nature research

Peer Review File

Manuscript Title: SARS-CoV-2 evolution during treatment of chronic infection with convalescent plasma

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

In this work, Gupta and colleagues characterize the shifts in SARS-CoV-2 virus population within an immune-suppressed individual over 101 days (and 23 time points analyzed), throughout multiple different medical treatments. These data revealed a mutation/deletion combination in the virus population (spike D796H S2 mutation and 69-70 NTD deletion) which correlated with treatment with convalescent plasma (CP). In vitro experiments with pseudovirus support that this SARS-CoV-2 mutation/deletion variant has decreased sensitivity to sera from recovered patients, including the CP administered to this patient. Given that the 69-70 deletion has arisen multiple times globally in sequenced SARS-CoV-2 isolates, and that deletions are unlikely to be reverted, understanding the implications of this deletion for efficacy of polyclonal sera produced by infection or vaccination is of importance. Overall, the correlation of the 69-70 deletion with the CP administration is notable, but the data presented here are not sufficient to clearly demonstrate that the 69-70 deletion confers resistance to polyclonal sera (as well as the reason for the fatal outcome of the infection).

Here are some additional comments:

1. In the Introduction, the authors mention that RNA viruses have inherently higher rates of mutation than DNA viruses, but there is no mention that SARS-CoV-2 has actually a quite modest error rate for an RNA virus (compared to, for example, influenza) and that overall, the evolution of SARS-CoV-2 during the pandemic is relatively slow.

2. More information should be provided regarding the Euroimmun assay used to assess the antibody titers of the CP, e.g. what is the antigen used for this assay and a very brief summary of the method.

3. Supp Fig 4 refers to 10 samples in the legend but there are 14 data points for each experiment.

4. Regarding the three persistent-shedding local individuals used for controls in the phylogenetic analysis: The treatment histories of these patients should also be outlined in brief (e.g. presumably they did not receive CP therapy).

5. Regarding Figure 3:

a. It would be helpful if the blue arrows indicating treatments in panel B were also included in panel A.
b. In panel A the duplicate genome annotation should be merged (there is one above and one below the plot).
c. Please provide a key for all lines in Panel B, and/or provide the underlying data in a supplementary table.

d. In particular, all variants discussed in the main text should be annotated in this figure.
e. The variant frequency scale should be better defined (based on the main text, I assume that 1.0 is 100%?)

6. A better explanation is needed for the results of consensus level sequencing (Fig 3A) as compared to deep sequencing results (Fig 3B). How can T39I reach 77% frequency on Day 44 (text lines 170-171, presumably this is the olive green line in Fig 3B) without being detected by consensus level sequencing?

7. Supp Fig 5 is missing data and the legend for Supp Fig 6 is cut off and not possible to read.

8. At line 175, the authors refer to the "near fixation" of D796H and the deletion mutant. It appears this term is being used to indicate that close to 100% of recovered samples contained this mutation. This is not an appropriate use of the word "fixation" (which implies subsequent persistence) considering that these mutations drop way down in frequency during the following several timepoints.

9. Regarding Fig 4A, presumably the black triangles indicate Ct values. This should be explained in the legend.

10. The experiments to evaluate in vitro fitness of the D796H + 69-70 deletion are not strong. There is no information in the main text, figure legend or Methods indicating how the RT activity and infectivity were performed and quantified. What are the error bars for RT activity? It appears that infectivity is a single experiment. How were the viral titers normalized prior to conducting these experiments? These results are important for evaluating the relative ID50 in the neutralization experiments.

11. The y-axes for Fig. 5C-D are mislabeled, this should be ID50, not IC50

12. The data shown in Fig. 5C-D are uninterpretable without more rigorous evaluation of infectivity between these two pseudoviruses, or performance of neutralization experiments with a monoclonal antibody not expected to be affected by these mutations (such as an RBD-targeting mAb). As is, it is not possible to evaluate whether the enhanced susceptibility of the WT pseudovirus is really due to the lack of the mutation/deletion combination, or to some other property of the pseudovirus preparation. The neutralization curves for these data should be provided as supplementary information.

13. The mAbs used in panel 5E should be better defined: What is known about their epitopes? Are they all RBD targeting mAbs (the most common category of neutralizing mAbs)?

14. Why was a different set of pseudoviruses used for the serum neutralization vs the mAb neutralization experiments? Results would be a lot stronger if experiments were repeated with the same set of pseudoviruses for both. (Which would ideally include both the mutation/deletion combination and the mutation and deletion as separate pseudovirus variants.)

15. To better understand the sera results, would it be possible to perform serology with these samples? To profile their relative response to different regions of spike, especially NTD and RBD.

16. The titer and specificity of antibodies present in the CP used should be investigated. Was the CP sample used particularly enriched for NTD-neutralizing Abs?

17. Did the Authors try to isolate viruses at the different time points? In vitro competition experiments between WT virus and the mutated one may help understanding the level of fitness of the D796H S2 mutation and 69-70 NTD deletion

18. Line 240 states that the 69-70 deletion is "close to the binding site of the polyclonal Abs derived from COV57 plasma." It is not clear where there is data in either of those two citations to support this statement, as the 69-70 deletion is in the spike N-terminal domain, whereas all the antibodies profiled in the two citations are RBD-binding Abs.

19. It would be interesting to speculate in the Discussion about the differences in viral evolution between the case patient and the patients in Avanzato et al and Choi et al, the former exhibiting very little viral divergence despite treatment with CP, and the latter showing much higher divergence (prior to any mAb treatment, and this patient did not receive CP).

Referee #2 (Remarks to the Author):

The authors describe the detailed virological and phylogenetic investigation a COVID-19 infection in a B cell / antibody deficient patient who had been treated with convalescent plasma therapy. This allowed the authors to document the emergence and re-emergence of viruses with spike delH69/delV70 plus D796H as likely escape from polyclonal antibody administered as convalescent plasma therapy. The authors then introduce these mutations together in a spike pseudotyped virus to demonstrate that these two mutations associate a reduction of susceptibility to neutralization with polyclonal immune serum. These findings have important public health implications as one of these mutations (delH69/V70) is not infrequently found among the GSAID database.
General comments:
a) It would be interesting to have an assessment of the likely T cell function in this patient. Was his underlying disease likely to have impaired T cell function? Was there any objective assessment of this? If T cell function was largely intact, it would be interesting to discuss why T cells would not contribute to containment of virus replication?
b) The phenotype of reduced neutralization with polyclonal sera was investigated with the spike Del69/70+D796H mutant. There was no analysis of the effects of each of these mutations (del69/70 vs. D796H) on the phenotype of neutralization. It would be desirable to make spike pseudotypes with each of these mutations to define the mutation critical for the neutralization phenotype. Del69/70 seems for more commonly found than D796H. So the distinction is important, although it appears that the two mutations co-appear and decline in this particular patient.
At different points in the text, the authors give the impression that delH69/V70 is the escape mutation (see line 89-91) but there is no evidence for this?
c) It would be of great interest to investigate the fitness of the del69/70+D796H mutant compared to wild type. The data presented here may suggest that this is a less fit virus that is replaced by wild-type in the absence of selective pressure of the convalescent plasma. This should be discussed. Ideally, reverse genetics reconstitution of this mutant virus will allow experimental investigation of fitness of this mutation vs wild type in vitro and also in vivo.
d) Are attempts being made to culture the mutant viruses? Some of the specimens have reasonable viral loads that may allow virus culture.
e) This antibody neutralization escape mutation/s arose in the immunocompromised patient after convalescent antibody therapy. In the discussion it is important to emphasise that it may not be as likely for the mutation to appear in immune-competent individuals because there is a major distinction between putting antibody selection pressure in this immunocompromised patient who had a well established infection with lots of virus in his respiratory tract when the convalescent serum was used with i. infection / exposure of someone with prior antibody elicited by natural infection of vaccination when the antibody is facing a lower virus diversity at the start, or ii. an immunocompetent individual who has the T cell arm to support the convalescent plasma. It is clear that the authors are well congnizant of this issue, but it is important to be explicit so that the reader also appreciates these issues.

Other points for clarification:
Line 145-149: Three additional patients analysed. Specify that these patients did not have convalescent antibody?
Line 239: delH69/V70 is close to the binding site of the polyclonal antibodies in COV57 plasma. How do you define the binding site of a polyclonal antibody? By definition, there will be multiple antibody specificities in a polyclonal serum.
Line 306-312: "our data suggest caution in use of CP in patients with immune suppression of both T cell and B cell arms." Do you have data to show that this patient was T cell deficient?

Referee #3 (Remarks to the Author):

In the manuscript "Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion" by Kemp et al., the authors describe genomic changes in the SARS-CoV-2 genome in a single, immunocompromised patient over the course of several COVID-19 treatments. They find that two changes in the SARS-CoV-2 spike protein emerge each time the patient is treated with convalescent plasma, and that these changes in the virus diminish serum neutralization. Understanding the selective pressures imposed by different treatment regimes is crucial to selecting appropriate long-term disease treatment. Though the work is interesting and obviously very timely, I have significant concerns with some of the points and discussion.

Major comments:

1. Overall, though the work reads well, I suggest editing the results section to avoid unnecessary medical jargon i.e. the significance and key points of the CRP test (inflammation mark) and respiratory rate (high especially on O2). It would be helpful to describe the patients' history and treatment course in a more straightforward manner. This feeds into Figure 1 - much greater clarity is suggested here. I strongly suggest these values for Ct be included as a supplemental table, along with which were sequenced, from what sample type, and with which technology. This was difficult to trace throughout the paper.
2. The same is true for when treatments were administered and when antibody testing was done. For example, on Line 121 the authors say they measure antibodies "over the course of admission", but then only describe testing on days 44, 50, 68, 90, 101.
3. Line 407: Which variant caller and version were used for this analysis? There is some evidence that the variant caller selected can affect which variants are called.
4. Sequencing data should be made available, consensus _and_ raw data if at all possible, especially given the minor variations discussed in the work.
5. Figure 2A is largely incomprehensible just due to the number of samples plotted. I suggest strongly that subsampling as was done in part in 2B is more useful, especially with regional samples of the virus as well as a reasonable global subsample. As it stands 2A has little utility except to show that different patients in the region/hospital system have virus of significantly different clades, better shown by Figure 2B.
6. In figure 2B phone and call bell look significantly different in number of mutations? The authors assert that they cluster with the respiratory samples - true, but the number of additional mutations acquired seems high, a larger distance than separating control patient 1 and case patient (from initial samples).
7. Figure 2B would benefit from an indication of variants versus time, especially for the case patient, but also for the "persistent shedders" which make up the other local patients. Is there any reason the authors can delineate why the persistent shedders _were_ persistent shedders, and why not chose some patients who had a "normal" course of disease, with expected rise and fall of viral titer?
8. In supplementary figure 4, why are only 10 samples shown? Again - a table of which samples

