nature

https://doi.org/10.1038/s41586-021-03291-y

Accelerated Article Preview

SARS-CoV-2 evolution during treatment of chronic infection

Received: 1 December 2020

Accepted: 26 January 2021

Accelerated Article Preview Published online 5 February 2021

Cite this article as: Kemp, S. A. et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nature* https://doi.org/10.1038/s41586-021-03291-y (2021).

Steven A. Kemp, Dami A. Collier, Rawlings P. Datir, Isabella A. T. M. Ferreira, Salma Gayed, Aminu Jahun, Myra Hosmillo, Chloe Rees-Spear, Petra Mlcochova, Ines Ushiro Lumb, David J. Roberts, Anita Chandra, Nigel Temperton, The CITIID-NIHR BioResource COVID-19 Collaboration, The COVID-19 Genomics UK (COG-UK) Consortium, Katherine Sharrocks, Elizabeth Blane, Yorgo Modis, Kendra Leigh, John Briggs, Marit van Gils, Kenneth G. C. Smith, John R. Bradley, Chris Smith, Rainer Doffinger, Lourdes Ceron-Gutierrez, Gabriela Barcenas-Morales, David D. Pollock, Richard A. Goldstein, Anna Smielewska, Jordan P. Skittrall, Theodore Gouliouris, Ian G. Goodfellow, Effrossyni Gkrania-Klotsas, Christopher J. R. Illingworth, Laura E. McCoy, & Ravindra K. Gupta

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Article

SARS-CoV-2 evolution during treatment of chronic infection

https://doi.org/10.1038/s41586-021-03291-y

Received: 1 December 2020

Accepted: 26 January 2021

Published online: 5 February 2021

Steven A. Kemp[1,19], Dami A. Collier[1,2,3,19], Rawlings P. Datir[2,3,19], Isabella A. T. M. Ferreira[2,3], Salma Gayed[4], Aminu Jahun[5], Myra Hosmillo[5], Chloe Rees-Spear[1], Petra Mlcochova[2,3], Ines Ushiro Lumb[6], David J. Roberts[6], Anita Chandra[2,3], Nigel Temperton[7], The CITIID-NIHR BioResource COVID-19 Collaboration*, The COVID-19 Genomics UK (COG-UK) Consortium*, Katherine Sharrocks[4], Elizabeth Blane[3], Yorgo Modis[8], Kendra Leigh[8], John Briggs[8], Marit van Gils[9], Kenneth G. C. Smith[2,3], John R. Bradley[3,10], Chris Smith[11], Rainer Doffinger[13], Lourdes Ceron-Gutierrez[13], Gabriela Barcenas-Morales[13,14], David D. Pollock[15], Richard A. Goldstein[1], Anna Smielewska[5,11], Jordan P. Skittrall[4,12,16], Theodore Gouliouris[4], Ian G. Goodfellow[5], Effrossyni Gkrania-Klotsas[4], Christopher J. R. Illingworth[12,17], Laura E. McCoy[1] & Ravindra K. Gupta[2,3,18]✉

SARS-CoV-2 Spike protein is critical for virus infection via engagement of ACE2[1], and is a major antibody target. Here we report chronic SARS-CoV-2 with reduced sensitivity to neutralising antibodies in an immune suppressed individual treated with convalescent plasma, generating whole genome ultradeep sequences over 23 time points spanning 101 days. Little change was observed in the overall viral population structure following two courses of remdesivir over the first 57 days. However, following convalescent plasma therapy we observed large, dynamic virus population shifts, with the emergence of a dominant viral strain bearing D796H in S2 and ΔH69/ΔV70 in the S1 N-terminal domain NTD of the Spike protein. As passively transferred serum antibodies diminished, viruses with the escape genotype diminished in frequency, before returning during a final, unsuccessful course of convalescent plasma. *In vitro*, the Spike escape double mutant bearing ΔH69/ΔV70 and D796H conferred modestly decreased sensitivity to convalescent plasma, whilst maintaining infectivity similar to wild type. D796H appeared to be the main contributor to decreased susceptibility but incurred an infectivity defect. The ΔH69/ΔV70 single mutant had two-fold higher infectivity compared to wild type, possibly compensating for the reduced infectivity of D796H. These data reveal strong selection on SARS-CoV-2 during convalescent plasma therapy associated with emergence of viral variants with evidence of reduced susceptibility to neutralising antibodies.

Clinical case history of SARS-CoV-2 infection in setting of immune-compromised host

A septuagenarian male was admitted to a tertiary hospital in summer of 2020 and had tested positive for SARS-CoV-2 RT-PCR 35 days previously on a nasopharyngeal swab (Day 1) at a local hospital (Extended data 1 and 2). His past medical history was significant for marginal B cell lymphoma diagnosed in 2012, with previous chemotherapy including vincristine, prednisolone, cyclophosphamide and anti-CD20 B cell depletion with rituximab. It is likely that both chemotherapy and underlying lymphoma contributed to B and T cell combined immunodeficiency (Extended data 2 and 3, Supplementary Table 1). Computed tomography (CT) of the chest showed widespread abnormalities consistent with COVID-19 pneumonia (Supplementary Figure 1). Treatment included two 10-day courses of remdesivir with a five day gap in between (Extended data 1). Two units of convalescent plasma were administered on days 63 and 65 (Extended data 3). Following clinical deterioration, remdesivir and a unit of convalescent plasma were administered on day 95, but the individual unfortunately died on day 102 (Supplementary text).

[1]Division of Infection and Immunity, University College London, London, UK. [2]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Cambridge, UK. [3]Department of Medicine, University of Cambridge, Cambridge, UK. [4]Department of Infectious Diseases, Cambridge University NHS Hospitals Foundation Trust, Cambridge, UK. [5]Department of Pathology, University of Cambridge, Cambridge, UK. [6]NHS Blood and Transplant, Oxford and BRC Haematology Theme, University of Oxford, Oxford, UK. [7]Viral Pseudotype Unit, Medway School of Pharmacy, University of Kent, Canterbury, UK. [8]Medical Research Council Laboratory of Molecular Biology, Cambridge, UK. [9]Department of Medical Microbiology, Academic Medical Center, University of Amsterdam, Amsterdam Institute for Infection and Immunity, Amsterdam, The Netherlands. [10]NIHR Cambridge Clinical Research Facility, Cambridge, UK. [11]Department of Virology, Cambridge University NHS Hospitals Foundation Trust, Cambridge, UK. [12]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [13]Department of Clinical Biochemistry and Immunology, Addenbrooke's Hospital, Cambridge, UK. [14]FES-Cuautitlán, UNAM, Cuautitlán Izcalli, Mexico. [15]Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, USA. [16]Clinical Microbiology and Public Health Laboratory, Addenbrooke's Hospital, Cambridge, UK. [17]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. [18]Africa Health Research Institute, Durban, South Africa. [19]These authors contributed equally: Steven A. Kemp, Dami A. Collier, Rawlings P. Datir. *Lists of authors and their affiliations appear in the Supplementary Information. ✉e-mail: rkg20@cam.ac.uk

Virus genomic comparative analysis of 23 sequential respiratory samples over 101 days

The majority of samples were respiratory samples from nose and throat or endotracheal aspirates during the period of intubation (Supplementary Table 3). Ct values ranged from 16-34 and all 23 respiratory samples were successfully sequenced by standard single molecule sequencing approach as per the ARTIC protocol implemented by COG-UK; of these 20 additionally underwent short-read deep sequencing using the Illumina platform (Supplementary table 4). There was general agreement between the two methods (Extended data 4). However due to the higher reliability of Illumina for low frequency variants, this was used for formal analysis[2,3]. Additionally, single genome amplification and sequencing of Spike using extracted RNA from respiratory samples was used as an independent method to detect mutations observed (Extended data 4). Finally, we detected no evidence of recombination, based on two independent methods.

Maximum likelihood analysis of patient-derived whole genome consensus sequences demonstrated clustering with other local sequences from the same region (Figure 1). The infecting strain was assigned to lineage 20B bearing the D614G Spike variant. Environmental sampling showed evidence of virus on surfaces such as telephone and call bell. Sequencing of these surface viruses showed clustering with those derived from the respiratory tract (Extended data 2). All samples were consistent with having arisen from a single underlying viral population. In our phylogenetic analysis, we included sequential sequences from three other local patients identified with persistent viral RNA shedding over a period of 4 weeks or more as well as two long term immunosuppressed SARS-CoV-2 'shedders' recently reported[4,5], (Extended data 2, Supplementary Table 2). While the sequences from the three local patients as well as from Avanzato et al[5] showed little divergence with no amino acid changes in Spike over time, the case patient showed significant diversification. The Choi et al report[4] showed similar degree of diversification as the case patient. Further investigation of the sequence data suggested the existence of an underlying structure to the viral population in our patient, with samples collected at days 93 and 95 being rooted within, but significantly divergent from the original population (Extended data 5 and 6). The relationship of the divergent samples to those at earlier time points argues against superinfection.

SARS-CoV-2 viral diversity

All samples tested positive by RT-PCR and there was no sustained change in Ct values throughout the 101 days following the first two courses of remdesivir (days 41 and 54), or the first two units of convalescent plasma with polyclonal antibodies (days 63 and 65, Extended data 3). Of note we were not able to culture virus from stored swab samples. Consensus sequences from short read deep sequence Illumina data revealed dynamic population changes after day 65, as shown by a highlighter plot (Extended data 6). In addition, we were also able to follow the dynamics of virus populations down to low frequencies during the entire period (Figure 2, Supplementary Table 4). Following remdesivir at day 41 the low frequency variant analysis allowed us to observe transient amino acid changes in populations at below 50% abundance in Orf 1b, 3a and Spike, with a T39I (C27509T) mutation in ORF7a reaching 79% on day 45 (Figure 2, pink, supplementary information). At day 66 we noted I513T in NSP2 (T2343C) and V157L (G13936T) in RdRp had emerged from undetectable at day 54 to almost 100% frequency (Figure 2, red and green dashed lines), with the polymerase being the more plausible candidate for driving this sweep. Notably, spike variant N501Y, which can increase the ACE2 receptor affinity[6], and which is present in the new UK B1.1.7 lineage[7], was observed on day 55 at 33% frequency, but was eliminated by the sweep of the NSP2/RdRp variant.

In contrast to the early period of infection, between days 66 and 82, following the first two administrations of convalescent sera, a shift in the virus population was observed, with a variant bearing D796H in S2 and ΔH69/ΔV70 in the S1 N-terminal domain (NTD) becoming the dominant population at day 82. This was identified in a nose and throat swab sample with high viral load as indicated by Ct of 23 (Figure 3A). The deletion was detected transiently at baseline according to short read deep sequencing. ΔH69/ΔV70 was due to an out of frame six nucleotide deletion resulting in the sequence of codon 68 changing from ATA to ATC.

On Days 86 and 89, viruses obtained from upper respiratory tract samples were characterised by the Spike mutations Y200H and T240I, with the deletion/mutation pair observed on day 82 having fallen to frequencies of 10% or less (Figure 2 and 3). The Spike mutations Y200H and T240I were accompanied at high frequency by two other non-synonymous variants with similar allele frequencies, coding for I513T in NSP2, V157L in RdRp and N177S in NSP15 (Figure 2A). Both of these were also previously observed at >98% frequency in the sample on day 66 (Figure 2A, red and green lines), arguing that this new lineage emerged out of a previously existing population.

Sequencing of a nose and throat swab sample at day 93 identified viruses characterised by Spike mutations P330S at the edge of the RBD and W64G in S1 NTD at close to 100% abundance, with D796H along with ΔH69/ΔV70 at <1% abundance and the variants Y200H and T240I at frequencies of <2%. Viruses with the P330S variant were detected in two independent samples from different sampling sites, arguing against the possibility of contamination. The divergence of these samples from the remainder of the population (Figure 2, 3B and Extended data 5 and 6) suggests the possibility that they represent a compartmentalised subpopulation.

Patterns in the variant frequencies suggest competition between virus populations carrying different mutations, viruses with the D796H/ΔH69/ΔV70 deletion/mutation pair rising to high frequency during CP therapy, then being outcompeted by another population in the absence of therapy. Specifically, these data are consistent with a lineage of viruses with the NSP2 I513T and RdRp V157L variant, dominant on day 66, being outcompeted during therapy by the mutation/deletion variant. With the lapse in therapy, the original strain, having acquired NSP15 N1773S and the Spike mutations Y200H and T240I, regained dominance, followed by the emergence of a separate population with the W64G and P330S mutations.

In a final attempt to reduce the viral load, a third course of remdesivir (day 93) and third dose of CP (day 95) were administered. We observed a re-emergence of the D796H + ΔH69/ΔV70 viral population (Figure 2, 3). The inferred linkage of D796H and ΔH69/ΔV70 was maintained as evidenced by the highly similar frequencies of the two variants, suggesting that the third unit of CP led to the re-emergence of this population under renewed positive selection. In further support of our proposed idea of competition, noted above, frequencies of these two variants appeared to mirror changes in the NSP2 I513T mutation (Figure 2), suggesting these as markers of opposing clades in the viral population. Ct values remained low throughout this period with hyperinflammation, eventually leading to multi-organ failure and death at day 102. The repeated increase in frequency of the viral population with CP therapy strongly supports the hypothesis that the deletion/mutation combination conferred selective advantage.

