Reviewer comments:

Reviewer's Responses to Questions

Summary and Strengths

Summary of the main conceptual contribution(s) and strengths of the paper.

Reviewer #1: Please see reviewer comments

Reviewer #2: This paper considers evidence for adaptive evolution of SARS-CoV-2 over the short period of time since it entered the human population. It proposes a method to detect this, by extending a familiar approach when handling more distantly related taxa to shorter periods. This is interesting and worthwhile. We see extremely clear evidence that over time the Spike gene (esp S1) has accumulated more non-synonymous variation than would be expected given rates in the rest of the genome, that this has been correlated with the most successful clades, has been increasing over time, and has often involved convergent evolution. A deletion in nsp6 (ORF1a:3675-3677del) is especially interesting, having been involved in multiple successful lineages. This points to some mutations being especially important.

Support for Conclusions

Assessment of whether the current experiments and analysis support the conclusions of the paper, noting any technical concerns with the experiments/ data analysis and how this impacts the conclusions. Please highlight key issues clearly or indicate if there are no major concerns.

Reviewer #1: Please see reviewer comments

Reviewer #2: The paper claims that these findings reflect adaptive evolution. And certainly some of the changes in spike are significant and have been selected. But it does not mean that _all_ of them have been. The dynamics of the changing viral population over the last year are pivotal to this work. Starting in roughly November 2020 multiple distinct lineages, since termed ‘variants’, emerged as more transmissible and, more contentiously, better able to evade prior immunity. Many ‘variants’ have been identified (prior to the global selective sweep of Delta). Of these the most serious were Alpha (B.1.1.7), Gamma (P.1) and Lambda. These both contained an excess of mutations in comparison with other circulating lineages. They are widely considered to have evolved in an unsampled, different evolutionary context such as a long-term infection in an immunocompromised host. Indeed, several of the mutations held up as convergent evolution in this paper have also been observed in such patients.

We agree that not all observed mutations in S1 have been the result of positive selection at the between-host level. As the reviewer notes, some S1 substitutions likely arose within long-term infections due to a selective advantage within-host and, within the global population of viruses, are hitchhiking with other mutations that are advantageous between-host. Therefore, it is certainly possible (and likely) that some successful clades contain some S1 substitutions that are advantageous within a single host and

evolutionarily-neutral (or slightly deleterious) at the between-host level. However, this does not alter the finding that, on average, viruses with more S1 substitutions are more successful.

We have added the following paragraph to the Discussion to clarify this:

> **"Our inference of adaptive evolution is based on a correlation between S1 substitution accumulation and clade success that falls well outside the null expectation (Figure 1C). It is important to emphasize that these results speak to the average evolutionary effect of S1 substitutions. This does not mean that every S1 substitution is selectively advantageous, and it is likely that some mutations have larger effects on fitness than others. In fact, it is possible that successful viruses contain some S1 substitutions that do not contribute at all to their evolutionary success. These hitchhiking mutations could have arisen during long-term infections where they were advantageous within a single host. For instance, S1 mutations 484K and 501Y have been observed to arise from continued evolution within a single host [20]. It is therefore possible that the parallel evolution of these particular mutations is due to a selective advantage at a within-host, rather than between-host, level. Phylogenetically, hitchhiking mutations that are beneficial within-host are especially likely to be present on long branches with many mutations. We observe that the majority of S1 mutation accumulation occurs on just one or two branches preceding some VOCs (like Gamma and Kappa), but is spread over many branches leading to others (like Alpha and Beta) (Figure S6). However, the context in which S1 mutations accumulated does not affect our finding that, on average, viruses with more nonsynonymous mutations in S1 are more successful within the global population of SARS-CoV-2 viruses."**

So far as I can see, the results of the current work are the consequence of the success of these variants. They will display 'clade success' as measured by the current metric, as they cause a larger proportion of cases over time, over a greater geographical area. This leads to the overall findings of the paper, that those successful lineages have had an excess of non synonymous mutations in Spike. The difficulty I have with the paper overall and its interpretation is the causal role attributed to this. These variants are successful AND they have an excess of mutations. It is not necessarily so that they are successful BECAUSE of that excess. In fact there are several lineages with a comparatively divergent Spike that have displayed much less capacity for transmission (Beta B.1.351 springs to mind). The evidence for convergent evolution certainly does suggest importance for certain of these, such as N501Y or E484K but these are clearly not sufficient as E484K has arisen multiple times (including on backgrounds with other advantages like Alpha, but where this has occurred the resulting viruses have not been notably successful). And the simple observations of convergent evolution have been made by many other investigators, limiting the nature and scale of the advance.

