reproducibility. The samples were collected by two geographically separate hospital trusts located 60 km apart: Oxford University Hospitals and Basingstoke and North Hampshire Hospital. Using veSEQ, an RNA-Seq protocol based on a quantitative targeted enrichment strategy (*16*), which we previously validated for other viruses (*16*–*19*), we characterized the full spectrum of within-host diversity in SARS-CoV-2 and analyzed it in the context of the consensus phylogeny.

We observed low levels of intrahost diversity in high-viral-load samples, with evidence of within-host evolutionary constraint genome wide, including S. Although within-host variants could be observed in multiple individuals in the same phylogenetic cluster, some of whom resided in the same household, most viral variants were either lost, or occasionally fixed, at the point of transmission, with a narrow transmission bottleneck. These results suggest that during early infection, when viral loads are high and transmission is most likely (*20*–*22*), mutations that increase transmissibility or potential vaccine- or therapy-escape mutations may rarely emerge and subsequently transmit. Nonetheless, we identified variants present in multiple individuals that could affect receptor binding or neutralization by antibodies. Because the fitness advantage of escape mutations in populations that are highly vaccinated or have high levels of natural immunity could be substantial, and because mutational effects can depend on the genetic background on which they are found, these findings underline the need for continued vigilance and monitoring.

Detection of variants is influenced by viral load

Reliable estimation of variant frequencies requires quantitative sequencing such that the number of reads is proportional to the amount of corresponding sequence in the sample of interest. The veSEQ protocol has been shown previously to be quantitative for a number of different pathogens (*17*), including respiratory viruses such as respiratory syncytial virus (RSV) (*18*). We demonstrated here that the same quantitative relationship holds for SARS-CoV-2. The number of uniquely mapped sequencing reads that we obtained rose log-log linearly with the number of RNA copies in serial dilutions of synthetic RNA controls (r^2 = 0.87; fig. S1A) and was consequently correlated with cycle threshold (Ct) values of clinical samples (fig. S1B), indicating that veSEQ reads can be

[1]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. [2]Department of Zoology, University of Oxford, Oxford OX1 3SZ, UK. [3]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Old Road Campus, Oxford OX3 7BN, UK. [4]Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK. [5]Division of Medical Virology, Stellenbosch University, Stellenbosch, South Africa. [6]Hampshire Hospitals NHS Foundation Trust, Basingstoke and North Hampshire Hospital, Basingstoke RG24 9NA, UK. [7]School of Biosciences and Medicine, University of Surrey, Guildford GU2 7XH, UK. [8]Gibraltar Health Authority, Gibraltar, UK. [9]School of Medicine, University of Southampton, Southampton SO17 1BJ, UK. [10]Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK. [11]Peter Medawar Building for Pathogen Research, University of Oxford, Oxford OX1 3SY, UK. [12]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK. [13]Big Data Institute, Nuffield Department of Public Health, University of Oxford, Old Road Campus, Oxford OX3 7FL, UK. [14]Pathogen Genomics Unit, Public Health Wales Microbiology, Cardiff CF10 4BZ, UK. [15]Cardiff University School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK. [16]MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, UK. [17]Wellcome Sanger Institute, Cambridge CB10 1SA, UK.
***Corresponding author. Email: Tanya.Golubchik@bdi.ox.ac.uk (T.G.); Katrina.Lythgoe@bdi.ox.ac.uk (K.A.L.); Matthew.Hall@bdi.ox.ac.uk (M.H.)**
†These authors contributed equally to this work.
‡The full list of the OVSG members is provided in the supplementary materials.
§The full list of names and affiliations of COG-UK members is provided in the supplementary materials.

considered a representative sample of viral sequences within the input RNA.

To understand within-host diversity, we quantified the number of intrahost single-nucleotide variants (iSNVs) in the full set of 1390 genomes, testing different thresholds for identifying variants of between 2 and 5% minor allele frequency (MAF). A minimum depth of at least 100 reads was also required to call an iSNV, and all sites with MAF greater than the threshold were included (Fig. 1A).

For all thresholds, we observed a nonlinear relationship between sample viral load (estimated by total unique mapped reads) and the number of detected iSNVs, with the highest number of iSNVs detected at intermediate viral loads (~2000 mapped reads). However, the mean MAF per sample did not vary with viral load when no threshold was applied (P = 0.291, linear regression; Fig. 1B). This indicates that as the number of mapped reads decreases, the variance in the observed MAF increases, whereas the mean stays the same. This effect is at least partially caused by the inverse relationship of the binomial distribution between the total number of draws and the variance in the proportion of successes observed among those draws. In Fig. 1C, we demonstrate this effect by down-sampling from high-depth samples: The increasing variance associated with sparser sampling causes the number of threshold-crossing iSNVs to increase until eventually so few reads are sampled that no iSNVs are detected.

This sampling effect of low viral load does not preclude the existence of biological mechanisms also contributing to greater intrahost diversity in low-viral-load samples. After the initial peak, viral loads typically decrease as infection progresses (*20*), whereas genetic diversity may increase, as observed in other viral infections such as HIV (*23*). RNA damage (*24*) as infection progresses could also contribute to the observed increased diversity in low-depth samples.

Within-host variant frequencies are reproducible

To calibrate our variant calling and to minimize false discovery rates, we compared iSNVs in resequenced controls with data for the stock RNA sequenced and provided by the manufacturer (Twist Bioscience) and masked sites vulnerable to in vitro generation of variants (table S2). We also masked a further 18 sites that were observed to be variant (>3% MAF) in 20 or more high-viral-load samples (table S3 and fig. S3). Most had consistently low MAFs among samples, and some showed evidence of strand bias and/or low reproducibility between technical replicates (fig. S2), suggesting that they were not true genomic variants. Among the excluded sites was 11083, which was observed in 46 samples and is globally ubiquitous in GISAID (Global Initiative on Sharing All Influenza Data) data. From manual examination of mapped reads in our dataset, this appeared to be caused by a common miscalling of a within-host polymorphic deletion upstream at site 11082 occurring in a poly-T homopolymeric stretch. If genuine, then this homopolymer stutter may have a structural or regulatory role; however, methodological issues in resolving this difficult-to-map region cannot be ruled out.

Establishing reliable variant calling thresholds for clinical samples in which true variant frequencies are unknown ideally requires resequencing of multiple samples from RNA to test for concordance. Working within the constraints of small volumes of remnant RNA from laboratory testing, we resequenced 76 high-viral-load samples, of which 27 replicate pairs generated sufficient read numbers (>50,000 unique mapped reads) for reliable minor variant detection. iSNVs with <2% MAF were generally indistinguishable from noise, whereas those with ≥3% MAF were highly concordant between replicates (Fig. 2A and fig. S2).

Within-host variants vary during infection

We also compared iSNV frequencies and consensus changes at different time points for the 41 multiply sampled individuals, with the duration between sampling ranging between 1 and 20 days apart (median 6 days; Fig. 2, B and C). Because viral loads tend to fall as infection progresses, we considered all samples rather than limiting ourselves to those with >50,000 unique mapped reads. Among the 41 individuals, we observed little concordance in minor variant frequencies across time points within individuals. Our observations, consistent with other studies (*24*–*26*), suggest a dynamic within-host landscape but also reflect the inherent stochasticity associated with low-viral-load samples.

The transmission bottleneck size within households is small

The transmission bottleneck size is a key component in determining the likelihood that new within-host variants will spread in the population (*27*). Estimating bottleneck size is difficult for SARS-CoV-2 because it requires sufficient genetic diversity to differentiate distinct viruses that may be transmitted in known source-recipient pairs (*28*–*31*) and confidence that transmission is the cause of variants observed in both source and recipients. The inclusion of variants that are not shared by transmission can greatly increase transmission bottleneck size estimations (*29*). We identified 16 households in which two individuals had a first positive sample within 2 weeks of each other, and assumed direct transmission if the consensus sequences in the individuals had fewer than three differences (thus excluding one household). A further household was