were taken when and from which days the plots are generated would be helpful. I suggest rather than box and whiskers, the correlation between Illumina and nanopore variant calls is what's helpful here - I would suggest generating a scatter plot per variant for Illlumina versus nanopore called frequency.
9. Figure 4B: I am quite surprised by the long branch lengths separating the sequences from days 93 and 95 from the rest of the patient samples. Using the scale provided, it looks to be ~12-15 mutations away from the other sequences. I suppose this could be explained by rapid mutation / escape following the emergence of the deletion mutations, but then these mutations are abruptly and completely lost again by day 99? I would compare this to the results from Choi et al which the authors plotted in 2B (data recently published from NEJM) with a patient with antiphospholipid syndrome on immunosuppressive drugs (e.g. cyclophosphamide) who also died of a long course (~154 days) of COVID19. It doesn't seem that the same type of behavior occurred in Choi et al with a complete loss _and regain_ of a relatively dominant mutation (in this case the del H69/V70). Though I understand the selection pressure being applied through the use of CP, it still seems surprising that the levels of these variants could change so much, and that if lost, the mutation could be reacquired.
10. It should be pointed out by the authors in the results and discussion that the nature of the ARTIC method makes it very difficult to assign quantitative levels to the levels of "mutation prevalence" i.e. Fig 4A - because of the amount of PCR amplification being employed, and the concerns of jackpotting especially when relatively high Ct samples were employed (i.e. in the 66-81 day range). At the very least technical replicates going _back_ to the RNA source should be employed - it seems that the authors used the same amplicon source for testing of Illumina v. Nanopore technical artifact, but I am more concerned about artifact arising from the RT and PCR.
11. The authors report the deletion at "<10% abundance" at day 93, but from their graph it looks more like 0? (Again, I recommend the data used to generate the graph be included as supplemental table/data to make interpretation clearer). What is the hypothesis here - not clearly delineated by the authors. I suppose they suggest that this mutation, already present but undetectable before CP treatment was then strongly selected for as other virus was largely eliminated by CP. But then _why_ given the data in Figure 5A showing little to no impairment of infectivity by the mutation, is the mutation lost in samples from day 86-95? Why wouldn't _additional_ mutations be acquired instead? And then how likely is it that the mutation is reacquired in the exact same way post-day 95? Isn't a more likely scenario superinfection by the Y200H and T240I mutations while the del H69/V70 lineage remains persistent in the patient? I understand it's difficult to rule this out without _extensive_ local sequencing of virus, but the authors state "the relationship of divergent samples to those at earlier time points rules out the possibility of superinfection" - I disagree. Examining data from local sequencing efforts including work from New York (Gonzalez-Reiche, Science 2020), Baltimore (Thielen, bioRxiv 2020) and Houston (Long et al mBio 2020) to say nothing of the COG-UK efforts suggests virus from different patients in the same area can fall within the same cluster at the level reported in this work,
12. The authors assert from Supplemental Figure 6 that they observe an increase in frequency of the T39I over the first period of the infection - but the fit is not great. It could easily be argued there is a peak in the data around day ~45, with a decline prior to use of CP, perhaps resulting from remdesevir treatment? I would suggest this actually goes to my earlier point that trying to pull quantitative numbers directly from coverage with a method with so many cycles of PCR is not ideal.
13. Perhaps an orthogonal method - qPCR with Taqman probes for the specific SNPs of interest could be used to improve quantitation. This is especially important for the deletion mutation.
14. Further, when mutations were lost and seemingly regained (del H69/V70) was the mutation _exactly the same_ i.e. the exact same loss of those 6 NT? Were there any other mutations in linkage that could be used to determine if it was a reacquisition or the same lineage? A comparison of the consensus, as well as a careful inspection of the aligned nanopore and illumina reads to these regions should be provided.
15. Supp Fig 7 is not particularly convincing. If you have achieved near fixation of the deletion, why would that mutation have disappeared and reappeared exactly without superinfection. Isn't it as likely that you have a superinfection event that wasn't detected from lack of local sampling?

Comparing against global SARS-CoV-2 data is not nearly as convincing as comparing against high local sampling would be.
16. The pseudotyping experiments are nice - showing that the deletion doesn't matter for infection, but "partial immunity" - though I argue >2 fold difference is more than partial immunity. But the authors do not discuss why the different CPs show such a large difference in IC50 within the different CP and between the other sera (S1-S5).
17. Fig 5E - I'm confused about this - basically the authors are arguing that natural CP is preferentially targeting epitopes that they didn't chose in their 7 neutralizing abs. This is particularly confusing as the paper cited identified mabs from patients with SARS-CoV-2? (https://science.sciencemag.org/content/369/6504/643)

Minor points:

1. Sup Fig 1 would benefit from alpha transparency.
2. Sup Fig 3 y-axis should have units or some sort of scaling to give context.
3. Figure 4A - black line with triangles is not annotated?
4. Sup Fig 8 - which received no reference in the text that I could find - is interesting to examine the prevalence of this mutation in GISAID, but the plot would perhaps be more interesting if instead of a radial plot a plot was generated for the rise of these mutations over time?
5. Line 103-104: Does repeat SARS-CoV-2 RT-PCR mean multiple tests on the same day? From the same swab? Please specify the time period and if these are biological or technical replicates.
6. Line 115: It would be helpful if the authors provided a little bit of background on monoclonal antibody therapies such as tocilizumab, mentioned here. Several different mono- and polyclonal antibody treatments were mentioned and discussed throughout in different contexts, and I found it very difficult to keep track of each (and which ones were for COVID-19 and which were administered for other purposes, e.g., rituximab on line 100).
7. The color labels in Figure 3B are difficult to interpret and I suggest points be plotted on the lines to demonstrate the frequency of sampling more clearly.

Referee #4 (Remarks to the Author):

Kemp and colleagues present a case of persistent SARS-CoV-2 infection and analyze the molecular evolution that unfolded within the host in detail. This case is not dissimilar from two recently described cases (Choi et al (10.1056/NEJMc2031364), Avanzato et al (10.1016/j.cell.2020.10.049)). In contrast to these previous cases, Kemp et al investigate within-host evolution using deep sequencing and trace the frequencies of different variants through time. They characterize three diverged variants with different mutations and deletions in the spike protein, some of which reduce neutralization titers of convalescent plasma in a pseudo-typed lentivirus. Overall, the work in this paper is well performed and it provides convincing evidence of in-vivo antibody escape.

I have a number of suggestions to improve the presentation, strengthen the conclusions, and to remove/tone down parts that might be misleading.

* Fig 2A: The radial tree in Figure 2 is rather unhelpful. The labels are hardly readable and distances between samples are very hard to judge from the radial presentation. A rectangular tree indicating major clades and the different within-host samples would be better.

* Fig 2B & 4B: I think the figure would be improved by changing the scale bar to correspond to one or two mutations (currently the scale bar is given in mutations per site and is roughly 6 mutations in 2B, 2 mutations in 4B). Zero-length branches in the ML trees should be collapsed into polytomies. Bootstrap values on SARS-CoV-2 trees are pretty useless. Better to label the branches with (number of) mutations that fall on the branch in a parsimony or ML reconstruction. This has a

one-to-one correspondence to bootstrap values and is more interpretable.

* The purple line in Fig 3B suggests an iSNV at frequency 30% on day one that persists at a frequency around 30% until day 82. This iSNV doesn't seem to be affected by the fixation of other iSNVs at time points 66 or 82 days. Would be good to look into this. It could indicate population structure which would imply parallel evolution. Or it could be an artifact (more likely). Either way, this should be looked at and discussed.

* To understand the rapid shifts in dominating variants better, it would be helpful to include a discussion of their frequencies when they are rare. It makes a difference to the interpretation if the minor variants are present at 10%, 1%, or 0.1%. The reader currently has to piece this together from supplementary table 3 and there are some discrepancies: The S:64G variant seems to be very rare after day 95 (not detected by high coverage Illumina) while the linked S330S is still picked up (at high frequency in low coverage data??). Mutations 200H,240I,258S are missing from supplementary table 3.

* Fig 5 would be more useful on a logscale. Bar charts should be avoided, individual data points need to be shown.

* the description of how evolutionary rates are estimated from within-host data is very short. I would caution against over-interpreting these estimates for two reasons: (i) Phylogenetic estimates are done with consensus sequences and thus ignore minor variation. (ii) rate estimates likely depend a lot on how the within-host variation is rooted and how the root height is constrained. The error of the mean rates (table S2) seems way too small in some cases (1% of the main) calling the entire procedure into question. I would cut this as I don't think this is reliable and it is not central to the paper.

* Similarly, the logistic fit to T39I in ORF7a (Supp Fig 6) is not evidence for selection. I don't see what this figure adds that is not visible in Fig 3. All that Fig S6 shows is that the variant was rare at day one and then bounced around frequency 0.5 between day 30 and 60. There is no reason to fit a logistic and insinuate selection.

* the distances presented in Fig S5 seem rather large (two-fold larger than what I would have guessed from the tree).

* accession numbers for consensus sequences and reads need to be provided.

Minor comments:
* Fig 4: unclear how shaded areas relate to the lines. Same in Fig S1. I have a hard time parsing the intersecting areas.
* line 133: while you are using a long-read sequencer, the amplicons are short. Better call it single-molecule sequencer.
* lines 160-162: I doubt genetic drift has a major effect here.
* caption Fig 3: the "spike genome"?
* Supplementary Fig 8 is completely unreadable**iewer Reports on the Initial Version:**

Author Rebuttals to Initial Comments:

Referees' comments:

Referee #1 (Remarks to the Author):

In this work, Gupta and colleagues characterize the shifts in SARS-CoV-2 virus population within an immune-suppressed individual over 101 days (and 23 time points analyzed), throughout multiple different medical treatments. These data revealed a mutation/deletion combination in the virus population (spike D796H S2 mutation and 69-70 NTD deletion) which correlated with treatment with convalescent plasma (CP). In vitro experiments with pseudovirus support that this SARS-CoV-2 mutation/deletion variant has decreased sensitivity to sera from recovered patients, including the CP administered to this patient. Given that the 69-70 deletion has arisen multiple times globally in sequenced SARS-CoV-2 isolates, and that deletions are unlikely to be reverted, understanding the implications of this deletion for efficacy of polyclonal sera produced by infection or vaccination is of importance.

Overall, the correlation of the 69-70 deletion with the CP administration is notable, but the data presented here are not sufficient to clearly demonstrate that the 69-70 deletion confers resistance to polyclonal sera (as well as the reason for the fatal outcome of the infection).