We agree with the reviewer that successful variants are driving the results we present in this paper. In fact, the aim of this paper is to more rigorously investigate the genomic distribution of mutations possessed by successful variants (clades) that avoids the 1-off specific variant investigations that have predominated in the literature. We have added the following sentence to the introduction to stress this point:

> **"With this method, we aim to present a rigorous quantification of the evolutionary process during this time and to show that the observed success of variant viruses is a result of adaptive, not neutral, evolution."**

We also agree with the reviewer that the results we present are correlations. This is in keeping with other tests of adaptive evolution (like dN/dS and McDonald-Kreitman) that have been applied throughout the field of population genetics to show evidence of adaptive evolution in Drosophila, influenza, etc. A true investigation of causation is very difficult because it requires laboratory experiments, which do not look directly at the natural evolution of the virus and come with a huge number of caveats (model system, feasibility of reproducing observed viral diversity, etc). So, we agree that correlation is not causation, but we view a scenario of S1 substitutions increasing virus fitness as the most parsimonious explanation for our results. We also note that the origin of these S1 substitutions still could be due to chronic infections and within-host pressure. The origin of these mutations is tangential to the observed increase in logistic growth of clade's bearing, on average, more S1 mutations than their competitors.

Also, as the reviewer notes, Beta is an example of a variant with many S1 mutations that is not as fit as would be expected based on the number of mutations. This could be due to a variety of possible factors including presence of deleterious mutations or competition with other more fit variants. Nevertheless, the finding that S1 substitution count correlates with clade success greatly exceeds the null expectation. This says that, *on average*, viruses with more S1 substitutions are more successful. Our manuscript does not aim to predict or explain the fitness of specific variants, but rather, to quantify and describe the evolution of the population of SARS-CoV-2 viruses as a whole.

The comments that these convergent changes have "giving rise to successful viral clades each time" are not borne out by the data. Some of these mutations, discussed above, have happened but not reliably produced a successful strain. This is an interesting question, which I would like to see discussed more. It is certainly true that N501Y, E484K and del69/70 seem important -especially in combination – but each alone is not sufficient.

The reviewer is correct that the effect we see is that some mutations occur in more successful viral clades, on average. We have removed the phrase "giving rise to successful viral clades each time". There are a variety of factors that could influence why a clade with a given mutation is more successful than another clade with the same mutation, and our analysis does not aim to differentiate between them. We acknowledge this in the Results:

> **“The specific mutations identified by this analysis will vary over time and depend on a multitude of factors (genetic, epidemiological, and otherwise) that determine clade success.”**

And in the Discussion:

> **“The fitness effect of a mutation is not an absolute quality – it depends on a multitude of influences including genetic background of the viral lineage, other co-circulating lineages, existing host immunity, and epidemiological factors (such as geographically heterogeneous mitigation efforts).”**

It is not clear to me how much the acceleration in figure 2 is due to the emergence of the variants, and how much to accelerating dN within them (and remnant founder virus populations). The former would be not so interesting, but the latter very much so. I am painfully aware that given the closely related nature of the samples, the emergence of some lineages is quite poorly resolved (B.1.526 for instance) and it is not clear from the text how this could impact the results (does this impact what is an internal branch, or how many descending tips they have?).

We calculate accumulation of dN/dS on internal branches, so the increasing ratio in S1 indicates that viruses are becoming increasingly diverged from the root and that this divergence is enriched in nonsynonymous change. This is mostly due to emergence of the variants, and less so to accelerating dN within them. We have updated the Results and Methods text to clarify how the dN/dS calculations are done. We have also added panel A to Figure 2, which diagrams the branches on which divergence accumulation is computed to further explain the methodology.

