SARS-CoV-2 genomics beyond the consensus sequence: evidence for circulating mixed viral populations

Katrina A. Lythgoe*+1, Matthew Hall*+1, Luca Ferretti[1], Mariateresa de Cesare[1,2], George MacIntyre-Cockett[1,2], Amy Trebes[2], Monique Andersson[3], Newton Otecko[1], Emma L. Wise[4,6], Nathan Moore[4], Jessica Lynch[4], Stephen Kidd[4], Nicholas Cortes[4], Matilde Mori[7], Anita Justice[3], Angie Green[2], M. Azim Ansari[5], Lucie Abeler-Dörner[1], Catrin E. Moore[1], Tim E. A. Peto[3], Robert Shaw[3], Peter Simmonds[5], David Buck[2], John A. Todd[2] on behalf of OVSG Analysis Group, David Bonsall[1,2], Christophe Fraser[1,2], Tanya Golubchik+1,2

*Equal contribution
+Corresponding authors
Tanya.Golubchik@bdi.ox.ac.uk
Katrina.Lythgoe@bdi.ox.ac.uk
Matthew.Hall@bdi.ox.ac.uk

[1]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7FL, UK
[2]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Old Road Campus, Oxford OX3 7BN, UK
[3]Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK
[4]Hampshire Hospitals NHS Foundation Trust, Basingstoke and North Hampshire Hospital, Basingstoke, RG24 9NA, UK
[5]Peter Medawar Building for Pathogen Research, University of Oxford, OX1 3SY, UK
[6]School of Biosciences and Medicine, University of Surrey, Guildford, GU2 7XH, UK
[7]School of Medicine, University of Southampton, Southampton, SO17 1BJ, UK

The OVSG Analysis Group membership comprises John A Todd, Tanya Golubchik, David Bonsall, Christophe Fraser, Derrick Crook, Tim Peto, Monique Andersson, Katie Jeffries, David Eyre, Timothy Walker, Robert Shaw, Peter Simmonds, Katrina Lythgoe, Luca Ferretti, Matthew Hall, Mariateresa de Cesare, Paolo Piazza, Richard Cornall.

Summary

SARS-CoV-2, the causative agent of COVID-19, emerged in late 2019 causing a global pandemic, with the United Kingdom (UK) one of the hardest hit countries. Rapid sequencing and publication of consensus genomes have enabled phylogenetic analysis of the virus, demonstrating SARS-CoV-2 evolves relatively slowly[1], but with multiple sites in the genome that appear inconsistent with the overall consensus phylogeny[2]. To understand these discrepancies, we used veSEQ[3], a targeted RNA-seq approach, to quantify minor allele frequencies in 413 clinical samples from two UK locations. We show that SARS-CoV-2 infections are characterised by extensive within-host diversity, which is frequently shared among infected individuals with patterns consistent with geographical structure. These results were reproducible in data from other sequencing locations around the UK, where we find evidence of mixed infection by major circulating lineages with patterns that cannot readily be explained by artefacts in the data. We conclude that SARS-CoV-2 diversity is transmissible, and propose that geographic patterns are generated by co-circulation of distinct viral populations. Co-transmission of mixed populations fundamentally changes our understanding of transmission of SARS-CoV-2 and could prove significant for treatment and vaccine design, as well as opening opportunities for resolving clusters of transmission and understanding pathogenesis.

Introduction

The nature of the ongoing evolution of the SARS-CoV-2 coronavirus has been the topic of considerable speculation as the pandemic has unfolded. Studies have raised concerns that new mutations may confer selective advantages on the virus, hampering efforts at control[4,5]. To date, attention has been focused on mutations observed in viral consensus genomes, which represent the dominant variants within infected individuals. Understanding the full underlying within-host diversity of the virus is of crucial importance for vaccine design, and understanding pathogenesis and patterns of transmission. Of particular interest are loci or genetic regions that are diverse in multiple individuals, since shared diversity may reveal signatures of host adaptation, or indicate the presence of co-transmitted lineages.

Phylogenetic analyses of consensus genomes have enabled tracking of patterns of viral spread, both regionally[6] and across the globe[7]. Clear lineage-defining single nucleotide polymorphisms (SNPs) have emerged[8], and it has been postulated that some of these might represent mutations that increase the fitness of the virus, raising significant concern for public health. Of specific importance are mutations in the spike (S) protein of the virus, at least one of which, D614G (genome position 23403) has been suggested to increase transmissibility[4,9,5]. SNPs at this and many other positions appear to have arisen multiple times on different lineages[4,9]. The presence of such a large number of homoplasies against a background of low overall genetic diversity is puzzling, and could be the consequence of recurrent mutation and selection, susceptibility of specific sites to RNA editing, mixed infections of multiple variants, or to artefacts arising from sequencing and/or processing errors[9]. Untangling these possible explanations is vital, as homoplasies can bias phylogeographic analyses, giving a misleading representation of how the virus evolves and spreads.

The United Kingdom (UK) has experienced one of the most severe waves of infection, with 11% of the reported global death toll as of 26th May 2020[10]. Multiple independent SARS-CoV-2 introductions have contributed to substantial viral diversity[7], making the UK an important setting for understanding SARS-CoV-2 evolution and transmission. In this study, we collected and analysed SARS-CoV-2 samples from 405 symptomatic individuals who tested positive for COVID-19 within two geographically-separate hospital trusts (Oxford University Hospitals and Basingstoke and North Hampshire Hospital, located 37 miles (60 km) apart; Supplementary Table 1). Using veSEQ, a sequencing protocol based on a quantitative targeted enrichment strategy[3], which we previously validated for other viruses[3,11,12], we characterised the full spectrum of within-host diversity in SARS-CoV-2 and contextualised our findings within other high-quality, publicly available deep-sequencing datasets from the UK generated on the high-fidelity Illumina platform[13,14]. All genomic data has been made publicly available as part of the COVID-19 Genomics UK (COG-UK) Consortium [cogconsortium.uk] via GISAID[15] and the European Nucleotide Archive (ENA) study PRJEB37886.

Within-host diversity is extensive and shared between individuals

To examine patterns of within-host diversity, we first considered the distribution of minor allele frequencies (MAFs) in the mapped reads at every position along the genome. This analysis was supported by data curation to ensure that only high-confidence variants were examined, which included analysis of in-batch quantification controls as well as a stringent computational clean-up to eliminate any residual cross-mapping[16], previously validated for targeted metagenomics[11] (see Methods and Supplementary Text for a full description). In combination with unique dual indexing (UDI), these procedures generated highly robust minority variant calls, which were reproducible in independent replicates (Supplementary Figure 1) and distinguishable from methodological noise above a threshold of 2% of reads at a given position (Supplementary Figure 2). The distribution of MAFs was a close fit to that expected under the model of an exponentially growing population (Supplementary Figure 3), giving high confidence that MAFs generated by veSEQ were representative of true within-sample diversity. We likewise observed a good fit for data from other UK centres, with slightly reduced goodness of fit consistent with some bias in MAF recovered by amplicon sequencing.

We used a conservative threshold of 5% to define an initial set of intrahost single nucleotide variant (iSNV) sites, and subsequently examined all samples at these iSNV sites for variants with MAF of at least 2%. We found extensive within-host diversity in SARS-CoV-2 samples from Oxford and Basingstoke (Figure 1), consistently higher than those observed between-host (Supplementary Figure 4), with a median of 11 iSNV sites per sample (range 0 - 137), consistent with previously reported levels[17]. Strikingly, intrahost diversity was shared by multiple individuals (Figure 1 and Supplementary Table 2), and this observation was robust to batch effects (Supplementary Figure 5). Of the 1227 shared high-confidence iSNVs in the Oxford/Basingstoke data, 9.3% (114/1227) coincided with single nucleotide polymorphisms (SNPs) on the SARS-CoV-2 consensus phylogeny. Furthermore, of the 11% (180/1655) global SNPs that were homoplasic (changed multiple times along the tree, Supplementary Table 3), 20 corresponded with shared iSNV sites in the Oxford/Basingstoke data, increasing to 59 sites when considering the complete dataset including samples from other UK centres (Supplementary Table 4). That is, 33% (59/180) of homoplasic sites are also shared iSNV sites, despite only 7% of the genome consisting of shared iSNV sites. This suggests within-host diversity is associated with a considerable proportion of the homoplasies observed in SARS-CoV-2, and which in turn may interfere with correct inference of the phylogeny and downstream analyses.

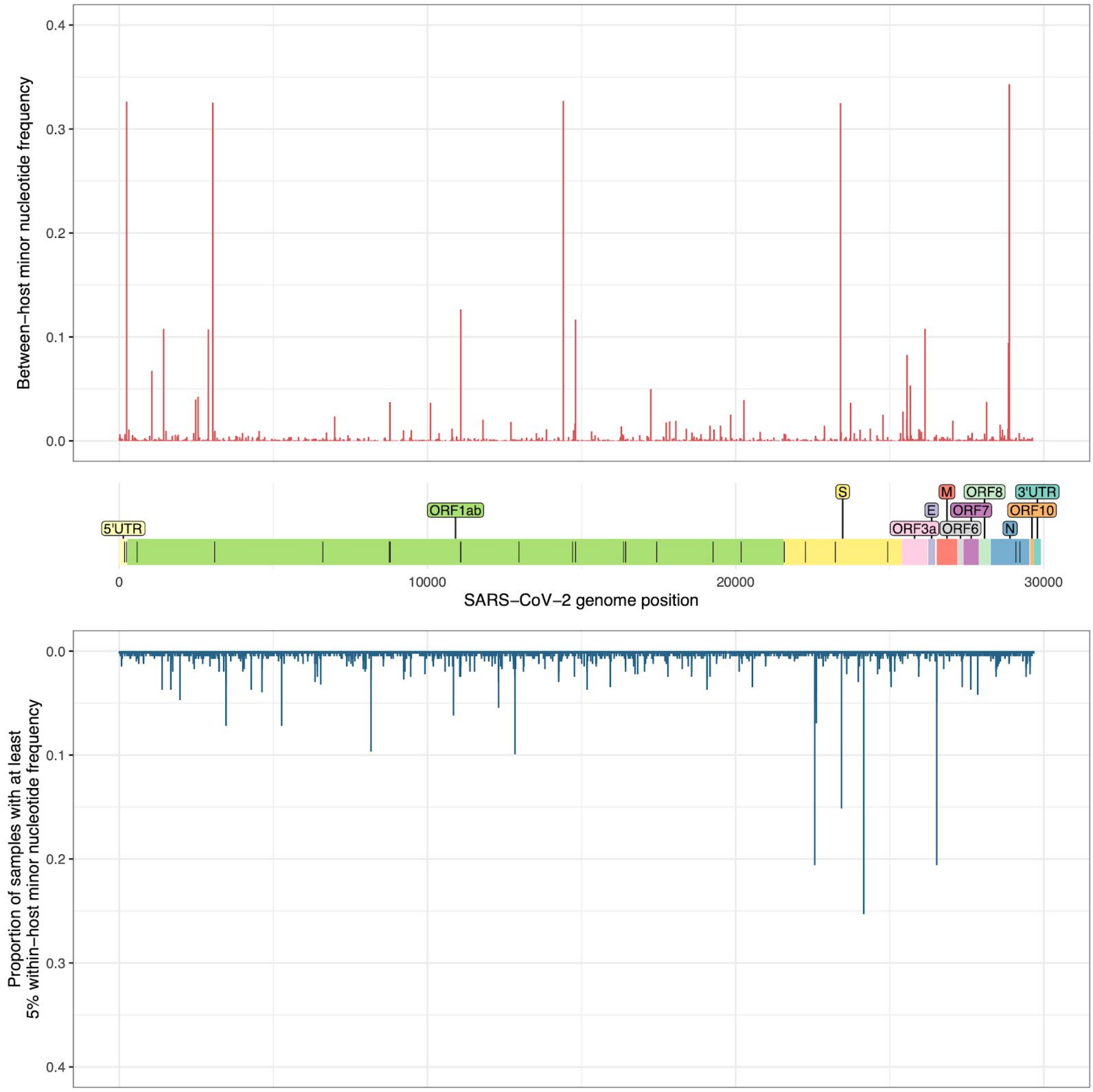


Figure 1: **With-host diversity of SARS-CoV-2 is more widespread than between-host diversity**. Above: proportion of consensus sequences in the global alignment with minority variants at each genome position. Below: proportion of genomes in the Oxford/Basingstoke dataset with within-host diversity (minor variant frequency of at least 5%), at each genome position. The x-axis is annotated with a map of the reading frames in the viral genome, with homoplasic sites marked by black vertical lines. Although homoplasic sites often correspond with shared within-host variable sites (iSNV sites), the most commonly shared iSNV sites do not correspond with homoplasic sites or the most diverse between-host sites.

We identified 37 iSNV sites shared by over ten individuals, with the four most common iSNVs found in over 50 individuals each (Figure 1, Supplementary Table 2). Three of these, genome positions 24156 (L865P), 22565 (L335V), and 23434 (synonymous), fall within the S gene encoding the spike protein, which mediates cell entry, is a target of antibodies, and is also the focus for new vaccines, making mutations in this region a specific concern[4,18,19]. The fourth site, 26524 (M1K/R/T), lies in the M gene, which encodes the membrane protein. Overall, the elevated ratio of 1st/2nd versus 3rd codon position variants in sites shared by 15 or more individuals gives the appearance of positive selection by standard methods (p<0.05, binomial test; Supplementary Figure 6). They may, for example, represent adaptive changes in SARS-CoV-2, associated with its recent change in host. However, a recent study has concluded homoplasies are typically neutral or mildly deleterious for SARS-CoV-2 fitness, arguing against positive selection at homoplasic sites at least[20]. Moreover, the appearance of homoplasies may be the result of host RNA editing of viral RNA at certain favoured contexts in the SARS-CoV-2 genome[21,22]. Overall, the high number of individuals in which we see identical iSNVs suggests their *de novo* generation in most individuals is unlikely.