Response: in our revised manuscript we present data that virus mutation leads to loss of CP susceptibility and that D796H is primarily responsible for this. However the infectivity of D796H is compromised in contrast to WT. The 69-70 deletion by itself increases infectivity of Spike in a single round infection. We therefore are not suggesting that 69-70 deletion confers resistance, rather that it is compensatory in this individual. We have gone through the text to clarify this point.

Here are some additional comments:

1. In the Introduction, the authors mention that RNA viruses have inherently higher rates of mutation than DNA viruses, but there is no mention that SARS-CoV-2 has actually a quite modest error rate for an RNA virus (compared to, for example, influenza) and that overall, the evolution of SARS-CoV-2 during the pandemic is relatively slow.

Response: This is a good point and we have now added text to this effect in the introduction

2. More information should be provided regarding the Euroimmun assay used to assess the antibody titers of the CP, e.g. what is the antigen used for this assay and a very brief summary of the method.

Response: We have provided this in the methods

3. Supp Fig 4 refers to 10 samples in the legend but there are 14 data points for each experiment.

Response: We have now rectified this discrepancy

4. Regarding the three persistent-shedding local individuals used for controls in the phylogenetic analysis: The treatment histories of these patients should also be outlined in brief (e.g. presumably they did not receive CP therapy).

Response: we have added treatment histories to supp table 2

5. Regarding Figure 3:

a. It would be helpful if the blue arrows indicating treatments in panel B were also included in panel A.

Response: we thank the reviewer for this comment and we have now done this

b. In panel A the duplicate genome annotation should be merged (there is one above and one below the plot).

Response: We have now done this.

c. Please provide a key for all lines in Panel B, and/or provide the underlying data in a supplementary table.

Response: We have now provided a table in supplementary figures of the output of the variant caller with all nucleotide positions that were used to produce the figure.

d. In particular, all variants discussed in the main text should be annotated in this figure.

Response: All of the variants discussed in the main text are shown in the revised figure.

e. The variant frequency scale should be better defined (based on the main text, I assume that 1.0 is 100%?)

Response: We have adapted the figure to reflect this change, and thank the reviewer for the suggestion.

6. A better explanation is needed for the results of consensus level sequencing (Fig 3A) as compared to deep sequencing results (Fig 3B). How can T39I reach 77% frequency on Day 44 (text lines 170-171, presumably this is the olive green line in Fig 3B) without being detected by consensus level sequencing?

Response: An error with the highlighter plot (Figure 3A) led to this being omitted by mistake. We have now amended this and thank the reviewer for this comment. There are now further observations about low level variants in the paper.

7. Supp Fig 5 is missing data and the legend for Supp Fig 6 is cut off and not possible to read.

Response: We apologise for this. We have now provided corrected supp figures and legends. Supp fig 6 has been removed due to comments from another reviewer

8. At line 175, the authors refer to the “near fixation” of D796H and the deletion mutant. It appears this term is being used to indicate that close to 100% of recovered samples contained this mutation. This is not an appropriate use of the word “fixation” (which implies subsequent persistence) considering that these mutations drop way down in frequency during the following several timepoints.

Response: We agree with this and have corrected the wording.

9. Regarding Fig 4A, presumably the black triangles indicate Ct values. This should be explained in the legend.

Response: CT values indicated by the right y-axis, and is clarified in the figure legend.

10. The experiments to evaluate in vitro fitness of the D796H + 69-70 deletion are not strong. There is no information in the main text, figure legend or Methods indicating how the RT activity and infectivity were performed and quantified. What are the error bars for RT activity? It appears that infectivity is a single experiment. How were the viral titers normalized prior to conducting these experiments? These results are important for evaluating the relative ID50 in the neutralization experiments.

Response We have expanded the methods section to describe how RT activity was used to normalise the infectivity data. We have outlined the use of the pseudotyping system in both results, methods and the legend. We also normalised virus inputs into infectivity assays and found very similar results, indicating the normalisation process was valid (supp figure X). Infectivity was repeated twice and representative data shown.

11. The y-axes for Fig. 5C-D are mislabeled, this should be ID50, not IC50

Response: we have now corrected this

12. The data shown in Fig. 5C-D are uninterpretable without more rigorous evaluation of infectivity between these two pseudoviruses, or performance of neutralization experiments with a monoclonal antibody not expected to be affected by these mutations (such as an RBD-targeting mAb). As is, it is not possible to evaluate whether the enhanced susceptibility of the WT pseudovirus is really due to the lack of the mutation/deletion combination, or to some other property of the pseudovirus preparation. The neutralization curves for these data should be provided as supplementary information.

Response:
We have put equal volumes of virus supernatant and then corrected the infectivity data retrospectively to take into account the differing amount of particles put in using the RT assay (figure 6). We have also used a second approach : normalise infectivity by modifiying the input virus amount such that the same amount of RT activity goes into each experiment (Supp figure 7). Results were very similar.

The neutralisation assays are done as standard in the field by using TCID50 normalisation.

A panel of RBD targeting (and one non RBD targeting) mAbs have also now been tested against the WT, D796H/Δ69/70 viruses, as well as single mutants. RBD targeting mAbs are largely unaffected by these mutations. Neutralisation curves have now been provided as SI.

13. The mAbs used in panel 5E should be better defined: What is known about their epitopes? Are they all RBD targeting mAbs (the most common category of neutralizing mAbs)?

Response: These mAbs were described in a previous paper and to keep to the word count this information was not highlighted in our original submission. This omission has now been corrected by: including a summary table in the supplementary, highlighting the epitope clusters and RBD specificity in the figure, and amending the main text as follows:
"mAbs isolated from three donors were previously identified to neutralize SARS-CoV-2. To establish if the mutations incurred in vivo resulted in a global change in neutralization sensitivity we tested neutralising mAbs targeting the 7 major epitope clusters previously described (excluding non-neutralising clusters II, V and small (n =<2) clusters IV, X). The seven RBD-specific mAbs exhibited no major change in neutralisation potency. The non-RBD specific COVA1-21 showed a reduction in potency against the del69/70 and del69/70+D796H."

mAb	Binding cluster	Target
COVA1-18	I	RBD
COVA2-29	I	RBD
COVA1-16	III	RBD
COVA2-07	III	RBD
COVA2-17	IX	RBD
COVA1-12	VI	RBD
COVA2-02	VII	RBD
COVA1-21	XI	Non-RBD

Nb. Clusters II, V contain only non-neutralising mAbs, smaller neutralising mAb clusters IV (n=2) and X (n=1) were not tested.

14. Why was a different set of pseudoviruses used for the serum neutralization vs the mAb neutralization experiments? Results would be a lot stronger if experiments were repeated with the same set of pseudoviruses for both. (Which would ideally include both the mutation/deletion combination and the mutation and deletion as separate pseudovirus variants.)
Response: This was due to mAbs and sera being evaluated in different laboratories due to MTA/Ethical permissions and the speed at which this project progressed. The UK restrictions have further added to complexity.

15. To better understand the sera results, would it be possible to perform serology with these samples? To profile their relative response to different regions of spike, especially NTD and RBD.
Response: we agree that it would be informative to be able to assess the relative responses to different regions of spike by for example measuring semi-quantitative titres for spike e ctodomaim vs RBD vs S1. We have now provided such data in figure 3A.

16. The titer and specificity of antibodies present in the CP used should be investigated. Was the CP sample used particularly enriched for NTD-neutralizing Abs?
Response: Currently, as described above, it is not feasible to assess what proportion of neutralizing sera target a particular epitope or subunit on Spike. This would require either well defined mutants that knock out a particular neutralizing epitope (as have been described for major HIV and influenza neutralizing epitopes) or recombinant subunits that do not impact viral infectivity which can be added in excess to serum neutralization assays to specifically block activity against that subunit (e.g. the D368R version of HIV gp120 which has been used to assess the proportion of gp120-specific neutralization). Even the polyclonal EM mapping technique used by Barnes et al. Cell 2020 to describe the COV57 plasma (see response below) would not give information on levels of neutralizing antibodies – and it requires a large amount of plasma not available in our study. We have provided the titre of CP antibodies in figure 3A for Spike trimer, RBD and N.

17. Did the Authors try to isolate viruses at the different time points? In vitro competition experiments between WT virus and the mutated one may help understanding the level of fitness of the D796H S2 mutation and 69-70 NTD deletion

Response: Viral culture was attempted but unsuccessful, despite favourable Ct values. We agree the competition experiments suggested would have been informative.

18. Line 240 states that the 69-70 deletion is "close to the binding site of the polyclonal Abs derived from COV57 plasma." It is not clear where there is data in either of those two citations to support this statement, as the 69-70 deletion is in the spike N-terminal domain, whereas all the antibodies profiled in the two citations are RBD-binding Abs.

Barnes et al., 2020, Cell 182, 828–842 contains a lot of complex data due to their use of a relatively new polyclonal serum mapping as originally describe by Bianchi et al. in Immunity in 2018. They show in Figure 4C that predominant Fab response in COV57 targest the S1a domain which is also referred to as the NTD. Individual 2D class averages on which this is based are shown in figure S4. The finding that both RBD and S1a domain reactivity is found is also listed as a "highlight" bullet point on the first page of the online version.

19. It would be interesting to speculate in the Discussion about the differences in viral evolution between the case patient and the patients in Avanzato et al and Choi et al, the former exhibiting very little viral divergence despite treatment with CP, and the latter showing much higher divergence (prior to any mAb treatment, and this patient did not receive CP).

Response: we have now added discussion on this point into the discussion.

Referee #2 (Remarks to the Author):

The authors describe the detailed virological and phylogenetic investigation a COVID-19 infection in a B cell / antibody deficient patient who had been treated with convalescent plasma therapy. This allowed the authors to document the emergence and re-emergence of viruses with spike delH69/delV70 plus D796H as likely escape from polyclonal antibody administered as convalescent plasma therapy. The authors then introduce these mutations together in a spike pseudotyped virus to demonstrate that these two mutations associate a reduction of susceptibility to neutralization with polyclonal immune serum. These findings have important public health implications as one of these mutations (delH69/V70) is not infrequently found among the GSAID database.
General comments:

a) It would be interesting to have an assessment of the likely T cell function in this patient. Was his underlying disease likely to have impaired T cell function? Was there any objective assessment of this? If T cell function was largely intact, it would be interesting to discuss why T cells would not contribute to containment of virus replication?