We feel that quantifying selective pressures on SARS-CoV-2 is largely lacking in the literature. Because of this, we think the result of dN/dS>1 in S1 is important to present. As a better point of comparison, we have also updated Figure S5 to present the exact same analysis for H3N2, H1N1pdm and OC43, beginning in 2009 when H1N1pdm emerged. This shows that H1N1pdm HA1 also had an elevated dN/dS ratio in years directly after its emergence, indicating that this may be a common phenomenon in the receptor-binding subunit of a virus after switching hosts. However, in H1N1pdm, dN/dS reached a maximum of 0.72, which emphasizes how remarkable the dN/dS of 1.8 that we observe in SARS-CoV-2 S1 is.

The paper is careful (and impressively so) to resist easy conclusions, but I want to suggest that even more caution is applied to whether these properties reflect selection for evasion of neutralizing antibodies in the population (distinct from nAbs in a long term infection treated with monoclonals). Further, this is now remarkably of historical interest. The variants in question have, all of them, been outcompeted by Delta, which has a MRCA at around the same time as Alpha (roughly), but which does not contain a markedly more divergent Spike to explain why it has driven Alpha close to extinction. And it is good that the discussion notes that while Delta

exhibits some of the mutations previously noted as important, it certainly doesn't have those that were considered most concerning over the pandemic up to its emergence. In other words the statement that some mutations appear more important than others is fair, but we knew that already and the emergence of delta makes it clear than we should be cautious about taking what we had seen in spike up to that point and assuming it would continue in future.

We have removed the reference to immune evasion that was in the introduction and have updated the paragraph in the Discussion that talks about selective pressures:

> **"Together, the results presented in Figures 1-3 offer phylogenetic evidence that SARS-CoV- 2 is evolving adaptively and that the primary locus of this adaptation is in S1. Our results are consistent with experimental demonstration of phenotypic changes conferred by VOC spike mutations [14,16,18,20]. Adaptive evolution in the S1 subunit during the period we focus on (December 2019 to March 2021) is likely driven by selection to adapt to a new host by increasing infectivity of human cells. However, the amount of immunity to SARS-CoV-2 is rising globally, increasing the selection for antibody escape. Given the virus's demonstrated propensity for adaptive change in S1, antigenic drift will likely begin to sculpt the evolution of SARS-CoV-2. The potential antigenic impact of adaptive S1 mutations, which are accruing at pace over 4 times that of influenza H3N2 (Figure 2, Figure S5), suggests that it may become necessary to update the SARS-CoV-2 vaccine strain."**

We agree that some specific mutations are more important than others and that their importance will vary over the continued evolution of the virus. We state this in the Results and the Discussion:

> **"The fitness effect of a mutation is not an absolute quality – it depends on a multitude of influences including genetic background of the viral lineage, other co-circulating lineages, existing host immunity, and epidemiological factors (such as geographically heterogeneous mitigation efforts)."**

We also agree that the pace of evolution we observe during the time period considered in this paper (Dec 2019 to May 2021) is not necessarily the pace at which evolution will continue. We state explain this in the Discussion:

> **"An initially high rate of protein-coding changes is consistent with the idea that, soon after a spillover event, there are many evolutionarily-accessible mutations that are advantageous in the new host environment. This was observed in the influenza H1N1 pandemic virus (H1N1pdm). Su et al [23] report that H1N1pdm had elevated genome-wide dN/dS rates for 2 years following its emergence in 2009, and evolution during this period is thought to largely have been adaptation to a new host, including increased transmission in humans. From 2011 onward, the adaptive evolution of H1N1pdm has been dominated by antigenic changes [23]. In agreement with these results, we observe that the accumulation of dN/dS in the HA1 subunit of H1N1pdm peaks around a year after the virus' emergence at 0.72,**

roughly twice the mean dN/dS ratio between 2009 and 2021 (Figure S5). It is possible that SARS-CoV-2 is following a similar trajectory of adaptive evolution, with initially high dN/dS due to host adaptation to be followed by sustained antigenic drift."

Additionally, we have updated the Discussion to state that the high number of S1 substitutions present in Omicron is consistent with our presentation of the importance of S1 substitutions in the adaptive evolution of SARS-CoV-2 (a conclusion gathered from previous lineages).