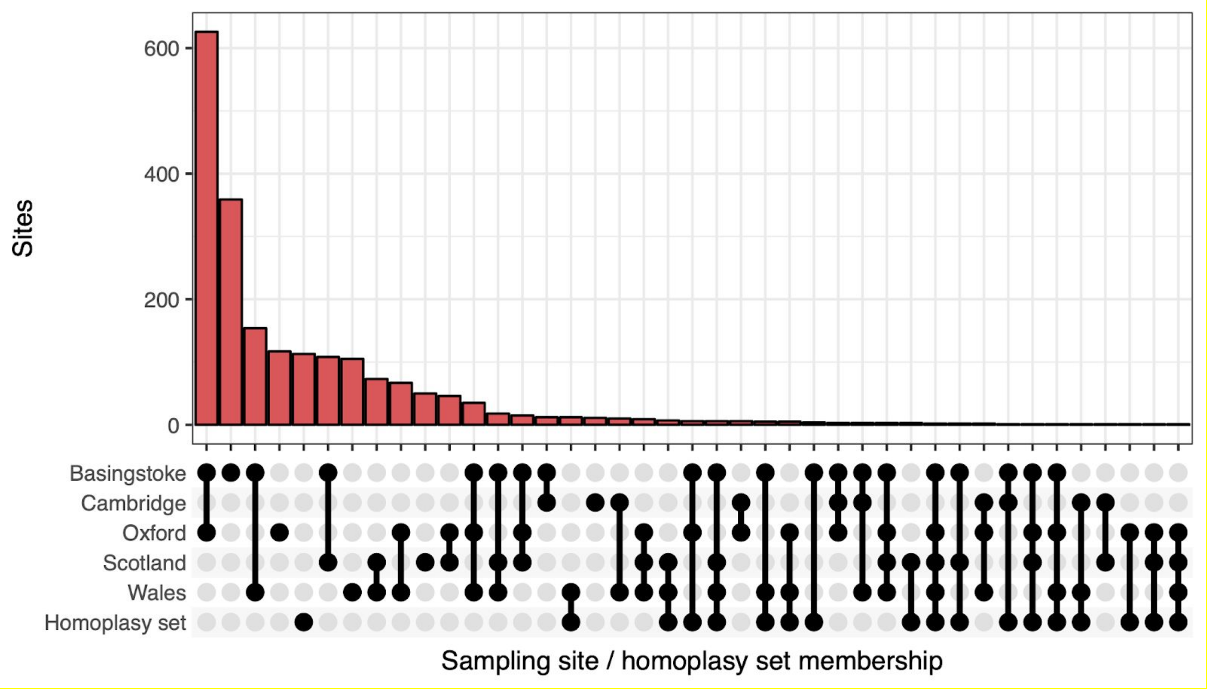


Figure 2: **Strong geographical patterns in the distribution of variant sites.** The upset plot shows the number of iSNV sites, with minor allele frequency (MAF) >0.05, that are shared among the locations indicated at the bottom of figure (or only found in a single location). The set of 180 homoplasic sites is also included. Primer binding sites in the non-Oxford/Basingstoke data were masked to avoid confounding variation within the primer sequences. Number of samples with iSNVs with MAF>0.05 and read depth>100 are 218 (Basingstoke), 159 (Oxford), 825 (Wales), 344 (Scotland), 34 (Cambridge).

Within-host diversity shows strong geographical patterns

The most commonly shared iSNV sites, and specifically those which are closely spaced on the genome, tend to be clustered within individuals, with a disproportionate number of individuals variant at either all sites, or no sites. For the four most commonly shared iSNV sites in Oxford/Basingstoke, 87 out of 405 individuals had iSNVs at all four sites at MAF >0.02, whilst 78 individuals had no iSNVs at these sites (p<0.001, binomial test). An even stronger pattern is observed in Welsh samples from the COG-UK data, where the ten most commonly shared iSNV sites are all in the S (Spike) gene, with 309 out of 827 samples sharing all iSNV sites and 128 individuals none (MAF>0.05, p<0.001). These linkage blocks of commonly shared iSNVs tend to be associated with specific locations, leading to strong geographical patterns in the distributions of shared variation (Figure 2; Supplementary Table 4)

To better understand the geographical patterns of within-host diversity, we considered the location from which the samples were collected. We found 31 iSNV sites that were significantly more likely to be within-host diverse in samples from Oxford compared to those from Basingstoke (p<0.05, Fisher exact test with Holm correction) and two sites more likely to be diverse in samples from Basingstoke (p<0.05; Supplementary Table 5). Since these Oxford and Basingstoke samples were sequenced in the same lab, using the same methodology across multiple batches, and within-host diversity at the same sites was observed amongst samples from multiple batches (Supplementary Figure 5), these geographical effects are unlikely to be artefactual.

We next investigated whether phylogeny can explain these patterns; that is, whether iSNVs are associated with phylogenetic lineages determined at the consensus level. Using a parsimony approach, we found that diversity at only five of the 31 sites (357, 22565, 25628, 25807, and 28469), is significantly associated with phylogenetic topology (p<0.05 by tip randomisation with Bonferroni correction; Figure 3). However, closer inspection reveals that even at these positions, phylogenetic structure is confounded by geography, with diversity at specific sites regularly occurring in numerous distinct clades that are specific to either Oxford or Basingstoke. As it is unlikely that diversity spontaneously appeared or disappeared in multiple lineages as they moved between UK locations, these results suggest that the consensus-based phylogeny lacks the resolution to uncover the geographical patterns that we observe in minor variant diversity.

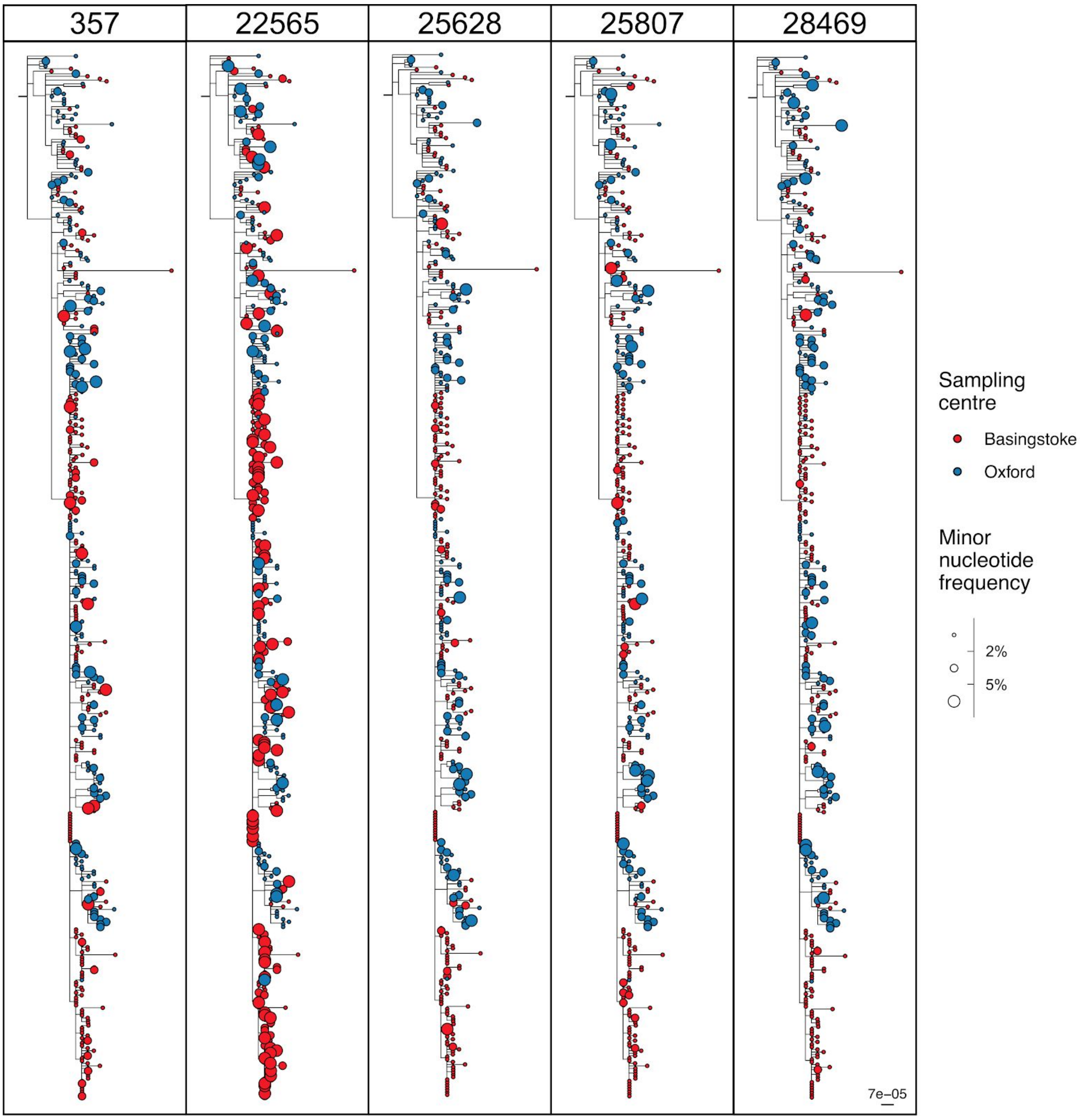


Figure 3: Consensus-level phylogenies cannot explain geographical patterns of within-host diversity. Consensus phylogenies illustrating within-host diversity in Oxford and Basingstoke samples, for sites in the SARS-CoV-2 genome where an association was detected both between minor allele frequency (MAF) >= 2% and sampling location, and between MAF >= 2% and phylogeny. Tips are sized by MAF at the genomic site in each panel's header and coloured by sampling location. We see a large number of independent clades with shared within-host diversity, each drawn largely from the same sampling location. This suggests that within-host diversity is not a trait that has emerged a limited number of times on specific tree branches, but rather that it is primarily associated with geography and that this confounds the apparent statistical association with the consensus phylogeny. In the absence of a host effect or sequencing artefact, the most parsimonious explanation is that the genomic differences between the Oxford and Basingstoke viral populations are not visible if the analysis is limited to the consensus genome.The single long branch is due to a complex variant at position 20716 - 20726 in individual OXON-ADD0E. A variant at the same position has been previously identified[17], suggesting either susceptibility of this locus to mutation, or a cryptic recombination event.

Within-host transmission can help resolve infection clusters

We explored whether consideration of within-host diversity, in addition to the consensus genome, could provide further resolution for identification of transmission clusters. We identified a group of 46 samples from the Oxford/Basingstoke data with indistinguishable consensus genomes (differing only by the presence or absence of the N ambiguity code; Figure 4A, red dots). We selected the 15 most commonly-shared iSNV sites in all Oxford/Basingstoke samples (using a threshold of 2% and a minimum coverage of 100 to identify diversity) and calculated the proportion of those sites (identical at the consensus level) at which each pair from the 46 samples shared within-host diversity, finding a median of 0.067 (IQR 0-0.2). A distinct clustering effect was evident, most clearly in a group of 9 of the 33 samples from Basingstoke (Figure 3), including a highly correlated triplet of samples (AE81B, AE417, AE893; mean proportion of shared iSNV sites 0.489), which on investigation we found had been collected from individuals closely linked by employment and sampled on consecutive days. Another triplet of clustered samples from Oxford (ACA62, AEFF8 and AEFBC; mean proportion of shared iSNV sites 0.267) had been collected from the same hospital within two days of each other. Further studies are needed to confirm whether this additional diversity can be reliably used to help resolve transmission clusters.