Spike mutants emerging post convalescent plasma impair neutralising antibody potency

Using lentiviral pseudotyping we generated wild type, ΔH69/ΔV70 + D796H and single mutant Spike proteins in enveloped virions in order to measure neutralisation activity of CP against these viruses (Figure 4). This system has been shown to give generally similar results to replication competent virus[8,9]. Spike protein from each mutant was detected in pelleted virions (Figure 4A). We also probed with an HIV-1 p24 antibody to monitor levels of lentiviral particle production (Figure 4A, Supplementary Figure 2). We then measured infectivity of the pseudoviruses, correcting for virus input using reverse transcriptase activity measurement, and found that ΔH69/ΔV70 appeared to have two-fold higher infectivity over a single round of infection compared to wild type (Figure 4B, Extended data 7). By contrast, the D796H single mutant had significantly lower infectivity as compared to wild type and double mutant had similar infectivity to wild type (Figure 4B, Extended data 7).

We found that D796H alone and the D796H + ΔH69/ΔV70 double mutant were less sensitive to neutralisation by convalescent plasma samples (Figure 4C-E, Extended data 7). By contrast the ΔH69/ΔV70 single mutant did not reduce neutralisation sensitivity. In addition, patient derived serum from days 64 and 66 (one day either side of CP2 infusion) similarly showed lower potency against the D796H + ΔH69/ΔV70 mutants (Figure 4F, G).

A panel of nineteen monoclonal antibodies (mAbs) isolated from three donors was previously identified to neutralize SARS-CoV-2. To establish if the mutations incurring *in vivo* (D796H and ΔH69/ΔV70) resulted in a global change in neutralization sensitivity we tested neutralising mAbs targeting the seven major epitope clusters previously described (excluding non-neutralising clusters II, V and small [n =<2] neutralising clusters IV, X). The eight RBD-specific mAbs (Extended data 8) exhibited no major change in neutralisation potency and non-RBD specific COVA1-21 showing 3-5 fold reduction in potency against ΔH69/ΔV70+D796H and ΔH69/ΔV70, but not D796H alone[9] (Extended data 8). We observed no differences in neutralisation between single/double mutants and wild type, suggesting that the mechanism of escape was likely outside these epitopes in the RBD. These data confirm the specificity of the findings from convalescent plasma and suggest that mutations observed are related to antibodies targeting regions outside the RBD. Interestingly, ΔH69/ΔV70 containing viruses showed reduced neutralisation sensitivity to the mAb COVA1-21, targeting an as yet undefined epitope outside the RBD[10].

To understand how the ΔH69/ΔV70 and D796H might confer antibody resistance, we assessed how they might affect the Spike structure (Extended data 9). We based this analysis primarily on a structure lacking stabilising modifications (PDB 6xr8)[11], but also referred to stabilised structures determined at different pH values[12]. ΔH69/ΔV70 is located in a disordered, glycosylated loop at the distal surface of the NTD, near the binding site of polyclonal antibodies derived from COV57 plasma[13,14] (Extended data 9). As this loop is flexible and highly accessible, ΔH69/V70 could in principle affect antibody binding in this region. D796 is located near the base of Spike, in a surface loop that is structurally somewhat disordered in the prefusion conformation and becomes part of a large disordered region in the post fusion S2 trimer[11] (Extended data 9). The loop containing residue 796 is proposed to be targeted by antibodies[15], despite mutations at position 796 being relatively uncommon (Extended data 9). In the RBD-down Spike structures[11,12], D796 forms contacts with residues in the neighbouring protomer, including the glycosylated residue N709 (Extended data 9).

Discussion

Here we have documented a repeated evolutionary response by SARS-CoV-2 in the presence of antibody therapy during the course of a persistent infection in an immunocompromised host. The observation of potential selection for specific variants coinciding with the presence of antibodies from convalescent plasma is supported by the experimental finding of two-fold reduced susceptibility of these viruses to convalescent plasma containing polyclonal antibodies. In this case the emergence of the variant was not the primary reason for treatment failure.

We have noted in our analysis signs of compartmentalised viral replication based on the sequences recovered in upper respiratory tract samples. Both population genetic and small animal studies have shown a lack of reassortment between influenza viruses within a single host during an infection, suggesting that acute respiratory viral infection may be characterised by spatially distinct viral populations[16,17]. In the analysis of data, it is important to distinguish genetic changes which occur in the primary viral population from apparent changes that arise from the stochastic observation of spatially distinct subpopulations in the host. While the samples we observe on days 93 and 95 of infection are genetically distinct from the others, the remaining samples are consistent with arising from a consistent viral population. We note that Choi et al reported the detection in post-mortem tissue of viral RNA not only in lung tissue, but also in the spleen, liver, and heart[4]. Mixing of virus from different compartments, for example via blood, or movement of secretions from lower to upper respiratory tract, could lead to fluctuations in viral populations at particular sampling sites.

This is a single case report and therefore limited conclusions can be drawn about generalisability.

An important limitation is that the data were derived from sampling from the upper respiratory tract and not the lower tract, thus limiting the inferences that can be drawn regarding viral populations in this single case.

In addition to documenting the emergence of SARS-CoV-2 Spike ΔH69/ΔV70 *in vivo*, we show that this mutation modestly increases infectivity of the Spike protein in a pseudotyping assay. The deletion was observed contemporaneously with the rare S2 mutation D796H after two separate courses of CP, with other viral populations emerging. D796H, but not ΔH69/ΔV70, conferred reduction in susceptibility to polyclonal antibodies in the units of CP administered, though we cannot speculate as to their individual impacts on sera from other individuals. It is intriguing that the ΔH69/ΔV70 + D796H double mutant diminished in between CP courses, suggesting that there were other selective forces at play in the intervening period, possibly driven by the inflammation observed in the individual. This includes the possibility that the haplotype with ΔH69/ΔV70 + D796H may have carried mutations in other regions deleterious during that intervening period. Although ΔH69/V70 is expanding at a high rate[18], D796 mutations are also increasing. D796H has been documented in 0.02% of global sequences and D796Y appears in 0.05% of global sequences (Extended data 9).

The effects of CP on virus evolution seen here are unlikely to apply in immune competent hosts where viral diversity is likely to be lower due to better immune control. Our data highlight that infection control measures may need to be tailored to the needs of immunocompromised patients and also caution in interpretation of CDC guidelines that recommend 20 days as the upper limit of infection prevention precautions in immune compromised patients who are afebrile[19]. Due to the difficulty with culturing clinical isolates, use of surrogates are warranted[20]. However, where detection of ongoing viral evolution is possible, this serves as a clear proxy for the existence of infectious virus. In our case we detected environmental contamination whilst in a single occupancy room and the patient was moved to a negative-pressure high air-change infectious disease isolation room.

Clinical efficacy of convalescent plasma in severe COVID-19 has not been demonstrated[21], and its use in different stages of infection and disease remains experimental; as such, we suggest that it should be reserved for use within clinical trials, with rigorous monitoring of clinical and virological parameters. The data from this single case report might warrant caution in use of convalescent plasma in patients with immune suppression of both T cell and B cell arms; in such cases, the antibodies administered have little support from cytotoxic T cells, thereby reducing chances of clearance and theoretically raising the potential for escape mutations. Whilst we await further data, where clinical trial enrolment is not possible, convalescent plasma administered for clinical need in immune suppression should ideally only be considered as part of observational studies, undertaken preferably in single occupancy rooms with enhanced infection control precautions, including SARS-CoV-2 environmental sampling and real-time sequencing. Understanding of viral dynamics and characterisation of viral evolution in response to different selection pressures in the immunocompromised host is necessary not only for improved patient management but also for public health benefit.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03291-y.

1. Hoffmann, M. *et al*. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280 e278, https://doi.org/10.1016/j.cell.2020.02.052 (2020).
2. Kim, K. W. *et al*. Respiratory viral co-infections among SARS-CoV-2 cases confirmed by virome capture sequencing. (2020).
3. Bull, R. A. *et al*. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun* **11**, 6272, https://doi.org/10.1038/s41467-020-20075-6 (2020).
4. Choi, B. *et al*. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *The New England journal of medicine* **383**, 2291-2293, https://doi.org/10.1056/NEJMc2031364 (2020).
5. Avanzato, V. A. *et al*. Case Study: Prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised cancer patient. *Cell* (2020).
6. Starr, T. N. *et al*. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310 e1220, https://doi.org/10.1016/j.cell.2020.08.012 (2020).
7. Rambaut A., L. N., Pybus O, Barclay W, Carabelli A. C., Connor T., Peacock T., Robertson D. L., Volz E., 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*, <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> (2020).
8. Schmidt, F. *et al*. Measuring SARS-CoV-2 neutralizing antibody activity using pseudotyped and chimeric viruses. 2020.2006.2008.140871, https://doi.org/10.1101/2020.06.08.140871 %J bioRxiv (2020).
9. Brouwer, P. J. M. *et al*. Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* **369**, 643-650, https://doi.org/10.1126/science.abc5902 (2020).
10. Zussman, M. E., Bagby, M., Benson, D. W., Gupta, R. & Hirsch, R. Pulmonary vascular resistance in repaired congenital diaphragmatic hernia vs. age-matched controls. *Pediatr Res* **71**, 697-700, https://doi.org/10.1038/pr.2012.16 (2012).
11. Cai, Y. *et al*. Distinct conformational states of SARS-CoV-2 spike protein. *Science*, https://doi.org/10.1126/science.abd4251 (2020).
12. Zhou, T. *et al*. Cryo-EM Structures of SARS-CoV-2 Spike without and with ACE2 Reveal a pH-Dependent Switch to Mediate Endosomal Positioning of Receptor-Binding Domains. *Cell Host & Microbe* **28**, 867-879.e865, https://doi.org/10.1016/j.chom.2020.11.004 (2020).
13. Robbiani, D. F. *et al*. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437-442, https://doi.org/10.1038/s41586-020-2456-9 (2020).
14. Barnes, C. O. *et al*. Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* **182**, 828-842 e816, https://doi.org/10.1016/j.cell.2020.06.025 (2020).
15. Shrock, E. *et al*. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science*, https://doi.org/10.1126/science.abd4250 (2020).
16. Sobel Leonard, A. *et al*. The effective rate of influenza reassortment is limited during human infection. *PLoS Pathog* **13**, e1006203, https://doi.org/10.1371/journal.ppat.1006203 (2017).
17. Richard, M., Herfst, S., Tao, H., Jacobs, N. T. & Lowen, A. C. Influenza A Virus Reassortment Is Limited by Anatomical Compartmentalization following Coinfection via Distinct Routes. *J Virol* **92**, https://doi.org/10.1128/JVI.02063-17 (2018).
18. Kemp, S. *et al*. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion H69/V70. *bioRxiv*, 2020.2012.2014.422555, https://doi.org/10.1101/2020.12.14.422555 (2021).
19. CDC. *Discontinuation of Transmission-Based Precautions and Disposition of Patients with COVID-19 in Healthcare Settings (Interim Guidance)*, <https://www.cdc.gov/coronavirus/2019-ncov/hcp/disposition-hospitalized-patients.html> (2020).
20. Boshier, F. A. T. *et al*. Remdesivir induced viral RNA and subgenomic RNA suppression, and evolution of viral variants in SARS-CoV-2 infected patients. *medRxiv*, 2020.2011.2018.20230599, https://doi.org/10.1101/2020.11.18.20230599 (2020).
21. Simonovich, V. A. *et al*. A Randomized Trial of Convalescent Plasma in Covid-19 Severe Pneumonia. *N Engl J Med*, https://doi.org/10.1056/NEJMoa2031304 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