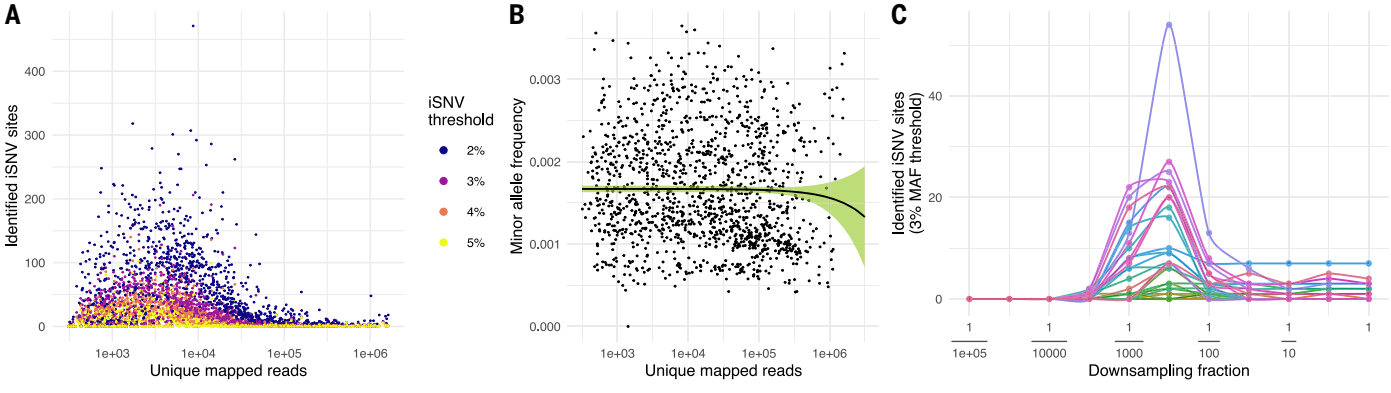

Fig. 1. Characterization of iSNV frequencies. (**A**) Distribution of the number of identified iSNV sites in each sample against the number of unique mapped reads. The colors represent different MAF thresholds. An iSNV site is identified within a sample if the MAF is greater than the threshold. (**B**) Distribution of the mean MAF in each sample against the number of unique mapped reads, with no MAF threshold applied. The black line is the estimated mean value by linear regression. The green ribbon is the 95% CI. (**C**) Distribution of the number of identified iSNV sites at the 3% MAF threshold when subsampling from high-depth samples. Each color represents a different high-depth sample.

excluded because the assumed source individual had no variants with >3% MAF.

Using the exact beta-binomial method (*28*), we estimated maximum likelihood bottleneck sizes between one and eight among the 14 household transmission pairs (Fig. 3A and table S4). These observations are consistent with the small bottleneck sizes observed for influenza (*30*–*32*) and SARS-CoV-2 (*33*–*37*) but considerably lower than estimates in a recent Austrian study (*25*). The reasons for the discrepancies are unclear but could reflect differences in how variants were selected for analysis (*37*) or how closely the observed diversity represents the diversity of virus both available for transmission and successfully transmitted. An association between the route of exposure and the transmission bottleneck has been demonstrated experimentally for influenza (*32*), so genuine differences in bottleneck sizes in different settings cannot be ruled out.

Within-host variants are present in most SARS-CoV-2 samples

To further characterize iSNV sites within individuals, we identified a set of 563 high-confidence iSNV sites that were observed (i) in high-viral load samples with at least 50,000 unique mapped reads (462 samples, 160 from Oxford

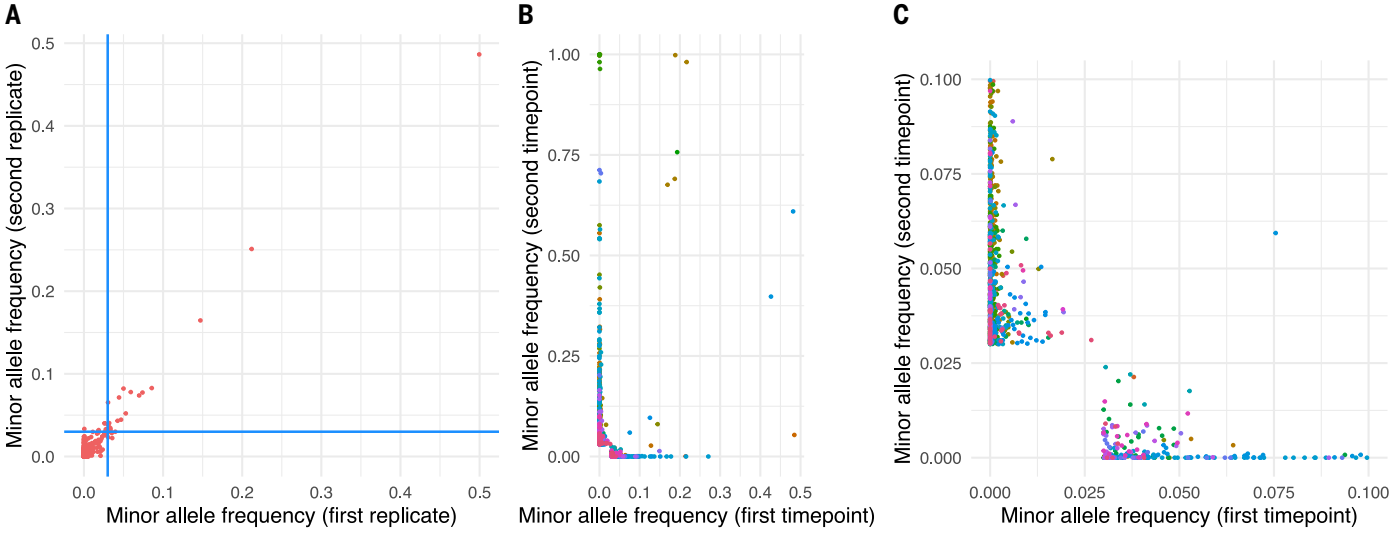

Fig. 2. Comparison of allele frequencies between sequencing replicates of the same sample and multiple time points from the same individual. (**A**) Comparison of MAFs from 27 replicate pairs resequenced from RNA, with each point representing a single genomic position in a pair of replicates. The plot represents all MAF frequency comparisons for the 27 samples where both replicates had >50,000 unique mapped reads, limited to genomic sites with MAF >0.02 in at least one of the 54 replicates. The blue lines are the threshold value of 0.03. (**B** and **C**) Comparison of allele frequencies from 41 individuals sampled on different days, with each point representing a genomic position in a pair of samples from the same individual. Each individual is represented by a different color, and for each individual, all genomic positions are considered where the MAF >0.03 at either sampling time point and/or a change in consensus was observed. In all cases, the poly-A tail and sites variable in RNA synthetic controls were excluded, as were sites observed to be variable in >20 samples at MAF >3% because these are unlikely to represent genomic variants. (C) is an enlargement of the region of (B) near the origin.

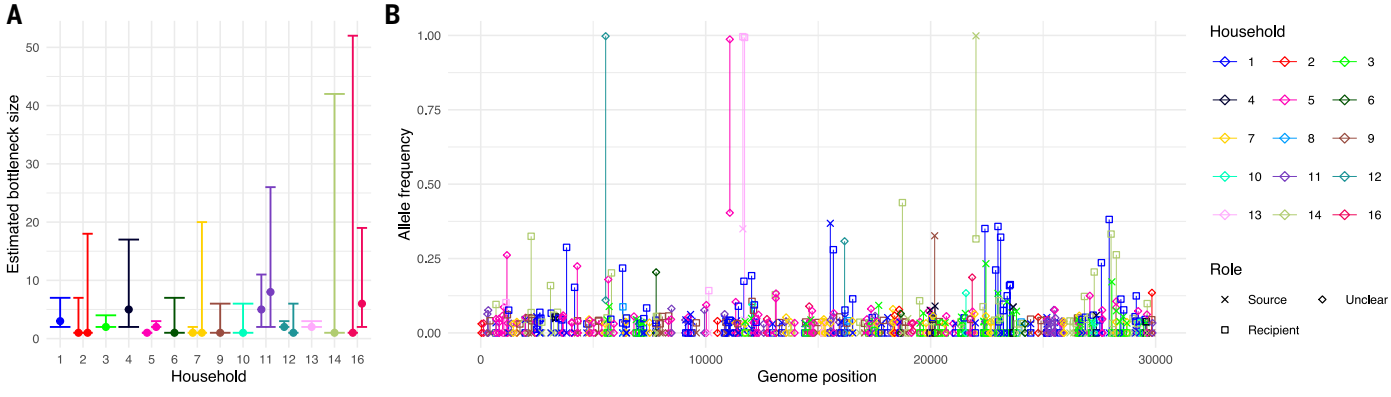

Fig. 3. Small transmission bottleneck size within households. (**A**) Estimated bottleneck size in 14 households calculated using the exact beta-binomial method described in (*28*). Bottleneck size for both combinations of potential source and recipient were calculated if the first positive samples from each individual in the household were collected within a week of each other. No estimate was recorded if there were no identified iSNVs >3% MAF in the source individual (household 8) or if the two individuals in the household had more than two consensus differences (household 15). The error bars represent the 95% CI determined by the likelihood ratio test. (**B**) Fate of the identified iSNVs within households. Each line links the allele frequency of a given variant in one household member with that in the second member. Points and lines are colored by household. Each was identified as an iSNV in at least one individual but not necessarily (and usually not) both. Where the dates of sample collection differed by at least a week, we also indicate the assumed source and recipient members of the household.

and 302 from Basingstoke), (ii) at a depth of at least 100 reads, (iii) with a MAF of at least 3%, and (iv) not observed to vary in synthetic RNA controls or to appear at low frequency in a large number of samples (table S3). All 1313 samples were included in our analysis under the assumption that by ascertaining on a small set of predefined sites, it is less likely that we included sites that only reach >3% MAF in low-viral-load samples because of the stochastic sampling effects described above.