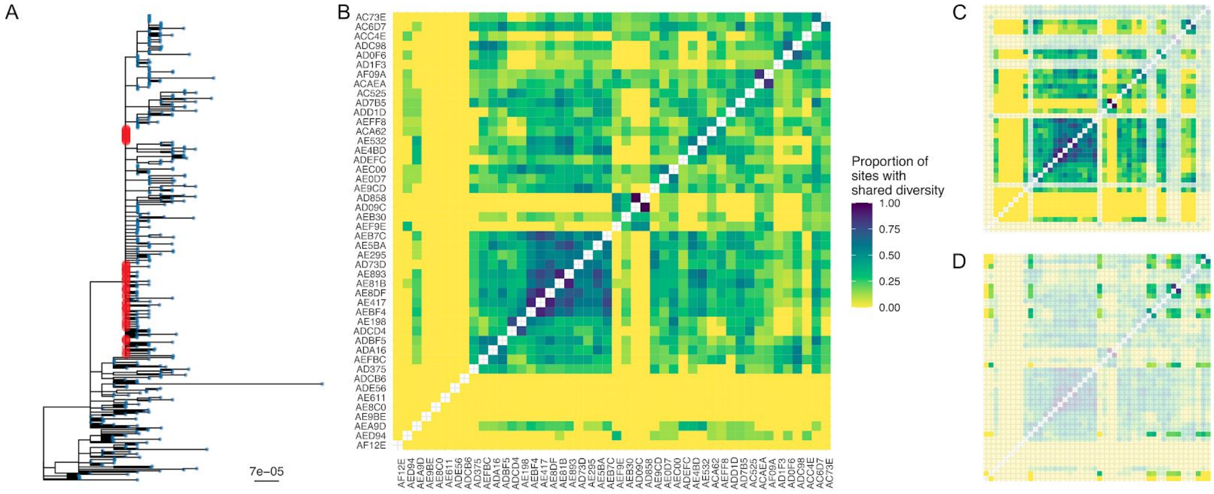


Figure 4: Minor variants can identify epidemiological clusters indistinguishable at the consensus level. *A) We selected a group of 46* Oxford and Basingstoke samples that are indistinguishable in the consensus phylogeny, highlighted by the red dots. B) For each pair of samples, we calculated the proportion of the 15 most commonly within-host diverse sites in Oxford and Basingstoke that showed diversity in both samples of at least 2%. This demonstrates structure in the genetic data by considering minor variants even though the consensus genomes are indistinguishable. C) and D) highlight pairs of samples which are coming both from Basingstoke (C) or both from Oxford (D). The largest cluster is exclusively from Basingstoke.

Major UK lineages appear to be co-transmitted

To support our observation of co-transmission of variants from the Oxford and Basingstoke samples, we analysed a larger dataset including an additional 1220 samples whose sequences have been made available by other COG-UK collaborating centres. We again found a large number of loci diverse in multiple individuals, many with a strong geographical distribution (Supplementary Table 4). To identify patterns beyond the regional level, we focussed on three common polymorphic sites in the UK. The first is the D614G (A-to-G base change at locus 23403) that arises on the B.1 lineage[8], and has been speculated to be linked to higher rates of transmission[4,5]. The other two polymorphic sites, 241 and 14408, we chose since they span a large part of the genome and have previously been identified as having linked variants with 23403[23]. At these loci, 89% of consensus genomes in Oxford and Basingstoke with coverage at these sites have haplotype T-T-G (sites represented in ascending order; lineage B.1), with the rest having haplotype C-C-G (representing lineages A.2, B.2, and B.3). Only 21 individuals had iSNVs at any of these sites above 2% frequency, and none at all three loci, suggesting we do not see mixed infections of lineage B.1 with any of the other lineages among the Oxford/Basingstoke samples.

Including all of the samples from COG-UK, we see a markedly different pattern. Overall, 1612 of 1634 samples have coverage at all three sites, of which 69% are T-T-G at consensus, and 31% C-C-A (plus one with T-C-A and one C-C-G haplotype, potentially representing earlier recombination events as neither had minor alleles >2% at these sites). Of the 20 individuals with minor variants above 2% frequency at all three sites,15 are c-c-a (with T-T-G as the major haplotype) and 5 t-t-g (with C-C-A as the major haplotype). Within each sample, the frequencies at these three sites are remarkably similar, suggesting that these sites are far from linkage equilibrium even within patients (Supplementary Figure 7). These patterns are clearly evident in the consensus phylogenetic tree (Figure 5), implying extensive co-transmission of these two lineages. Co-transmission is strongly supported by the observation that samples that share any of these iSNV sites are phylogenetically closer on the tree than would be expected by chance (Supplementary Figure 8).

Since we do not have linkage information, we are determining putative haplotypes based on the major and minor allele frequencies, and therefore cannot rule out within-host recombination. However, the small number of possible recombinants at the consensus level requires explanation. Wide transmission bottlenecks could act to keep minor variants at low frequency along transmission chains, regardless of recombination. Alternatively, epistatic effects, including within-host spatial structuring due to different selective environments[24], could help to maintain the separate lineages. The patterns we see are consistent with consensus genomes circulating in specific locations. The B.1 lineage is dominant in Oxford and Basingstoke and it is therefore unsurprising that mixed infections including other lineages are not observed.

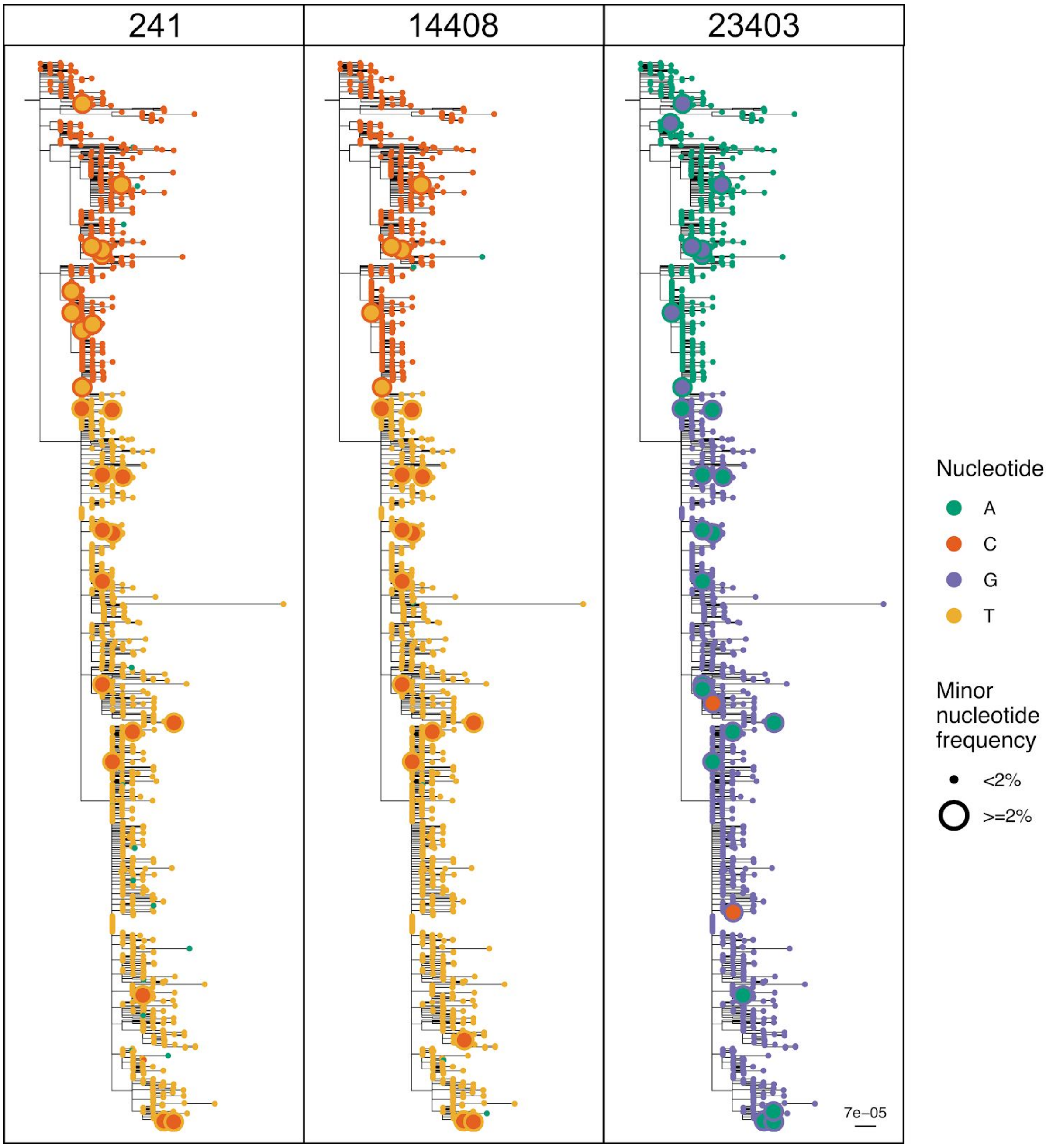


Figure 5: Major and minor variants representative of major UK lineages cluster together on the phylogenetic tree. Distribution of major and minor nucleotide variants at genome positions 241,14408 and 23403, on the consensus phylogeny of 1633 British SARS-CoV-2 sequences. Single-coloured tips are isolates for which there was no or little within-host nucleotide diversity, with at least 98% of reads having the major nucleotide or an overall depth of less than 100. These are coloured by the major nucleotide. Larger, bi-coloured tips represent isolates with a depth of at least 100 where less than 98% of reads showed the major nucleotide. In those cases the colour of the outer ring represents the major nucleotide, and the inner circle the most common minor nucleotide. Some samples are diverse at one or two sites, but not all three; 20 samples are diverse at all three sites.

Discussion

We uncovered a consistent pattern of extensive within-host SARS-CoV-2 diversity, with shared iSNV sites showing strong geographical patterns. Throughout, we went to great lengths to avoid and rule out sequencing artefacts or sample contamination (Supplementary Methods), leading to confidence that the patterns we observe are real.

Wide transmission bottlenecks enabling multi-variant transmission, with these variants perpetuated along transmission chains, provides the most parsimonious explanation for these observations (Figure 6). Transmission of minor variants has previously been reported in a small number of SARS-CoV-2 transmission pairs[17,24], providing extra support to our hypothesis. Wide transmission bottlenecks are consistent with the high transmissibility of SARS-CoV-2[25], particularly in light of the lack of pre-existing immunity to this novel pathogen, which could enable more viral particles to establish infection. We expect transmitted variant lineages to be gradually eroded through within-host evolution, genetic drift, stochastic loss during onward transmission, and recombination, leaving a residual amount of underlying diversity but without a clear phylogenetic structure. Restricted movement due the UK-wide lockdown imposed on 23rd March 2020 might also explain some of the more striking geographical patterns, and which we predict will be eroded as lockdown conditions are relaxed.

Although within-host selection and RNA editing may be responsible for the original generation of within-host diversity, the sheer number of individuals sharing identical iSNVs, and which are often closely linked on the genome and show strong geographical patterns, suggests they are unlikely to have arisen *de novo* in most individuals. In addition, superinfection from multiple transmission events is likely to occur in SARS-CoV-2[26,27] and could enable the generation of some within-host diversity. For example, at least one superinfection event is likely responsible for the pattern of mixed infections of the B.1 lineage with other major UK lineages. However, superinfection events are unlikely to be responsible for the bulk of the shared diversity we observe since shared iSNVs often represent variants not present among individuals at the consensus level (94%, Figure 1), For superinfection to drive this pattern, it would be expected substitutions present as minority variants in some samples would also be present as majority variants in others, but this was seen in only 9% of within-host SNPs in the Oxford/Basingstoke data.

Mixed SARS-CoV-2 infections could provide an explanation for many of the homoplasies observed on phylogenies due to difficulty in resolving the consensus genome at sites that are highly diverse. The presence of mixed infection may also result in significant discrepancies between consensus phylogenies and the true transmission tree, potentially obscuring transmission clusters. Our observations provide evidence that accounting for all of the diversity within individuals could prove to be a better route for defining clusters than relying on the consensus sequences alone, as has been demonstrated in other viruses[28,29,30]. Given the existence of within-host evolution and population structure[24], it is remarkable that we see structuring of minor variants among epidemiologically linked individuals. Transmission of this variation as a result of a weak or absent population bottleneck is the most plausible explanation.

The consequences of this variability for immune-based approaches, including vaccine development, could be substantial. For example, significant effort is being directed to identifying therapeutic neutralising monoclonal antibodies, and it is possible that some of the variants we have defined and enumerated, and others yet to be identified, may affect the binding of antibodies and therefore their efficacy. Our observations may also be of important clinical relevance, particularly if infection by diverse viral populations leads to more severe and/or longer-lasting infections, as has been suggested for other viruses[31,32,33].

It is important to recognise that our sampling, as well as that of the majority of the UK sampling of SARS-CoV-2 at corresponding calendar dates, was dominated by symptomatic individuals presenting at hospitals, with moderate to severe infection. If mixed infections are more likely to lead to severe infection, it may be no coincidence that we find extensive evidence for high degrees of diversity in these samples, and these individuals might also have been exposed to high infectious doses. It will be important to compare these findings with those obtained from mild or asymptomatic infection, which are now increasingly becoming available due to widespread testing rollout, as well as those with a broad range of other clinical outcomes. Our results emphasise the power of open data, and the importance of integrating genomic, clinical, and epidemiological information, to gain rapid understanding of SARS-CoV-2, and no doubt other emerging pathogens in the future.