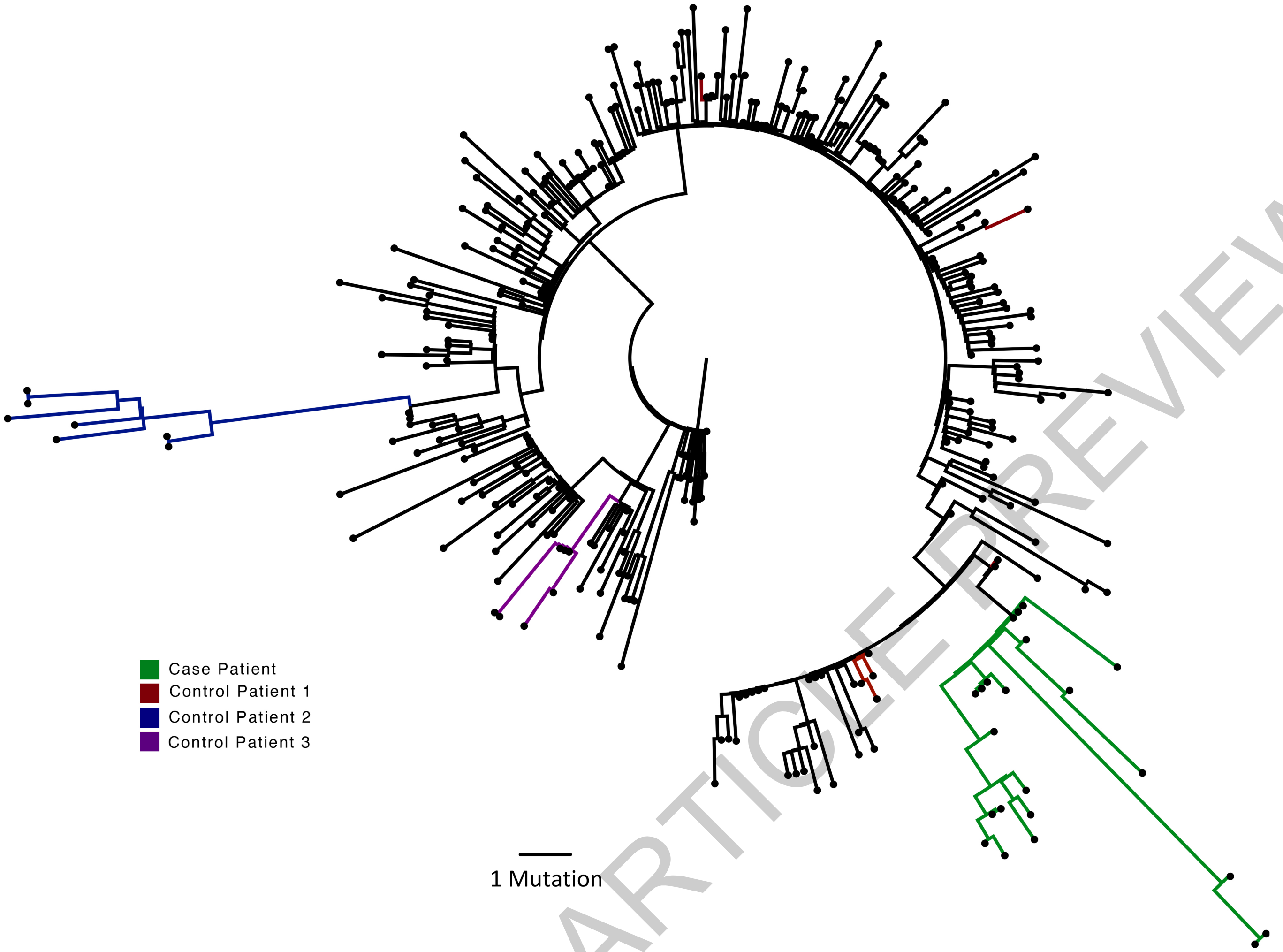


Fig. 1 | Analysis of 23 Patient derived whole SARS-CoV-2 genome sequences. in context of local sequences and other cases of chronic SARS-CoV-2 shedding. Circularised maximum-likelihood phylogenetic tree rooted on the Wuhan-Hu-1 reference sequence, showing a subset of 250 local SARS-CoV-2 genomes from GISAID. This diagram highlights significant diversity of the case patient (green) compared to three other local patients with prolonged shedding (blue, red and purple sequences). All "United Kingdom / English" SARS-CoV-2 genomes were downloaded from the GISAID database and a random subset of 250 selected as background.

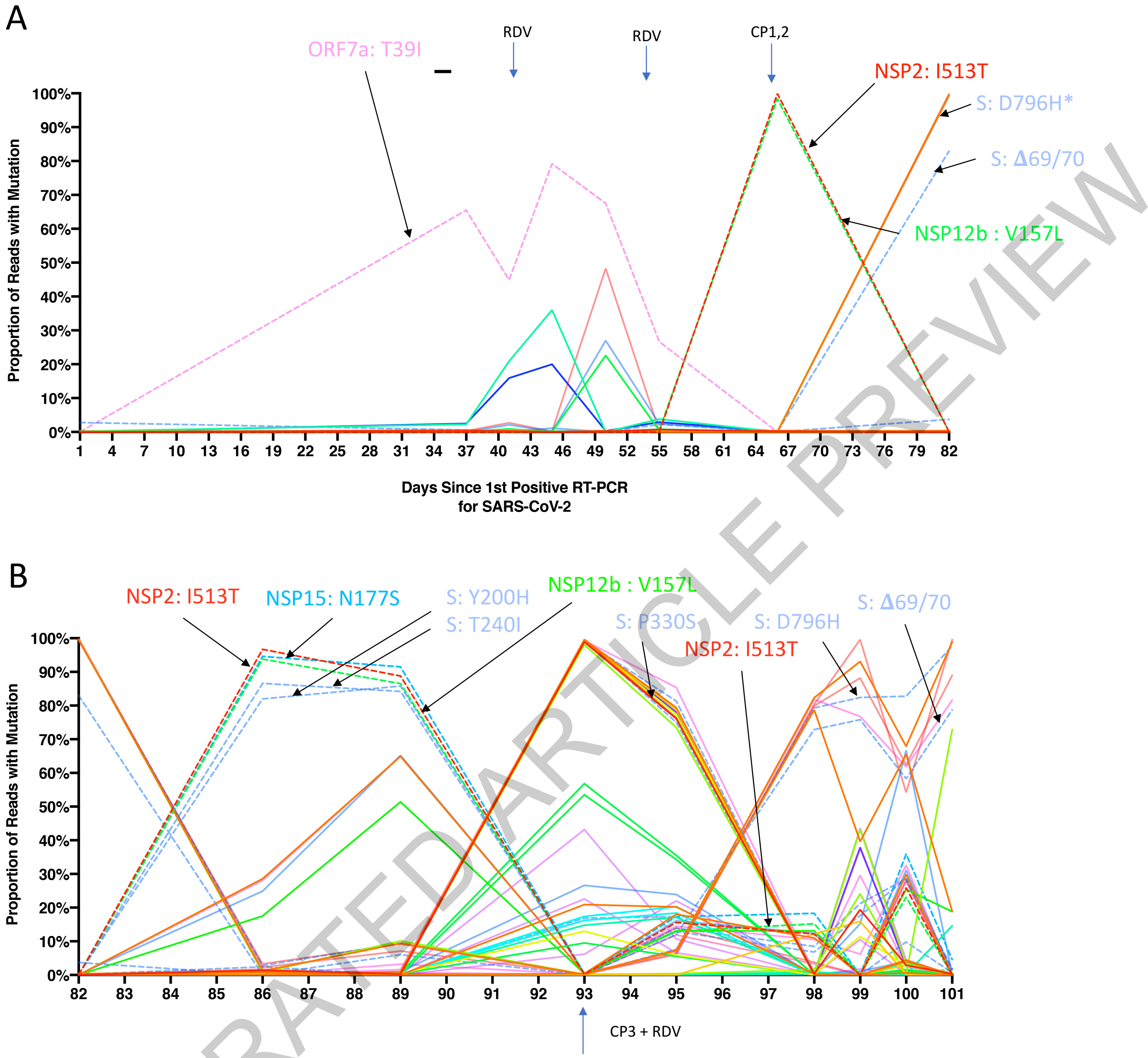


Fig. 2 | Whole genome variant trajectories showing amino acids and relationship to treatments. Data based on Illumina short read ultra deep sequencing at 1000x coverage. Variants shown reached a frequency of at least 10% in at least 2 samples. Treatments indicated are convalescent plasma (CP) and Remdesivir (RDV). Variants described in the text are designated by labels using the same colouring as the position in the genome. Variants labelled are represented by dashed lines. **A.** Variants detected in the patient from days 1-82. *D796H (light blue) is at the same frequency as NSP3 K902N (orange) therefore it is hidden beneath **B.** Variants detected in the patient from days 82-101.

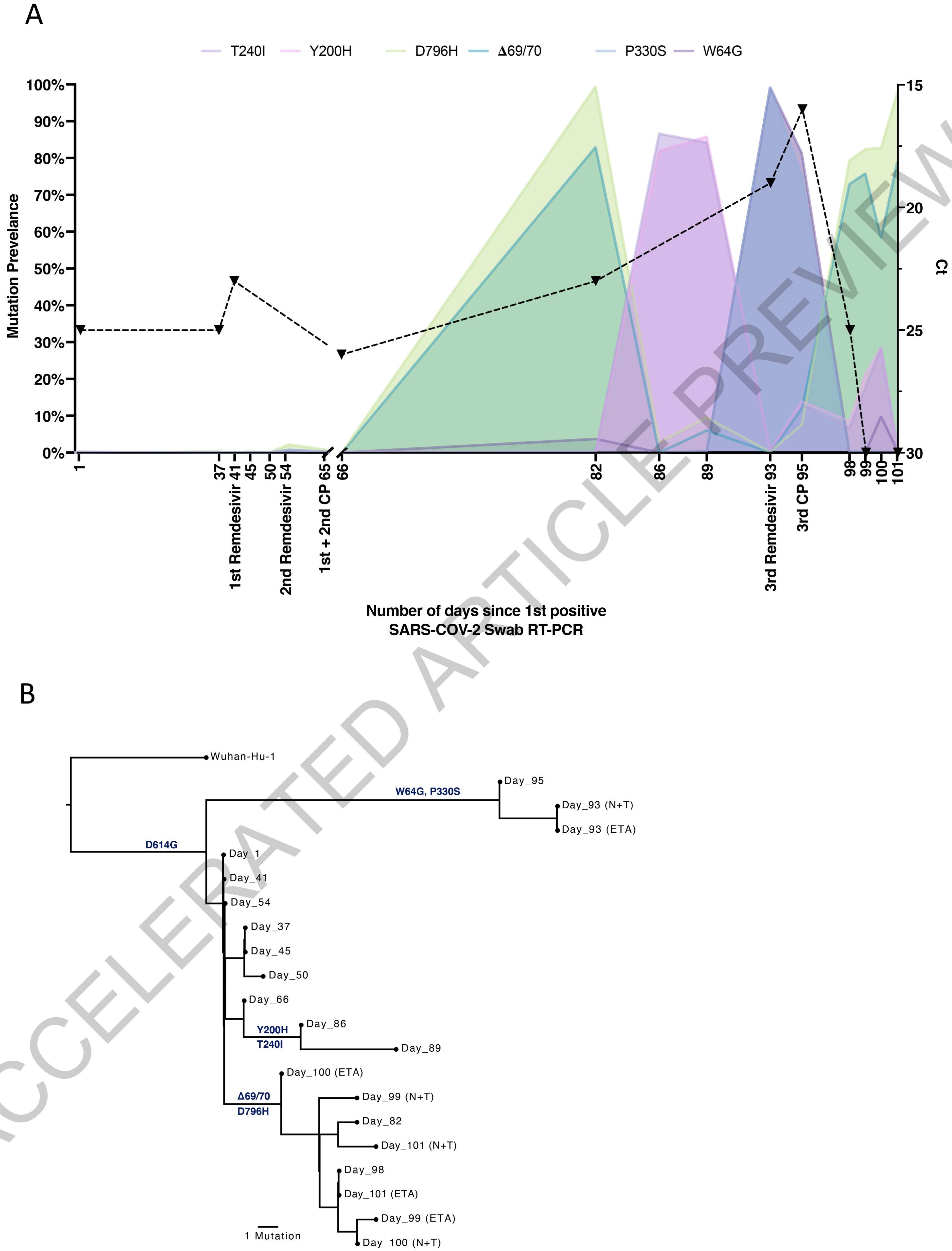


Fig. 3 | See next page for caption.

Fig. 3 | Longitudinal variant frequencies and phylogenetic relationships for virus populations bearing six Spike (S) mutations A. At baseline, all six S variants (Illumina sequencing) except for ΔH69/V70 were absent (<1% and <20 reads). Approximately two weeks after receiving two units of convalescent plasma (CP), viral populations carrying ΔH69/V70 and D796H mutants rose to frequencies >80% but decreased significantly four days later. This population was replaced by a population bearing Y200H and T240I, detected in two samples over a period of 6 days. These viral populations were then replaced by virus carrying W64G and P330S mutations in Spike, which both dominated at day 93. Following a 3rd course of remdesivir and an additional unit of convalescent plasma, the ΔH69/V70 and D796H virus population re-emerged to become the dominant viral strain reaching variant frequencies of >75%. Pairs of mutations arose and disappeared simultaneously indicating linkage on the same viral haplotype. CT values from respiratory samples are indicated on the right y-axis (black dashed line and triangles). Where there were duplicate readings on the same day, to remain consistent, N+T samples were plotted **B**. Maximum likelihood phylogenetic tree of the case patient with day of sampling indicated. Spike mutations defining each of the clades are shown ancestrally on the branches on which they arose. On dates where multiple samples were collect, these are indicated as endotracheal aspirate (ETA) and Nose + throat swabs (N+T).