> **"Indeed, the emergence of the Omicron Variant of Concern demonstrated a SARS-CoV-2 virus with an extraordinarily high number of S1 substitutions [25] that spread rapidly across the world and showed significantly reduced neutralization titers relative to preceding variants [26]. With Omicron, we now know that significant antigenic variants can emerge with highly modified S1 domains. However, the observed pace of adaptive evolution in S1 perhaps should have suggested the potential for emergence of such a variant."**

The comparison with HA from influenza should, in my view, be removed. This is a different gene in a different virus, and so putting it in here is misleading. A comparison with eg OC43 would be worthwhile, but if the sequences are not available to permit it, that doesn't mean it is reasonable to use flu as an alternative. It is certainly not an "equivalent subunit". Further, while much is made of the observation that dN/dS in spike is higher than HA in flu, less is made of the fact that the rdrp gene shows a dN/dS almost as high as a major surface antigen in another virus!

We have removed the line showing average dN/dS for H3N2 HA1 from Figure 2. However, we do think it is important to provide some point of reference for the "accumulation of dN/dS" calculations we present in Figure 2. In other words, we would like to give the reader context to answer: where does dN/dS of ~1.8 stand relative to the textbook example of viral adaptive evolution? To this end, we have revised Figure S5 to include the same calculation of dN/dS within the HA1 subunit of H3N2 and H1N1pdm, and the S1 subunit of seasonal coronavirus OC43. Figure S5 presents this ratio over time (as in done in Figure 2 for SARS-CoV-2) starting in January 2009 in order to capture the period of time just after H1N1pdm emerged. This offers an interesting point of comparison for the high dN/dS values observed in Figure 2. Figure S5 shows that dN/dS was elevated within the HA1 subunit of H1N1pdm for a couple years after its emergence, peaking roughly a year after. The peak H1N1pdm dN/dS was roughly twice the average dN/dS in the following years, but still only reached 0.7. This is in contrast to a peak dN/dS of 1.8 in SARS-CoV-2 S1.

Recommendations for revisions

Summary of revisions that would be needed for publication in particular journals, highlighting whether the requests are needed to extend the current conclusions or to support the main points of the study as it stands.

Reviewer #1: Please see reviewer comments

Reviewer #2: I would like to see a thorough examination of how the sampling approach might skew conclusions. It is really hard to do this right. The authors should also consider making it clear how much of the signal is due to a few variants with excess dN in spike becoming much more common, and how much there is evidence of dN/dS accelerating _within_ those variants. A subanalysis of a well sampled location might be a possible route to this.

We agree that sampling bias can be a huge issue for phylogenetic analyses and we have completed 3 additional analyses to assess sampling bias and added a paragraph to the Discussion to talk about this. Figure S10 addresses the issue of observing skewed results due to too few samples. In Figure S10, the results presented in the primary figures are repeated using a tree with twice as many samples (n = 19,694), giving very similar results. Figure S11 addresses the fact that, at a continental level, different lineages have circulated during 2021. The main text and figures present results from phylogeny that contains a roughly equal number of samples from each geographic region over time. For Figure S11, we built separate phylogenies for each geographic region (Africa, Asia, Europe, North America, Oceania, and South America), each containing about 10,000 samples. Figure S11 shows that the correlation between S1 substitution count and clade growth is not solely driven by a specific region. Figure S12 considers how consistent results we present are over time. For this figure, we built 13 phylogenies with end dates spanning a year of time from Nov 15, 2020 to Nov 15, 2021. Because we calculate logistic growth rate during the 6 weeks preceding the end date of the tree, this allows us to look at how the correlation between S1 substitutions and growth rate changes over time. Figure S11 presents the correlation coefficient *r* for each of these time points, showing that the correlation is highest between January 2021 and September 2021. The new paragraphs in the Discussion describing this read:

> **“Phylogenetic inferences of evolution can be biased by the samples included in the analysis. To reduce sampling biases, our study is based on a phylogeny of 9544 SARS-CoV-2 genomes sampled evenly over space and time. The strong correlation between S1 mutation accumulation and clade growth rate persists if the number of genomes included in the phylogeny is doubled (Figure S10), indicating that our results are not biased by the number of samples included in the analysis. We also find that global adaptive evolution in S1 is not driven solely by certain geographic regions. Using phylogenies that only include samples from a particular geographic region, we observe that clade success strongly correlates with S1 substitutions in Asia, Europe, North America, Oceania and South America (Figure S11). The only region where this correlation is not observed at**

this timepoint is Africa, where decline in frequency of a particular clade of Beta drives an overall lack of correlation.