Response: The follicular lymphoma is associated with reduced T cell functionality as was prior chemo with vincristine, cyclophosphamide and prednisone, in addition to rituximab. We have added this to the text in addition to results from whole blood T cell functional testing at two time points (supp figure). This showed poor T cell responses.

b) The phenotype of reduced neutralization with polyclonal sera was investigated with the spike Del69/70+D796H mutant. There was no analysis of the effects of each of these mutations (del69/70 vs. D796H) on the phenotype of neutralization. It would be desirable to make spike pseudotypes with each of these mutations to define the mutation critical for the neutralization phenotype. Del69/70 seems for more commonly found than D796H. So the distinction is important, although it appears that the two mutations co-appear and decline in this particular patient.
At different points in the text, the authors give the impression that delH69/V70 is the escape mutation (see line 89-91) but there is no evidence for this?

Response: We have now performed the experiments indicated and have clarified that we do not believe delH69/V70 is the primary escape mutation.

c) It would be of great interest to investigate the fitness of the del69/70+D796H mutant compared to wild type. The data presented here may suggest that this is a less fit virus that is replaced by wild-type in the absence of selective pressure of the convalescent plasma. This should be discussed. Ideally, reverse genetics reconstitution of this mutant virus will allow experimental investigation of fitness of this mutation vs wild type in vitro and also in vivo.

Response: We have now shown single round infectivity data of spike pseudotyped lentiviruses with single and double mutants. In the time frames we have it has not been possible to do reverse genetics reconstitution of the mutant though agree this would be ideal. There are also safety issues surrounding the generation of highly fit /escape mutant viruses.

d) Are attempts being made to culture the mutant viruses? Some of the specimens have reasonable viral loads that may allow virus culture.

Response: Virus culture has been attempted without success, despite reasonable Ct values.

e) This antibody neutralization escape mutation/s arose in the immunocompromised patient after convalescent antibody therapy. In the discussion it is important to emphasise that it may not be as likely for the mutation to appear in immune-competent individuals because there is a major distinction between putting antibody selection pressure in this immunocompromised patient who had a well established infection with lots of virus in his respiratory tract when the convalescent serum was used with i. infection / exposure of someone with prior antibody elicited by natural infection of vaccination when the antibody is facing a lower virus diversity at the start, or ii. an immunocompetent individual who has the T cell arm to support the convalescent plasma. It is clear that the authors are well congnizant of this issue, but it is important to be explicit so that the reader also appreciates these issues.

Response: we fully agree with this point and have amended the discussion to take this into account.

Other points for clarification:
Line 145-149: Three additional patients analysed. Specify that these patients did not have convalescent antibody?
We have now clarified this in the supplemental material. One of the control patients did receive CP but as consolidation therapy after virus clearance.

Line 239: delH69/V70 is close to the binding site of the polyclonal antibodies in COV57 plasma. How do you define the binding site of a polyclonal antibody? By definition, there will be multiple antibody specificities in a polyclonal serum.

This is a very new and innovatve technique originally described by Bianchi M et al. 2018, Immunity. The process of collecting negative stain EM data for mAbs relies on capturing individual antigen+fab complexes and generating class averages from these to resolve the final structure. Bianchi et al. adapted this for for fabs purified from polyclonal sera. Instead of averaging all individual images, they "binned" similar images to assess the relative proportion of antibodies targeting each epitope. This was feasible in the first case as a set of mAbs to the antigen had already been defined (McCoy et al. 2016, Cell Reports), and the technique is now starting to be applied and used in de novo situations where mAbs to a given antigen have not been mapped. Thus, Barnes et al., 2020, Cell used the same technique to establish the predominant Fab reactivity found in COV57 as described in response to reviewer 1 above, however, it should be noted this does not give information on whether antibodies are neutralizing or not.

Line 306-312: "our data suggest caution in use of CP in patients with immune suppression of both T cell and B cell arms." Do you have data to show that this patient was T cell deficient?

Response: Yes we now show these data in the supplementary figures

Referee #3 (Remarks to the Author):

In the manuscript "Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion" by Kemp et al., the authors describe genomic changes in the SARS-CoV-2 genome in a single, immunocompromised patient over the course of several COVID-19 treatments. They find that two changes in the SARS-CoV-2 spike protein emerge each time the patient is treated with convalescent plasma, and that these changes in the virus diminish serum neutralization. Understanding the selective pressures imposed by different treatment regimes is crucial to selecting appropriate long-term disease treatment. Though the work is interesting and obviously very timely, I have significant concerns with some of the points and discussion.

Major comments:

1. Overall, though the work reads well, I suggest editing the results section to avoid unnecessary medical jargon i.e. the significance and key points of the CRP test (inflammation mark) and respiratory rate (high especially on O2). It would be helpful to describe the patients' history and treatment course in a more straightforward manner. This feeds into Figure 1 - much greater clarity is suggested here. I strongly suggest these values for Ct be included as a supplemental table, along with which were sequenced, from what sample type, and with which technology. This was difficult to trace throughout the paper.

Response: we have now removed as much jargon as possible. We have now also added a supp table with the information suggested

2. The same is true for when treatments were administered and when antibody testing was done. For example, on Line 121 the authors say they measure antibodies “over the course of admission”, but then only describe testing on days 44, 50, 68, 90, 101.

Response: We have now also show this information in figure 3A for Luminex assays.

3. Line 407: Which variant caller and version were used for this analysis? There is some evidence that the variant caller selected can affect which variants are called.

Response: we used SAMFIRE designed by Chris Illingworth. More details have been added to the methodology section. We

4. Sequencing data should be made available, consensus _and_ raw data if at all possible, especially given the minor variations discussed in the work.

Response: FASTQ files have been deposited in the NCBI SRA database – accession numbers are detailed in the manuscript. Consensus sequences are available in GISAID

5. Figure 2A is largely incomprehensible just due to the number of samples plotted. I suggest strongly that subsampling as was done in part in 2B is more useful, especially with regional samples of the virus as well as a reasonable global subsample. As it stands 2A has little utility except to show that different patients in the region/hospital system have virus of significantly different clades, better shown by Figure 2B.

Response: We have heavily down sampled and adapted the figure to be clearer.

6. In figure 2B phone and call bell look significantly different in number of mutations? The authors assert that they cluster with the respiratory samples - true, but the number of additional mutations acquired seems high, a larger distance than separating control patient 1 and case patient (from initial samples).

Response: We have since remade the phylogeny (with corrected Illumina data). We note that the Patient’s call bell has a total of 19 nucleotide substitutions, resulting in 11 amino acid substitutions (N:R203K,N:G204R,ORF14:G50N,ORF1a:E452G,ORF1a:T1246I,ORF1a:G3278S,ORF1b:V33L,ORF1b:P314L,ORF7a:S81P,S:D614G,S:A899S). The patient’s mobile phone has a total of 14 nucleotide substations resulting in 8 amino acid substitutions (N:R203K,N:G204R,ORF14:G50N,ORF1a:E452G,ORF1a:G3278S,ORF1b:P314L,ORF7a:T39I,S:D614G). This distance is accurately reflected on the new phylogeny with the adjusted scale bar. Thank you for drawing this to our attention.

7. Figure 2B would benefit from an indication of variants versus time, especially for the case patient, but also for the “persistent shedders” which make up the other local patients. Is there any reason the authors can delineate why the persistent shedders _were_ persistent shedders, and why not chose some patients who had a “normal” course of disease, with expected rise and fall of viral titer?

Response: The persistent shedders were 1. Renal transplant recipient on immune suppression 2. Cardiac disease 3. XLA immune deficiency. These data are in supp table 2. It is true that given the spectrum of shedding duration the delineation is difficult. Nonetheless we thought addition of these individuals would be helpful. The third patient has been published and there were very few variants.

8. In supplementary figure 4, why are only 10 samples shown? Again - a table of which samples were taken when and from which days the plots are generated would be helpful. I suggest rather than box and whiskers, the correlation between Illumina and nanopore variant calls is what's helpful here - I would suggest generating a scatter plot per variant for Illlumina versus nanopore called frequency.

Response: Thank you for this suggestion. We have added a supp table of sampling and removed the box and whisker plots, instead replacing with a more complete table comparing the methods.

9. Figure 4B: I am quite surprised by the long branch lengths separating the sequences from days 93 and 95 from the rest of the patient samples. Using the scale provided, it looks to be ~12-15 mutations away from the other sequences. I suppose this could be explained by rapid mutation / escape following the emergence of the deletion mutations, but then these mutations are abruptly and completely lost again by day 99? I would compare this to the results from Choi et al which the authors plotted in 2B (data recently published from NEJM) with a patient with antiphospholipid syndrome on immunosuppressive drugs (e.g. cyclophosphamide) who also died of a long course (~154 days) of COVID19. It doesn't seem that the same type of behavior occurred in Choi et al with a complete loss _and regain_ of a relatively dominant mutation (in this case the del H69/V70). Though I understand the selection pressure being applied through the use of CP, it still seems surprising that the levels of these variants could change so much, and that if lost, the mutation could be reacquired.

Response: we agree that this is surprising, though the shifts in populations was also seen in the Avanzato et al report. We have added text to the discussion to explore these issues.

10. It should be pointed out by the authors in the results and discussion that the nature of the ARTIC method makes it very difficult to assign quantitative levels to the levels of "mutation prevalence" i.e. Fig 4A - because of the amount of PCR amplification being employed, and the concerns of jackpotting especially when relatively high Ct samples were employed (i.e. in the 66-81 day range). At the very least technical replicates going _back_ to the RNA source should be employed - it seems that the authors used the same amplicon source for testing of Illumina v. Nanopore technical artifact, but I am more concerned about artifact arising from the RT and PCR.

Response: Thank you for the suggestion. Additionally we have gone back to the stored RNA and conducted single genome amplification of the spike genome in order to further validate our findings (Supp table 5).

11. The authors report the deletion at "<10% abundance" at day 93, but from their graph it looks more like 0? (Again, I recommend the data used to generate the graph be included as supplemental table/data to make interpretation clearer). What is the hypothesis here - not clearly delineated by the authors. I suppose they suggest that this mutation, already present but undetectable before CP treatment was then strongly selected for as other virus was largely eliminated by CP. But then _why_ given the data in Figure 5A showing little to no impairment of infectivity by the mutation, is the mutation lost in samples from day 86-95? Why wouldn't _additional_ mutations be acquired instead? And then how likely is it that the mutation is reacquired in the exact same way post-day 95? Isn't a more likely scenario superinfection by the Y200H and T240I mutations while the del H69/V70 lineage remains persistent in the patient? I understand it's difficult to rule this out without _extensive_ local sequencing of virus, but the authors state "the relationship of divergent samples to those at earlier time points rules out the possibility of superinfection" - I disagree. Examining data from local sequencing efforts including work from New York (Gonzalez-Reiche, Science 2020), Baltimore (Thielen, bioRxiv 2020) and Houston (Long et al mBio 2020) to say nothing of the COG-UK

efforts suggests virus from different patients in the same area can fall within the same cluster at the level reported in this work,

Response: we describe a number of variants that change in frequency that largely argue against superinfection but we take the reviewer's point.