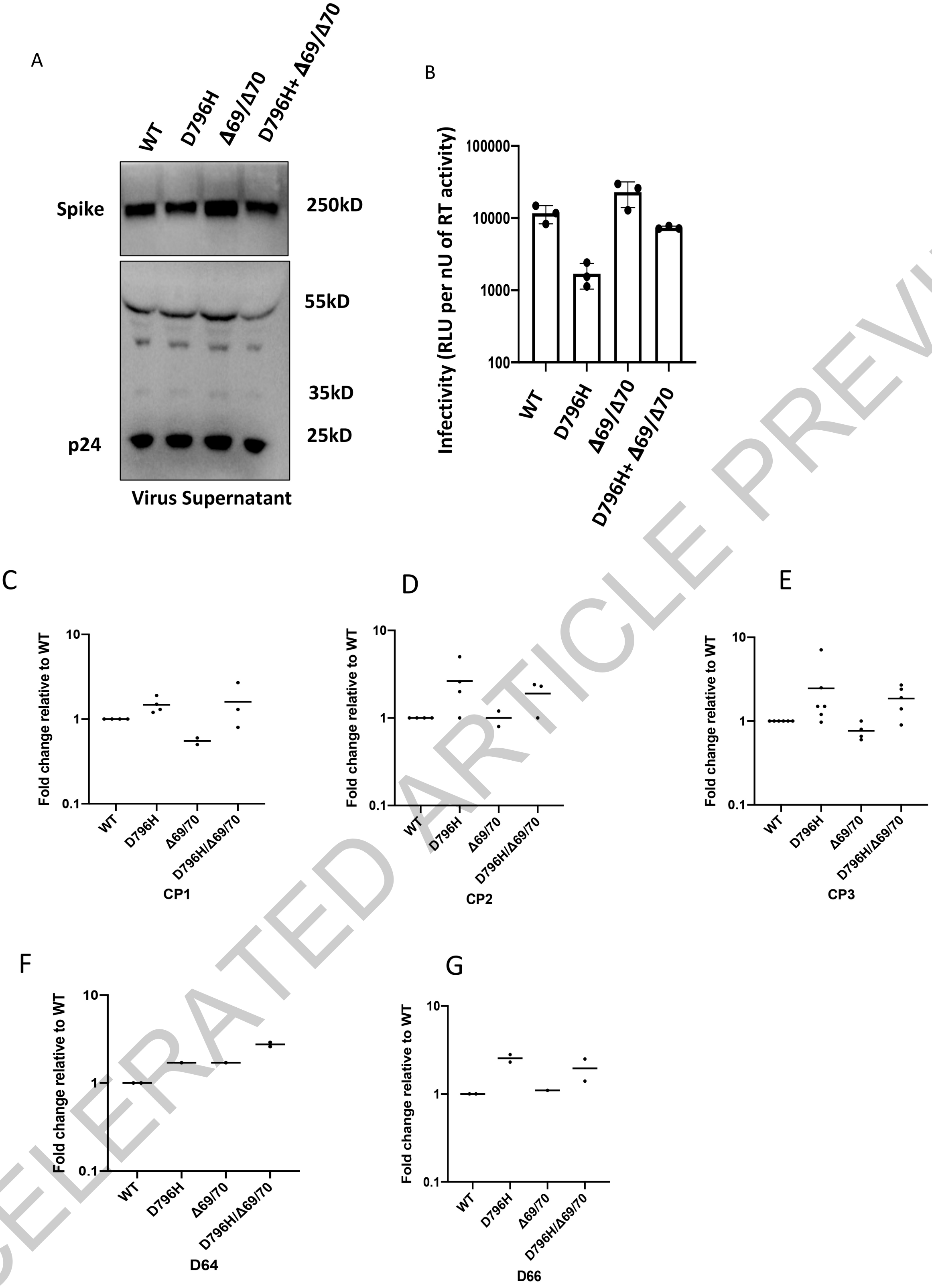


Fig. 4 | Spike mutant D796H + ΔH69/V70 infectivity and sensitivity convalescent plasma (CP). A. western blot of virus pellets after centrifugation of supernatants from cells transfected with lentiviral pseudotyping plasmids including Spike protein. Blots are representative of two independent transfections. **B**. Single round Infectivity of luciferase expressing lentivirus pseudotyped with SARS-CoV-2 Spike protein (WT versus mutant) on 293T cells co-transfected with ACE2 and TMPRSS2 plasmids. Infectivity is corrected for reverse transcriptase activity in virus supernatant as measured by real time PCR. Data points represent technical replicates (n=3) with mean and error bars representing standard error of mean; data are representative of two independent experiments **C-E**. convalescent plasma (CP units 1-3) neutralization potency against pseudovirus virus bearing Spike mutants D796H, ΔH69/V70 and D796H + ΔH69/V70 **F, G** patient serum neutralisation potency against pseudovirus virus bearing Spike mutants D796H, ΔH69/V70 and D796H + ΔH69/V70. Patient serum was taken at indicated Day (D). Indicated is serum dilution required to inhibit 50% of virus infection (ID50), expressed as fold change relative to WT. Data points represent means of technical replicates and each data point is an independent experiment (n=2-6). Mean of data points in C-G is shown by horizontal bars.

Methods

Clinical Sample Collection and Next generation sequencing

Serial samples were collected from the patient periodically from the lower respiratory tract (sputum or endotracheal aspirate), upper respiratory tract (throat and nasal swab), and from stool. Nucleic acid extraction was done from 500μl of sample with a dilution of MS2 bacteriophage to act as an internal control, using the easyMAG platform (Biomerieux, Marcy-l'Étoile) according to the manufacturers' instructions. All samples were tested for presence of SARS-CoV-2 with a validated one-step RT q-PCR assay developed in conjunction with the Public Health England Clinical Microbiology[22]. Amplification reaction were all performed on a Rotorgene™ PCR instrument. Samples which generated a CT of ≤36 were considered to be positive.

Sera from recovered patients in the COVIDx study[23] were used for testing of neutralisation activity by SARS-CoV-2 mutants.

SARS-CoV-2 serology by multiplex particle-based flow cytometry (Luminex)

Recombinant SARS-CoV-2 N, S and RBD were covalently coupled to distinct carboxylated bead sets (Luminex; Netherlands) to form a 3-plex and analyzed as previously described (Xiong et al. 2020). Specific binding was reported as mean fluorescence intensities (MFI).

Whole blood T cell and innate stimulation assay

Whole blood was diluted 1:5 in RPMI into 96-well F plates (Corning) and activated by single stimulation with phytohemagglutinin (PHA; 10 μg/ml; Sigma-Aldrich), or LPS (1 μg/ml, List Biochemicals) or by co-stimulating with anti-CD3 (MEM57, Abcam, 200 ng/ml, 1:1000) and IL-2 (Immunotools, 1430U/ml, 1:1000). Supernatants were taken after 24 hours. Levels (pg/ml) are shown for IFNg, IL17, IL2, TNFa, IL6, IL1b and IL10. Cytokines were measured by multiplexed particle based Flow cytometry on a Luminex analyzer (Bio-Plex, Bio-Rad, UK) using an R&D Systems custom kit (R&D Systems, UK).

For viral genomic sequencing, total RNA was extracted from samples as described. Samples were sequenced using MinION flow cells version 9.4.1 (Oxford Nanopore Technologies) following the ARTICnetwork V3 protocol (https://doi.org/10.17504/protocols.io.bbmuik6w) and BAM files assembled using the ARTICnetwork assembly pipeline (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). A representative set of 10 sequences were selected and also sequenced using the Illumina MiSeq platform. Amplicons were diluted to 2 ng/μl and 25 μl (50 ng) were used as input for each library preparation reaction. The library preparation used KAPA Hyper Prep kit (Roche) according to manufacturer's instructions. Briefly, amplicons were end-repaired and had A-overhang added; these were then ligated with 15mM of NEXTflex DNA Barcodes (Bio Scientific, Texas, USA). Post-ligation products were cleaned using AMPure beads and eluted in 25 μl. Then, 20 μl were used for library amplification by 5 cycles of PCR. For the negative controls, 1ng was used for ligation-based library preparation. All libraries were assayed using TapeStation (Agilent Technologies, California, USA) to assess fragment size and quantified by QPCR. All libraries were then pooled in equimolar accordingly. Libraries were loaded at 15nM and spiked in 5% PhiX (Illumina, California, USA) and sequenced on one MiSeq 500 cycle using a Miseq Nano v2 with 2x 250 paired-end sequencing. A minimum of ten reads were required for a variant call.

Bioinformatics Processes

For long-read sequencing, genomes were assembled with reference-based assembly and a curated bioinformatics pipeline with 20x minimum coverage across the whole-genome[24]. For short-read sequencing, FASTQs were downloaded, poor-quality reads were identified and removed, and both Illumina and PHiX adapters were removed using TrimGalore v0.6.6[25]. Trimmed paired-end reads were mapped to the National Center for Biotechnology Information SARS-CoV-2 reference sequence MN908947.3 using MiniMap2-2.17 with arguments -ax and sr[26]. BAM files were then sorted and indexed with samtools v1.11 and PCR optical duplicates removed using Picard (http://broadinstitute.github.io/picard). A consensus sequences of nucleic acids with a minimum whole-genome coverage of at least 20× were generated with BCFtools using a 0% majority threshold.

Variant calling

Variant frequencies were validated using custom code as part of the *AnCovMulti* package (github.com/PollockLaboratory/AnCovMulti). The main idea behind this validation was to identify and remove consistent potential amplification errors and mutability near the end of Illumina reads. Furthermore, stringent filtering was applied to remove biased amplification of early laboratory-induced mutations or very low copy variations.

Filtering consisted of requiring exact initiation at a primer within two bp of the start of a read, a minimum of 247 bp length read, fewer than four well-separated sites divergent from the reference sequence, a maximum insertion size of three nucleotides, a maximum deletion size of 11 bp, and resolution of conflicting signal from different primers.

Single Genome Amplification and sequencing

Viral RNA extracts were reverse transcribed from each sample to sufficiently capture the diversity of the viral population without introducing resampling bias. SuperScript IV (Thermofisher Scientific) and the gene specific primers were used for reverse transcription. Template RNA was degraded with RNAse H (Thermofisher Scientific). All primers used were 'in-house' primers designed using the multiple sequence alignment of the patient's consensus NGS sequences. Partial Spike (amino acids 21- 800) was amplified as 1 continuous length of DNA (Spike ~ 1.8 kb) by nested PCR. Terminally diluted cDNA was PCR- amplified using Platinum® Taq DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA) so that 30% of reactions were positive[27]. By Poisson statistics, sequences were deemed ≥80% likely to be derived from HIV-1 single genomes. We obtained between 20–60 single genomes at each sample time point to achieve 90% confidence of detecting variants present at ≥8% of the viral population in vivo[28,29]. Partial spike amplicons obtained from terminal dilution PCR amplification were Sanger sequenced to form a contiguous sequence using another set of 8 in-house primers. Sanger sequencing was provided by Genewiz UK and manual sequence editing was performed using DNA Dynamo software (Blue Tractor Software Ltd, UK).

Phylogenetic Analysis

All available full-genome SARS-CoV-2 sequences were downloaded from the GISAID database (http://gisaid.org/)[30] on 16th December. Duplicate and low-quality sequences (>5% N regions) were removed, leaving a dataset of 212,297 sequences with a length of >29,000bp. All sequences were sorted by name and only sequences sequenced with United Kingdom / England identifiers were retained. From this dataset, sequences were de-duplicated and where background sequences were required in figures, randomly subsampled using seqtk (https://github.com/lh3/seqtk). All sequences were aligned to the SARS-CoV-2 reference strain MN908947.3, using MAFFT v7.475 with automatic flavour selection[31]. Major SARS-CoV-2 clade memberships were assigned to all sequences using both the Nextclade server v0.9 (https://clades.nextstrain.org/) and Phylogenetic Assignment Of Named Global Outbreak Lineages (pangolin)[32].

Maximum likelihood phylogenetic trees were produced using the above curated dataset using IQ-TREE v2.1.2[33]. Evolutionary model selection for trees were inferred using ModelFinder[34] and trees were estimated using the GTR+F+I model with 1000 ultrafast bootstrap replicates[35]. All trees were visualised with Figtree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/), rooted on the SARS-CoV-2 reference sequence and nodes arranged in descending order. Nodes with bootstraps values of <50 were collapsed using an in-house script.