We observe temporal structure in the adaptive evolution of SARS-CoV-2. We find that the correlation between clade success and S1 substitutions changes over time, though shows strong signals of adaptive evolution from Jan to Sep 2021 (Figure S12). Enrichment of the ratio of nonsynonymous to synonymous divergence (d_N/d_S) in S1 also increases over time (Figure 2). Additionally, substitutions within S1 cluster temporally (Figure 3), rather than accruing at a steady rate. This temporal structure potentially indicates a changing evolutionary landscape: either through the emergence of new selective pressure, and/or through the occurrence of permissive mutations that made adaptive mutations more accessible."

We agree with the reviewer that an analysis of within-lineage evolution would be very interesting. We have done a cursory analysis of this type and see that most divergence accumulation precedes lineage emergence, and that dN/dS is actually fairly flat within-lineage. A proper examination of within-lineage evolution would require careful consideration of sampling strategies, and is beyond the scope of this paper.

Presentation and interpretation

Feedback on the scholarship of the manuscript, clarity of the figures, and whether the conclusions are appropriately contextualized and discussed, including whether a specific "limitations" section is recommended.

Reviewer #1: Please see reviewer comments

Reviewer #2: There is no specific limitations section. The figures are in many cases very hard to understand as a result of similar colors being used for different clades. Labels might improve this. I think that the conclusions could be improved by noting questions about the origin of variants, and that if they spent time in a different evolutionary context (long term infection, maybe another species) before starting to transmit H2H again this might explain the different dN/dS. They are certainly selected. And some mutations are being selected, but the dN/dS is not causal and instead correlated with the fact that the different evolutionary context allows the virus to reach a fitness peak for H2H transmission.

We have changed the color scheme in Figure 1 and supplements that use colors to identify different clades. The color scheme now has a color for each of the prominent VOCs Alpha, Beta, Delta and Gamma and groups all other VOIs into another color.

We have added the following paragraph to the Discussion, which lists limitations of this study:

“In addition to sampling biases, there are several limitations to our approach presented here. Firstly, our analysis intentionally considers the average effect of mutations in different regions of the genome on viral fitness, with the goal of taking a populations genetics approach to quantify adaptive evolution of SARS-CoV-2. This means that, while we observe a significant correlation, given a correlation coefficient of r = 0.46, our results cannot predict the fitness of specific variants solely based on S1 mutation counts. Similarly, as mentioned above, this means it is likely that some successful viral clades contain S1 substitutions that are not advantageous, but rather, are hitchhiking along with positively- selected mutations. Additionally, our analysis focuses on the period of VOC and VOI emergence from December 2019 to May 2021. So, while we can speculate how our findings of high adaptive potential in S1 will translate to future evolution of the virus, we cannot directly predict how the pace of adaptive evolution will change over time.”

Transparency of reporting

Degree to which manuscript is transparent about data sharing and methods highlighting concerns on resource availability upon publication and reproducibility of study.

Reviewer #1: Please see reviewer comments

Reviewer #2: Excellent

GENERAL COMMENTS

Reviewer #1: This is an interesting paper on aspects of adaptive evolution in SARS-CoV-2. Although some of this story has been told before - particular the strong selection on S1 and the widespread convergence - there were some novel aspects, particularly the convergent evolution of the three amino acid deletion in nsp6.

I do have a number of technical issues/comments. In no particular order.

1. Although comparison to H3N2 influenza virus is useful is also a little misleading as this virus has been circulating in humans for >50 years. Could a temporal comparison be done on this

virus, perhaps taking 5 year chunks? Was the rate of adaptation highest in the early years (e.g. 1968-1972)?