12. The authors assert from Supplemental Figure 6 that they observe an increase in frequency of the T39I over the first period of the infection - but the fit is not great. It could easily be argued there is a peak in the data around day ~45, with a decline prior to use of CP, perhaps resulting from remdesevir treatment? I would suggest this actually goes to my earlier point that trying to pull quantitative numbers directly from coverage with a method with so many cycles of PCR is not ideal.

Response: We agree that the fit is not appropriate on reflection and have removed this figure

13. Perhaps an orthogonal method - qPCR with Taqman probes for the specific SNPs of interest could be used to improve quantitation. This is especially important for the deletion mutation.

Response: We have used SGA to corroborate findings (supp table 5)

14. Further, when mutations were lost and seemingly regained (del H69/V70) was the mutation _exactly the same_ i.e. the exact same loss of those 6 NT? Were there any other mutations in linkage that could be used to determine if it was a reacquisition or the same lineage? A comparison of the consensus, as well as a careful inspection of the aligned nanopore and illumina reads to these regions should be provided.

Response: indeed it is the same out of frame 6 nt deletion in each case. The deletion does not disappear completely however, providing added confidence that this is indeed a virus population in decline following washout of neutralising antibodies. We have now provided an alignment of the deletion region in figure 4.

15. Supp Fig 7 is not particularly convincing. If you have achieved near fixation of the deletion, why would that mutation have disappeared and reappeared exactly without superinfection. Isn't it as likely that you have a superinfection event that wasn't detected from lack of local sampling? Comparing against global SARS-CoV-2 data is not nearly as convincing as comparing against high local sampling would be.

Response: We have provided a tree with Cambridge sequences and we also have variant dynamics to show that superinfection is highly improbable

16. The pseudotyping experiments are nice - showing that the deletion doesn't matter for infection, but "partial immunity" - though I argue >2 fold difference is more than partial immunity. But the authors do not discuss why the different CPs show such a large difference in IC50 within the different CP and between the other sera (S1-S5).

Response: we thank the reviewer for the comment. CP are from different patients and time between COVID-19 and CP donation unknown. Severity, associated with titres, is also unknown. Heterogeneity in titre/quality of serum responses across people is emerging as the defining feature in many studies. Given we have now shown data on individual mutants for 3 CP units we have removed the other sera data.

17. Fig 5E - I'm confused about this - basically the authors are arguing that natural CP is preferentially targeting epitopes that they didn't chose in their 7 neutralizing abs. This is particularly confusing as the paper cited identified mabs from patients with SARS-CoV-2? (https://science.sciencemag.org/content/369/6504/643)

Response: Our data shows that escape is not mediated via classical RBD-specific mAbs. Of 84 mAbs in the paper cited only 19 are neutralising, we picked representative neutralising mAbs from 6 different RBD epitope clusters and 1 non-RBD neutralising epitope cluster to see if the escape observed in patient could be attributed to changes in sensitivity in RBD-specific or if it implied a non-canoncial epitope.

Minor points:

1. Sup Fig 1 would benefit from alpha transparency.
   Response: We have not been able to corrected this and hope that this is ok.

2. Sup Fig 3 y-axis should have units or some sort of scaling to give context.

Response: The axis is a ratio so we are unsure of the point being made.

3. Figure 4A - black line with triangles is not annotated?

Response: We have now corrected this.

4. Sup Fig 8 - which received no reference in the text that I could find - is interesting to examine the prevalence of this mutation in GISAID, but the plot would perhaps be more interesting if instead of a radial plot a plot was generated for the rise of these mutations over time?

Response: We have removed this figure given a follow up paper on the repeated emergence of the deletion.

5. Line 103-104: Does repeat SARS-CoV-2 RT-PCR mean multiple tests on the same day? From the same swab? Please specify the time period and if these are biological or technical replicates.

Response: this means on the same day from different sites eg ETA and nose throat swab

6. Line 115: It would be helpful if the authors provided a little bit of background on monoclonal antibody therapies such as tocilizumab, mentioned here. Several different mono- and polyclonal antibody treatments were mentioned and discussed throughout in different contexts, and I found it very difficult to keep track of each (and which ones were for COVID-19 and which were administered for other purposes, e.g., rituximab on line 100).

Response: given the length of paper and the fact that other reviewers asked for less clinical detail we have not added anything here.

7. The color labels in Figure 3B are difficult to interpret and I suggest points be plotted on the lines to demonstrate the frequency of sampling more clearly.

Response: we have reformatted the figure for clarity.

Referee #4 (Remarks to the Author):

Kemp and colleagues present a case of persistent SARS-CoV-2 infection and analyze the molecular evolution that unfolded within the host in detail. This case is not dissimilar from two recently described cases (Choi et al (10.1056/NEJMc2031364), Avanzato et al (10.1016/j.cell.2020.10.049)). In contrast to these previous cases, Kemp et al investigate within-host evolution using deep sequencing and trace the frequencies of different variants through time. They characterize three diverged variants with different mutations and deletions in the spike protein, some of which reduce neutralization titers of convalescent plasma in a pseudo-typed lentivirus. Overall, the work in this paper is well performed and it provides convincing evidence of in-vivo antibody escape.

I have a number of suggestions to improve the presentation, strengthen the conclusions, and to remove/tone down parts that might be misleading.

* Fig 2A: The radial tree in Figure 2 is rather unhelpful. The labels are hardly readable and distances between samples are very hard to judge from the radial presentation. A rectangular tree indicating major clades and the different within-host samples would be better.

Response: We have now corrected this.

* Fig 2B & 4B: I think the figure would be improved by changing the scale bar to correspond to one or two mutations (currently the scale bar is given in mutations per site and is roughly 6 mutations in 2B, 2 mutations in 4B). Zero-length branches in the ML trees should be collapsed into polytomies. Bootstrap values on SARS-CoV-2 trees are pretty useless. Better to label the branches with (number of) mutations that fall on the branch in a parsimony or ML reconstruction. This has a one-to-one correspondence to bootstrap values and is more interpretable.

Thank you for the suggestion, we have now altered the scale bars to represent a length of 1 amino acid mutation. We have also removed the bootstrap support number from both trees. The relevant amino acids changes we wish to draw attention to are on the branches in Figure 4b.

* The purple line in Fig 3B suggests an iSNV at frequency 30% on day one that persists at a frequency around 30% until day 82. This iSNV doesn't seem to be affected by the fixation of other iSNVs at time points 66 or 82 days. Would be good to look into this. It could indicate population structure which would imply parallel evolution. Or it could be an artifact (more likely). Either way, this should be looked at and discussed.

Upon further inspection, this iSNV was an artefact – it appears that there was significant amplification at the boundary of an amplicon. As such, we have now removed this variant.

* To understand the rapid shifts in dominating variants better, it would be helpful to include a discussion of their frequencies when they are rare. It makes a difference to the interpretation if the minor variants are present at 10%, 1%, or 0.1%. The reader currently has to piece this together from supplementary table 3 and there are some discrepancies: The S:64G variant seems to be very rare after day 95 (not detected by high coverage Illumina) while the linked S330S is still picked up (at high frequency in low coverage data??). Mutations 200H,240I,258S are missing from supplementary table 3.

Response: We have added comments to the text following the progress of the dominating variants. For example, the three variants across the genome which reach high frequencies at day 81 are all less than 5% on day 86 and less than 10% on day 89. Y200H and T240I were at less than 2% frequency in the samples on day 93.

* Fig 5 would be more useful on a logscale. Bar charts should be avoided, individual data points need to be shown.
Response: We have now used log scales and shown data points

* the description of how evolutionary rates are estimated from within-host data is very short. I would caution against over-interpreting these estimates for two reasons: (i) Phylogenetic estimates are done with consensus sequences and thus ignore minor variation. (ii) rate estimates likely depend a lot on how the within-host variation is rooted and how the root height is constrained. The error of the mean rates (table S2) seems way too small in some cases (1% of the main) calling the entire procedure into question. I would cut this as I don't think this is reliable and it is not central to the paper.

Response: We have now removed the estimate of mutation rates as suggested.

* Similarly, the logistic fit to T39I in ORF7a (Supp Fig 6) is not evidence for selection. I don't see what this figure adds that is not visible in Fig 3. All that Fig S6 shows is that the variant was rare at day one and then bounced around frequency 0.5 between day 30 and 60. There is no reason to fit a logistic and insinuate selection.

Response: We have now removed the Supp Fig 6.

* the distances presented in Fig S5 seem rather large (two-fold larger than what I would have guessed from the tree).

This is correct. It is a result of the way in which the distances are calculated. In this figure we calculate distances as sums of the absolute changes in allele frequencies. To give a simple example, a variant might be at 10% frequency in sample A, and 20% frequency in sample B. In a phylogenetic metric, which looks at consensus sequences, this would count as a distance of zero, as the difference is wiped out by taking the consensus. In our metric, it would count as a distance of 0.1, leading to, in general, a larger distance.

While unorthodox, we believe that the distance metric we use in this Figure has some mathematical advantages. For example, a variant at 45% in sample A and 55% in sample B would also count as a distance of 0.1 under our metric, but would equate to a distance of 1 when consensus sequences were taken.

* accession numbers for consensus sequences and reads need to be provided.

Response: We have now provided these

Minor comments:
* Fig 4: unclear how shaded areas relate to the lines. Same in Fig S1. I have a hard time parsing the intersecting areas.
Response: we apologise for this. We have variant frequencies in supp table 4

* line 133: while you are using a long-read sequencer, the amplicons are short. Better call it single-molecule sequencer.

Response: We have changed the text to reflect this, using the term single molecule sequencing

* lines 160-162: I doubt genetic drift has a major effect here.

* caption Fig 3: the "spike genome"?

Response: we have corrected this

* Supplementary Fig 8 is completely unreadable

Response: this has been removed as is the subject of another analysis

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Authors have fulfilled most of the points that were raised and overall improved the quality of the manuscript. Text revision helped strengthen and clarify the paper and dealt with several issues brought up by the reviewers.

There are a few more issues - and these are listed below.

It would be important to mention in the text that viral culture was attempted, but unsuccessfully. As to Figure 3A it would had been important to include testing of NTD (in addition to N, S and RBD). This type of reagent is readily available from commercial suppliers.