Among the iSNV sites taken forward for variant analysis, most were only observed in one or two of the 1313 samples (Fig. 4A), but most samples with >50,000 unique reads (305/462, 66%) harbored at least one iSNV (Fig. 4B). These low levels of SARS-CoV-2 within-host diversity during acute infection are consistent with other reported levels (*26*, *33*) but lower than in some other studies (*24*, *25*), likely reflecting how variants were identified.

Two samples had a particularly high number (15 and 18) of iSNVs, each with high and correlated MAFs consistent with coinfection by two diverse variant haplotypes (*38*). For one of these samples, laboratory contamination was unlikely because we could not identify any samples that could be the source. We could not distinguish between coinfection and contamination in the other sample because both variant haplotypes within it represented common genotypes in our study.

In general, however, the low level of genetic diversity of the virus makes identifying coinfection or contamination—and distinguishing between them—difficult. If sites where a large number of SNPs is present (mutations that distinguish common lineages in our dataset) are only observed to be variant within host because of coinfection or contamination, then we estimate that between ~1 and 2% of samples are potentially affected by coinfection or contamination (table S2). As a precaution against contamination or batch effects, we sequenced known epidemiologically linked samples in different batches where possible (fig. S4).

We hypothesized that a proportion of the observed within-host variation could have been due to coinfection with seasonal coronaviruses, which has been reported in 1 to 4% of SARS-CoV-2 infections (*39*, *40*). Specifically, closely matching reads from similar viruses could be mapped to SARS-CoV-2 and appear as mixed-base calls. To understand the impact of coinfection, we recaptured and analyzed a random subset of 180 samples spanning the full range of observed SARS-CoV-2 viral loads (Ct 14 to 33, median 19.8) using the Castanet multipathogen enrichment panel (*17*), which contains probes for all known human coronaviruses with the exception of SARS-CoV-2. Among the 111 samples that yielded both SARS-CoV-2 and Castanet data, we identified one sample that was also positive for another betacoronavirus, human coronavirus OC43 (fig. S5). Within the SARS-CoV-2 genome from this sample, which was complete and high-depth, we observed only a single iSNV at position 28580 and no evidence of mixed-base calls at any other genomic position. This suggests that even when coinfection was present, it did not affect the estimation of SARS-CoV-2 within-host diversity in our protocol. However, whether coinfection with OC43 or other coronaviruses exerts a selective pressure on SARS-CoV-2 remains an open question.

Distribution of iSNVs across the genome

We next considered the distribution of the identified high-confidence iSNV sites across the genome. Even excluding the untranslated regions (UTRs), which have a highly elevated density of iSNV sites, there was considerable variability across the genome, with open-reading frames (ORFs) 3a, 7a, and 8 and nucleocapsid (N) showing the highest densities (Table 1). In addition, we calculated ratio of nonsynonymous to synonymous substitutions (*dN/dS*) values under the assumption that each iSNV appeared de novo in each individual in which it was observed (Table 1). Consistent with other studies (*24*, *33*), most areas of the genome appeared to be under purifying selection, with *dN/dS* values <1, including S. Without a full model incorporating within-host evolutionary dynamics and transmission, it is difficult to draw strong conclusions. However, we obtained similar results assuming that each iSNV was only generated once de novo and then subsequently transmitted (table S5). These patterns are also broadly consistent with *dN/dS* values calculated for SNPs among SARS-CoV-2 consensus genomes (*41*), suggesting that evolutionary forces at the within-host level are reflected at the between-host level, at least for within-host variant sites in high-viral-load samples.

Within-host variant sites are phylogenetically associated

We sought to gain a better understanding of SARS-CoV-2 evolution and to determine whether iSNVs could be used to help resolve phylogenies and transmission clusters. For the 1390 genomes in our study, we constructed a phylogeny using the robust procedure outlined by (*42*) (Fig. 5A). Viral phylogenies are based on the consensus sequence for each sample, with branches indicating differences in the consensus sequence among samples. Given the inferred narrow transmission bottleneck size, we hypothesized that consensus changes on the phylogeny arise because of the emergence of within-host variants that either reach consensus within the individual in which they emerged or fail to reach consensus but are then transmitted and result in a consensus change in the recipient. In a sufficiently densely sampled

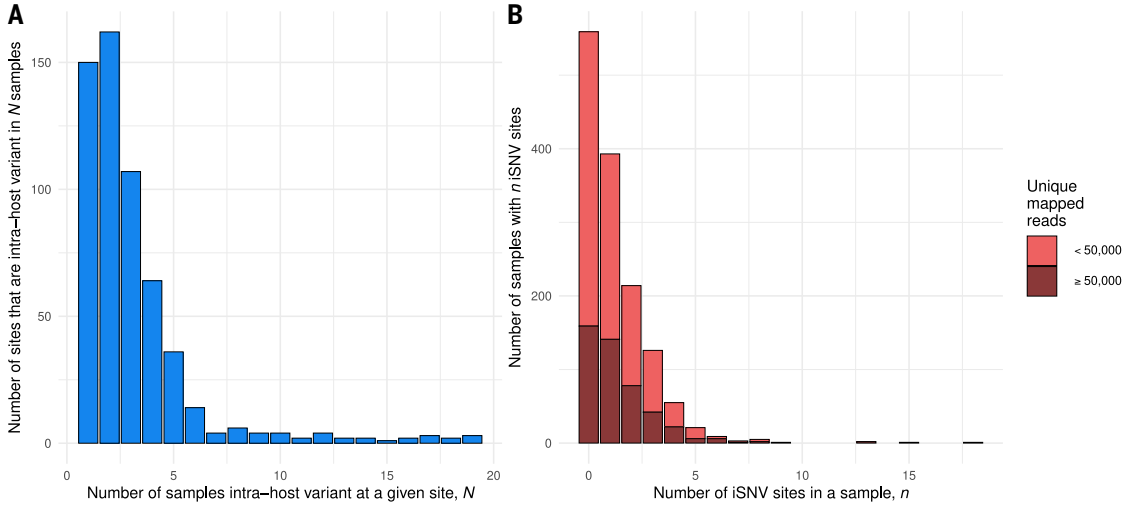

Fig. 4. iSNV sites were often found in multiple samples and most samples had at least one iSNV. (**A**) Histogram showing the number iSNV sites that were found in *N* samples. All samples in our dataset are included. (**B**) Stacked histogram showing the number of samples that had *n* iSNV sites for all samples with >50,000 mapped reads (dark red) and samples with <50,000 mapped reads (light red). All 563 sites identified for variant analysis were included (see main text), including sites in the 3′UTR and 5′UTR but excluding the polyA tail and the 18 sites variable in 20+ individuals.

Table 1. iSNVs and *dN/dS* by gene and over the whole genome.

Gene	Length	iSNVs			Mean iSNVs per 100 sites	*dN/dS* (95% CI)
		Total	NS	S		
5′UTR	265	82	-	-	0.0223	-
ORF1a	13218	572	369	203	0.0031	0.51 (0.43, 0.61)
nsp1	540	54	39	15	0.0072	0.79 (0.44, 1.47)
nsp2	1914	105	65	40	0.0039	0.46 (0.31, 0.69)
nsp3	5835	175	108	67	0.0022	0.45 (0.33, 0.61)
nsp4	1500	101	61	40	0.0048	0.44 (0.3, 0.66)
nsp5A	918	25	22	3	0.002	2.08 (0.72, 8.77)
nsp6	870	62	42	20	0.0051	0.58 (0.35, 1.01)
nsp7	249	6	2	4	0.0017	0.14 (0.02, 0.73)
nsp8	594	13	7	6	0.0016	0.32 (0.11, 0.98)
nsp9	339	15	9	6	0.0032	0.46 (0.17, 1.37)
nsp10	417	16	14	2	0.0028	1.99 (0.56, 12.67)
nsp12*	2795	122	69	53	0.0031	0.34 (0.24, 0.49)
ORF1b	8088	349	212	137	0.0031	0.42 (0.34, 0.52)
nsp13	1803	59	33	26	0.0024	0.37 (0.22, 0.63)
nsp14	1581	92	59	33	0.0042	0.48 (0.31, 0.74)
nsp15	1038	31	21	10	0.0021	0.57 (0.27, 1.26)
nsp16	894	45	30	15	0.0036	0.54 (0.29, 1.03)
S	3822	190	129	61	0.0036	0.6 (0.45, 0.82)
ORF3a	828	108	96	12	0.0094	2.29 (1.31, 4.4)
E	228	13	4	9	0.0041	0.15 (0.04, 0.47)
M	669	32	20	12	0.0034	0.51 (0.25, 1.08)
ORF6	186	10	8	2	0.0039	0.97 (0.24, 6.43)
ORF7a	366	41	34	7	0.0081	1.43 (0.67, 3.52)
ORF7b	132	8	8	0	0.0044	∞ (0.93, ∞)
ORF8	366	49	19	30	0.0096	0.17 (0.09, 0.3)
N	1260	145	106	39	0.0083	0.81 (0.56, 1.18)
ORF10	117	11	6	5	0.0068	0.32 (0.09, 1.09)
3′UTR	229	74	-	-	0.0232	-
All coding regions†	29260	1526	1009	517	0.0038	0.55 (0.49, 0.61)
Full genome	22903	1708	-	-	0.0041	-

All genome positions are relative to the Wuhan-Hu-1 reference sequence. iSNVs at the 18 "highly shared" sites and those identified from the synthetic controls are excluded, as are those in the poly-A tail (positions 29865 to 29903). The "mean iSNVs per 100 sites" column is the mean number in each gene over all 1390 sequenced genomes. Note that because of gene overlap and noncoding intergenic regions, the total number of iSNVs (1708) cannot be obtained as the sum of any column in this table, even if the rows for nonstructural proteins in ORF1ab are excluded. *nsp12 overlaps the boundary between ORF1a and ORF1b. †Intergenic regions are excluded from this row.

population of infected individuals, we should therefore be able to observe a phylogenetic association between samples containing iSNVs with branches on the tree leading to a change in consensus at the same locus.