This is a good point. We have revised Figure S5 to offer a comparison between H3N2, H1N1pdm, and seasonal coronavirus OC43 between the years 2009 and 2021. At this point, H3N2 and OC43 had both been circulating in humans for over 50 years and both appear to have roughly consistent dN/dS ratios with an average of 0.36 (H3N2) and 0.48 (OC43). However, H1N1pdm emerged in humans in 2009 and Figure S5 shows that dN/dS was elevated within the HA1 subunit of H1N1pdm for a couple years after its emergence, peaking roughly a year after. The peak H1N1pdm dN/dS was roughly twice the average dN/dS in the following years, but still only reached 0.7. This is in contrast to a peak dN/dS of 1.8 in SARS-CoV-2 S1.

2. The authors state that the adaptive pattens they describe reflect adaptation to "a partially immune host population". But I don't see any evidence for immune selection and the increase in the adaptive rate in S1 clearly occurred when most of the global population were still naive to the virus. Just because most adaptation occurs in S1 does not mean that it is due to immune selection. Indeed, the D614G mutation swept globally early in the epidemic without any evidence of immune selection. Also, the most antigenically distinct strain - beta- was clearly outcompeted by that with the greatest infectivity - delta. Both these VOCs had many spike mutations. So, the statements on immune selection should be toned down.

Thank you for pointing this out- we have removed the mention of "partially immune host population" from the introduction. We have also revised the paragraph in the Discussion that mentions selective pressures acting on S1 mutations to emphasize that adaptive evolution during the time period we consider (Dec 2019 to May 2021) is likely driven by adaptation to a new host, but that the adaptive potential of S1 suggests that, as immunity mounts, selection to evade antibodies will likely result in antigenic drift. That paragraph now reads:

> **"Together, the results presented in Figures 1-3 offer phylogenetic evidence that SARS-CoV- 2 is evolving adaptively and that the primary locus of this adaptation is in S1. Our results are consistent with experimental demonstration of phenotypic changes conferred by VOC spike mutations [15,17,19,21]. Adaptive evolution in the S1 subunit during the period we focus on (December 2019 to March 2021) is likely driven by selection to adapt to a new host by increasing infectivity of human cells. However, the amount of immunity to SARS-CoV-2 is rising globally, increasing the selection for antibody escape. Given the virus's demonstrated propensity for adaptive change in S1, antigenic drift will likely begin to sculpt the evolution of SARS-CoV-2. Given the virus's demonstrated propensity for adaptive change in S1, antigenic drift will likely begin to sculpt the evolution of SARS-CoV-2. The potential antigenic impact of adaptive S1 mutations, which are accruing at pace over 4 times that of influenza H3N2 (Figure 2, Figure S5), suggests that it may become necessary to update the SARS-CoV-2 vaccine strain.**

Indeed, the emergence of the Omicron Variant of Concern demonstrated a SARS-CoV-2 virus with an extraordinarily high number of S1 substitutions [26] that spread rapidly across the world and showed significantly reduced neutralization titers relative to preceding variants [27]. With Omicron, we now know that significant antigenic variants can emerge with highly modified S1 domains. However, the observed pace of adaptive evolution in S1 perhaps should have suggested the potential for emergence of such a variant."