In-depth allele frequency variant calling

The SAMFIRE package version 1.06[36] was used to call allele frequency trajectories from BAM file data. Reads were included in this analysis if they had a median PHRED score of at least 30, trimming the ends of reads to achieve this if necessary. Nucleotides were then filtered to have a PHRED score of at least 30; reads with fewer than 30 such reads were discarded. Distances between sequences, accounting for low-frequency variant information, was also conducted using SAMFIRE. The sequence distance metric, described in an earlier paper[37], combines allele frequencies across the whole genome. Where L is the length of the genome, we define q(t) as a 4 x L element vector describing the frequencies of each of the nucleotides A, C, G, and T at each locus in the viral genome sampled at time t. For any given locus i in the genome we calculate the change in allele frequencies between the times t_1 and t_2 via a generalisation of the Hamming distance

$$d(q_i(t_1), q_i(t_2)) = \frac{1}{2} \sum_{a \in \{A,C,G,T\}} |q_i^a(t_1) - q_i^a(t_2)|$$

where the vertical lines indicate the absolute value of the difference. These statistics were then combined across the genome to generate the pairwise sequence distance metric

$$D(\boldsymbol{q}(t_1), \boldsymbol{q}(t_2)) = \sum_i d(q_i(t_1), q_i(t_2))$$

The Mathematica software package was to conduct a regression analysis of pairwise sequence distances against time, leading to an estimate of a mean rate of within-host sequence evolution. In contrast to the phylogenetic analysis, this approach assumed the samples collected on days 93 and 95 to arise via stochastic emission from a spatially separated subpopulation within the host, leading to a lower inferred rate of viral evolution for the bulk of the viral population.

All variants were indecently validated using custom code as part of the AnCovMulti package, found at https://github.com/PollockLaboratory/AnCovMulti.

Western blot analysis

Forty-eight hours after transfection of cells with plasmid preparations, the culture supernatant was harvested and passed through a 0.45-μm-pore-size filter to remove cellular debris. The filtrate was centrifuged at 15,000 rpm for 120 min to pellet virions. The pelleted virions were lysed in Laemmli reducing buffer (1 M Tris-HCl [pH 6.8], SDS, 100% glycerol, β-mercaptoethanol, and bromophenol blue). Pelleted virions were subjected to electrophoresis on SDS–4 to 12% bis-Tris protein gels (Thermo Fisher Scientific) under reducing conditions. This was followed by electroblotting onto polyvinylidene difluoride (PVDF) membranes. The SARS-CoV-2 Spike proteins were visualized by a ChemiDoc® MP imaging system (Biorad) using anti-Spike S2 (Invitrogen at 1:1000 dilution) and anti-p24 Gag antibodies (NIH AIDS Reagents 1:1000 dilution).

Recombination Detection

All sequences were tested for potential recombination, as this would impact on evolutionary estimates. Potential recombination events were explored with nine algorithms (RDP, MaxChi, SisScan, GeneConv, Bootscan, PhylPro, Chimera, LARD and 3SEQ), implemented in RDP5 with default settings[38]. To corroborate any findings, ClonalFrameML v1.12[39] was also used to infer recombination breakpoints. Neither programs indicated evidence of recombination in our data.

Structural Viewing

The Pymol Molecular Graphics System v2.4.0 (https://github.com/schrodinger/pymol-open-source/releases) was used to map the location of the four spike mutations of interested onto a SARS-CoV-2 spike structure visualised by Wrobel et al (PDB: 6ZGE)[40].

Testing of convalescent plasma for antibody titres

The Anti-SARS-CoV-2 ELISA (IgG) assay used to test CP for *antibody titres* was Euroimmun Medizinische Labordiagnostika AG. This indirect ELISA based assay uses a recombinant structural spike 1 (S1) protein of SARS-CoV-2 expressed in the human cell line HEK 293 for the detection of SARS-CoV2 IgG.

Generation of Spike mutants

Amino acid substitutions were introduced into the D614G pCDNA_SARS-CoV-2_Spike plasmid as previously described[41] using the QuikChange Lightening Site-Directed Mutagenesis kit, following the manufacturer's instructions (Agilent Technologies, Inc., Santa Clara, CA).

Pseudotype virus preparation

Viral vectors were prepared by transfection of 293T cells by using Fugene HD transfection reagent (Promega). 293T cells were transfected with a mixture of 11ul of Fugene HD, 1μg of pCDNAΔ19Spike-HA, 1ug of p8.91 HIV-1 gag-pol expression vector[42,43], and 1.5μg of pCSFLW (expressing the firefly luciferase reporter gene with the HIV-1 packaging signal). Viral supernatant was collected at 48 and 72h after transfection, filtered through 0.45um filter and stored at -80˚C. The 50% tissue culture infectious dose ($TCID_{50}$) of SARS-CoV-2 pseudovirus was determined using Steady-Glo Luciferase assay system (Promega).

Standardisation of virus input by SYBR Green-based product-enhanced PCR assay (SG-PERT)

The reverse transcriptase activity of virus preparations was determined by qPCR using a SYBR Green-based product-enhanced PCR assay (SG-PERT) as previously described[44]. Briefly, 10-fold dilutions of virus supernatant were lysed in a 1:1 ratio in a 2x lysis solution (made up of 40% glycerol v/v 0.25% Trition X-100 v/v 100mM KCl, RNase inhibitor 0.8 U/ml, TrisHCL 100mM, buffered to pH7.4) for 10 minutes at room temperature.

12µl of each sample lysate was added to thirteen 13µl of a SYBR Green master mix (containing 0.5µM of MS2-RNA Fwd and Rev primers, 3.5pmol/ml of MS2-RNA, and 0.125U/µl of Ribolock RNAse inhibitor and cycled in a QuantStudio. Relative amounts of reverse transcriptase activity were determined as the rate of transcription of bacteriophage MS2 RNA, with absolute RT activity calculated by comparing the relative amounts of RT to an RT standard of known activity.

Serum/plasma pseudotype neutralization assay

Spike pseudotype assays have been shown to have similar characteristics as neutralisation testing using fully infectious wild type SARS-CoV-2[8].Virus neutralisation assays were performed on 293T cell transiently transfected with ACE2 and TMPRSS2 using SARS-CoV-2 Spike pseudotyped virus expressing luciferase[45]. Pseudotyped virus was incubated with serial dilution of heat inactivated human serum samples or convalescent plasma in duplicate for 1h at 37˚C. Virus and cell only controls were also included. Then, freshly trypsinized 293T ACE2/TMPRSS2 expressing cells were added to each well. Following 48h incubation in a 5% CO_2 environment at 37 °C, the luminescence was measured using Steady-Glo Luciferase assay system (Promega).

mAb pseudotype neutralisation assay

Virus neutralisation assays were performed on HeLa cells stably expressing ACE2 and using SARS-CoV-2 Spike pseudotyped virus expressing luciferase as previously described[46]. Pseudotyped virus was incubated with serial dilution of purified mAbs[9] in duplicate for 1h at 37˚C. Then, freshly trypsinized HeLa ACE2- expressing cells were added to each well. Following 48h incubation in a 5% CO_2 environment at 37 °C, the luminescence was measured using Bright-Glo Luciferase assay system

(Promega) and neutralization calculated relative to virus only controls. IC50 values were calculated in GraphPad Prism.

Ethics

The study was approved by the East of England – Cambridge Central Research Ethics Committee (17/EE/0025). Written informed consent was obtained from both the patient and family. Additional controls with COVID-19 were enrolled to the NIHR BioResource Centre Cambridge under ethics review board (17/EE/0025).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Long-read sequencing data that support the findings of this study have been deposited in the NCBI SRA database with the accession codes SAMN16976824 - SAMN16976846 under BioProject PRJNA682013. Short reads and data used to construct figures were deposited at https://github.com/Steven-Kemp/sequence_files. All data are also available from the corresponding author. Source data are provided with this paper.

Code availability

The SAMFIRE package Version 1.06 was used for filtering and calling variants from the Illumina data. It is available at https://github.com/cjri/samfire/ for review. Additional code was used to validate the variant frequencies and can be found at https://github.com/PollockLaboratory/AnCovMulti .

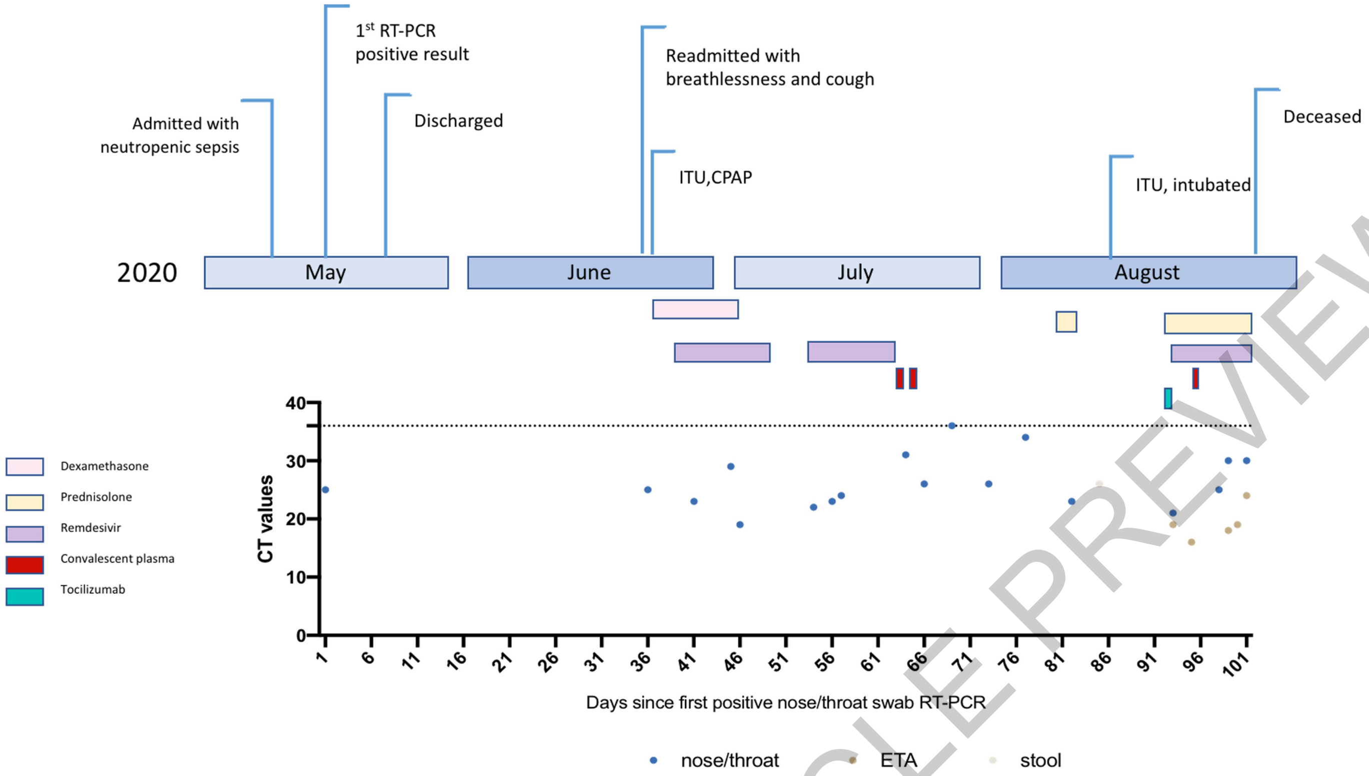
22. Meredith, L. W. *et al*. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *The Lancet Infectious Diseases* **20**, 1263-1272, https://doi.org/10.1016/S1473-3099(20)30562-4 (2020).
23. Collier, D. A. *et al*. Point of Care Nucleic Acid Testing for SARS-CoV-2 in Hospitalized Patients: A Clinical Validation Trial and Implementation Study. *Cell Rep Med*, 100062, https://doi.org/10.1016/j.xcrm.2020.100062 (2020).
24. Loman, N., Rowe, W. & Rambaut, A. (v1, 2020).
25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094-3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
27. Jordan, M. R. *et al*. Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations. *J Virol Methods* **168**, 114-120, https://doi.org/10.1016/j.jviromet.2010.04.030 (2010).
28. Palmer, S. *et al*. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *Journal of clinical microbiology* **43**, 406-413, https://doi.org/10.1128/JCM.43.1.406-413.2005 (2005).
29. Keele, B. F. *et al*. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 7552-7557 (2008).
30. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **22**, 30494, https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).
31. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772-780, https://doi.org/10.1093/molbev/mst010 (2013).
32. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* **5**, 1403-1407, https://doi.org/10.1038/s41564-020-0770-5 (2020).
33. Minh, B. Q. *et al*. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534, https://doi.org/10.1093/molbev/msaa015 (2020).
34. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587-589, https://doi.org/10.1038/nmeth.4285 (2017).
35. Minh, B. Q., Nguyen, M. A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188-1195, https://doi.org/10.1093/molbev/mst024 (2013).
36. Illingworth, C. J. SAMFIRE: multi-locus variant calling for time-resolved sequence data. *Bioinformatics* **32**, 2208-2209, https://doi.org/10.1093/bioinformatics/btw205 (2016).
37. Lumby, C. K., Zhao, L., Breuer, J. & Illingworth, C. J. A large effective population size for established within-host influenza virus infection. *Elife* **9**, https://doi.org/10.7554/eLife.56915 (2020).
38. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution* **1** (2015).
39. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* **11**, e1004041 (2015).
40. Wrobel, A. G. *et al*. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nature Structural & Molecular Biology* **27**, 763-767, https://doi.org/10.1038/s41594-020-0468-7 (2020).
41. Gregson, J. *et al*. HIV-1 viral load is elevated in individuals with reverse transcriptase mutation M184V/I during virological failure of first line antiretroviral therapy and is associated with compensatory mutation L74I. *Journal of Infectious Diseases* (2019).
42. Naldini, L., Blomer, U., Gage, F. H., Trono, D. & Verma, I. M. Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. *Proc Natl Acad Sci U S A* **93**, 11382-11388, https://doi.org/10.1073/pnas.93.21.11382 (1996).
43. Gupta, R. K. *et al*. Full length HIV-1 gag determines protease inhibitor susceptibility within in vitro assays. *AIDS* **24**, 1651 (2010).
44. Vermeire, J. *et al*. Quantification of reverse transcriptase activity by real-time PCR as a fast and accurate method for titration of HIV, lenti- and retroviral vectors. *PloS one* **7**, e50859-e50859, https://doi.org/10.1371/journal.pone.0050859 (2012).
45. Mlcochova, P. *et al*. Combined point of care nucleic acid and antibody testing for SARS-CoV-2 following emergence of D614G Spike Variant. *Cell Rep Med*, 100099, https://doi.org/10.1016/j.xcrm.2020.100099 (2020).
46. Seow, J. *et al*. Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nat Microbiol* **5**, 1598-1607, https://doi.org/10.1038/s41564-020-00813-8 (2020).