The finding that partial immune escape is primarily driven by the D796H mutation is unexpected, and does not help explaining if the observed shifts in the virus population is the result of immune escape or not. The level of fold change relative to WT is overall modest (2 to 3 fold difference) to make a strong point out of this, in particular in light of the more convincing effect on reduced infectivity.
Is this modest reduced susceptibility the real driver for in vivo selection of this variant by CP treatments?
The effect of this S2 mutation on S requires more functional and structural studies, that seems out of the scope of this study. It is clear, however, that the 69-70 deletion did not affect neutralization by plasma samples as well as by several RBD mAbs (excluding the possible allosteric effect of this deletion on RBD). These results may also be influenced by the type of assay used where ACE2 was over-expressed. As suggested by other studies auxiliary receptors, such as DC-SIGN or L-SIGN may also play a role on SARS-CoV-2 entry and antibodies directed to NTD may have an impact here (like shown for the NTD-specific neutralizing mAb 4A8). This should be at least mentioned in the discussion.

Mutations at position D796 have now been reported (and quite recently) from multiple countries (as of today a total of 282 mutants are reported in GISAID), of which 70 have the D796H mutation (and 144 D796Y). Authors should comment on what is known about this mutation and its prevalence. This is clearly less prevalent as compared to the 69-70 deletion (more than 10,000 counts), but yet not so unreported like a few months ago.

What is the specificity of the COVA1-21 mAb used? it is mentioned to be non-RBD. Is this NTD-specific? this should be clarified.

Referee #2 (Remarks to the Author):

Line 120: "each from one donor". Please be clear. Do you mean both plasma units were from the same donor or you mean one unit each from two donors?
Line 181 and Figure 5A: Why is the RT-PCR CT values not presented prior to the convalescent therapy at day 65/66?
Line 247 and Figure 6 legend: IN line 247 it is stated that the psudotyped viruses were probed with HIV-p24. In the legend it is stated that the virus input was standardised using the RT activity in the virus supernatant. How exactly were the different pseudovirus preps standardized? Using p24 or RT activity or both?

Referee #3 (Remarks to the Author):

1. Figure 1 seems to be missing from the revision - as this was one the points we requested greater clarity from, it would be helpful to see any revision to this figure. However, Supp Table 3 is helpful in aiding clarity.

2. Ok, thanks.

3. SAMFIRE details suggest a PHRED q score of 30 was used to filter, but this is unlikely given that ONT q scores rarely get that high. Can the authors clarify _exactly_ what they did? Code should preferably be deposited in github. Further, the authors are not clear on the _consensus_ variant caller used, presumably SAMFIRE is only used for the allele frequency calling. This could be related to the apparent sentence fragment in the response.

4. Great, thank you

5. In Fig 2A - much improved, but I'm curious why the different control patient cases seem so far separated from the case patient? Was the downsampling done to be representative of the region?

6. Why is the scale according to AA mutations instead of nt mutations? For genetic epi isn't nt mutations as or more useful than the AA mutations? I understand looking at AA for selection/resistance purposes, but for clustering purposes it seems to me nt is more appropriate.

7. So of these persistent shedders, it seems that only control 3 (XLA immune deficiency) had significant mutational changes over time? Why is that one different than the case, or than the other controls?

8. Supp Table 4 is extremely useful, but actually reveals more cause for concern in places. Consider for example the spike deletion (21763) - and the fact that on day 93 - one of the samples Illumina and nanopore conflict - Illumina detecting 0.0% and Nanopore 9.2%. There are quite significant differences in this table between the two technologies with the allele frequency methods they are calculating. I am _surprised by these differences when in non homopolymer methods and still would like either A) stats on validation B) explanation for differences and/or C) more clear methods on how variants and consensus were calculated (see point 3). Both Thielen et al (medrxiv 2020) and Bull et al (Nat Commun 2020) examined Illumina vs nanopore carefully and found accurate consensus level data, so some of the number reported in this table are surprising.

9. Focusing further on this point and again looking at day 93 - There are several Illumina/Nanopore mismatches on Day 93. Given the long branch lengths to Day 93 and other samples, these genomes should be examined more closely. If indeed it is an issue of jackpotting causing certain variants to display a higher frequency than others I would think that these mismatches should be interrogated a little more closely? It's especially surprising to me given that the Ct values for 93 were pretty reasonable (21).

My read of the hypothesis of the authors is that the CP introduced a pressure to result in (at least) the deletion mutation). But I also don't see a reason for that deletion to be lost, if CP was an effective therapy - i.e. why the 69/70 del is gone at that point. The authors try to address this with their "compartmentalization" theory, but why wouldn't the CP have wiped out the non-mutated virus, precluding anything that doesn't have the del? In my mind this is still the sticking point.

10. This is interesting, but I'm confused by the dates the authors selected? The 0s from days 1 and 37 are not especially interesting, I would argue that Day 98 is but I need a way of correlating the SGS result to the plot in Figure 4 - and/or what is preferable are times sampled in this active area around days 83-100?

11. The authors . . . . didn't really answer my point? I think the table helps, but did they address in the text or elsewhere the apparent loss and regain of certain mutations? I'm not asking them to rule in or out superinfection - and their argument in the discussion about "compartments" of virus or different possible lineages is not implausible, but they need to be clear in the results that the nature of sampling precludes a clear picture of the virus in the patient.

12. Ok

13. The SGA data is nice, but the days selected were . . . inappropriate. I suggest it would have been more useful to probe in samples drawn from the "active" part of the viral timeline.

14. Great.

15. I suppose that Supp Fig 8 is the figure they are referring to (Cambridge Sequences) - this is pretty useful and does in fact suggest that the patient, though their initial infection was from the community at large, is likely all derived from the initial infection.

16. The pseudoviral figure has changed substantially, but I'm still confused at the significant spread in the points within a sample/condition(i.e. the replicates). Isn't it surprising that the spread is so large for example in the double mutant for CP1? Further, the authors assert that the "single mutant did not impact neutralization (255-256). But it _does_ seem to in the patient derived sera, as well as actually having a _stronger_ neutralizing effect in CP1?

17. Ok, it seems to have been edited in the results as well as the response, so this is fine.

Referee #4 (Remarks to the Author):

The authors have addressed most of my comments (changed tree diagrams, removed artifactual SNPs, removed things I thought were distracting or not robust). However, the authors have also added more data in response to comments by the other referees, some of which seem inconsistent with the previous version of the manuscript or cast serious doubts on some inferences.

The main additions are more data to disentangle the potential effects of D796H and the 69/70 deletion. The main message of the paper now is that 69/70 deletion is a fitness-enhancing or

compensatory mutation while D796H is responsible for neutralization escape. Previously more emphasis was put on the 69/70 deletion. I am not convinced, however, that such specific statements can be made with the available data. 69/70 is a deletion that occurred repeatedly across the globe, suggesting that it is a mutation that is "easy" for the virus to discover. If this had a substantial effect on virus fitness, you'd expect that it also appeared without the D796H mutations and persisted between times CP2 and CP3 treatment. Furthermore, the IgG levels from doses 1 and 2 don't change much between day 70 and day 98 and don't rise substantially after CP3 is administered making it less plausible that CPs are driving the evolution.

Major:
- I am trying to link up data in the current figure 6 with the previous figure 5. You previously reported similar RT activity between WT and the double mutant (about 10% difference, previous Fig 5A) and substantially increased infectivity (+50%, previous Fig 5B). In the new figure 6B, the infectivity of the double mutant is slightly lower than WT. What changed?

- Fig 6 C-G: no attempt is made to assess the statistical significance of the differences between genotypes. Looking at Fig S7 (the corresponding neutralization curves I think) suggests that the fits are pretty bad and often not only IC50 but the cooperativity seems to change. Some data points have wide CIs, others don't seem to have any CIs at all (some points are negative!!!). While these data are suggestive, I don't think they conclusively show that D796H or 69/70del have a consistent effect.

- https://github.com/Steven-Kemp/sequence_files/tree/main/figure_data is empty

- line 188: do you have any evidence that this is a "driven" sweep?

- line 193: is this "sweep" any more dramatic than the one described in the previous paragraph?

- line 235: Fig 4 suggests that NSP2:I513T and 69/70del are simultaneously at very high frequency at days 98 and 99, directly contradicting the statements that they are in opposing clades. Overall, these data don't seem to match up with what is reported in the supplementary data. The deletion (position 21768) is at low frequency in the table for days 99-101 but at high frequency in Fig 4.

- The supplementary table with the frequency data and Fig S5 suggests that SNP frequencies are often very discordant between nanopore and Illumina data. Although the caption of Fig S5 states "There was good concordance for mutations between the two methods, and no significant difference between the proportion of reads measured by both methods." the data are all over the place and either the plot or the caption is completely wrong. If the frequencies are really as unreliable as the plot suggests, none of the frequency dynamics in the paper can be trusted.

Minor:

- Fig 2: is the scale bare one amino acid or one mutation? I think it should be one mutation.
- Fig 4C: not sure this is useful. The nature of this deletion has been discussed a lot and showing ten genomes doesn't really add information. ALL 69/70 deletions in GISAID are like this. Maybe better point out that codon 68 changes from ATA to ATC with this deletion.
- Fig 5C-E: please show on logscale with consistent y-axis between panels.
- Fig 7: I would clarify which Spike mutations in the figure title.
- Fig S7: all three panels in S7 have different axis labels.

Author Rebuttals to First Revision:

Response to reviewer comments

Reviewer 1
(Remarks to the Author)
Authors have fulfilled most of the points that were raised and overall improved the quality of the manuscript. Text revision helped strengthen and clarify the paper and dealt with several issues brought up by the reviewers.

Response: we thank the reviewer for taking the effort to review the revised paper.

There are a few more issues - and these are listed below.

It would be important to mention in the text that viral culture was attempted, but unsuccessfully.

Response: we have now added text to reflect this (line 181)
As to Figure 3A it would had been important to include testing of NTD (in addition to N, S and RBD). This type of reagent is readily available from commercial suppliers.

Response: whilst this would have been desirable, the point still remains that antibodies can be non-neutralising and therefore NTD testing would have had limited value. We have been unable to therefore do this.

The finding that partial immune escape is primarily driven by the D796H mutation is unexpected, and does not help explaining if the observed shifts in the virus population is the result of immune escape or not. The level of fold change relative to WT is overall modest (2 to 3 fold difference) to make a strong point out of this, in particular in light of the more convincing effect on reduced infectivity.
Is this modest reduced susceptibility the real driver for in vivo selection of this variant by CP treatments?

Response: Non RBD targeting antibodies are increasingly appreciated so we do not think it is so unexpected, given that D796H forms an exposed loop, as in the manuscript. As the reviewer kindly points out below there are increasing numbers of mutations at 796 being reported, thus supporting a role for adaptation against immunity.