Of the 563 high-confidence iSNV sites, we identified 153 sites that were present in at least two samples and in which we also observed differences in the consensus among samples (SNPs). We call these sites iSNV-SNPs. We examined the proximity of tips with the iSNVs to the position of consensus changes (between the two most common bases at the site of the iSNV) on the phylogeny. A highly significant negative association (one-sided Mann-Whitney *U* test, $P < 3 \times 10^{-16}$; fig. S6A) was found between the presence of an iSNV at a given site in a sample and the patristic distance to the nearest example of a consensus change at the same site; that is, intrahost variation clustered on the tree with branches supported by the same variant as consensus. When we tested sites where we had identified at least two iSNVs individually, six showed a significant association after Benjamini-Hochberg correction ($P < 0.05$), reducing to five if only one sample from each individual was included. Repeating this procedure on each of 1000 phylogenetic bootstrap replicates yielded a universally very strong association when taking sites across the whole genome (maximum $P = 2.46 \times 10^{-10}$), whereas every bootstrapped tree had between one and nine significant iSNV-SNPs (median seven, IQR five to seven).

In Fig. 5B, we show the example of site 28580 (significant in 85.8% of bootstrap replicates), with the red clade representing change from the global consensus G to A (a nonsynonymous change D103N in N) and nearby iSNVs occurring both as minor As in the nodes ancestral to the change branch and as minor Gs in the branch's immediate descendants. Based on corresponding epidemiological data, this represents a health care-associated cluster with onward transmission to close contacts. In Fig. 5C, we give the further example of site 20796 (significant in 98.4% of bootstrap replicates), a synonymous substitution L6843 in ORF1a. Trees for the other significant sites after Benjamini-Hochberg correction are shown in fig. S7. Supporting this relationship between SNPs and iSNVs, we note that in the household transmission pairs that we examined, for the five consensus differences in which there was sufficient depth, all were within-host variant in one of the two individuals (Fig. 4B).

For the 261 iSNVs that were present in at least two individuals but never reached consensus, we analyzed the association with the phylogeny of each iSNV as a discrete trait using two statistics: the association index (*34*) and the mean patristic distance between iSNV tips. After adjustment for multiple testing, no sites showed a *P*-value <0.05 for a phylogeny-iSNV association for either statistic. Similarly, if we simply compared the distance to the nearest iSNV tip among iSNV and non-iSNV tips across all 261 iSNV sites, there was also no evidence of phylogenetic association (one-sided Mann-Whitney *U* test, $P \approx 1$; fig. S6B). Nevertheless, some individual sites did show patterns suggestive of iSNV transmission, with diversity maintained after transmission (22 with $P < 0.05$ before adjustment for multiple testing for at least one of the two statistics; the nine with $P < 0.025$ are shown in fig. S7), suggesting that we may lack the power to statistically detect some associations. Among the 15 household transmission pairs, we observed only one iSNV shared in two individuals within the same household. This iSNV was specific to these two individuals in our dataset, demonstrating a likely example of transmitted viral diversity (Fig. 3B).

Taken together, our observations suggest that the transmission bottleneck can be wide enough to permit cotransmission of multiple genotypes in some instances but narrow enough that multiple variants do not persist after a small number of subsequent transmissions. In the cases in which transmission culminated in a consensus change on the phylogeny, these patterns were readily observable, but in most cases, we suggest that patterns of cotransmission were drowned out by the high proportion of iSNVs that failed to transmit or were transmitted but then lost. Analysis of transmission events over multiple generations is needed to fully elucidate these patterns.

Variants occurring repeatedly but without phylogenetic association could indicate sites under selection in distinct individuals (*43*). Of

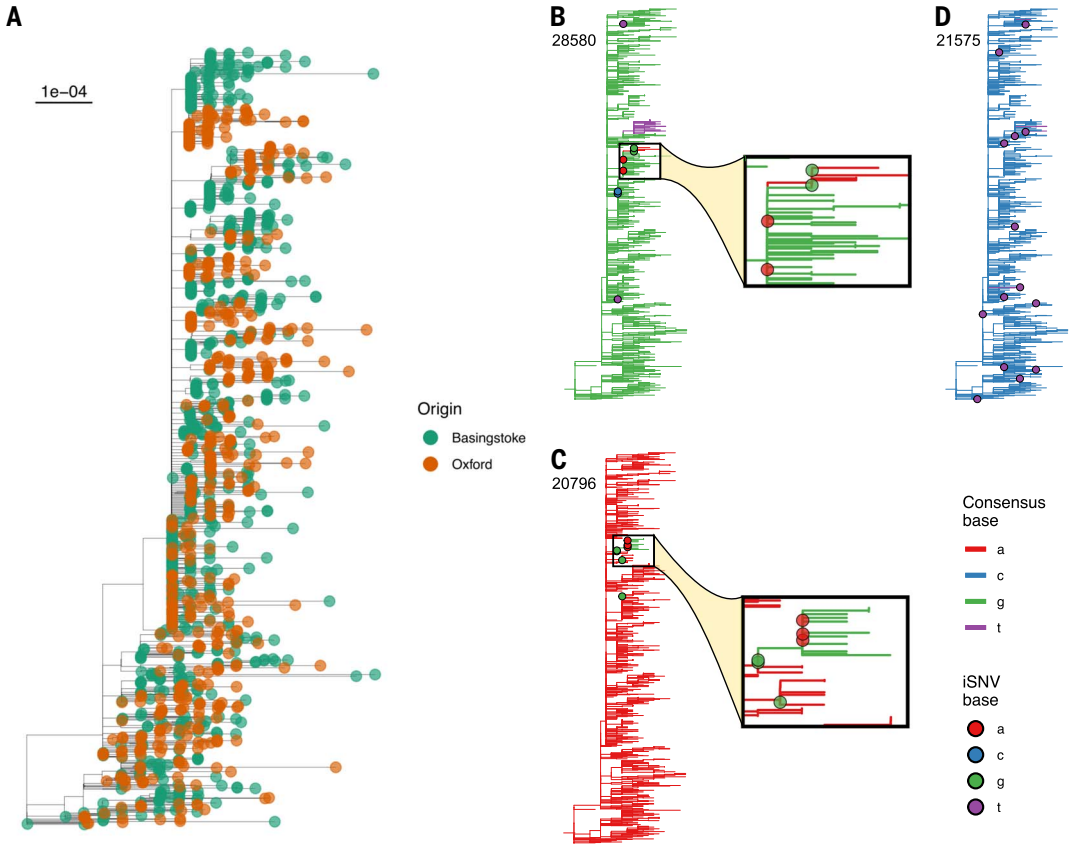

Fig. 5. Consensus phylogeny of all isolates. In (**A**), tips are colored by sampling center (Oxford = orange; Basingstoke = green). The tree scale is in substitutions per site. (**B** to **D**) Distribution of samples with iSNVs at three loci. The genomic coordinate (with respect to the Wuhan-Hu-1 reference sequence) appears in the top left. Tree branches are colored by the consensus base at that position, and filled circles indicate iSNVs present at a minimum of 3% frequency in samples with depth of at least 100 at that position, and are colored by the most common minor variant present. For sites 28580 (B) and 20796 (C), an inset panel enlarges the section of the phylogeny where a consensus change is in close proximity to iSNVs with the relevant pair of nucleotides involved. The highlighted samples were prepared in separate batches and the patterns were not caused by contamination. (D) Variants at site 21575 (L5F) occurred in 14 samples but with no phylogenetic association with consensus changes at this site, which may represent independent emergence of this variant in multiple individuals. The phylogeny was constructed by maximum likelihood according to the robust procedure outlined by Morel *et al*. (*42*).

particular note are the variants that we observed at three sites in S: 21575 (L5F), 22899 (G446V), and 24198 (A879V), with G446V lying within the receptor-binding domain. The minor variant F5 was observed in 14 samples and represented SNPs in eight samples but did not have phylogenetic association in our iSNV-SNP analysis (P = 0.771 before multiple testing adjustment; Fig. 5D). This L5F mutation has been shown to increase infectivity in vitro (*44*) and has previously been identified as a potential site subject to selection (*45*). This variant has repeatedly been observed in global samples, including as minority variant, but appears to be increasing in frequency slowly if at all, suggesting that it is only advantageous within a small subset of individuals, with the variant either "reverting" in subsequent infections [as seen in HIV (*46*)] or failing to transmit at all. Similarly, we observed the minor variants V446 and V879 in four and six individuals, respectively. Both variants have previously been shown to reduce sensitivity to convalescent sera in vitro (*44*), and V446 strongly reduces binding of one of the antibodies (REGN10987) in the REGN-Cov2 antibody cocktail (*47*), suggesting that these may represent antibody escape mutations. We did not observe N501Y or E484K, both mutations of concern, in any of our samples (*48*).