Acknowledgements We are immensely grateful to the patient and his family. We would also like to thank the staff at CUH and the NIHR Cambridge Clinical Research Facility. We would like to thank Dr Ruthiran Kugathasan and Professor Wendy Barclay for helpful discussions and Dr Martin Curran, Dr William Hamilton, and Dr. Dominic Sparkes. We would like to thank Prof Andres Floto and Prof Ferdia Gallagher. We thank Dr James Voss for the kind gift of HeLa cells stably expressing ACE2. We would like to thank James Nathan for RBD protein and Leo James for N protein. 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. RKG is supported by a Wellcome Trust Senior Fellowship in Clinical Science (WT108082AIA). LEM is supported by a Medical Research Council Career Development Award (MR/R008698/1). SAK is supported by the Bill and Melinda Gates Foundation via PANGEA grant: OPP1175094. DAC is supported by a Wellcome Trust Clinical PhD Research Fellowship. CJRI acknowledges MRC funding (ref: MC_UU_00002/11). This research was supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, the Cambridge Clinical Trials Unit (CCTU) and by the UCL Coronavirus Response Fund and made possible through generous donations from UCL's supporters, alumni, and friends (LEM). JAGB is supported by the Medical Research Council (MC_UP_1201/16). IG is a Wellcome Senior Fellow and supported by the Wellcome Trust (207498/Z/17/Z). DDP is supported by NIH GM083127.

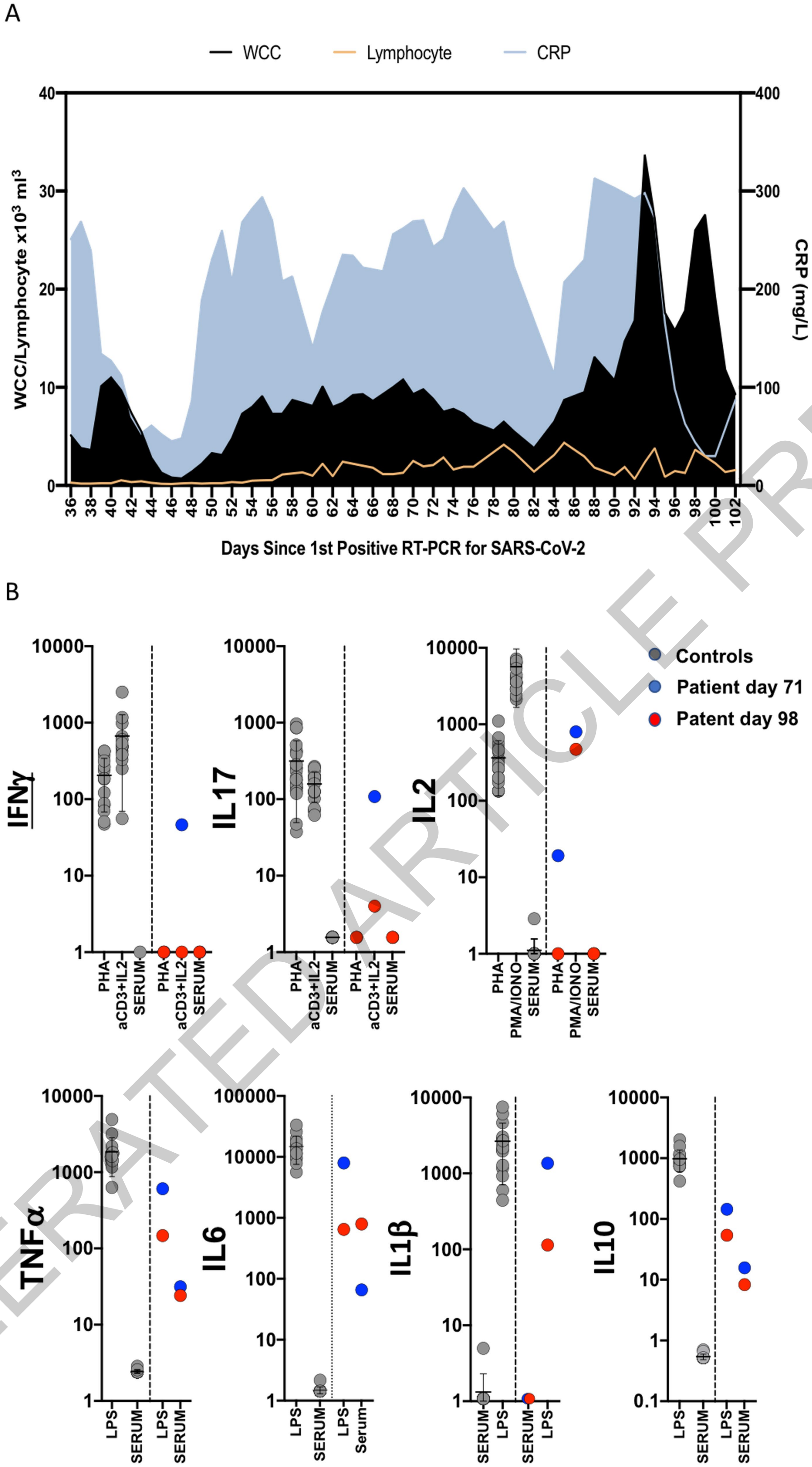
Author contributions Conceived study: RKG, SAK, DAC, AS, TG, EGK. Designed experiments: RKG, SAK, DAC, LEM, JAGB, EGK, AC, NT, AC, CS, RD, RG, DDP, YM. Performed experiments: SAK, DAC, LEM, RD, CRS, AJ, IATMF, KS, TG, CJRI, BB, JS, MJvG, LGC, GBM, LK. Interpreted data: RKG, SAK, DAC, PM, LEM, JAGB, PM, SG, KS, TG, JB, KGCS, IG, CJRI, JAGB, IUL, DR, JS, BB, RAG. DDP, RD, LCG, GBM.

Competing interests The authors declare no competing interests.

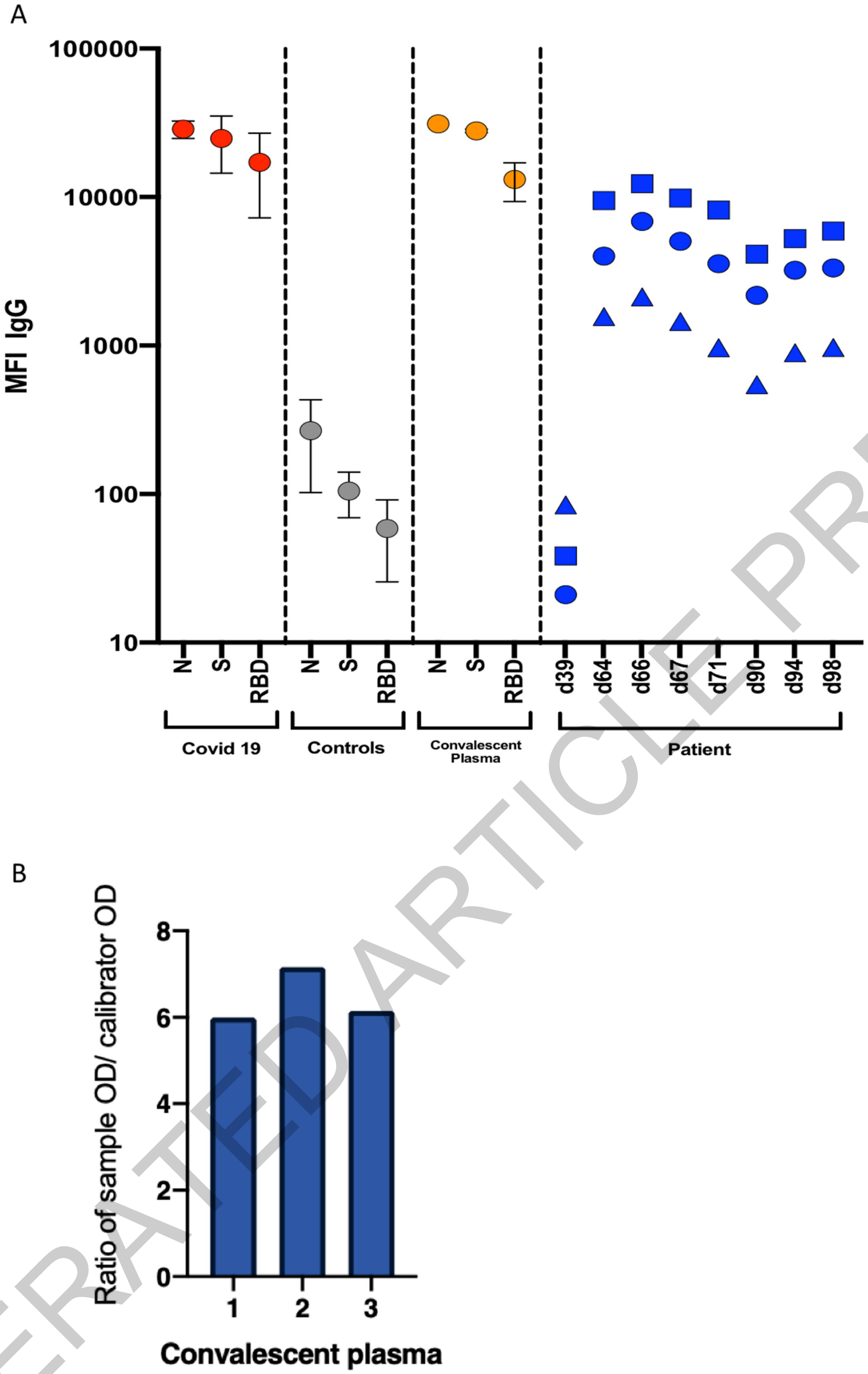
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03291-y.
Correspondence and requests for materials should be addressed to R.K.G.
Peer review information *Nature* thanks Richard Neher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



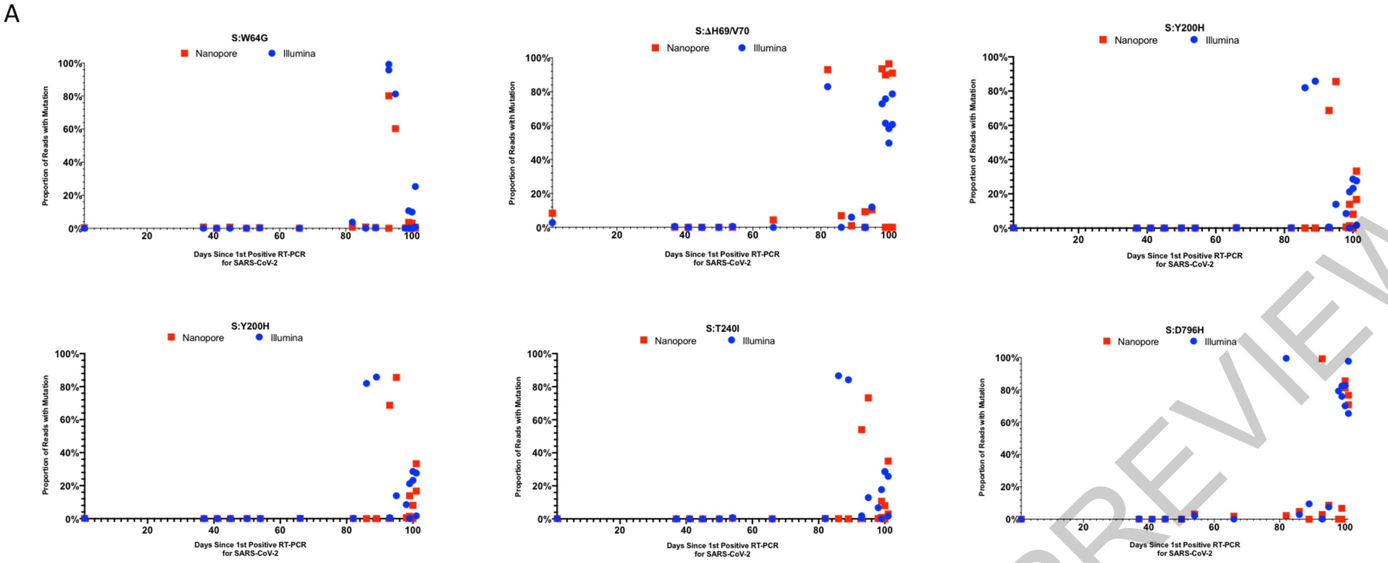
Extended Data Fig. 1 | Clinical time line of events with longitudinal respiratory sample CT values. CT – cycle threshold.