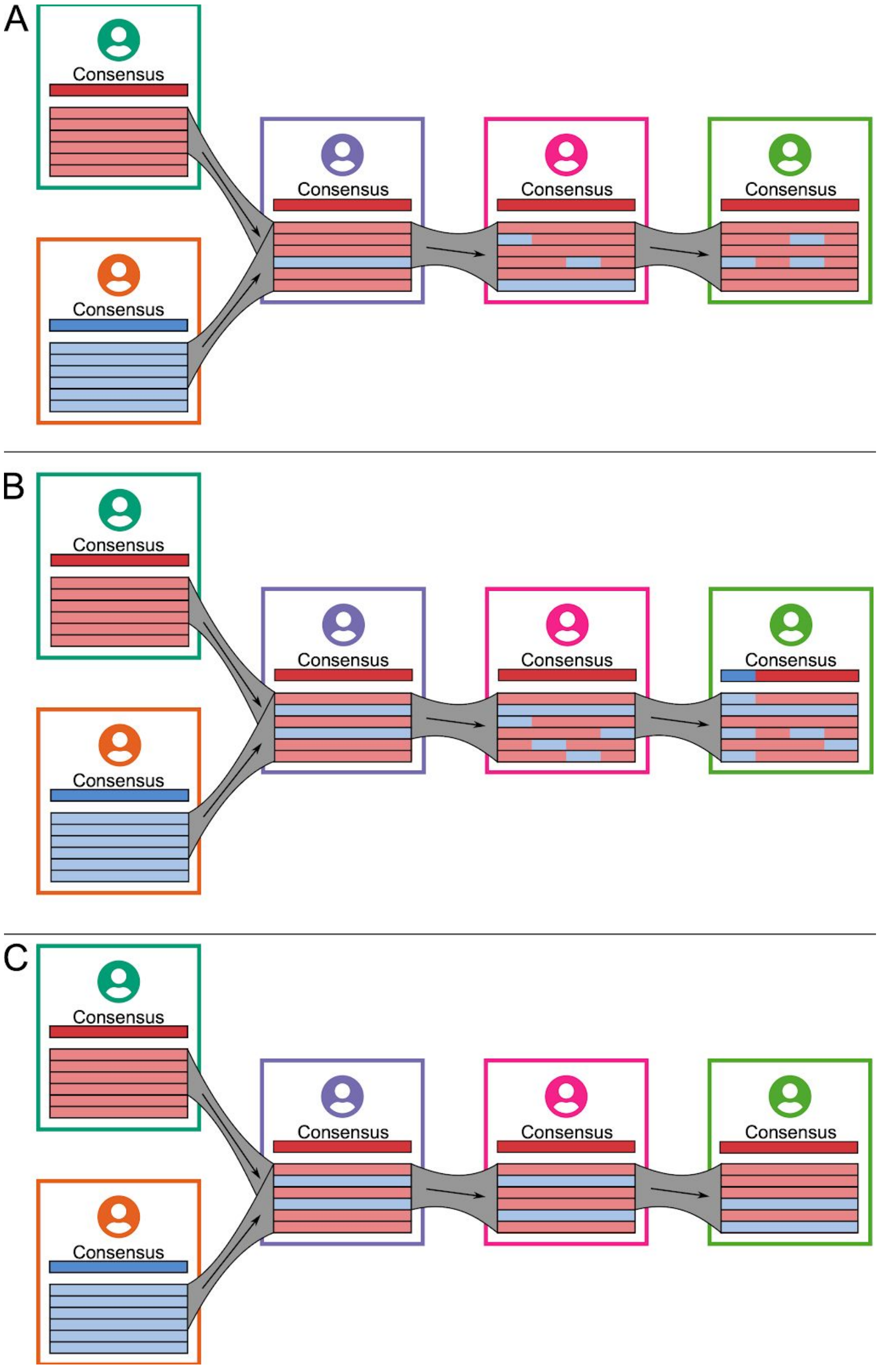


Figure 6: Co-transmission of mixed viral populations. Superinfection from multiple source individuals, or within-host diversification, can lead to mixed viral populations in an infected individual. We propose that much of this within-host diversity is transmitted between infected individuals, leading to the co-circulation of lineages. In this diagram we show three, non-exclusive scenarios: A) The rare blue minor variant lineage is gradually eroded through recombination, drift, and/or partial bottlenecking at transmission (even if the transmission bottleneck is large. B) Two high frequency lineages recombinine with one another, with alleles from both lineages remaining. The consensus sequence may reflect different lineages at different sites due to fluctuations in allele frequencies. As in A) variants can gradually be lost. C) Two high frequency lineages co-exist, but epistasis, within-host structuring, or other processes maintain the two distinct lineages without recombination.

Acknowledgments
We gratefully acknowledge the UK COVID-19 Genomics Consortium (COG UK) for funding and UK-wide sequence data availability. The research was supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z with funding from the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. We are deeply grateful to Robert Esnouf and the BMRC Research Computing team for unfailing assistance with computational infrastructure. We also thank Benjamin Carpenter and James Docker for assistance in the laboratory, and Lorne Lonie, Maria Lopopolo, Chris Allen, John Broxholme and the WHG high-throughput genomics team for sequencing and quality control. The HIV clone p92BR025.8 was obtained through the Centre For AIDS Reagents from Drs Beatrice Hahn and Feng Gao, and the UNAIDS Virus Network (courtesy of the NIH AIDS Research and Reference Reagent Program).
KAL is supported by a Sir Henry Dale Fellowship awarded by The Wellcome Trust and The Royal Society (107652/Z/15/Z). MH, LF, MdC, GMC, NO, LAD, DB, CF and TG are supported by Li Ka Shing Foundation funding awarded to CF. PS is supported by a Wellcome Investigator Award (WT103767MA).

References

1. Taiaroa, G. *et al.* Direct RNA sequencing and early evolution of SARS-CoV-2. *bioRxiv* 2020.03.05.976167 (2020) doi:10.1101/2020.03.05.976167.
2. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* 104351 (2020).
3. Bonsall, D. *et al.* A comprehensive genomics solution for HIV surveillance and clinical monitoring in a global health setting. *bioRxiv* 397083 (2018) doi:10.1101/397083.
4. Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020.04.29.069054 (2020) doi:10.1101/2020.04.29.069054.
5. Bhattacharyya, C. *et al.* Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. *bioRxiv* 2020.05.04.075911 (2020) doi:10.1101/2020.05.04.075911.
6. Lu, J. *et al.* Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China.

medRxiv 2020.04.01.20047076 (2020).

7. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

8. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020.04.17.046086 (2020) doi:10.1101/2020.04.17.046086.

9. NicolaDeMaio, Pond, S., Maclean, O., Parker, M. & Shaw, L. Issues with SARS-CoV-2 sequencing data. *Virological* http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (2020).

10. Download today's data on the geographic distribution of COVID-19 cases worldwide. *European Centre for Disease Prevention and Control* https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide (2020).

11. Goh, C. *et al.* Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection. *bioRxiv* 716902 (2019) doi:10.1101/716902.

12. Bonsall, D. *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).

13. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 1–14 (2018).

14. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* **14**, 265–279 (2016).

15. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

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

17. Shen, Z. *et al.* Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients.

Clin. Infect. Dis. (2020) doi:10.1093/cid/ciaa203.

18. Yuan, M. *et al.* A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630–633 (2020).
19. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
20. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Genomics* 501 (2020).
21. Simmonds, P. Rampant C->U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. *Microbiology* 50 (2020).
22. Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci.Adv.* **17**, eabb5813 (2020).
23. Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* (2020) doi:10.1016/j.ygeno.2020.04.016.
24. Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* 1–5 (2020).
25. Chen, J. Pathogenicity and transmissibility of 2019-nCoV—A quick overview and comparison with other emerging viruses. *Microbes and Infection* **22**, 69–71 (2020).
26. Su, S. *et al.* Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends in MIcrobiology* **24**, 490–502 (2016).
27. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).
28. Wymant, C. *et al.* PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Mol. Biol. Evol.* **35**, 719–733 (2018).
29. De Maio, N., Worby, C. J., Wilson, D. J. & Stoesser, N. Bayesian reconstruction of

transmission within outbreaks using genomic variants. *PLoS Comput. Biol.* **14**, e1006117 (2018).

30. Worby, C. J., Lipsitch, M. & Hanage, W. P. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. *Am. J. Epidemiol.* **186**, 1209–1216 (2017).

31. Janes, H. *et al.* HIV-1 infections with multiple founders are associated with higher viral loads than infections with single founders. *Nat. Med.* **21**, 1139–1141 (2015).

32. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348 (2006).

33. Cao, L. *et al.* Coexistence of hepatitis B virus quasispecies enhances viral replication and the ability to induce host antibody and cellular immune responses. *J. Virol.* **88**, 8656–8666 (2014).

Methods

RNA extraction. Residual RNA from COVID-19 RT-qPCR-based testing was obtained from Oxford University Hospitals, extracted on the QIASymphony platform with QIAsymphony DSP Virus/Pathogen Kit (QIAGEN), and from Basingstoke and North Hampshire Hospital, extracted with one of: Maxwell RSC Viral total nucleic acid kit (Promega); Reliaprep blood gDNA miniprep system (Promega); or Prepito NA body fluid kit (PerkinElmer). An internal extraction control was added to the lysis buffer prior to extraction to act as a control for extraction efficiency (genesig qRT-PCR kit, #Z-Path-2019-nCoV for Basingstoke, MS2 bacteriophage[1] 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 pers. comm, 6 April 2020). The MS2 RNA likewise has no SARS-CoV-2 similarity[1]. Neither control RNA interfered with sequencing.

Targeted metagenomic sequencing. Sequencing libraries were constructed from remnant volume of nucleic acid after clinical testing, ranging from 5 to 45 µl (median 30µl) for each sample depending on the available amount of eluate. These volumes represented 1-15% of the starting material (swab). Libraries were constructed following the veSEQ protocol[2] 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 USA, California, US) with no fragmentation of the RNA. An equal volume of library from each sample was pooled for capture. Size selection was performed on the pool to eliminate short fragments below 400nt which otherwise may be preferentially amplified and sequenced. Target enrichment of SARS-CoV-2 libraries in the pool was obtained through a custom xGen Lockdown Probes panel (IDT, Coralville, USA), using the SeqCap EZ Accessory Kits v2 and SeqCap Hybridization and Wash Kit (Roche, Madison, US) for hybridization of the probes and removal of unbound DNA. PCR of 12 cycles was carried out for post-capture amplification and the final product was purified by Agencourt AMPure XP (Beckman Coulter, California, US). Sequencing was performed on the Illumina MiSeq or NovaSeq 6000 platform (Illumina, California, US) at the Oxford Genomics Centre (OGC), generating 150-bp or 250-bp paired-end reads.

Avoiding cross-contamination. Next-generation sequence data produced at scale typically necessitates batching of large pools of samples together to make the process cost effective. We sought to avoid batch effects or contamination during library preparation or sequencing, which could otherwise obscure the true signal of sequence diversity. All samples had unique dual indexing (UDI) to prevent cross-detection of reads in the same pool (known as index misassignment). Across all sequencing runs, only 36 pairs of samples have colliding indices: these pairs were processed one month apart, sequenced on different instruments (MiSeq and NovaSeq 6000), and share fewer iSNVs than average. To guard against contamination, every batch of 90 samples was sequenced together with a series of controls. In addition, one sample was split and sequenced in two batches (as OXON-AF346 and OXON-AF179), with ~50x difference in read depth. For all iSNV sites present in at least one batch at MAF>2%, we found a strong correlation in frequencies (no MAF cutoff, linear regression, $p<0.001$; Supplementary Figure 1).The controls comprised: a negative buffer in-capture control; a standard curve consisting of a dilution series of a positive SARS-CoV-2 control (Twist Synthetic SARS-CoV-2 RNA Control 1 (MT007544.1),Twist Bioscience) from 100 through to 0.5 million copies per reaction; and 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 (NIBSC)[3] . As additional negative controls, we sequenced 6 matched clinical samples from non-COVID-19 patients. No SARS-CoV-2 sequences were detected in any negative controls or negative clinical samples in any pool; no HIV reads were detected in the SARS-CoV-2 samples, and the expected log-linear relationship between the number of reads and viral copy number was observed in the standard curve (Supplementary Figure 9). As previously reported[2], the veSEQ method is quantitative, and the number of sequenced reads is expected to correlate with the number of input copies. We were therefore satisfied that all sequenced runs were clean. To further minimise any concerns about residual contamination, we performed an additional stringent computational cleanup of the read data. For all reads in a pool, we identified any optical duplicates that shared the same mapping coordinates (start of reads 1 and 2, and template length), and in each case removed the duplicate cluster from all samples except the one containing the greatest number of these reads (see Bioinformatics processing). In this way, no two samples within a run shared any duplicate reads. All reads with similarity to human sequences or known kit contaminants were removed prior to mapping, as detailed below.

Bioinformatics processing. De-multiplexed sequence read pairs were classified by Kraken v2[4] using a custom database containing the human genome (GRCh38 build) and all RefSeq bacterial and viral genomes. Sequences identified as either human or bacterial were removed using the filter_keep_reads.py script in the Castanet[5] workflow (https://github.com/tgolubch/castanet). Remaining reads, comprised of viral and unclassified reads, were trimmed to remove adapter sequences using Trimmomatic v0.36[6], with the ILLUMINACLIP options set to “2:10:7:1:true MINLEN:80”, using the set of Illumina adapters supplied with the software. The trimmed reads were mapped to the SARS-CoV-2 RefSeq genome of isolate Wuhan-Hu-1 (NC_045512.2), using the shiver_map procedure from the shiver pipeline[7], without deduplication, using either smalt[8] or bowtie2[9] as the mapper. Both mappers generated comparable results. To remove any possibility of index misassignment (“index hopping”) that may result from sequencing multiple samples in a single pool[10], the BAM alignments were deduplicated by pool, for each set of mapping coordinates (start of read 1, start of read 2 and mapped length) retaining only the read pairs from the sample with the greatest number of reads at these coordinates, using the Castanet scripts process_pool_grp.py and filter_bam.py (https://github.com/tgolubch/castanet). The median depth was 2,300x across the genome (IQR 800 - 8,600) (Supplementary Figure 10). For analysis of consensus genomes, consensus calls required a minimum of 2 unique deduplicated reads per position, to avoid calling consensus from optical duplicates. Analysis of within-host diversity was restricted only to positions with minimum minor allele frequency (defined as 1 - major allele frequency) of 2% and a minimum depth of 10, to focus on high-confidence variants.