Concluding remarks

We uncovered a consistent and reproducible pattern of within-host SARS-CoV-2 diversity in a large dataset of >1000 individuals, with iSNV sites showing strong phylogenetic clustering patterns if they were also associated with a change in the consensus variant at the same site. However, most samples harbored few intrahost variants, and estimated transmission bottleneck sizes were very small, with maximum likelihood estimates between 1 and 8 among household transmission pairs. This means that if mutations do arise, they will be prone to loss at the point of transmission. The dense sampling and deep sequencing of SARS-CoV-2 has enabled us to witness "evolution in action," with variants generated in one individual, if transmitted, leading to a change in consensus and fixation in subsequently infected individuals. This suggests that within-host variants could be used, at least in some instances, to help better resolve patterns of transmission in a background of low consensus diversity.

Our observations indicate that the within-host emergence of vaccine- and therapeutic-escape mutations is likely to be relatively rare, at least during early infection, when viral loads are high. However, even in the absence of vaccine or therapeutic selection pressure, potential host-adaptive mutations are observable with sufficient frequency that even a rare transmission event combined with narrow bottleneck size could result in rapid spread. Here, we identified 30 nonsynonymous minor variants in S that were present in multiple individuals (table S2). Two of these (G446V and A879V) have previously been shown to escape antibody binding (*44*), and a third, L5F, has been shown to increase viral infectivity (*44*). We suggest that commonly occurring iSNVs, along with variants known to affect transmissibility, severity of infection, or immune responses, should be investigated and monitored, particularly as vaccines and therapeutics are rolled out more widely.

The emergence of new variants of concern, including B.1.1.7, B.1.351, and P.1, underscores the need for continued vigilance. A leading hypothesis is that these variants, characterized by a large number of nonsynonymous mutations, originated within individuals with long durations of infection during which the virus was subject to prolonged immune pressure (*7*, *8*), and that this was potentially facilitated

by the within-host emergence of deletions (*49*). However, the presence of multiple mutations on the same genetic background is not a necessary prerequisite for a new variant to be cause for concern. The single D614G S mutation spread globally after it emerged during the early stages of the pandemic, likely because of a transmission advantage (*50*). The potential for mutations including N439K and E484K, which may enable the virus to evade host-immune responses (*47*, *51*), to emerge on the highly transmissible B.1.1.7 background is also troubling, particularly as population immunity builds due to natural infection and vaccination.

Our work demonstrates that an essential requirement for incorporating intrahost variants in any analysis is an understanding of the observed intrahost diversity in the context of the methods used to produce the deep-sequencing data. Throughout this study, we aimed to minimize sequencing artifacts and sample contamination where possible. Moreover, our results emphasize the power of open data, large and rigorously controlled datasets, and the importance of integrating genomic, clinical, and epidemiological information to gain an in-depth understanding of SARS-CoV-2 as the pandemic unfolds.

Materials and methods

RNA extraction

Residual RNA from COVID-19 reverse transcription quantitative polymerase chain reaction (RT-qPCR)–based testing was obtained from Oxford University Hospitals (hereafter "Oxford"), extracted on the QIASymphony platform with QIAsymphony DSP Virus/Pathogen Kit (QIAGEN), and from Basingstoke and North Hampshire Hospital (hereafter "Basingstoke"), extracted with one of the following: the Maxwell RSC Viral total nucleic acid kit (Promega), the Reliaprep blood gDNA miniprep system (Promega), or the Prepito NA body fluid kit (PerkinElmer). An internal extraction control was added to the lysis buffer before extraction to act as a control for extraction efficiency [genesig qRT-PCR kit, #Z-Path-2019-nCoV in Basingstoke, MS2 bacteriophage (*52*) in Oxford]. The #Z-Path-2019-nCoV control is a linear, synthetic RNA target based on sequence from the rat *ptprn2* gene, which has no sequence similarity with SARS-CoV-2 (GENESIG PrimerDesign, personal communication, 6 April 2020). The MS2 RNA likewise has no SARS-CoV-2 similarity (*52*). Neither control RNA interfered with sequencing.

Targeted metagenomic sequencing

Samples with suspected epidemiological linkage, where this information was available before sequencing, were processed in different batches. Sequencing libraries were constructed from remnant volume of nucleic acid after clinical testing, ranging from 5 to 45 µl (median 30 µl) for each sample depending on the available amount of eluate. These volumes represented 1 to 15% of the original specimen (swab). Libraries were generated following the veSEQ protocol (*16*) with some modifications. Briefly, unique dual indexed (UDI) libraries for Illumina sequencing were constructed using the SMARTer Stranded Total RNA-Seq Kit v2 Pico Input Mammalian (Takara Bio) with no fragmentation of the RNA. An equal volume of library from each sample was pooled for capture. Size selection was performed on the captured pool to eliminate fragments shorter than 400 nucleotides (nt), which otherwise may be preferentially amplified and sequenced. Targeted enrichment of SARS-CoV-2 libraries in the pool was obtained through a custom xGen Lockdown Probes panel (IDT), using the SeqCap EZ Accessory Kits v2 and SeqCap Hybridization and Wash Kit (Roche) for hybridization of the probes and removal of unbound DNA. After 12 cycles of PCR for postcapture amplification, the final product was purified using Agencourt AMPure XP (Beckman Coulter). Sequencing was performed on the Illumina MiSeq (batches 1 and 2) or NovaSeq 6000 (batches 3 to 27) platform (Illumina) at the Oxford Genomics Centre, generating 150–base pair (bp) or 250-bp paired-end reads.

Quantification controls

A dilution series of in vitro–transcribed SARS-CoV-2 RNA [Twist Synthetic SARS-CoV-2 RNA Control 1 (MT007544.1), Twist Bioscience] was included in every capture pool of 90 samples starting from batch 3 and sequenced alongside the clinical samples. Control RNA was serially diluted into Universal Human Reference RNA (UHRR) to a final concentration of SARS-CoV-2 RNA of 500,000, 50,000, 5000, 500, 100, and 0 copies/reaction. From this, we produced a standard curve demonstrating linear association between viral load and read depth (fig. S1). For an experiment comparing iSNV presence with and without probe capture, we additionally sequenced two replicates of the Twist RNA control without capture, diluted into UHRR to give an expected concentration of 50,000 copies per reaction.

As an additional validation step, we compared iSNVs in resequenced controls with data for the stock RNA sequenced and provided by the manufacturer (Twist Bioscience). Six well-defined iSNVs, which were present in the manufacturer's data and presumably arose during in vitro transcription, were also recovered by our protocol (fig. S8). In addition, we identified 112 sites that appeared vulnerable to low-frequency intrahost variation in vitro (table S3), possibly as a result of structural variation along the genome or interaction with the sequencing protocol. We blacklisted vulnerable sites from further analysis.

In-run controls

In addition to the synthetic RNA standards described above, each batch included a non-SARS-CoV-2 in-run control consisting of purified, in vitro–transcribed HIV RNA from clone p92BR025.8 obtained from the National Institute for Biological Standards and Control (*53*). For batches 1 and 2, which were sequenced before synthetic RNA became available, we included negative buffer controls. As additional negative controls, we sequenced six matched clinical samples from non–COVID-19 patients distributed across different sequencing runs, and none contained any SARS-CoV-2 reads.

Minimizing risk of index misassignment

All samples had UDI to prevent cross-detection of reads in the same pool. The in-run HIV RNA controls were used to estimate index misassignment because this provided a sequence-distinct source of RNA: <3 SARS-CoV-2 reads were detected in any HIV control (median 0), and <10 HIV reads were detected in any SARS-CoV-2 control (median 0), suggesting that index misassignment, if present, occurred at extremely low levels.