Extended Data Fig. 2 | A. Blood parameters over time in patient case. : White cell count (WCC) and lymphocyte counts are expressed as x10^3 Cells/mm^3. CRP: C reactive protein. **B. Assessment of T cell and innate function**. Whole blood cytokines were measured in whole blood after 24 hours stimulation either after T-cell stimulation with PHA or anti CD3/IL2 or innate stimulation with LPS. Healthy controls are shown as grey circles (N=15), Patient at d71 and d98 is shown as blue circles or red circles respectively. Cytokine levels are shown as pg/ml stimulation. Mean is shown by line and whiskers representing standard deviation.



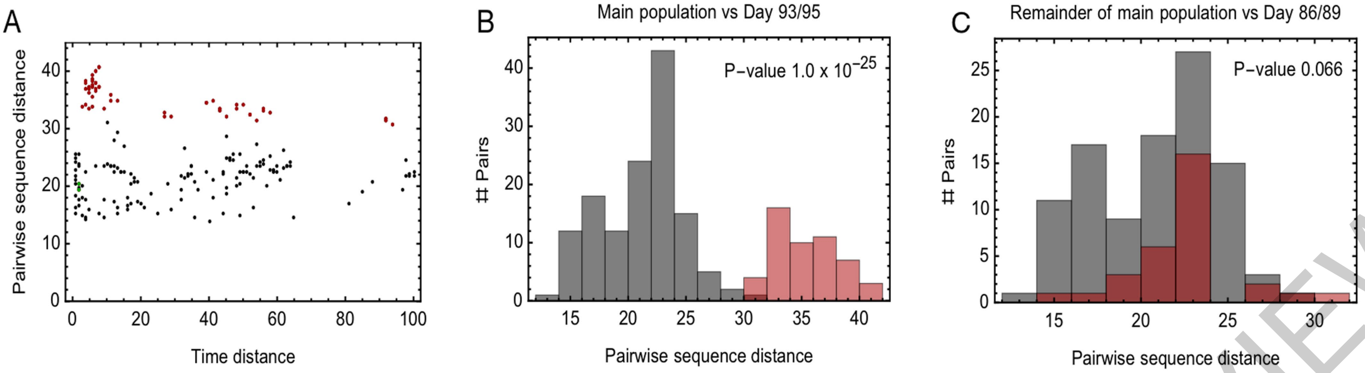
Extended Data Fig. 3 | A. Serum SARS-CoV-2 antibody levels and virus population changes in chronic SARS-CoV-2 infection. Anti SARS-CoV2 IgG antibodies in patient and pre/post convalescent plasma compared to RNA+ Covid19 patients and prepandemic healthy controls: Red, grey and gold. : IgG antibodies to SARS-CoV2 nucleocapsid protein (N), trimeric S protein (S) and the receptor binding domain (RBD) were measured by multiplexed particle based flow cytometry (Luminex) in RNA+ COVID-19 patients (N=20, red dots), Pre-pandemic healthy controls (N=20, grey dots) and in the convalescent donor plasma (orange dots); Results are shown as mean fluorescent intensity (MFI) +/- SD. **Patient sera over time in blue:** Anti SARS-CoV2 IgG to N (blue squares), S (blue circles) and RBD (blue triangles). Timing of CP units is also shown. **B. SARS-CoV-2 antibody titres in patient and in convalescent plasma**. Measurement of SARS-CoV-2 specific IgG antibody titres in three units of convalescent plasma (CP) by Euroimmun assay.



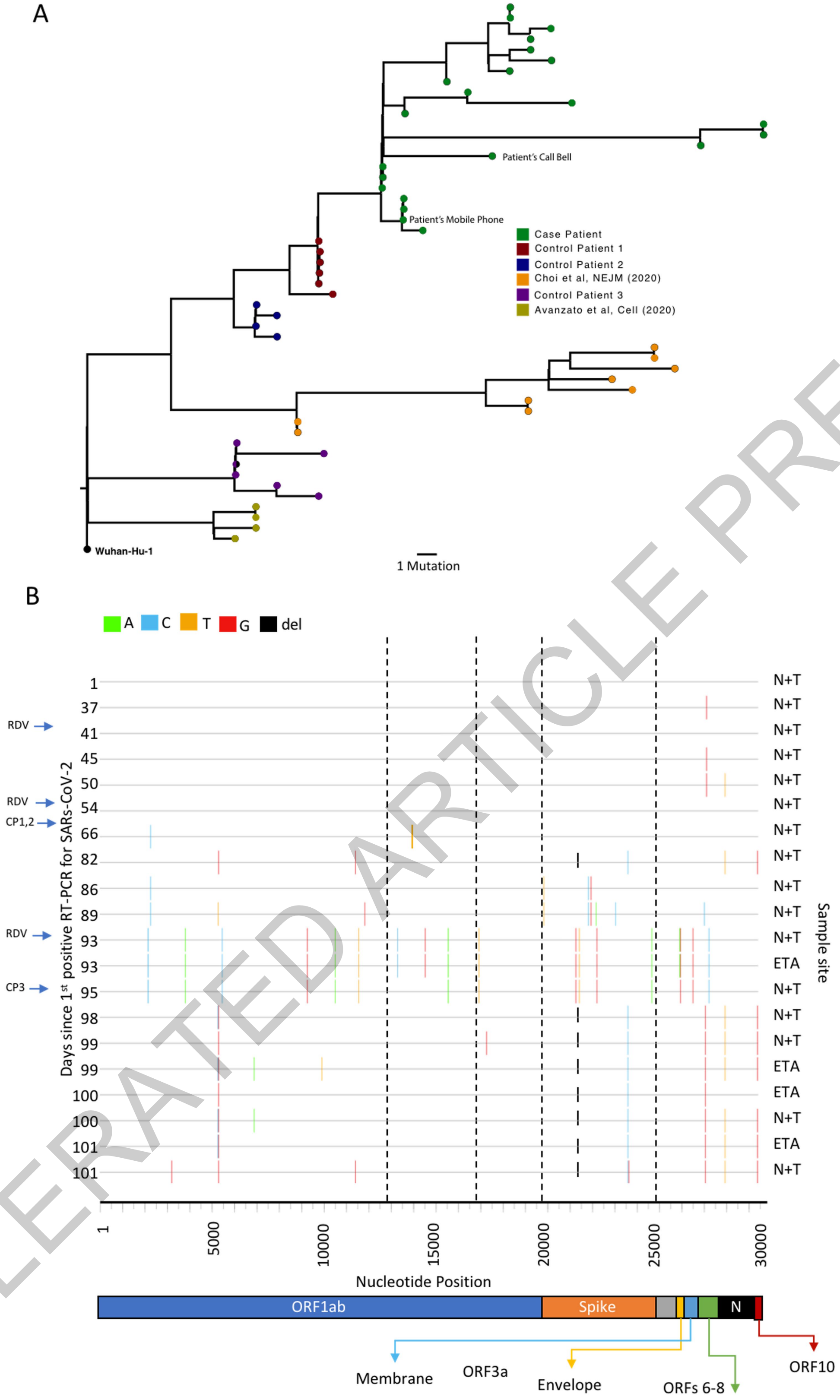
B

	W64G	P330S	ΔH69/V70	D796H	T200I	Y240H
Day 1 (n=7)	0	0	0	0	0	0
Day 37 (n=38)	0	0	0	0	0	0
Day 98 (n=21)	1 (4.8)	1 (4.8)	17 (81.0)	13*(68.4)	3 (14.3)	3 (14.3)

Extended Data Fig. 4 | Comparison between short-read (Illumina) and long-read single molecule (Oxford Nanopore) sequencing methods for the six observed Spike mutations. Concordance was generally good between the majority of timepoints, however due to large discrepancies in a number of timepoints, we suggest that due to the high base calling error rate, Nanopore is not yet suitable for calling minority variants. As such, all figures in the main paper were produced using Illumina data only. **B. Single genome sequencing (SGS) data from respiratory samples at indicated days.** Indicated are the number of single genomes obtained at each time point with the mutations of interest (identified by deep sequencing). *denominator is 19 as for 2 samples the primer reads were poor quality at amino acid 796 at day 98. Amino acid variant and corresponding nucleotide position: S:W64G = 21752, S: Δ69 = 21765-21770, S:Y200H = 22160, S:T240I = 22281, S:P330S = 22550, S:D795H = 23948.

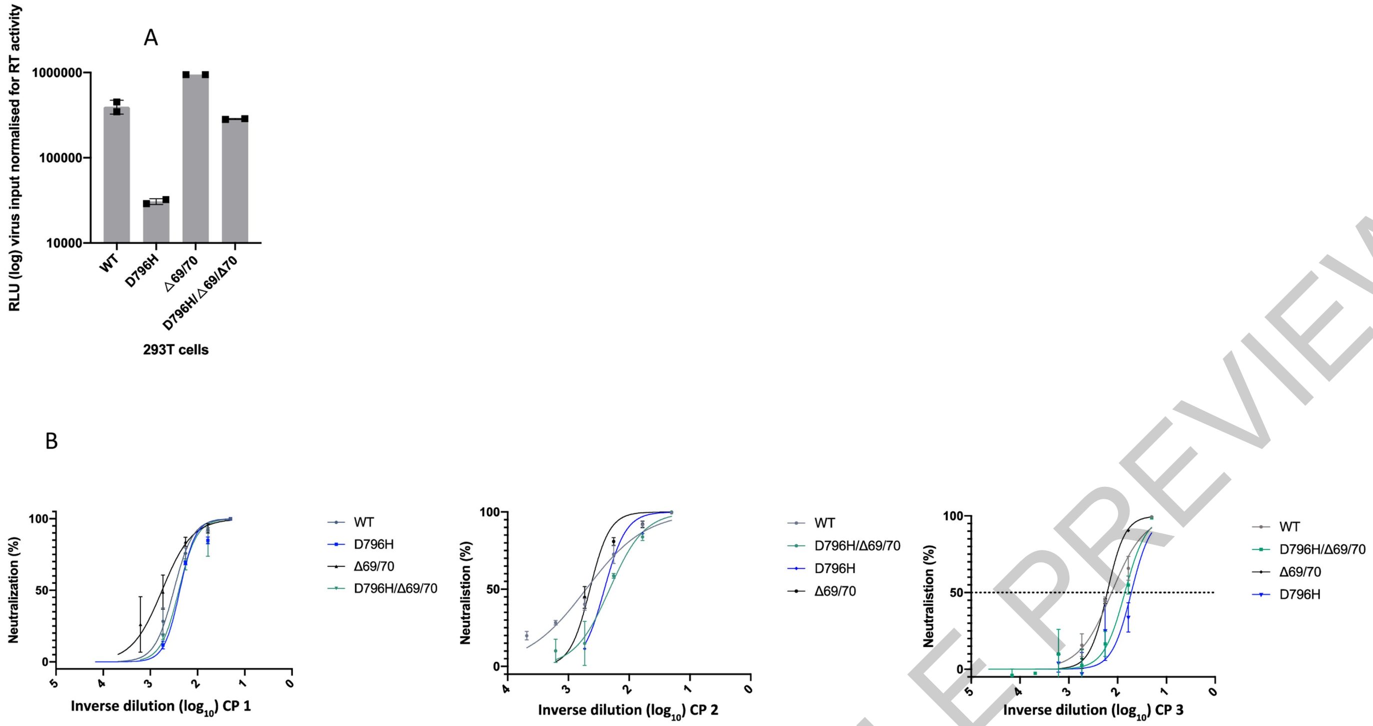


Extended Data Fig. 5 | Evidence for within-host cladal structure. A. Pairwise distances between samples measured using the all-locus distance metric plotted against pairwise distances in time (measured in days) between samples being collected. Internal distances between samples in the proposed main clade are shown in black, distances between samples in the main clade and samples collected on days 93 and 95 are shown in red, and internal distances between samples collected on days 93 and 95 are shown in green. **B**. Pairwise distances between samples in the larger clade (black) and between these samples and those collected on days 93 and 95 (red). The median values of the distributions of these values are significantly different according to a Mann Whitney test. **C**. Pairwise distances between samples in the main clade, once those collected on days 86, 89, 93, 95 have been removed (black) and between these samples and those collected on days 86 and 89 (red). The median values of the distributions of these values are not significantly different at the 5 level according to a Mann Whitney test.