Alignment. We separately generated sequence sets for just the Oxford and Basingstoke sequences, and for those sequences combined with the other UK data. To place these data into the global phylogenetic context, a collection of non-UK consensus sequences from the GISAID database[11] were also selected. Oxford and Basingstoke samples were selected if the consensus sequence (inferred from unique mapped reads) consisted of no more than 25% N characters. For COG UK (https://www.cogconsortium.uk) samples this was lowered to 5%. As an alignment to the reference sequence was already performed in *shiver*, no further alignment was necessary. All GISAID[12] sequences were downloaded from the database on the 26th April 2020 and filtered to remove sequences that were less than 29800 base pairs in length, were more than 1% Ns, or were from the United Kingdom (as this set had considerable overlap with our other data). The remaining sequences were clustered using CD-HIT-EST[13] using a similarity threshold of 0.995, and then one sequence per cluster picked. The resulting set, along with the reference genome NC_045512, were aligned using

MAFFT[14], with some manual improvement of the alorithmic alignment and removal of problematic sequences performed as a post-processing step.

Phylogenetics. Phylogenetic analysis was performed on both alignments using IQ-TREE version 1.6.12[15], using the GTR+F+I substitution model. The tree was then rooted with respect to the reference sequence RefSeq ID NC_045512. Association of the phylogenetic topology with within-host diversity was performed by, first, pruning the tree of the reference sequence and all GISAID sequences (as within-host diversity for that data is unknown). The parsimony score for the tree for a given site was calculated using the presence or not of within-host diversity at that site (a cumulative minor nucleotide frequency of at least 0.02 and a read depth of at least 100) as a trait. The same score was also calculated after permuting the tip labels of the tree 1000 times, and an approximate *p*-value calculated by comparing the true score to the distribution of shuffled scores. This value was adjusted by Bonferroni correction to account for the multiple testing imposed by applying this procedure to multiple genome positions. 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[16] and rooting the tree using the reference genome NC_045512.

Overrepresentation of shared iSNVs within individuals. For highly shared iSNVs on similar regions of the genome (top four shared iSNVs for Oxford/Basingstoke, MAF>0,02; nine of the ten top shared iSNVs for PHWC, MAF>0.05) we computed the number of individuals with none or all iSNVs, and determined the probability of each of these from 100,000 randomisations.

Geographical comparison of within-host diversity. To look for sites with an increased diversity in Oxford, we computed the Fisher exact test for the number of iSNVs in either Oxford or Basingstoke samples and either at a given site or elsewhere in the genome. We repeat the test for each genomic position, looking for an overrepresentation of iSNVs in Oxford samples at that site. Before applying multiple testing corrections, we ignored all sites that could not reach significance at level $p<0.05$ even if all diversity at the position would be concentrated in Oxford. We also performed the corresponding analysis to look for an overrepresentation of iSNVs in Basingstoke samples at each site.

Overrepresentation of shared iSNVs within clusters. For each cluster, we computed the average number of shared iSNVs (sharing the same mutation as well) between two random samples from the cluster. To assess its significance, we compared it with the average number of shared iSNVs from 1000 permutation of all variants among all clusters.

Frequency dissimilarity as a proxy for linkage disequilibrium. For this analysis, for each pair of sites A and B, we considered the dissimilarity in iSNV frequencies as a proxy for the cumulative amount of recombination shuffling the two sites. The overall dissimilarity is defined as the mean across all individuals with iSNVs in both A and B of the dissimilarity for the *i*th individual $D_i=(f_{A,i}-f_{B,i})^2/[(f_{A,i}+f_{B,i})(2-f_{A,i}-f_{B,i})]$, where $f_{A,i}$ and $f_{B,i}$ denote the frequencies of the alternative allele in A and B. The quantity D_i represents the square of the fluctuations in frequency, normalised by a term proportional to the variance of such fluctuations due to genetic drift. For weakly recombining neutral sequences, D_i grows approximately linearly with time and recombination rates. To assess the significance, we performed 1000 randomisations of the iSNVs among samples, while keeping the number of iSNVs per sample and the distribution of iSNV frequencies per site approximately constant. These randomisations simulate a complete reshuffling, i.e. linkage equilibrium.

Methods references

1. Zambenedetti, M. R. *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).
2. Bonsall, D. *et al.* A comprehensive genomics solution for HIV surveillance and clinical monitoring in a global health setting. doi:10.1101/397083.
3. Gao, F. *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).
4. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
5. Goh, C. *et al.* Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection. *bioRxiv* 716902 (2019) doi:10.1101/716902.
6. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
7. Wymant, C. *et al.* Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evol* **4**, vey007 (2018).
8. (webteam), W.-C. Search Tools and Software. https://www.sanger.ac.uk/science/tools.
9. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* vol. 9 357–359 (2012).
10. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
11. Mavian, C., Marini, S., Prosperi, M. & Salemi, M. A snapshot of SARS-CoV-2 genome availability up to April 2020 and its implications. *JMIR Public Health Surveill* (2020) doi:10.2196/19170.

12. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).

13. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

14. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* vol. 30 3059–3066 (2002).

15. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

16. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

Acknowledgements: The HIV clone p92BR025.8 was obtained through the Centre For AIDS Reagents from Drs Beatrice Hahn and Feng Gao, and the UNAIDS Virus Network (courtesy of the NIH AIDS Research and Reference Reagent Program)

SARS-CoV-2 genomics beyond the consensus sequence: Evidence for circulating mixed viral populations

Supplementary Text, Table and Figures

Katrina A. Lythgoe[*+1,] Matthew Hall[*+1], Luca Ferretti[1], Mariateresa de Cesare[1,2], George MacIntyre-Cockett[1,2], Amy Trebes[2], Monique Andersson[3], Newton Otecko[1], Emma L. Wise[4,6], Nathan Moore[4], Jessica Lynch[4], Stephen Kidd[4], Nicholas Cortes[4], Matilde Mori[7], Anita Justice[3], Angie Green[2], M. Azim Ansari[5], Lucie Abeler-Dorner[1], Catrin E. Moore[1], Tim E. A. Peto[3], Robert Shaw[3], Peter Simmonds[5], David Buck[2], John A. Todd[2] on behalf of OVSG Analysis Group, David Bonsall[1,2], Christophe Fraser[1,2], Tanya Golubchik[+1,2]

The OVSG Analysis Group membership comprises John A Todd, Tanya Golubchik, David Bonsall, Christophe Fraser, Derrick Crook, Timothy Peto, Monique Andersson, Katie Jeffries, David Eyre, Timothy Walker, Robert Shaw, Peter Simmonds, Katrina Lythgoe, Luca Ferretti, Matthew Hall, Mariateresa de Cesare, Paolo Piazza, Richard Cornall.

*Equal contribution
+Corresponding authors
Tanya.Golubchik@bdi.ox.ac.uk
Katrina.Lythgoe@bdi.ox.ac.uk
Matthew.Hall@bdi.ox.ac.uk

[1]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7FL, UK
[2]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Biomedical Research Centre, University of Oxford, Old Road Campus, Oxford OX3 7BN, UK
[3]Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK
[4]Hampshire Hospitals NHS Foundation Trust, Basingstoke and North Hampshire Hospital, Basingstoke, RG24 9NA, UK
[5]Peter Medawar Building for Pathogen Research, University of Oxford, OX1 3SY, UK
[6]School of Biosciences and Medicine, University of Surrey, Guildford, GU2 7XH, UK
[7]School of Medicine, University of Southampton, Southampton, SO17 1BJ, UK

A. Supplementary Text

Controlling and assessing possible artefacts

Throughout our sequencing and analysis, the possibility of sequencing artefacts and/or contamination was at the forefront. We first explain the protocols and controls we used to avoid and/or detect artefacts during sequencing and processing of data, and second how our main results are inconsistent with contamination of samples. Although artefacts can never be ruled out entirely, they are extremely unlikely to explain the broad patterns that we observe.

1. Lab protocols and data processing

i) Sample collection and extraction

Sample collection was carried out by several geographically separated hospitals, before being sent for RNA extraction and testing to either OUH (Oxford) or BSNH (Basingstoke) laboratories. Alongside SARS-CoV-2 positive samples, we processed and sequenced six SARS-CoV-2-negative samples from OUH laboratories, and in each case found zero SARS-CoV-2 reads. This strongly suggests the absence of laboratory-level contamination in OUH data. Similarly, from BSNH we collected a large number of very weakly positive SARS-CoV-2 samples (cycle threshold values > 36 using RdRp as the qPCR marker). These samples had near-zero SARS-CoV-2 reads mapped, again suggesting that laboratory contamination cannot be widespread or provide an explanation for the observed diversity in the data. While we cannot rule out laboratory-level contamination in other laboratories contributing samples to external COG UK datasets, if such contamination was present, it would have to have been pervasive in every hospital and/or sequencing centre across many weeks of sequencing and many sequencing runs to explain the observation of shared within-host diversity observed in these datasets.

ii) Library prep, bait capture and sequencing

Next-generation sequence data produced at scale typically necessitates batching of large pools of samples together to make the process cost effective. We sought to avoid batch effects or contamination during library preparation or sequencing, which could otherwise obscure the true signal of sequence diversity. To guard against contamination, every batch of 90 samples was sequenced together with a series of controls. The controls comprised: a negative buffer in-capture control; a standard curve consisting of a dilution series of a positive SARS-CoV-2 control (Twist Synthetic SARS-CoV-2 RNA Control 1 (MT007544.1),Twist Bioscience) from 100 through to 0.5 million copies per reaction; and 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 (NIBSC)[3]. As additional negative controls, we sequenced six matched clinical samples from non-COVID-19 patients. No SARS-CoV-2 sequences were detected in any negative controls or negative clinical samples in any pool; no HIV reads were detected in the SARS-CoV-2 samples, and the expected log-linear relationship between the number of reads and viral copy number was observed in the standard curve (Supplementary Figure 9). As previously reported (Bonsall et al. 2018 doi:10.1101/397083), the veSEQ method is quantitative, and the number of sequenced reads is expected to correlate with the number of input copies. We were therefore satisfied that all sequenced runs were clean.

iii) Minimising risk of index misassignment

All samples had unique dual indexing (UDI) to prevent cross-detection of reads in the same pool (known as index misassignment or index hopping). We avoided using the same UDI series in any run. Of the 413 sequenced samples, only 36 pairs of samples have identical indices: these pairs were processed one month apart, sequenced on different instruments (MiSeq and NovaSeq 6000), and share fewer iSNVs than average. The majority of the data were sequenced on the NovaSeq 6000 instrument which uses a patterned flowcell, which further reduces the chance of observing index hopping due to detection of hybrid optical clusters.

iv) Human read removal

To avoid mixed base calls that may appear as a result of mis-mapping of host or contaminant reads, we first screened all raw data for the presence of reads with sequence similarity to the human genome, the mitochondrial genome, or any bacterial genomes, and removed these reads prior to mapping (see Methods).

v) Non-SARS-CoV-2 coronaviruses

We considered the possibility of the presence of non-SARS-CoV-2 (seasonal) coronaviruses in the samples, which could cause closely-matching reads to be mapped to SARS-CoV-2 and appear as mixed base calls. To exclude this possibility, we analysed a subset of 90 samples (batch Cov8) for the presence of other coronaviruses using the Castanet bait enrichment panel (Goh et al. 2019 doi: 10.1101/716902). We did not find any samples positive for coronaviruses other than SARS-CoV-2.

vi) Post-mapping computational cleaning

To eliminate any residual index misassignment, we performed a stringent computational cleanup of the read data. For all reads in a sequencing run, consisting of up to three pools, we identified optical duplicates that shared the same mapping coordinates (start of reads 1 and 2, and template length), and in each case removed the duplicate cluster from all samples except the one containing the greatest number of these reads (see Methods: Bioinformatics processing). In this way, no two samples within a run shared any duplicate reads.

vii) Resequencing

One sample was split into two aliquots and sequenced in separate batches (as OXON-AF346 and OXON-AF179), with ~50x difference in read depth. We first analysed the pre-cleaned data. As expected, we find strong concordance between iSNV frequencies in the two replicates (Supplemental Figure 1), with a noisier distribution for the cleaned data. Although cleaning of the data reduces the number of reads, and consequently adds noise to iSNV frequencies, the importance of eliminating index misassignment outweighed concerns that we had of losing meaningful signal. Thus all of our results are conservative.