Bioinformatics processing

Demultiplexed sequence read pairs were classified by Kraken version 2 (*54*) using a custom database containing the human genome (GRCh38 build) and the full RefSeq set of bacterial and viral genomes (pulled May 2020). Sequences identified as either human or bacterial were removed using filter_keep_reads.py from the Castanet (*17*) workflow (*55*). Remaining reads, composed of viral and unclassified reads, were trimmed in two stages: first to remove the random hexamer primers from the forward read and SMARTer TSO from the reverse read, and then to remove Illumina adapter sequences using Trimmomatic version 0.36 (*56*), with the ILLUMINACLIP options set to "2:10:7:1:true MINLEN:80." Trimmed reads were mapped to the SARS-CoV-2 RefSeq genome of isolate Wuhan-Hu-1 (NC_045512.2) using *shiver* (*57*) version 1.5.7, with either smalt (*58*) or bowtie2 (*59*) as the mapper. Both mappers generated comparable results, and smalt was used for the final analysis. Only properly paired reads with insert size <2000 and with at least 70% sequence identity to the reference were retained. For analysis of consensus genomes, consensus calls required a minimum of two uniquely mapped (deduplicated) reads per position, equivalent to >15 raw reads per position. Analysis of within-host diversity was restricted only to positions with minimum raw depth of 100, except when examining diversity within presumed recipients of transmissions in the bottleneck analysis. MAFs were computed at every position using shiver (*57*) (tools/AnalysePileup.py), with the default settings of no BAQ and maximum pileup depth of

1000000. Lineages were assigned by the Pangolin web server (*60*) using the determined consensus genome for each sequenced sample.

Alignment

Oxford and Basingstoke samples were selected if the consensus sequence (inferred from unique mapped reads) consisted of no more than 25% N characters. As an alignment to the reference sequence was already performed in *shiver*, no further alignment was necessary. To place these data into the global phylogenetic context and to help resolve ancestry, a collection of non-UK consensus sequences from the GISAID database (*61*) were included in the set of sequences to be aligned. All GISAID (*62*) sequences were downloaded from the database on 26 April 2020 and filtered to remove sequences that were <29,800 base pairs in length, had >1% Ns, or were from the United Kingdom. The remaining sequences were clustered using CD-HIT-EST (*63*) using a similarity threshold of 0.995, and then one sequence per cluster picked. The resulting set, along with the reference genome Wuhan-Hu-1 (RefSeq ID NC_045512), were aligned using MAFFT (*64*), with some manual improvement of the algorithmic alignment and removal of problematic sequences performed as a postprocessing step. Indels with respect to Wuhan-Hu-1 in both the Oxford and/or Basingstoke and GISAID alignments were deleted, resulting in two alignments of 29,903 nucleotides that could be readily combined.

Demonstration of the effect of read down-sampling

To demonstrate the effect of read depth on estimated iSNV counts, we selected the 30 samples with the highest total number of mapped reads, chose a variety of down-sampling fractions for each, and removed all but that proportion of called bases from consideration. We then determined, for each sample and fraction, the number of iSNVs that would be identified at a threshold of 3% MAF at a minimum depth of 100 if only that fraction of called bases were available to us.

Transmission bottleneck analysis

Sixteen potential transmission pairs were identified by shared address (household) and first positive sample within 2 weeks. If samples from the two individuals in the household differed by fewer than three consensus differences (15 households), direct transmission was assumed. Apart from one genome position in household 6 and one in household 12, all sites associated with a consensus difference within a household were within-host variable in at least one member of the household pair, lending support to assumption of direct transmission (the exceptions are associated with low-read samples). Household 15 had six consensus differences and was therefore excluded from our bottleneck analysis, although we note that for all six positions, the site was within-host variable in one or other individual. This pattern is inconsistent with direct transmission but may represent transmission from a common source. When the first samples for each individual in the household were >1 week apart, we assumed that the earlier sampled individual was the source; otherwise, we considered both possible directions of transmission. If individuals had more than one sample or replicate sequences from the same sample, then we used the sample and/or replicate with the highest number of mapped reads.

Bottleneck size was calculated using the exact beta-binomial method described in (*28*). Because most samples in the analysis had <50,000 mapped reads, we considered all sites in the genome, including sites in the 3′ and 5′ UTR, but excluding the poly-A tail (positions 29865 to 29903), the 18 "highly shared" sites, and those identified from the synthetic controls. All sites with >3% MAF and >100 reads in the assumed source individual were used in the analysis. In the recipient, all reads at these sites were considered, with an error threshold of 0.5% MAF. Following (*28*), 95% confidence intervals (CIs) were calculated using a likelihood ratio test. No estimate was recorded for household 8 because there were no identified iSNVs >3% in the source.

Calculation of dN/dS

The total number of synonymous and nonsynonymous substitutions in the SARS-CoV-2 genome was estimated using the first method of (*65*) applied to the coding regions of the Wuhan-Hu-1 reference sequence. Overlapping reading frames were accounted for such that a substitution was considered nonsynonymous overall if it was nonsynonymous in either frame.

We took two approaches to this calculation, first by counting all iSNVs individually, and second by counting only unique iSNVs. In the latter case, where we detected iSNVs with different base changes at the same position, we included only the most frequent. The results of the former are the basis for Table 1, whereas those of the latter appear in table S5.

The *dN/dS* ratio for iSNVs over a genomic region G was then calculated as follows:

$$\frac{\sum_{p\in G} i_p^N}{T_G^N} \Big/ \frac{\sum_{p\in G} i_p^S}{T_G^S}$$

where i_p^N is the fraction of iSNVs at p that are nonsynonymous, or 0 if there are no iSNVs at p; T_G^N is the total number of potential nonsynonymous substitutions in G; and the denominator replaces N with S to represent synonymous substitutions. The 95% CIs for these estimates were obtained using the likelihood ratio test.

Phylogenetics

Phylogenetic reconstruction was performed on the alignment consisting of the 1390 consensus sequences, along with the GISAID set and the Wuhan-Hu-1 reference sequence. We followed the recommendations of Morel *et al.* (*42*), in which 100 separate maximum likelihood phylogenies were generated using RAxML-NG (*66*) and the GTR+G substitution model, such that each reconstruction used a different random starting parsimony tree. The final phylogeny was then obtained from this set using majority rule. This final tree was rooted with respect to the reference sequence, and then that and all GISAID isolates were pruned.

To identify homoplasic sites, we selected sites that changed state more than once along the tree after inferring the states at internal nodes using ancestral state reconstruction as implemented in ClonalFrameML (*67*) and rooting the tree using the reference genome NC_045512.

The recommendations of Morel *et al.* do not easily lend themselves to fast bootstrapping, so to explore phylogenetic uncertainty, we performed an additional phylogenetic reconstruction on the same alignment using the ultrafast bootstrap procedure in IQ-TREE (*68*). A total of 1000 bootstrap replicates were used.

Phylogenetic association of iSNVs and SNPs

Where an iSNV corresponded to a consensus SNP (by the base pair involved, not simply the site), we performed ancestral state reconstruction on the consensus trees using ClonalFrameML (*67*) to identify all branches upon which that substitution was involved. Tips derived from the same clinical sample were then pruned until only one (the one with the highest overall depth) remained. Then, for each tip in the tree, we calculated the patristic distance from that tip to the midpoint of the closest one of these branches and used a one-tailed Mann-Whitney U test to test for association between the iSNV existing in a sample and this distance. Multiple testing was controlled for using the Benjamini-Hochberg adjustment. As a sensitivity analysis, this was repeated such that all but one tip per infected individual, rather than per clinical sample, were pruned. These analyses were done both on an individual site level and across all sites of interest.

To confirm that the associations that we observed here were unaffected by phylogenetic uncertainty, we used the set of 1000 IQ-TREE bootstraps. We repeated the Mann-Whitney U tests above for each of these 1000 trees.

Phylogenetic association of iSNVs at consensus invariant positions

For the remaining iSNVs, we calculated the extent of association with the consensus

phylogeny by treating the presence of an iSNV as a discrete character and calculating the association index and the mean patristic distance between iSNV tips. Once again, the consensus tree was pruned such that tips corresponding to samples with read depth <100 at the position and all but one tip coming from the same individual were removed. A null distribution was generated by permuting the tip labels of this tree 10,000 times, and a one-sided permutation test *P*-value was calculated. Multiple testing was adjusted for as above. In addition, for each tip in the phylogeny at each site of interest, we calculated the minimum patristic distance to a different tip corresponding to an iSNV and used the Mann-Whitney *U* test again to compare the distribution of these distances between iSNV and non-iSNV tips.