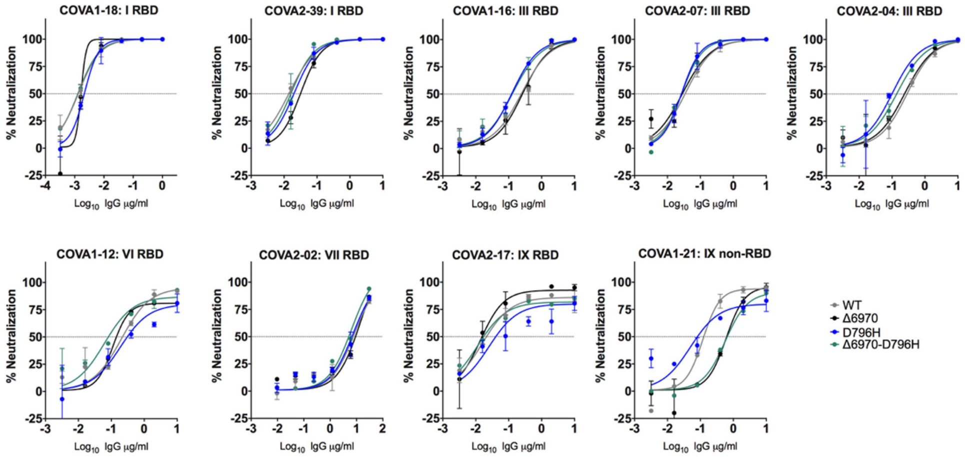
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | A. Close-view maximum-likelihood phylogenetic tree. indicating the diversity of the case patient and three other long-term shedders from the local area (red, blue and purple), compared to recently published sequences from Choi et al (orange) and Avanzato et al (gold). Control patients generally showed limited diversity temporally, though the Choi et al sequences were highly divergent. Environmental samples (patient's call bell, and patient's mobile phone) are indicated. Tree branched have been collapsed where bootstrap support was <60. **B. Highlighter plot indicating nucleotide changes at consensus level in sequential respiratory samples compared to the consensus sequence at first diagnosis of COVID-19**. Each row indicates the timepoint the sample was collected (number of days from first positive SARS-CoV-2 RT-PCR). Black dashed lines indicate the RNA-dependent RNA polymerase (RdRp) and Spike regions of the genome. There were few nucleotide substitutions between days 1-54, despite the patient receiving two courses of remdesivir. The first major changes in the spike genome occurred on day 82, following convalescent plasma given on days 63 and 65. The amino acid deletion in S1, ΔH69/V70 is indicated by the black lines. Sites: Endotracheal aspirate (ETA) or Nose/throat swabs (N+T).



Extended Data Fig. 7 | In vitro infectivity and neutralisation sensitivity of Spike pseudotyped lentiviruses. A. infection of target 293T cells expressing TMPRSS2 and ACE2 receptors using equal amounts of virus as determined by reverse transcriptase activity. Data points represent technical replicates (n=2), with mean shown with error bars representing standard deviation. Data are representative of n=2 independent experiments (n=2). **B**. Representative Inverse dilution plots for Spike variants against convalescent plasma units 1-3. Data points represent mean neutralisation of technical replicates and error bars represent standard error of the mean of replicates. Data are representative of two independent experiments (n=2).

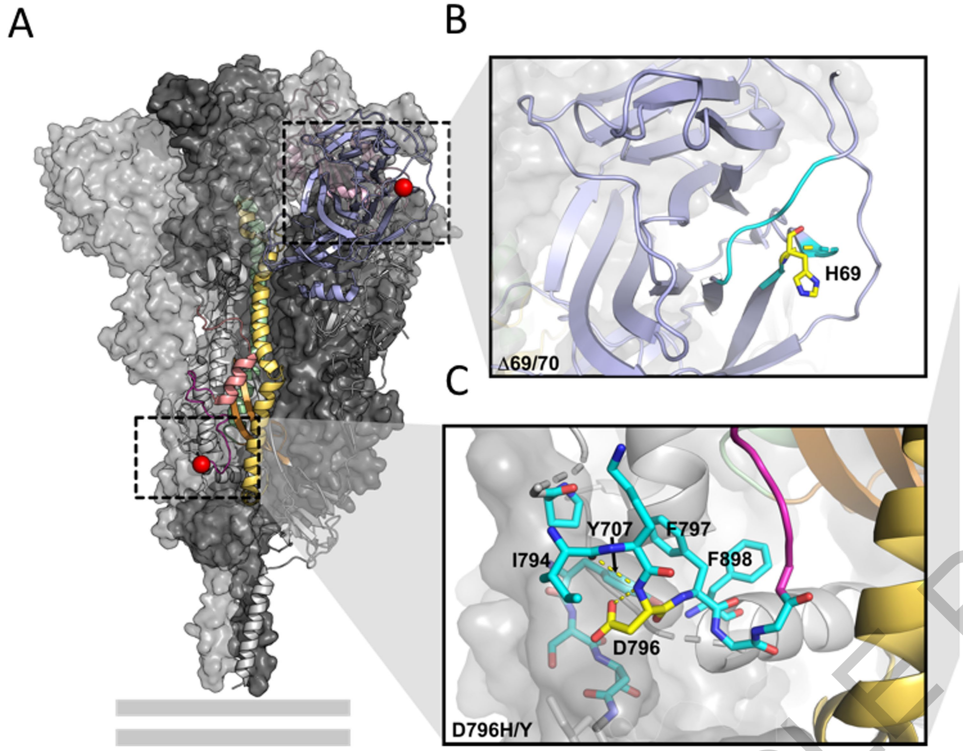
A



B

mAb	Cluster	Target		WT	D796H	Δ6970	Δ6970-D796H	Fold decrease		
COVA1-18	I	RBD		0.0014	0.0022	0.0016	0.0013	1.6	1.2	1.0
COVA2-39	I	RBD	Structure	0.0143	0.0203	0.0319	0.0163	1.4	2.2	1.1
COVA1-16	III	RBD	Structure	0.2441	0.1242	0.2651	0.1308	0.5	1.1	0.5
COVA2-07	III	RBD		0.0349	0.0269	0.0288	0.0272	0.8	0.8	0.8
COVA2-04	III	RBD	Structure	0.2887	0.1009	0.2425	0.1401	0.3	0.8	0.5
COVA2-17	IX	RBD		0.0156	0.0248	0.0139	0.0113	1.6	0.9	0.7
COVA2-02	VII	RBD		5.8590	4.9670	6.5680	3.5380	0.8	1.1	0.6
COVA1-12	VI	RBD		0.2007	0.1863	0.1105	0.0611	0.9	0.6	0.3
COVA1-21	XI	Non-RBD		0.1189	0.0498	0.6035	0.5682	0.4	5.1	4.8

Extended Data Fig. 8 | A. Neutralization potency of a panel of monoclonal antibodies targeting the RBD is not impacted by Spike mutations D796H or ΔH69/V70. Lentivirus pseudotyped with SARS-CoV-2 Spike protein: WT (D614G background), D796H, ΔH69/V70, D796H+ΔH69/V70 were produced in 293T cells and used to infect target Hela cells stably expressing ACE2 in the presence of serial dilutions of indicated monoclonal antibodies. Data are means of technical replicates with error bars representing SD. Data are representative of at least two independent experiments. RBD: receptor binding domain. **B. Classes of RBD binding antibodies and fold changes for Spike mutations D796H or ΔH69/V70** are indicated based Bouwer et al. Clusters II, V contain only non-neutralising mAbs, smaller neutralising mAb clusters IV (n=2) and X (n=1) were not tested. Red indicates significant fold changes.



D

Mutation	Number of Sequences	Global Prevalence (%)
W64G	0	0.00
ΔH69/V70	12883	4.32
Y200H	7	<0.01
T240I	77	0.02
P330S	167	0.06
D796H	65	0.02
D796Y	141	0.05

Extended Data Fig. 9 | Location of Spike mutations ΔH69/Y70 and D796H. A. The SARS-CoV-2 spike trimer (PDB ID: 6xr8) with two protomers represented as surfaces and one protomer represented as a ribbon. The NTD is coloured in light blue, the RBD in light pink, the fusion peptide in dark pink, the HR1 domain in yellow, the CH domain in pale green, and the CD domain in brown. The location of D796 and H69 are indicated by red spheres. The loop connecting D796 to the fusion peptide is coloured magenta to improve visibility. The double grey lines provide orientation relative to the membrane. **B**. A close-up of the region defined by the box around H69 in panel A. H69 is highlighted in yellow. Residues containing atoms that are within 6 Å of H69 are highlighted in cyan. **C**. A close-up of the region defined by the box around D796 in panel A. D796 is highlighted in yellow. Residues containing atoms that are within 6 Å of D796 are highlighted in cyan. Hydrogen bonds are indicated by dashed yellow lines. Hydrophobic residues in the vicinity of D796 have been labelled. Y707 is from the neighbouring protomer. **D. Global prevalence of selected spike mutations detailed in this paper**. All high coverage sequences were downloaded from the GISAID database on 6th January and aligned using MAFFT; as of this date there were 298254 sequences available. The global prevalence of each of the six spike mutations W64G, ΔH69/V70, Y200H, T240I, P330S and D796H were assessed by viewing the multiple sequence alignment in AliView, sorting by the column of interest, and counting the number of mutations.

nature research

Corresponding author(s): Ravindra K Gupta (rkg20@cam.ac.uk)

Last updated by author(s): Jan 4, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Graphad Prism v8 and v9 were used to produce figures. Mafft v7.475 was used for multiple sequence alignments. IQTREE and ModelFinder v2.1.2 was used to infer maximum-likelihood phylogenies. Figtree v1.4.4 was used to annotate and manipulate phylogeny trees. NextClade server v0.9 and Pangolin v2.12 were used to assign lineages to sequences.
Data analysis	Software versions and parameters used for sequence conservation analysis are reported in methods. For in-depth variant analyses, the SAMFIRE package v1.06 was used (https://github.com/cjri/samfire/). To validate frequency variants, custom code was used as part of the package AnCovMutlti (github.com/PollockLaboratory/AnCovMulti).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequences fro SARS-CoV-2 were obtained from GISAID database (https://gisaid.org/) using the filters and search parameters defined in the methods section.

Structural models were obtained from the Protein Data Bank (PDB) https://www.rcsb.org/.
Long-read sequencing data that support the findings of this study have been deposited in the NCBI SRA database with the accession codes SAMN16976824 - SAMN16976846 under BioProject PRJNA682013 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA682013). Short-read and consensus fasta files have been deposited and are available to view and download on GitHub (https://github.com/Steven-Kemp/sequence_files). Raw data used to create figures are available to view and download on GitHub (https://github.com/Steven-Kemp/sequence_files/tree/main/figure_data).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculation was undertaken as this case study report focused only on a single patient. By definition n=1
Data exclusions	No data were excluded.
Replication	All experimental data were reproducible and we have shown data representative of at least two independent experiments
Randomization	not applicable as this is a descriptive study of one patient
Blinding	not applicable as this is a descriptive study of one patient and blinding was not appropriate

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Panel of non-commercially available antibodies fully described in Science: DOI: 10.1126/science.abc5902. were provided by Brouwer et al at request. All antibiodies were diluted to a starting concentration of 10 ug/ml and then titrated 5-fold, except for COVA1-18 which was at a starting concentration of 1 ug/ml and COVA2-02 at a starting assay concentration of 30 ug/ml due to their different neutralization potency. Spike S2 antibody (Invitrogen,, Cat no: Pa1-41165), p24 antibody (NIH AIDS Reagents cat no: ARP465). Anti-CD3 antibody (MEM57, Abcam, 200 ng/ml, 1:1000, Cat no Ab8090)
Validation	Validated by providing lab: see supplementary: https://science.sciencemag.org/content/sci/suppl/2020/06/15/science.abc5902.DC1/abc5902-Brouwer-SM.pdf

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa cells were donated by kind request from James Voss as noted in the acknowledgments section. HEK 293T cells from ATCC were used for transfection work.
Authentication	None of the cell lines used were authenticated.

Mycoplasma contamination	All cell lines used were tested a(by PCR) and were mycoplasma free.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	A single septuagenarian male was treated at a local Cambridge hospital for COVID-19 symptoms.
Recruitment	As part of routine testing, nose + throat samples and endotracheal aspirates were collected from the patient at 23 time-points.
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.