Regarding the effect size, reviewer 3 noted that 2-3 fold was significant and we agree given that we are testing polyclonal sera with multiple antibodies with different specificities. Even a 2 fold reduction is likely important. One should also note that this effect size is over a single round of infection and the virus generation rate of SARS-CoV-2 is very high, thus massively amplifying modest effects.

Furthermore, in this key in vitro evolution paper (https://www.biorxiv.org/content/10.1101/2020.12.28.424451v1.full.pdf) the authors see SARS-CoV-2 virus become completely resistant to highly neutralising serum. The first mutation, a deletion in NTD conferred 2 fold reduced susceptibility. The second mutation, K484E in the RBD (one of the most concerning mutations in the new South African variant) conferred 4 fold reduced susceptibility.

Finally, antibody shifts of only 4-fold were observed in mink variants with 5 spike mutations (69/70 deletion, Y453F in the RBD and 3 other S mutations) compared to wild type (https://files.ssi.dk/Mink-cluster-5-short-report_AFO).

We have nonetheless toned down the language of effect size in the revised paper (highlighted areas in yellow) and indeed changed the title to reflect the reviewer's opinion.

The effect of this S2 mutation on S requires more functional and structural studies, that seems out of the scope of this study. It is clear, however, that the 69-70 deletion did not affect neutralization by plasma samples as well as by several RBD mAbs (excluding the possible allosteric effect of this deletion on RBD). These results may also be influenced by the type of assay used where ACE2 was over-expressed. As suggested by other studies auxiliary receptors, such as DC-SIGN or L-SIGN may also play a role on SARS-CoV-2 entry and antibodies directed to NTD may have an impact here (like shown for the NTD-specific neutralizing mAb 4A8). This should be at least mentioned in the discussion.

Response: we thank the reviewer for these comments. We agree that the effect of this S2 mutation on S requires more functional and structural studies outside the scope of the study. We have added a little more data on 796 from structural modelling in a modified figure 8. We now also acknowledge the limitation that an overexpression system was used and noted in discussion the potential role of auxiliary receptors, such as DC-SIGN/ L-SIGN and possible role for NTD directed antibodies (lines 368-70).

Mutations at position D796 have now been reported (and quite recently) from multiple countries (as of today a total of 282 mutants are reported in GISAID), of which 70 have the D796H mutation (and 144 D796Y). Authors should comment on what is known about this mutation and its prevalence. This is clearly less prevalent as compared to the 69-70 deletion (more than 10,000 counts), but yet not so unreported like a few months ago.

Response: we thank the reviewer for these comments and have now added text regarding D796 mutations and added to supplementary table 7 updated numbers for both D796H and D796Y, reflecting the increase in numbers of cases of detection.

What is the specificity of the COVA1-21 mAb used? it is mentioned to be non-RBD. Is this NTD-specific? this should be clarified.

Response: The only information from the primary paper and authors is that the mAb does bind spike but does not bind the monomeric RBD subunit. However, it is plausible that it may bind to RBD only the context of full Spike trimer, as has been observed for other trimeric viral entry proteins. Competition analysis in the original paper showed that the other mAbs do not compete with COVA1-21, however, pre-binding of COVA-1-21 prevented binding of all other mAb classes, suggesting COVA1-21 destabilises Spike and further studies are ongoing to elucidate this complex problem.

Reviewer 2:

Line 120: "each from one donor". Please be clear. Do you mean both plasma units were from the same donor or you mean one unit each from two donors?

Response: we intended to state one unit from two donors and thank the reviewer for pointing this out.

Line 181 and Figure 5A: Why is the RT-PCR CT values not presented prior to the convalescent therapy at day 65/66?

Response: this has now been corrected thank you for pointing this out

Line 247 and Figure 6 legend: IN line 247 it is stated that the pseudotyped viruses were probed with HIV-p24. In the legend it is stated that the virus input was standardised using the RT activity in the virus supernatant. How exactly were the different pseudovirus preps standardized? Using p24 or RT activity or both?

Response: yes we probed for p24 to verify particle generation but used RT activity exclusively for standardisation. We have now clarified this.

Reviewer 3:

1. Figure 1 seems to be missing from the revision - as this was one the points we requested greater clarity from, it would be helpful to see any revision to this figure. However, Supp Table 3 is helpful in aiding clarity.

Response: Apologies, this was in its own file - we have now included this.

2. Ok, thanks.

3. SAMFIRE details suggest a PHRED q score of 30 was used to filter, but this is unlikely given that ONT q scores rarely get that high. Can the authors clarify _exactly_ what they did? Code should preferably be deposited in GitHub. Further, the authors are not clear on the _consensus_ variant caller used, presumably SAMFIRE is only used for the allele frequency calling. This could be related to the apparent sentence fragment in the response.

Response: Initially, ONT was used as part of the COG-UK ARTIC pipeline. However we since re-sequenced 20 samples from the patient with Illumina. In the paper, only Illumina data has now been formally analysed due to known poorer performance of ONT at low variant frequencies. The initial PHRED score is therefore applicable to the Illumina data (but the reviewer is correct that it is not the case with nanopore). For the consensus-level variant calling, consensus sequences were called using BCFtools (as detailed in the methods section). We used additional custom code to validate all variant calls – this has now been added to the methods section as well as the code availability clause. Both sets of code are available to reviewers on GitHub (https://github.com/PollockLaboratory/AnCovMulti; https://github.com/cjri/samfire/).

4. Great, thank you

5. In Fig 2A - much improved, but I'm curious why the different control patient cases seem so far separated from the case patient? Was the downsampling done to be representative of the region?

Response: Thank you, we are glad the figure is improved. Downsampling was done due to low number of non-duplicate sequences and a random selection was chosen using seqtk to provide background. The patients were from the same region though may have acquired infection from diverse sources.

6. Why is the scale according to AA mutations instead of nt mutations? For genetic epi isn't nt mutations as or more useful than the AA mutations? I understand looking at AA for selection/resistance purposes, but for clustering purposes it seems to me nt is more appropriate.

Response: Thank you for pointing this out, we have now rectified this.

7. So of these persistent shedders, it seems that only control 3 (XLA immune deficiency) had significant mutational changes over time? Why is that one different than the case, or than the other controls?

Response: the XLA was immune suppressed to a significant extent and was treated with RDV and CP at the end of shedding. The other controls were shedders over a shorter period and did not have significant immune suppression as indicated in supplementary table 2, potentially explaining the differences in mutations.

8. Supp Table 4 is extremely useful, but actually reveals more cause for concern in places. Consider for example the spike deletion (21763) - and the fact that on day 93 - one of the samples Illumina and nanopore conflict - Illumina detecting 0.0% and Nanopore 9.2%. There are quite significant differences in this table between the two technologies with the allele frequency methods they are calculating. I am _surprised by these differences when in non-homopolymer methods and still would like either A) stats on validation B) explanation for differences and/or C) more clear methods on how variants and consensus were calculated (see point 3). Both Thielen et al (medrxiv 2020) and Bull et al (Nat Commun 2020) examined Illumina vs nanopore carefully and found accurate consensus level data, so some of the number reported in this table are surprising.

Response: As mentioned above, we did not use nanopore sequences for formal analyses, due to widely known inferior performance of ONT at low variant frequencies. Originally we wanted for transparency to be able to present the nanopore data obtained as the data were already available via COG-UK. (We think the reviewer is referring to position 21768?)
We have provided B: an explanation for differences: As the reviewer points out the references cited show agreement in consensus level (>20%) rather than close correlation in variant frequency calling at <20%. Illumina is regarded as the gold standard for low frequency variant detection and that is why we ultimately used this for our analyses for detection of low frequency variants.

And C: We utilised a piece of custom code (detailed in methods + code availability) for variant calling. Consensus was inferred by BCF tools using Illumina read data.

Here is the text we have inserted into the methods:

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.

9. Focusing further on this point and again looking at day 93 - There are several Illumina/Nanopore mismatches on Day 93. Given the long branch lengths to Day 93 and other samples, these genomes should be examined more closely. If indeed it is an issue of jackpotting causing certain variants to display a higher frequency than others I would think that these mismatches should be interrogated a little more closely? It's especially surprising to me given that the Ct values for 93 were pretty reasonable (21).

Response: We are no longer considering the nanopore variants in formal analysis given known shortcomings of ONT for low frequency variants, please see comment above.

My read of the hypothesis of the authors is that the CP introduced a pressure to result in (at least) the deletion mutation). But I also don't see a reason for that deletion to be lost, if CP was an effective therapy - i.e. why the 69/70 del is gone at that point. The authors try to address this with their "compartmentalization" theory, but why wouldn't the CP have wiped out the non-mutated virus, precluding anything that doesn't have the del? In my mind this is still the sticking point.

Response: we can infer from the data that the mutants provided a selective advantage in the presence of CP for viruses sampled in N+T. This is clear from the re-emergence of the same population and the in vitro mutagenesis work. We are not surprised that the mutant failed to persist in between CP as the pressure had gone, thereby creating a new set of selective pressures. These pressures were likely different from pre CP time points as the patient was sicker with more inflammation. Hence we saw different mutants in between CP doses. It is formally possible that the del6970 mutant may have carried deleterious mutations in other parts of the genome making it less fit in the absence of CP. Finally CP was not potent enough as seen by lack of change in viral load so 'wiping out' unlikely and persistence of non-69/70 deleted virus is likely.

10. This is interesting, but I'm confused by the dates the authors selected? The 0s from days 1 and 37 are not especially interesting, I would argue that Day 98 is but I need a way of correlating the SGS result to the plot in Figure 4 - and/or what is preferable are times sampled in this active area around days 83-100?

Response: The SGA data can be correlated by comparing supp table 4 with supp table 5. We had no reason to mistrust the Illumina variant data at later complex time points, and we wanted to show that the Spike mutations observed were not present prior to CP e.g. due to mixed infection, hence the intensive sampling at earlier time points. We thought this was important to uphold the inference that the mutations were selected by the CP, though we admit that there is room for debate here. Indeed our SGA data also vindicate us at later time point (day 98) where SGA clearly demonstrates the high proportions of genomes with del69/70 + D796H and lower proportion of genomes with P330S/W64G and T240I/Y200H. SGA is excellent for unbiased amplification and is extremely labour intensive - we feel that our data show that the Illumina reflects actual viral population frequency and do not believe further SGA data are warranted. However, of course we would have ideally wanted SGA at every time point, including the latter ones.

11. The authors . . . . didn't really answer my point? I think the table helps, but did they address in the text or elsewhere the apparent loss and regain of certain mutations? I'm not asking them to rule in or out superinfection - and their argument in the discussion about "compartments" of virus or

different possible lineages is not implausible, but they need to be clear in the results that the nature of sampling precludes a clear picture of the virus in the patient.