REFERENCES AND NOTES

1. A. Rambaut *et al*., A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020). doi: 10.1038/s41564-020-0770-5; pmid: 32669681
2. J. Hadfield *et al*., Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018). doi: 10.1093/bioinformatics/bty407; pmid: 29790939
3. J. Lu *et al*., Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997–1003.e9 (2020). doi: 10.1016/j.cell.2020.04.023; pmid: 32359424
4. B. Korber *et al*., Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19 (2020). doi: 10.1016/j.cell.2020.06.043; pmid: 32697968
5. E. Volz *et al*., Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64–75.e11 (2021). doi: 10.1016/j.cell.2020.11.020; pmid: 33275900
6. Y. J. Hou *et al*., SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020). pmid: 33184236
7. A. Rambaut *et al*., on behalf of COVID-19 Genomics Consortium UK (CoG-UK), Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations (2020); 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.
8. S. A. Kemp, W. T. Harvey, S. Lytras, The COVID-19 Genomics UK (COG-UK) consortium, A. M. Carabelli, D. L. Robertson, R. K. Gupta, Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion H69/V70. bioRxiv 422555 [Preprint] 9 February 2021. https://doi.org/10.1101/2020.12.14.422555.
9. E. Volz *et al*., Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. medRxiv 20249034 [Preprint]. 4 January 2021. https://doi.org/10.1101/2020.12.30.20249034.
10. H. Tegally *et al*., Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv 20248640 [Preprint]. 22 December 2020. https://doi.org/10.1101/2020.12.21.20248640.
11. N. R. Faria *et al*., Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings (2021); https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586).
12. C. K. Wibmer *et al*., SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. bioRxiv 427166 [Preprint]. 1 March 2021. https://doi.org/10.1101/2021.01.18.427166.
13. S. Cele *et al*., Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. medRxiv 21250224 [Preprint]. 27 February 2021. https://doi.org/10.1101/2021.01.26.21250224.
14. K. Wu *et al*., mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. bioRxiv 427948 [Preprint]. 25 January 2021. https://doi.org/10.1101/2021.01.25.427948.
15. L. du Plessis *et al*., Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* 10.1126/science.abf2946 (2021).
16. D. Bonsall *et al*., A comprehensive genomics solution for HIV surveillance and clinical monitoring in low-income settings. *J. Clin. Microbiol.* **58**, e00382-20 (2020). doi: 10.1128/JCM.00382-20; pmid: 32669382
17. C. Goh *et al*., Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection. bioRxiv 716902 [Preprint]. 28 July 2019. https://doi.org/10.1101/716902.
18. G.-L. Lin *et al*., Simultaneous viral whole-genome sequencing and differential expression profiling in respiratory syncytial virus infection of infants. *J. Infect. Dis.* **222**, S666–S671 (2020). doi: 10.1093/infdis/jiaa448; pmid: 32702120
19. D. Bonsall *et al*., ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Res.* **4**, 1062 (2015). doi: 10.12688/f1000research.7111.1; pmid: 27092241
20. S. M. Kissler *et al*., SARS-CoV-2 viral dynamics in acute infections. medRxiv 20217042 [Preprint]. 1 December 2020. https://doi.org/10.1101/2020.10.21.20217042.
21. L. Ferretti *et al*., The timing of COVID-19 transmission. medRxiv 20188516 [Preprint]. 16 September 2020. https://doi.org/10.1101/2020.09.04.20188516.
22. M. Marks *et al*., Transmission of COVID-19 in 282 clusters in Catalonia, Spain: A cohort study. *Lancet Infect. Dis.* (2021). doi: 10.1016/S1473-3099(20)30985-3
23. J. Raghwani *et al*., Evolution of HIV-1 within untreated individuals and at the population scale in Uganda. *PLOS Pathog.* **14**, e1007167 (2018). doi: 10.1371/journal.ppat.1007167; pmid: 30052678
24. G. Tonkin-Hill *et al*., Wellcome Sanger Institute COVID-19 Surveillance Team, Patterns of within-host genetic diversity in SARS-CoV-2. bioRxiv 424229 [Preprint]. 25 December 2020. https://doi.org/10.1101/2020.12.23.424229.
25. A. Popa *et al*., Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* **12**, eabe2555 (2020). doi: 10.1126/scitranslmed.abe2555; pmid: 33229462
26. A. L. Valesano *et al*., Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *bioRxiv* 2021.01.19.427330 (2021). doi: 10.1101/2021.01.19.427330; pmid: 33501443
27. M. P. Zwart, S. F. Elena, Matters of size: Genetic bottlenecks in virus infection and their potential impact on evolution. *Annu. Rev. Virol.* **2**, 161–179 (2015). doi: 10.1146/annurev-virology-100114-055135; pmid: 26958911
28. A. Sobel Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *J. Virol.* **91**, e00171-17 (2017). doi: 10.1128/JVI.00171-17; pmid: 28468874
29. A. Sobel Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Correction for Sobel Leonard et al., "Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus". *J. Virol.* **93**, e00936-19 (2019). doi: 10.1128/JVI.00936-19; pmid: 31409730
30. M. Ghafari, C. K. Lumby, D. B. Weissman, C. J. R. Illingworth, Inferring transmission bottleneck size from viral sequence data using a novel haplotype reconstruction method. *J. Virol.* **94**, e00014-20 (2020). doi: 10.1128/JVI.00014-20; pmid: 32295920
31. J. T. McCrone *et al*., Stochastic processes constrain the within and between host evolution of influenza virus. *eLife* **7**, e35962 (2018). doi: 10.7554/eLife.35962; pmid: 29683424
32. A. Varble *et al*., Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe* **16**, 691–700 (2014). doi: 10.1016/j.chom.2014.09.020; pmid: 25456074
33. Z. Shen *et al*., Genomic diversity of severe acute respiratory syndrome–coronavirus 2 in patients with coronavirus disease 2019. *Clin. Infect. Dis.* (2020). doi: 10.1093/cid/ciaa203
34. T. H. Wang, Y. K. Donaldson, R. P. Brettle, J. E. Bell, P. Simmonds, Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75**, 11686–11699 (2001). doi: 10.1128/JVI.75.23.11686-11699.2001; pmid: 11689650
35. K. M. Braun *et al*., Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. bioRxiv 384917 [Preprint]. 4 January 2021. https://doi.org/10.1101/2020.11.16.384917.
36. S. E. James *et al*., High resolution analysis of transmission dynamics of Sars-Cov-2 in two major hospital outbreaks in South Africa leveraging intrahost diversity. medRxiv 20231993 [Preprint]. 16 November 2020. https://doi.org/10.1101/2020.11.15.20231993.
37. M. A. Martin, K. Koelle, Reanalysis of deep-sequencing data from Austria points towards a small SARS-COV-2 transmission bottleneck on the order of one to three virions. bioRxiv 432096 [Preprint]. 22 February 2021. https://doi.org/10.1101/2021.02.22.432096.
38. M. Gelbart *et al*., Drivers of within-host genetic diversity in acute infections of viruses. *PLOS Pathog.* **16**, e1009029 (2020). doi: 10.1371/journal.ppat.1009029; pmid: 33147296
39. M. D. Nowak, E. M. Sordillo, M. R. Gitman, A. E. Paniz Mondolfi, Coinfection in SARS-CoV-2 infected patients: Where are influenza virus and rhinovirus/enterovirus? *J. Med. Virol.* **92**, 1699–1700 (2020). doi: 10.1002/jmv.25953; pmid: 32352574
40. D. Kim, J. Quinn, B. Pinsky, N. H. Shah, I. Brown, Rates of co-infection between SARS-CoV-2 and other respiratory pathogens. *JAMA* **323**, 2085–2086 (2020). doi: 10.1001/jama.2020.6266; pmid: 32293646
41. B. Dearlove *et al*., A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 23652–23662 (2020). doi: 10.1073/pnas.2008281117; pmid: 32868447
42. B. Morel *et al*., Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol.* msaa314 (2020). doi: 10.1093/molbev/msaa314; pmid: 33316067
43. L. van Dorp *et al*., Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020). doi: 10.1016/j.meegid.2020.104351; pmid: 32387564
44. Q. Li *et al*., The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**, 1284–1294.e9 (2020). doi: 10.1016/j.cell.2020.07.012; pmid: 32730807
45. S. Pond. "Natural selection analysis of global SARS-CoV-2/COVID-19 enabled by data from GISAID" (Datamonkey, 2021); https://observablehq.com/@spond/revised-sars-cov-2-analytics-page.
46. J. T. Herbeck *et al*., Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host. *J. Virol.* **80**, 1637–1644 (2006). doi: 10.1128/JVI.80.4.1637-1644.2006; pmid: 16439520
47. T. N. Starr *et al*., Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021). doi: 10.1126/science.abf9302; pmid: 33495308
48. A. J. Greaney *et al*., Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. bioRxiv 425021 [Preprint]. 4 January 2021. https://doi.org/10.1101/2020.12.31.425021.
49. K. R. McCarthy *et al*., Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* eabf6950 (2021). pmid: 33536258
50. E. M. Volz *et al*., Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. medRxiv 20166082 [Preprint]. 1 September 2020. https://doi.org/10.1101/2020.07.31.20166082.
51. E. C. Thomson *et al*., Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171–1187.e20 (2021). doi: 10.1016/j.cell.2021.01.037; pmid: 33621484
52. M. R. Zambenedetti *et al*., Internal control for real-time polymerase chain reaction based on MS2 bacteriophage for RNA viruses diagnostics. *Mem. Inst. Oswaldo Cruz* **112**, 339–347 (2017). doi: 10.1590/0074-02760160380; pmid: 28403327
53. F. Gao *et al*., A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol.* **72**, 5680–5698 (1998). doi: 10.1128/JVI.72.7.5680-5698.1998; pmid: 9621027
54. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019). doi: 10.1186/s13059-019-1891-0; pmid: 31779668
55. T. Golubch, "Castanet" (Github, 2020); https://github.com/tgolubch/castanet.
56. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). doi: 10.1093/bioinformatics/btu170; pmid: 24695404
57. C. Wymant *et al*., Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with *shiver*. *Virus Evol.* **4**, vey007 (2018). doi: 10.1093/ve/vey007; pmid: 29876136