2. Post-analysis considerations

If our main finding, of high within-host diversity that is geographically structured, is due to contamination of samples, or systematic sequencing errors. There are certain patterns in the data we would (and would not) expect to see.

i) Patterns of diversity across batches

If patterns of diversity in the Oxford and Basingstoke data are due to batch effects, we would expect diversity at specific sites to be clustered within batches. For example, for the five sites shown in Figure 3, we find no sign of batch effects on patterns of diversity (Supplementary Figure 5), and therefore batch effects are unlikely to be driving our observations. More generally, these iSNVs are highly represented in at least two different batches.

ii) Patterns of diversity across Oxford and Basingstoke

The Oxford and Basingstoke samples were sequenced in the same lab, by the same people (GMC, AT, MdC), using the same protocols. The identification of shared iSNV sites between Oxford and Basingstoke suggests that contamination at the point of sample collection and RNA extraction cannot explain the presence of these iSNVs. Conversely, the statistically significant differences in the distribution of some iSNVs between Oxford and Basingstoke suggests these do not arise from lab-based contamination or other sequencing artefacts.

iii) Shared iSNV sites are typically not polymorphic at the consensus level

If rare iSNVs arise from contamination, we would expect minor variant alleles to be present in other samples at high frequency. 94% of the iSNVs we detected in Oxford and Basingstoke are not polymorphic at the consensus (global) scale. This makes cross-contamination of samples unlikely.

iv) Patterns of shared iSNVs across locations

We find strong geographical patterns, with some iSNV sites only found in a single location, but others found in two, three, and four locations (Figure 2, main text). This is unlikely to be caused by contamination among locations, and if it were, we would expect more iSNV sites to be shared among locations. This is also unlikely to be due to systematic biases in sequencing methodologies. The only site that is repeatedly variant in samples is 11083, and is variant in most samples, not just a subset of samples. This site is a well-recognised homoplasy in SARS-CoV-2 in general, and appears to be due to a variable truncation of a long stretch of poly(T) at this position, which depending on the mapping software may present as a gap at the end of the homopolymer run or one position immediately afterward, which can present as a T/g iSNV. We cannot rule out systematic biases at other sites found in multiple (e.g. >3 locations), but these sites are few.

v) Mixed infections of major UK lineages

One of the most striking patterns we found was the repeated identification of minor variants at sites 241,14408, and 23403 in apparent linkage equilibrium in both Wales and Glasgow. That is haplotype T-T-G as the major variant and c-c-a as the minor variant, or *vice versa*. Contamination cannot be ruled out, but this would have to be at an unprecedented scale at two different locations. Moreover, samples that share any of these iSNV sites are phylogenetically closer on the tree than would be expected by chance (Supplementary Figure 8), strongly suggesting cross-sample contamination is not the cause.

vi) Patterns of selection at aminoacid level in shared iSNVs

There are strong patterns of selection on iSNVs that are not expected to arise from artefacts, as revealed by the localisation of shared iSNVs in codons and by their effect on aminoacid sequence. Widely shared iSNVs with frequency >5% are preferentially found in the 1st and 2nd base of each codon (see Supplementary Figure 6). This is significantly different from the uniform pattern that would be expected from artefacts, it is not observed in the controls, and it is a signature of positive selection at aminoacid level. Oxford/Basingstoke specific variants are also under positive selection at aminoacid level, as can be confirmed by a non-synonymous/synonymous comparison: the ratio of non-synonymous/synonymous polymorphisms pN/pS among this list of variants is about three times higher than pN/pS for random variants at >2% frequency (p=0.015 by Fisher exact test).

B. Supplementary Tables

Supplementary Table 1. Baseline characteristics of SARS-CoV-2 samples in our dataset collected from symptomatic patients attending hospitals in Oxford and Basingstoke, UK, between 8 March and 14 April 2020.

	Oxford	Basingstoke
Samples, n(%)	179(43.34)	234(56.65)
Proportion female	0.43	0.58
Age, median	53	46
(min - max)	(3 - 98)	(1 - 94)
Sampling date, median	04-Apr-2020	08-Apr-2020
(min - max)	(16-Mar-2020 - 08-Apr-2020)	(08-Mar-2020 - 14-Apr-2020)
SARS-CoV-2 lineage, n(%)		
A.2	1(0.60)	0
B	0	2(0.90)
B.1	123(68.70)	188(80.30)
B.1.1	13(7.30)	6(2.60)
B.1.10	2(1.10)	7(3.00)
B.1.11	5(2.80)	8(3.40)
B.1.13	4(2.20)	1(0.40)
B.1.16	0	1(0.40)
B.1.24	2(1.10)	0
B.1.5	4(2.20)	2(0.90)
B.1.6	0	1(0.40)
B.2	15(8.40)	2(0.90)
B.2.1	6(3.40)	7(3.00)
B.2.2	1(0.60)	5(2.10)
B.2.4	0	2(0.90)
B.2.5	0	1(0.40)
B.3	3(1.70)	1(0.40)
Ct value, median	22.62	22.21
(min - max)	(13 - 29)	(15 - 33)

Supplementary Table 2. Oxford and Basingstoke: Sites shared by more than ten individuals at minor allele frequency >5% and read depth >100

Genome position	Gene	Residue position	Number of variants by type[a]		Number of individuals[b]		
			Nonsynonymous	Consensus	OXFD	BSNH	All
24156	S	865	99	2	50	49	99
22565	S	335	81	1	15	66	81
26524	M	1	79	0	28	51	79
23434	S	624	0	0	18	40	58
8168	ORF1ab	2635	39	1	26	13	39
12842	ORF1ab	4193	36	0	20	16	36
3466	ORF1ab	1067	4	0	10	19	29
22618	S	352	0	0	16	12	28
5270	ORF1ab	1669	28	0	16	12	28
10845	ORF1ab	3527	25	0	2	23	25
12309	ORF1ab	4015	21	0	6	15	21
26523	M	1	18	0	6	12	18
26522	none	NA	0	0	3	15	18
1971	ORF1ab	569	18	0	7	11	18
22592	S	344	17	0	1	16	17
27856	ORF7	155	16	0	4	12	16
22616	S	352	16	0	8	8	16
4630	ORF1ab	1455	1	1	6	10	16
19071	ORF1ab	6269	15	0	8	7	15
27627	ORF7	78	0	0	6	8	14
27351	ORF6	50	0	0	5	9	14
25043	S	1161	14	0	9	5	14
20542	ORF1ab	6759	0	0	9	5	14
15931	ORF1ab	5222	0	0	7	7	14
15181	ORF1ab	4972	0	0	7	7	14
4281	ORF1ab	1339	14	0	5	9	14

1675	ORF1ab	470	6	0	6	8	14
1391	ORF1ab	376	13	0	5	8	13
6538	ORF1ab	2091	0	0	4	8	12
3468	ORF1ab	1068	12	0	3	9	12
23979	S	806	11	1	2	9	11
19072	ORF1ab	6269	0	0	2	9	11
14257	ORF1ab	4664	2	0	6	5	11
12308	ORF1ab	4015	11	0	6	5	11
9232	ORF1ab	2989	2	0	3	8	11
8167	ORF1ab	2634	5	1	7	4	11
6353	ORF1ab	2030	11	0	5	6	11

[a] Number of intrahost single nucleotide variant (iSNVs) by type at each genome position. Nonsynonymous: the minor and major alleles code for different amino acids; Consensus: the minor allele is the global consensus.
[b] Number of individuals with intrahost single nucleotide variants (iSNVs) at each genome position. OUH, Oxford University Hospitals; BSNH, Basingstoke and North Hampshire Hospital.
See Supplemental Figure 5 for the full list of shared sites for all COG-UK locations included in this study.

Supplementary Table 3. List of Homoplasic Sites Determined on the Global Consensus Phylogeny
https://github.com/katrinalythgoe/COVIDdiversity

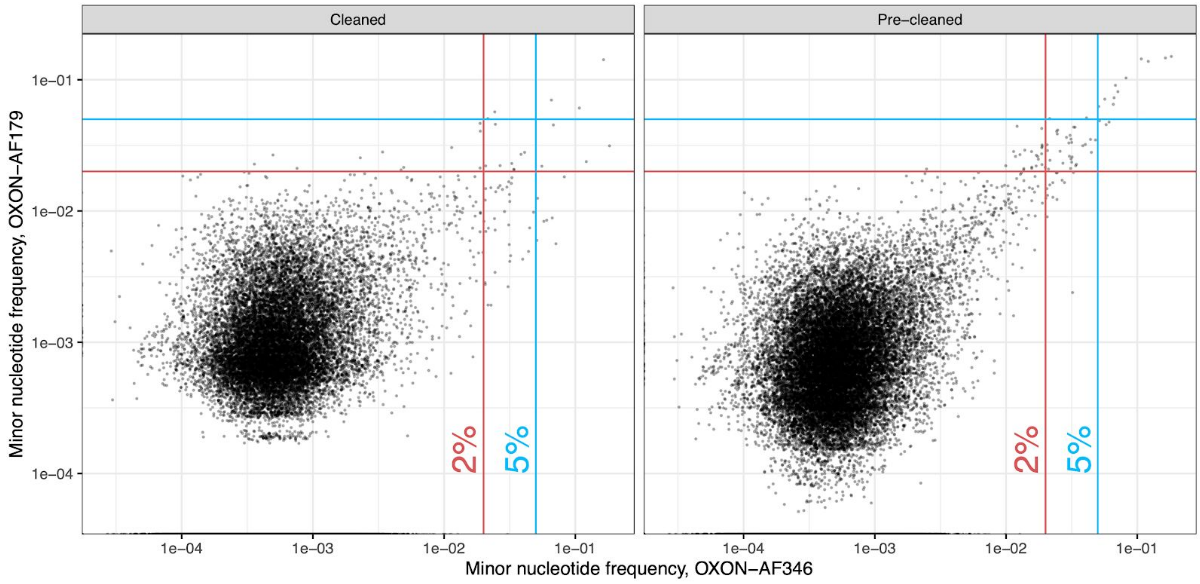
Supplementary Table 4. COG-UK sites shared by two or more individuals at minor allele frequency >5% and read depth >100. Sites shared by two or more samples across COG-UK sites at minor allele frequency >5% and read depth >100. Primer binding sites for individuals sequenced outside of Oxford/Basingstoke were masked to avoid confounding by variation within the primer sequences. reference.pos: nucleotide position in the genome; residue.position: amino acid (AA) position in the gene; n.individuals: total number of individuals with an iSNV at this site; n.synonymous: number of individuals where the most common minor variant codes for a different AA than the most common variant; n.consensus: the number of individuals where the most common minor variant nucleotide matches the population-level consensus; OXFD: Oxford; BSNH: Basingstoke; PHWC: Wales; GLAS: Scotland; CAMB: Cambridge; Nucleotides are represented by ABC/def, where the capital letters gives the most common variant in individuals in decreasing order of how frequently the they are observed as the most common variant in the COG-UK dataset. The lower-case letters represent the most common minor variant. Similarly for Amino acids.
https://github.com/katrinalythgoe/COVIDdiversity

Supplementary Table 5. Sites with different iSNV frequencies in Oxford and Basingstoke. Sites where iSNVs appear in significantly higher number of patients from Oxford than from Basingstoke, and vice versa.

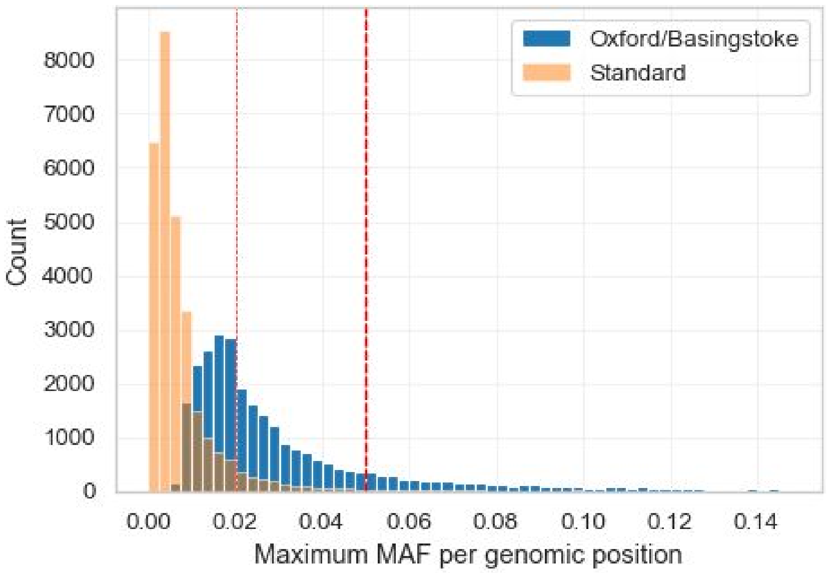
Genome position	p-value	count iSNVs in Oxford	count iSNVs in Basingstoke	ratio of counts
More frequent in Oxford:				
238	1.17E-03	24	3	8.00
356	5.86E-04	34	8	4.25
357	9.19E-07	52	13	4.00
369	6.91E-03	27	6	4.50
2490	1.47E-02	32	10	3.20
2738	2.08E-09	41	3	13.67
2949	2.71E-12	83	20	4.15
5270	1.26E-03	65	32	2.03
5322	1.05E-03	22	2	11.00
8168	4.70E-02	68	42	1.62
8569	1.34E-02	43	18	2.39
14331	5.23E-03	22	3	7.33
17545	6.03E-05	20	0	infinity
19330	1.49E-06	39	6	6.50
19393	2.80E-03	30	7	4.29
20989	3.25E-05	40	9	4.44
21147	1.08E-02	14	0	infinity
21180	1.50E-12	50	3	16.67
21236	4.99E-08	37	3	12.33
22691	0.0286	39	16	2.44
24225	0.0476	19	3	6.33
25202	3.66E-07	32	2	16
25216	4.50E-06	23	0	infinity
25296	2.50E-08	29	0	infinity
25628	1.09E-06	45	9	5
25807	1.67E-09	46	5	9.2
25949	2.56E-06	68	25	2.72
27152	2.35E-06	27	1	27

28469	5.25E-23	76	2	38
29360	0.044	15	1	15
29574	0.0045	15	0	infinity
More frequent in Basingstoke:				
10845	4.72E-08	6	69	11.5
22565	0.0024	46	134	2.91