Response: we have now made an explicit statement on the nature of sampling and limited inferences that can be made about virus populations in the text (lines 352-4).

12. Ok

13. The SGA data is nice, but the days selected were . . . inappropriate. I suggest it would have been more useful to probe in samples drawn from the "active" part of the viral timeline.

Response: we partly disagree here given the removal of nanopore low frequency variant data and high degree of confidence in Illumina based on literature and the comparison with SGA here. We chose early sample dates to verify that indeed there were not low level variants of our key spike mutations of interest that had been missed. Please see further details above.

14. Great.

15. I suppose that Supp Fig 8 is the figure they are referring to (Cambridge Sequences) - this is pretty useful and does in fact suggest that the patient, though their initial infection was from the community at large, is likely all derived from the initial infection.

Response: we have removed this following other reviewer comments

16. The pseudoviral figure has changed substantially, but I'm still confused at the significant spread in the points within a sample/condition(i.e. the replicates). Isn't it surprising that the spread is so large for example in the double mutant for CP1? Further, the authors assert that the "single mutant did not impact neutralization (255-256). But it _does_ seem to in the patient derived sera, as well as actually having a _stronger_ neutralizing effect in CP1?

Response: One has to remember that serum is a biological product polyclonal and subject to variation based on freeze thaw cycles amongst other factors. We also had to retrieve additional aliquots to complete the experiment. The data points are not from replicates but from separate experiments and different aliquots of stored CP where some variability is inevitable. Nonetheless the difference between D796H containing virus and WT is apparent, and a 2-3x effect against polyclonal sera from one mutation significant in the field and comparable to the effect seen with major RBD mutations. The reviewer is right to say that the single **Δ**69/70 mutant appears a little more sensitive to CP1 than WT and there may be an underlying structural reason for this, beyond the scope of the paper. We have now amended the sentence (line 255) and thank the reviewer for this point. The data are nonetheless consistent with our hypothesis that the role of the deletion lies in its enhancement of infectivity whilst D796H modestly reduces antibody neutralisation (whilst incurring significant infectivity defect). Nonetheless, we have toned down our claims regarding effect size throughout and modified the title.

17. Ok, it seems to have been edited in the results as well as the response, so this is fine.

Referee #4

The authors have addressed most of my comments (changed tree diagrams, removed artefactual SNPs, removed things I thought were distracting or not robust).

Response: we are pleased that these have been dealt with satisfactorily

However, the authors have also added more data in response to comments by the other referees, some of which seem inconsistent with the previous version of the manuscript or cast serious doubts on some inferences.

Response: whilst we have indeed added data to improve the paper, the inconsistencies have now been resolved.

The main additions are more data to disentangle the potential effects of D796H and the 69/70 deletion. The main message of the paper now is that 69/70 deletion is a fitness-enhancing or compensatory mutation while D796H is responsible for neutralization escape. Previously more emphasis was put on the 69/70 deletion. I am not convinced, however, that such specific statements can be made with the available data. 69/70 is a deletion that occurred repeatedly across the globe, suggesting that it is a mutation that is "easy" for the virus to discover. If this had a substantial effect on virus fitness, you'd expect that it also appeared without the D796H mutations and persisted between times CP2 and CP3 treatment.

Response: we take the point made by the reviewer and partly disagree. We have presented clear evidence that the deletion increases single round infectivity and that the S2 mutation reduces neutralisation, albeit modestly in single round experiments. We have nonetheless toned down the claims in the paper regarding 'escape' and also changed the title.

Whilst 69/70 deletions occur alone, their appearance of the deletion most commonly occurs with an RBD mutation globally e.g., N439K, Y453F, N501Y, suggestive of compensation. Of particular note, a new case has been reported in Russia of the deletion arising along with the RBD mutation Y453F that is unrelated to the mink cluster 5. We have kept our claims modest in the manuscript, however.

The fact that the double mutant disappears when antibodies wane is interesting, and we have now added some additional text (lines 362-4). We know that the virus with the deletion also has mutations in other regions that might alter its competitive advantage in a situation where the antibodies have disappeared. Furthermore this was a patient who was sick with significant inflammation – this could contribute to the balance of virus genotypes observed in nose and throat at particular time points. The critical finding is that the deletion and S2 mutant regain dominance following CP3.

Furthermore, the IgG levels from doses 1 and 2 don't change much between day 70 and day 98 and don't rise substantially after CP3 is administered making it less plausible that CPs are driving the evolution.

Response: Binding IgG levels do not equate to neutralisation activity (hence we do neutralisation assays) and therefore even low levels of neutralising antibody can drive adaptation over multiple rounds of replication with amplification of the mutation at each round.

Major:
- I am trying to link up data in the current figure 6 with the previous figure 5. You previously reported similar RT activity between WT and the double mutant (about 10% difference, previous Fig 5A) and substantially increased infectivity (+50%, previous Fig 5B). In the new figure 6B, the infectivity of the double mutant is slightly lower than WT. What changed?

Response: RT activity will vary between experiments and we showed the data as a representative example. In the revised paper we redid a whole new panel of experiments with individual and combined mutations and present all the data points to generate a mean. Inter-experimental variation can be of the order of 1.5x.

- Fig 6 C-G: no attempt is made to assess the statistical significance of the differences between genotypes. Looking at Fig S7 (the corresponding neutralization curves I think) suggests that the fits are pretty bad and often not only IC50 but the cooperativity seems to change. Some data points have wide CIs, others don't seem to have any CIs at all (some points are negative!!!). While these data are suggestive, I don't think they conclusively show that D796H or 69/70del have a consistent effect.

Response: It is standard when comparing nAb activity to consider fold change in 50% inhibitory values. This is evidenced by reviewing high-profile antibody characterisation literature in relation to SARS-CoV2 and the HIV bnAb field. As ID50s/IC50s are not means it is not common practice to use standard statistical tests designed for comparing means. It is also common to see non-perfect sigmoidal curves in neutralization assays due primarily to intrinsic viral heterogeneity as extensively discussed in our previous work for HIV (https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1005110). As such, negative values do occur and reflect the cell-based nature of the assays. Many groups choose to leave out curves altogether and present tables with no statistics e.g. Avanzato Cell 2020. We clearly show that there is a reproducible effect of D796H and that 69/70del alone does not impact neutralisation. In addition the monoclonal antibody figure clearly shows that RBD directed antibodies are equally effective against all mutants, demonstrating the specificity of our findings.

- https://github.com/Steven-Kemp/sequence_files/tree/main/figure_data is empty

Response: This has now been populated. Thank you for pointing this out.

- line 188: do you have any evidence that this is a "driven" sweep?

Response: We have modified the use of the term ‘driven’

- line 193: is this "sweep" any more dramatic than the one described in the previous paragraph?

Response: we have removed the term dramatic.

- line 235: Fig 4 suggests that NSP2:I513T and 69/70del are simultaneously at very high frequency at days 98 and 99, directly contradicting the statements that they are in opposing clades. Overall, these data don't seem to match up with what is reported in the supplementary data. The deletion (position 21768) is at low frequency in the table for days 99-101 but at high frequency in Fig 4.

The discrepancy in frequencies was due to there being samples from the same day from two different sampling sites. We have now resolved this by keeping the sampling site consistent for variant analysis purposes (nose and throat samples). The table and figure are now consistent with each other as a result.

- The supplementary table with the frequency data and Fig S5 suggests that SNP frequencies are often very discordant between nanopore and Illumina data. Although the caption of Fig S5 states

"There was good concordance for mutations between the two methods, and no significant difference between the proportion of reads measured by both methods." the data are all over the place and either the plot or the caption is completely wrong. If the frequencies are really as unreliable as the plot suggests, none of the frequency dynamics in the paper can be trusted.

Response: We apologise for the confusion – the caption was indeed misleading in that regard and has been modified. Nanopore is well known to be less reliable than illumina at low frequencies and is therefore was only left in the paper for transparency – all formal analysis used only Illumina data. We limited use of nanopore for confirming the consensus sequence inference from illumina, as supported by Bull et al Nature Comms, 2020. We had aimed to use 2 methods to provide reassurance and demonstration of a rigorous approach at the outset, but we should have taken into account the limitations of nanopore at the outset and directly used the gold standard method for low frequency variants.

Minor:

- Fig 2: is the scale bare one amino acid or one mutation? I think it should be one mutation.

Response: it should be mutation thank you for pointing this out.

- Fig 4C: not sure this is useful. The nature of this deletion has been discussed a lot and showing ten genomes doesn't really add information. ALL 69/70 deletions in GISAID are like this. Maybe better point out that codon 68 changes from ATA to ATC with this deletion.

Response: we had put this figure in due to another reviewer request but agree it could be removed. We have added text to say that codon 68 changes from ATA to ATC with this deletion

- Fig 5C-E: please show on logscale with consistent y-axis between panels.
Response: We have now used log scales and made the y axis consistent.

- Fig 7: I would clarify which Spike mutations in the figure title.

Response: we have now done this

- Fig S7: all three panels in S7 have different axis labels.

Response: this has now been harmonised

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Authors have tone down strong statements and the overall re-edited new version of the manuscript is improved. While the Authors have not fulfilled all the requests from the last round of revision it is clear the importance of reporting soon on these new findings.

Referee #3 (Remarks to the Author):

The authors have addressed all my concerned, especially with their additions of limitations/caveats in the discussion and their explicit focus on the Illumina data.

Referee #4 (Remarks to the Author):

Most of my points have been addressed. The focus on consistent sampling sites and only one sequencing methodology make the results more interpretable. But the discrepancies between technologies are still a source of concern. The degree to which the rapid shifts in frequencies are driven by alternating dominance of spatially structured population as opposed to adaptation and selection by CP remains unclear. In that light, moderation of the article's claims is welcome.

Author Rebuttals to Second Revision:

Response to reviewers

Referee #1 (Remarks to the Author):

Authors have tone down strong statements and the overall re-edited new version of the manuscript is improved. While the Authors have not fulfilled all the requests from the last round of revision it is clear the importance of reporting soon on these new findings.

Response : We have done as much as we could and thank the reviewer

Referee #3 (Remarks to the Author):

The authors have addressed all my concerned, especially with their additions of limitations/caveats in the discussion and their explicit focus on the Illumina data.

Response : we thank the reviewer

Referee #4 (Remarks to the Author):

Most of my points have been addressed. The focus on consistent sampling sites and only one sequencing methodology make the results more interpretable. But the discrepancies between technologies are still a source of concern. The degree to which the rapid shifts in frequencies are driven by alternating dominance of spatially structured population as opposed to adaptation and selection by CP remains unclear. In that light, moderation of the article's claims is welcome.

Response: we thank the reviewer and have suitably moderated the claims at the previous round of reviews.

Ravi Gupta