58. Sanger Institute, "Tools directory" (Sanger Institute, 2021); https://www.sanger.ac.uk/science/tools.
59. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286
60. A. O'Toole, V. Hill, J. T. McCrone, E. Scher, A. Rambaut, "Pangolin COVID-19 lineage assigner v2.1.7, lineages version 2021-01-20" (Centre for Genomic Pathogen Surveillance, 2021); https://pangolin.cog-uk.io/.
61. C. Mavian, S. Marini, M. Prosperi, M. Salemi, A snapshot of SARS-CoV-2 genome availability up to April 2020 and its implications. *JMIR Public Health Surveill.* **6**, e19170 (2020). doi: 10.2196/19170; pmid: 32412415
62. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). doi: 10.2807/1560-7917.ES.2017.22.13.30494; pmid: 28382917
63. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). doi: 10.1093/bioinformatics/bts565; pmid: 23060610
64. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002). doi: 10.1093/nar/gkf436; pmid: 12136088
65. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986). pmid: 3444411
66. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019). doi: 10.1093/bioinformatics/btz305; pmid: 31070718
67. X. Didelot, D. J. Wilson, ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLOS Comput. Biol.* **11**, e1004041 (2015). doi: 10.1371/journal.pcbi.1004041; pmid: 25675341
68. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018). doi: 10.1093/molbev/msx281; pmid: 29077904
69. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk, An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100 (2020). doi: 10.1016/S2666-5247(20)30054-9; pmid: 32835336
70. Data for: K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, R. Williams, G. Vernet, A. Justice, A. Green, S. M. Nicholls, M. A. Ansari, L. Abeler-Dörner, C. E. Moore, T. E. A. Peto, D. W. Eyre, R. Shaw, P. Simmonds, D. Buck, J. A. Todd on behalf of the Oxford Virus Sequencing Analysis Group (OVSG), T. R. Connor, S. Ashraf, A. da Silva Filipe, J. Shepherd, E. C. Thomson, The COVID-19 Genomics UK (COG-UK) Consortium, D. Bonsall, C. Fraser, T. Golubchik, SARS-CoV-2 within-host diversity and transmission. Zenodo (2021); https://doi.org/10.5281/zenodo.4570598.

ACKNOWLEDGMENTS

We thank R. Esnouf, A. Huffman, and the BMRC Research Computing team for unfailing assistance with computational infrastructure; B. Carpenter and J. Docker for assistance in the laboratory; and L. Lonie, M. Lopopolo, C. Allen, J. Broxholme, A. Lee, and the WHG high-throughput genomics team (Oxford Genomics Centre) for sequencing and quality control. The HIV clone p92BR025.8 was obtained through the Centre For AIDS Reagents from B. Hahn and F. Gao and the UNAIDS Virus Network (courtesy of the NIH AIDS Research and Reference Reagent Program). **Ethics:** The COVID-19 Genomics UK (COG-UK) consortium study protocol was approved by the Public Health England Research Ethics and Governance Group (reference: R&D NR0195) on the 8th of April 2020. For seasonal coronavirus screening, samples were collected with consent to assay for infectious causes of respiratory disease from patients admitted to Hampshire Hospitals NHS Foundation Trust to aid diagnosis and outbreak management and inform public health surveillance. **Funding:** We gratefully acknowledge the UK COVID-19 Genomics Consortium (COG UK) for funding. COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR), and Genome Research Limited, operating as the Wellcome Sanger Institute. The research was also supported by a Wellcome Core Award (203141/Z/16/Z) with additional funding from the NIHR Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. K.A.L. and M.A.A. were supported by The Wellcome Trust and The Royal Society (107652/Z/15/Z to K.A.L. and 220171/Z/20/Z to M.A.A.). M.H., L.F., M.d.C., G.M.C., N.O., L.A.D., D.B., C.F., and T.G. were supported by Li Ka Shing Foundation funding awarded to C.F. P.S. was supported by a Wellcome Investigator Award (WT103767MA). J.A.T. was supported by a Wellcome Core Award (203141/Z/16/Z). C.E.M. was supported by the Fleming Fund at the Department of Health and Social Care, UK; the Wellcome Trust (209142/Z/17/Z) and the Bill and Melinda Gates Foundation (OPP1176062). DWE is a Robertson Fellow and an NIHR Oxford BRC Senior Fellow. **Author contributions:** Conceptualization: T.G., K.A.L., M.H., L.F., C.F.; Data curation: D.W.E., N.M., T.G.; Formal analysis: M.H., K.A.L.; Funding acquisition: D. Bonsall, COGUK, C.F.; Investigation: K.A.L., M.H., L.F., T.G., M.d.C., A.T., G.M.-C.; Methodology: T.G., K.A.L., M.H., M.d.C., A.T., D. Bonsall; Project administration: D. Bonsall, T.G., A.T., A.G., L.A.-D., D. Buck; Resources: M.A., E.L.W., N.M., J.L., S.K., M.M., R.W., G.V., A.J., N.O., S.M.N., M.A.A., C.E.M., T.E.A.P., D.W.E., R.S., D. Buck, A.G., J.A.T., OVSG Analysis Group, P.S., J.S., S.A., A.d.S.F., COGUK; Software: T.G., M.H.; Supervision: T.G., D. Bonsall, C.F., E.C.T., T.R.C., J.A.T.; Validation: T.G., K.A.L., M.H., P.S.; Visualization: M.H., T.G., K.A.L.; Writing – original draft preparation: K.A.L., M.H., T.G.; Writing – review and editing: K.A.L., M.H., T.G., C.F., D. Bonsall, L.A.-D., P.S., J.A.T., COGUK. **Competing interests:** D.W.E. declares personal fees from Gilead outside the submitted work. M.A. is on the advisory board of Prenetics. The remaining authors declare no competing interests. **Data and materials availability:** All genomic data have been made publicly available as part of the COVID-19 Genomics UK (COG-UK) Consortium (*69*) through GISAID (*62*) and through the European Nucleotide Archive (ENA) study PRJEB37886. All other data are available in the main text, supplementary materials, or (*70*) (including the full alignment of consensus sequences and inferred tree).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/372/6539/eabg0821/suppl/DC1
Figs. S1 to S8
Tables S1 to S5
List of OVSG members
List of COG-UK consortium names and affiliations
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

9 December 2020; accepted 3 March 2021
Published online 9 March 2021
10.1126/science.abg0821

Science

SARS-CoV-2 within-host diversity and transmission

Katrina A. Lythgoe, Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, Newton Otecko, Emma L. Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M. Nicholls, M. Azim Ansari, Lucie Abeler-Dörner, Catrin E. Moore, Timothy E. A. Peto, David W. Eyre, Robert Shaw, Peter Simmonds, David Buck, John A. Todd, on behalf of the Oxford Virus Sequencing Analysis Group (OVSG), Thomas R. Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C. Thomson, The COVID-19 Genomics UK (COG-UK) Consortium, David Bonsall, Christophe Fraser and Tanya Golubchik

Science **372** (6539), eabg0821.
DOI: 10.1126/science.abg0821originally published online March 9, 2021

Patterns and bottlenecks

A year into the severe acute respiratory syndrome coronavirus 2 pandemic, we are experiencing waves of new variants emerging. Some of these variants have worrying functional implications, such as increased transmissibility or antibody treatment escape. Lythgoe *et al.* have undertaken in-depth sequencing of more than 1000 hospital patients' isolates to find out how the virus is mutating within individuals. Overall, there seem to be consistent and reproducible patterns of within-host virus diversity. The authors observed only one or two variants in most samples, but a few carried many variants. Although the evidence indicates strong purifying selection, including in the spike protein responsible for viral entry, the authors also saw evidence for transmission clusters associated with households and other possible superspreader events. After transmission, most variants fizzled out, but occasionally some initiated ongoing transmission and wider dissemination.

Science, this issue p. eabg0821

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.