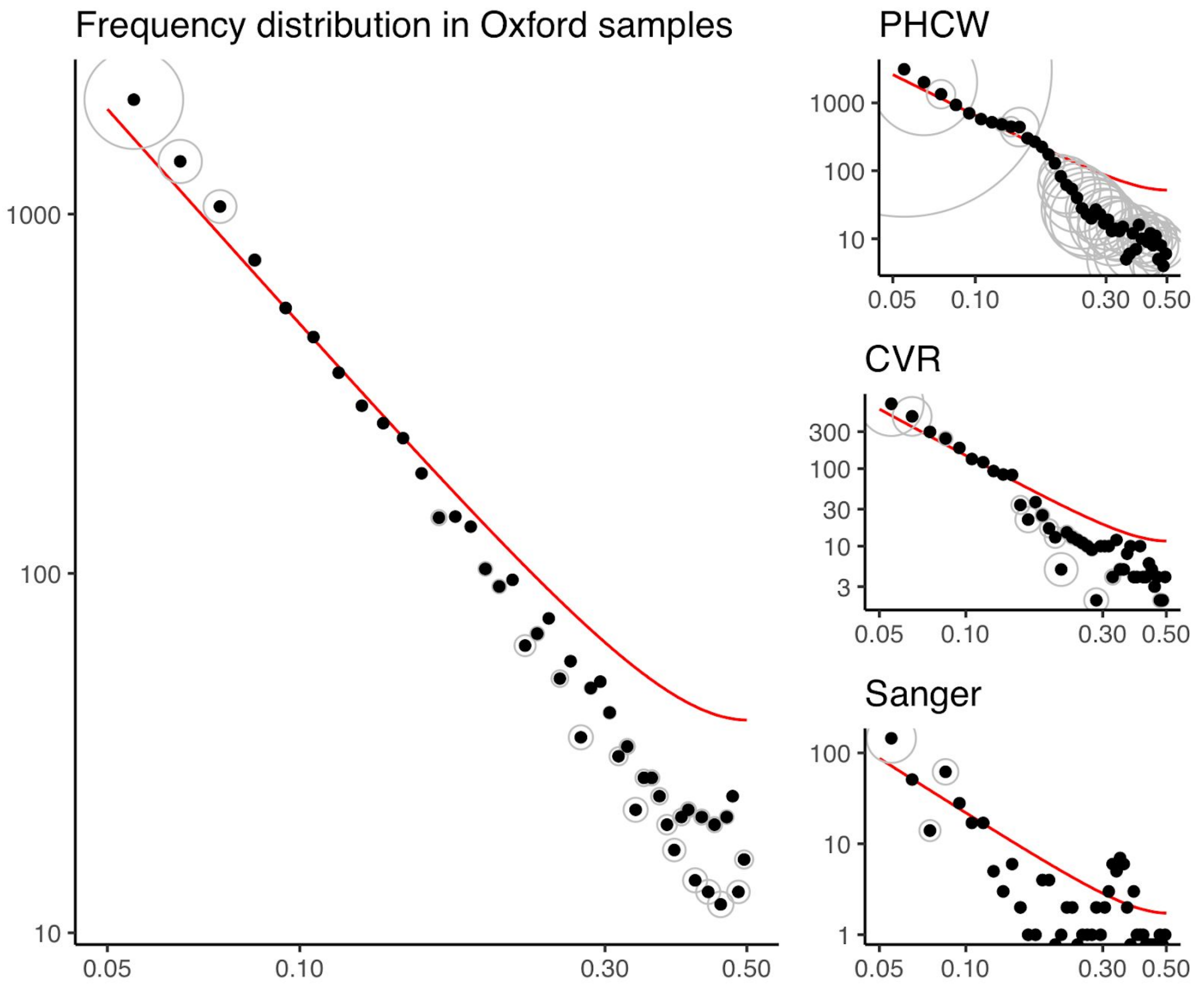
C. Supplementary Figures



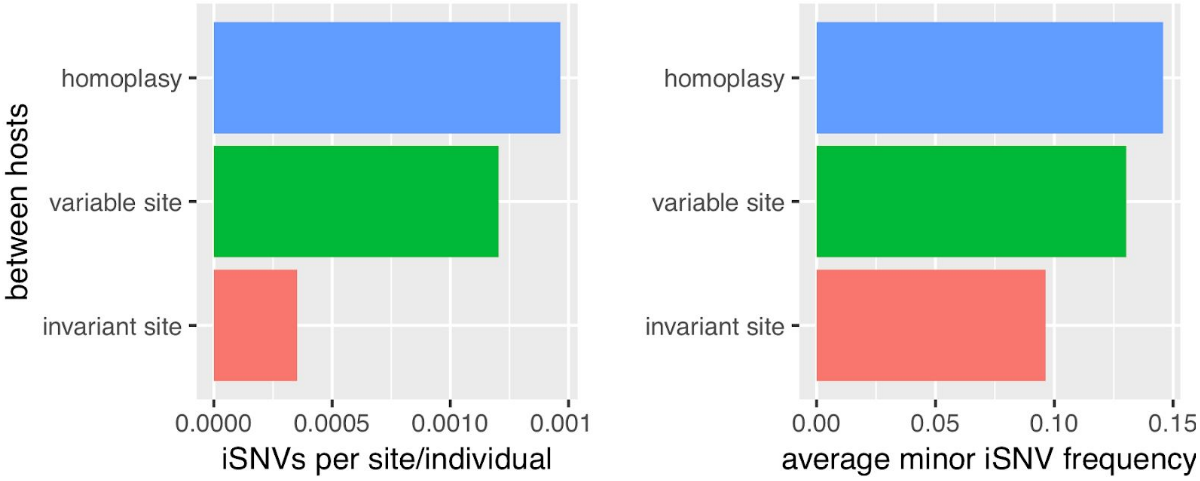
Supplementary Figure 1. Concordance of MAF for two replicates of the same specimen, prepared independently in independent captures, from two separate RNA aliquots. Sample OXON-AF346 was sequenced on a MiSeq instrument to an extremely high depth (median 37,000x) while sample OXON-AF179 was sequenced in a larger pool of 96 samples to a median depth of 1,700x. Plot shows Log10 MAF for the two replicates, with lines indicating MAF of 2% (red) and 5% (blue). The figure on the left gives the MAFs after post-mapping computational cleaning, whereas the figure on the right gives the MAFs before cleaning.



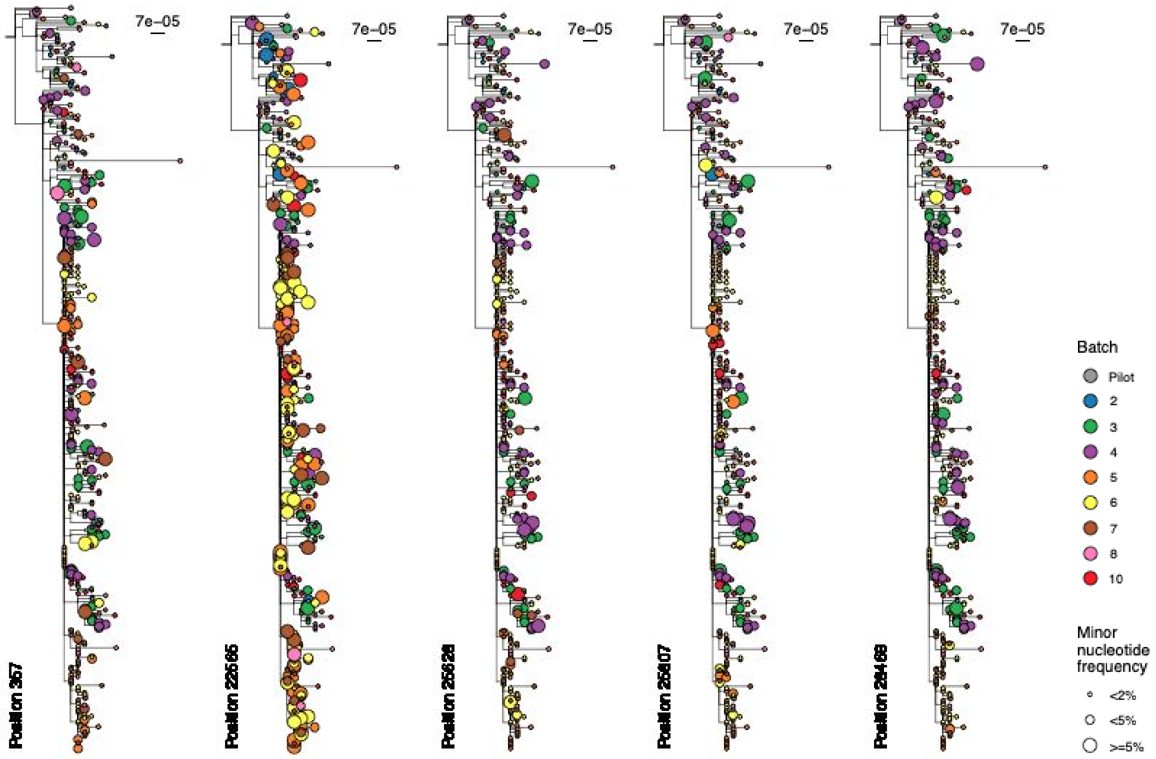
Supplementary Figure 2. Observed maximum minor allele frequency (MAF) per site in the clinical samples from Oxford/Basingstoke (blue) compared with those observed from the synthetic SARS-CoV-2 RNA Twist standards (Twist biosciences). Dashed lines indicate thresholds of 2% (thin) and 5% (thick), the latter being used for ascertainment of sites of interest. Low-level variation in the standards may arise as a result of sequencing and mapping errors, and we used this phenomenon to determine a threshold above which the within-host diversity in clinical samples was identifiable from potential methodological noise. Specifically, we typically used 5% minor allele frequency to ascertain on sites of interest, and 2% once sites had been identified, to maximise information contained in the sequence data.



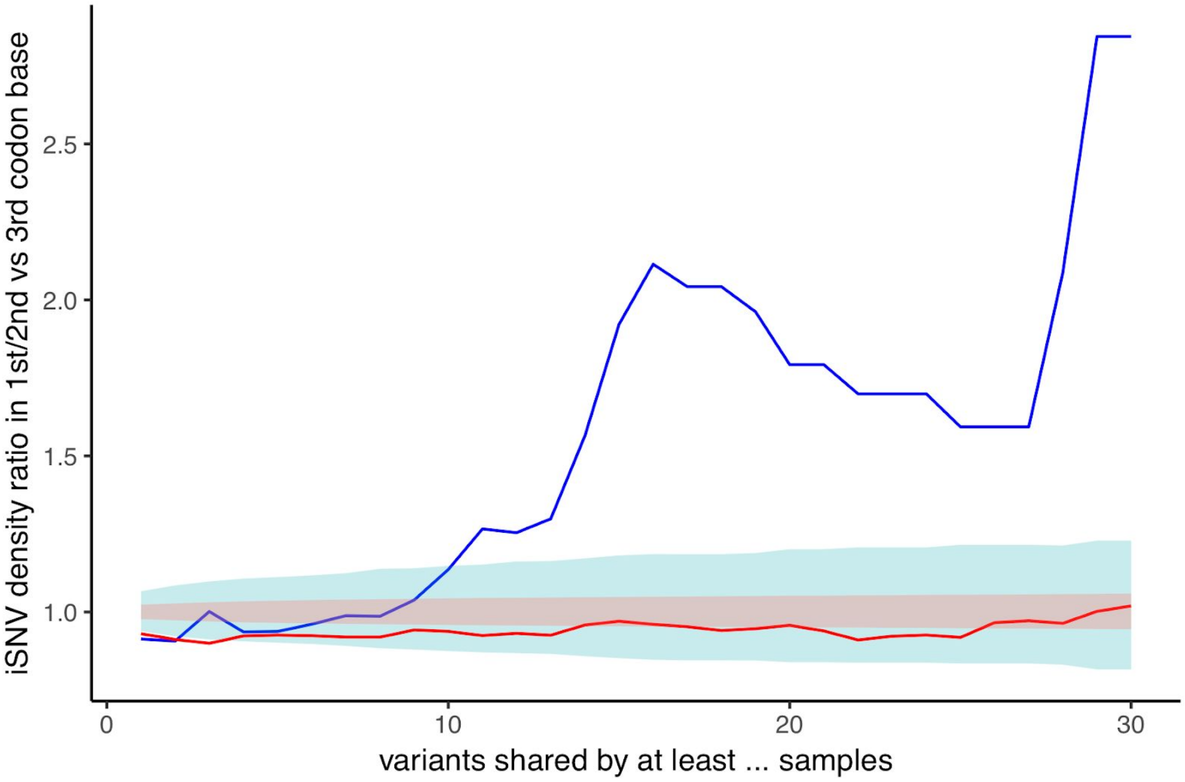
Supplementary Figure 3. Log-log plots of the distributions of minor allele frequencies, binned in intervals of 0.01. The red line shows the Poisson Maximum Likelihood fit of the expected distribution for neutrally evolving populations under rapid exponential growth, which is proportional to $1/f^2+1/(1-f)^2$. The size of the grey circles illustrate the deviation of each bin from the expected curve, in terms of χ^2-goodness of fit. PHCW: Wales, CVR: Scotland, Sanger: Cambridge. Goodness of fit (median of downsampling to 400 iSNVs), Oxford, -0.5; Wales -5.0; Scotland -3.7; Cambridge -42.9. Cramer-Von Mises p-value (median of downsampling to 400 iSNVs), Oxford 0.011; Wales, 5.3E-06; Scotland, 7.7E-07; Cambridge <2E-16.



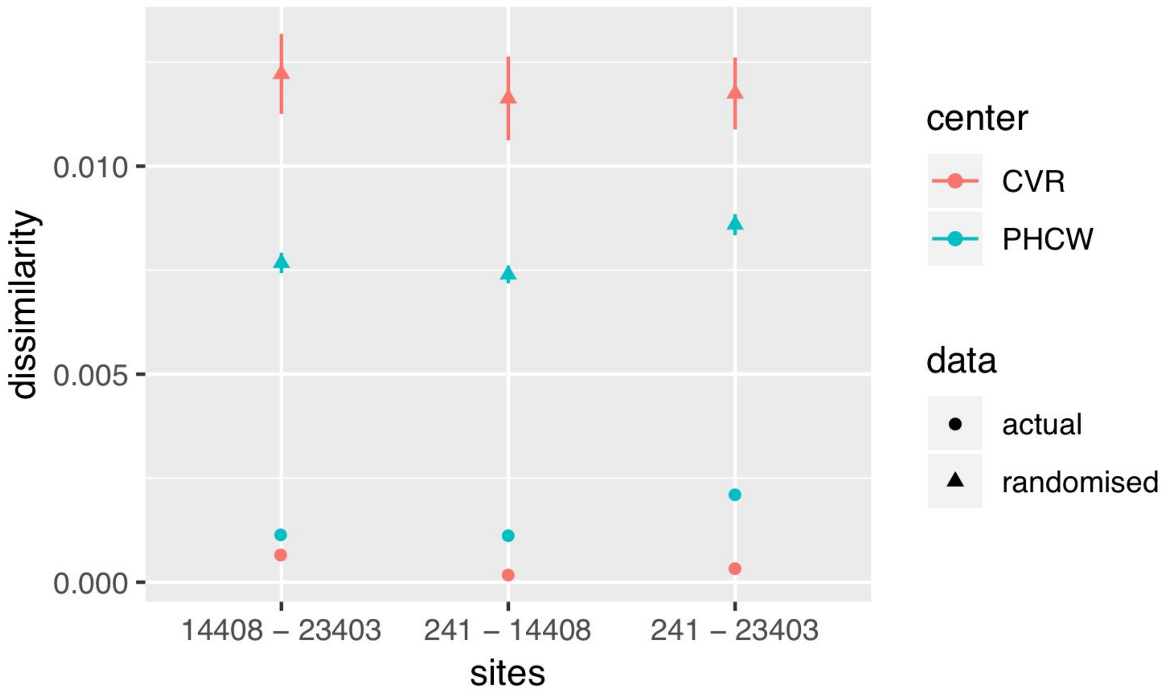
Supplementary Figure 4. Comparison of iSNV sites shared by 2 or more COG-UK samples at >5% frequency and between-host diversity at the same site in the genome. Sites are classified as invariant between hosts, variable (but not homoplasic) sites, and sites with homoplasic variants between hosts. Left: number of iSNVs per site per individual; right: average within-host minor frequency.



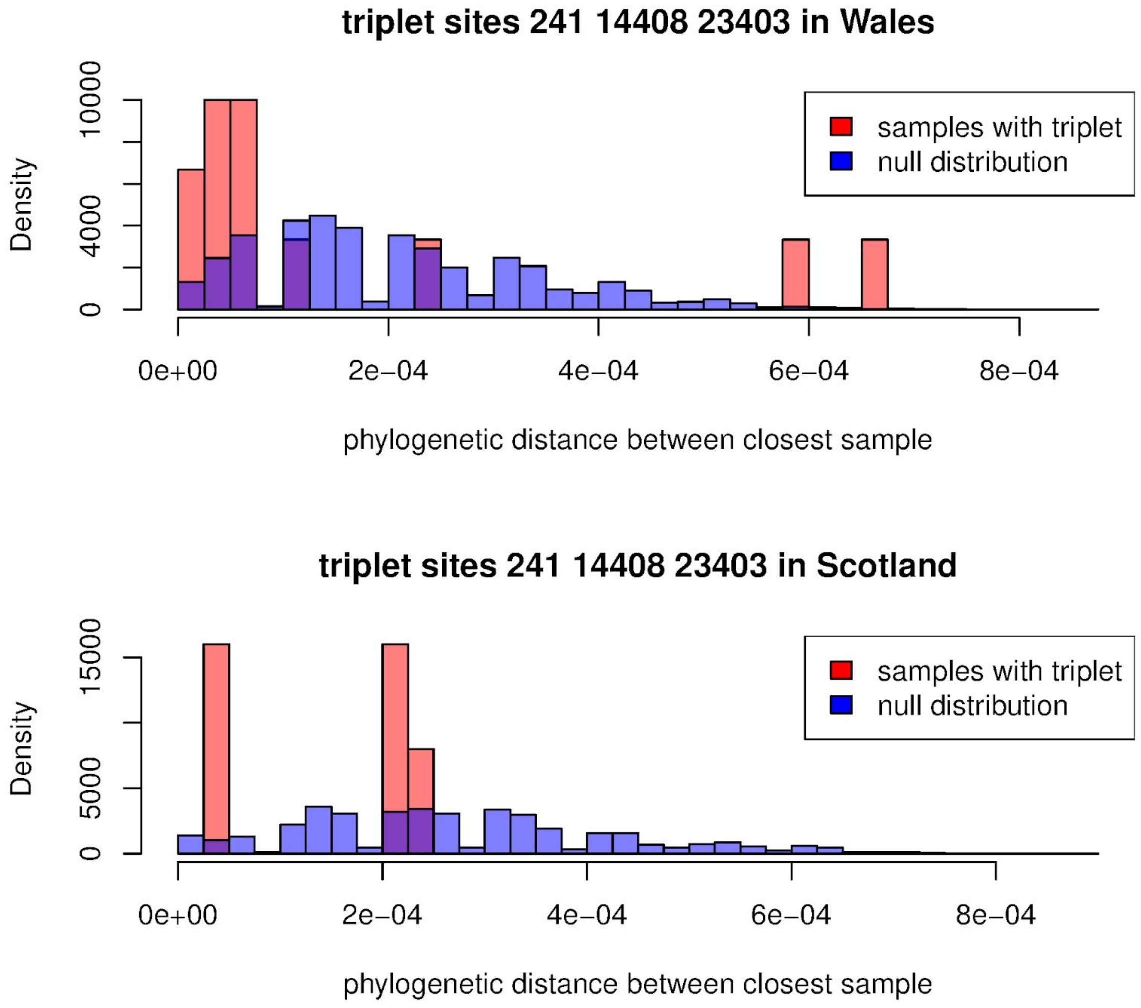
Supplementary Figure 5. The five trees from figure 2 (main text), with tips coloured by sequencing batch rather than location. For each genome position, samples with MAFs of at least 2%, and at least 5%, exist in at least three separate batches, demonstrating that these patterns are not the result of batch-specific contamination.



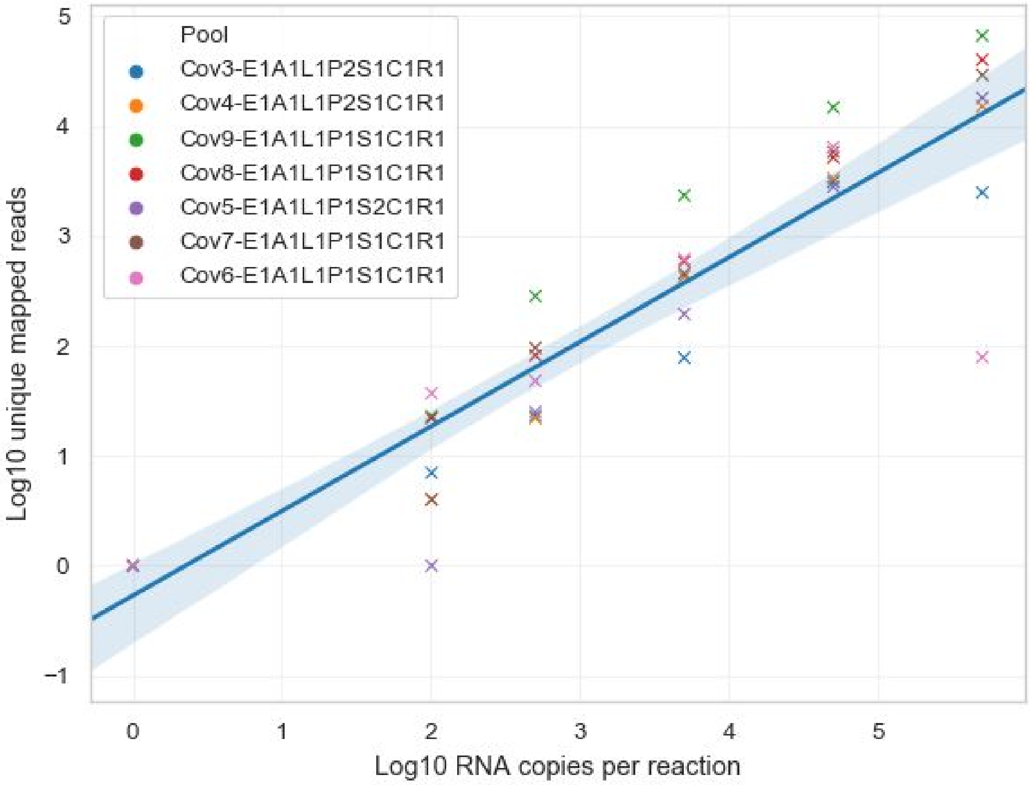
Supplementary Figure 6. Ratio of density of iSNVs in 1st/2nd versus 3rd codon position, computed on all variants with minor frequency >5% shared among multiple Oxford samples (in blue), as a function of the minimum number of samples sharing the variant. A ratio around 1 suggests neutrality, while higher ratios suggest that a fraction of the nonsynonymous variants are under positive selection and therefore appear more often above the 5% frequency threshold. The binomial confidence interval for neutrality at the 95% level is shown in light blue. For variants shared by more than 20 samples, the ratio of iSNV densities in 1st/2nd vs 3rd base is about 1.7; this suggests that at least 7 out of 17 shared variants in the 1st or 2nd base of the codon (i.e. about 32% of the sites shared by more than 20 samples) are under positive selection. The same results for variants with minor frequency >2% are shown in red.



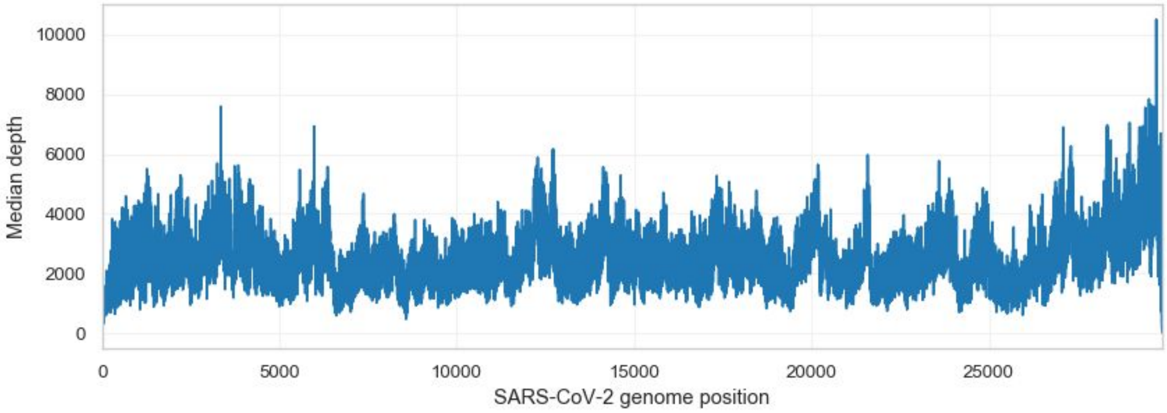
Supplementary Figure 7. Dissimilarity between frequencies at pairs of sites in the triplet 241, 14408, 23403 across multiple individuals (which corresponds to the cumulative amount of recombination between the two sites, and is inversely related to linkage disequilibrium), compared with a randomised sample with the same marginal distributions of frequencies. All differences between actual and randomised data are significant (p<0.001). CVR: Scotland; PHCW: Wales.



Supplementary Figure 8. For each sample containing an iSNV in the triplet of sites 241, 14408 and 23403, we report the phylogenetic distance from the closest sample in the same run containing a triplet iSNV. The null distribution is obtained from 10000 random permutations of iSNVs among samples. Corresponding p-values are computed using two-sided Mann-Whitney U-test (Wales p=0.016, Scotland p=0.07); combined p-value via Fisher's method: p<0.01 (p=0.0094).



Supplementary Figure 9. Correlation between number of SARS-CoV-2 unique reads and RNA copies/ml for a within-batch standard curves from a positive control. The synthetic RNA (Twist) was serially diluted into Universal Human Reference RNA (UHRR) to a final concentration of SARS-CoV-2 RNA of 5e05, 5e04, 5e03, 5e02, 1e02 and 0 copies/reaction. A standard curve was processed and sequenced alongside each batch of samples (batches Cov3-9 shown).



Supplementary Figure 10. Genome-wide median sequencing depth for samples from Oxford and Basingstoke.