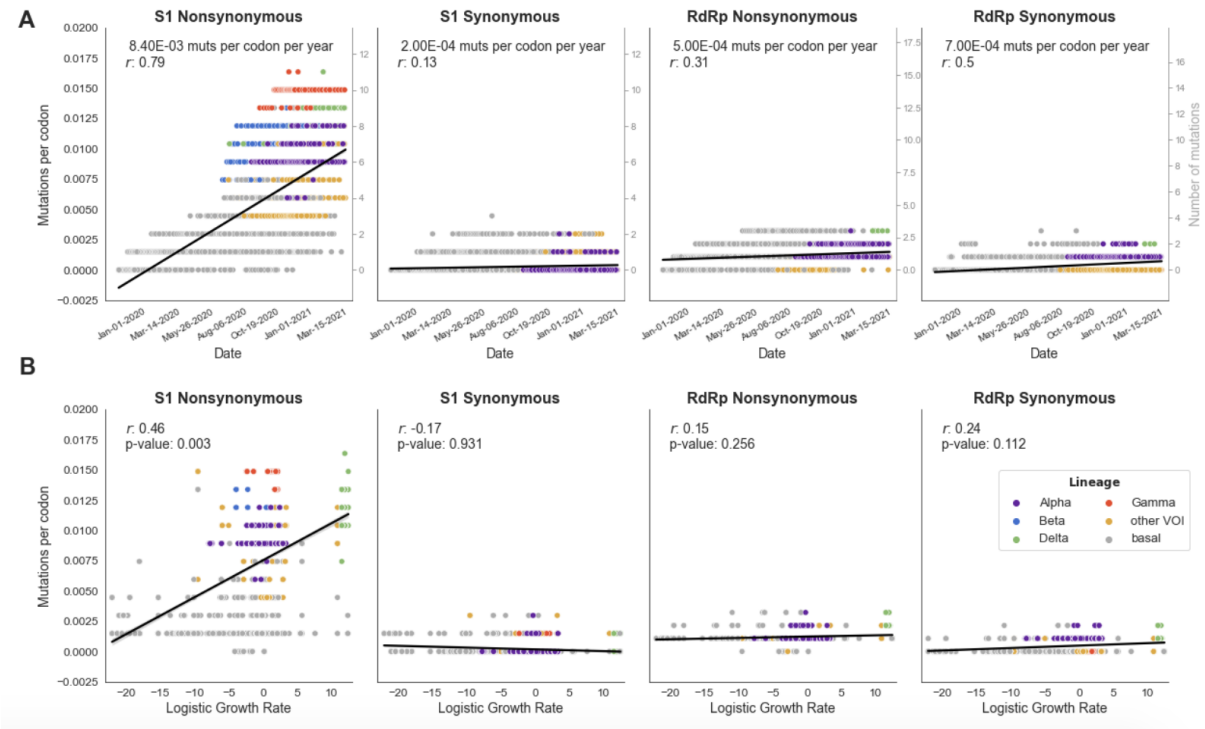
3. Is there are proper reference for substitution rate in SARS-CoV-2 rather than a Nextstrain link.

We have added a citation to Hadfield et al. The substitution rate given is taken directly from the tree at this link, and we think it is useful for the reader to see this if they want to, so we have also kept the Nextstrain link. It's unfortunate that there isn't a great canonical reference for SARS-CoV-2 evolutionary rate. There are early Virological.org posts that dive into this and there are papers that derive a rate for their own purposes, but there isn't a great reference (to our knowledge) for the substitution rate as observed over >1 year of evolution.

4. Figure 1. I think it would be useful to add RdRp synonymous. I'd like to see if the synonymous rate is constant across the genome. This raises a broader question: wouldn't you expect there to be some linkage in rates? If RdRp nonsynonymous is wouldn't RdRp synonymous be high as well because there in the same codons. Similarly, given the low rate of recombination in SARS-CoV-2 genome, if S nonsynonymous is increased by adaptive evolution wouldn't you expect higher rates in linked genes? Is that not a signature of selective sweeps?

Because the SARS-CoV-2 genome is linear, all genes are linked to S. However, this does not mean that we would expect different rates of neutral mutation in genomes that have positively-selected S1 mutations versus genomes that do not. This is often referred to in population genetics as "genetic draft" (summarized in Chapter 4.3 of *Population Genetics* by John Gillespie). The reasoning is as follows: a neutral mutation A has the probability of fixing equal to its frequency in the population (let's say it's at 30% current frequency). Because mutations occur at random, the probability of an adaptive mutation B co-occurring with mutation A is equal to the frequency of A in the population (30%). Therefore, in the selective sweep that drives B to fixation, there is a 30% probability that A hitchhikes along with it. Therefore, the presence of an adaptive mutation does not alter the probability of fixation of a neutral mutation and therefore divergence of neutral variation is not sped up or slowed down by adaptive evolution at linked loci (though polymorphism is affected).

Below we have included Figure 1 including RdRp synonymous mutations. The mutation rate is similar to that of S1 synonymous and the correlation between RdRp synonymous and clade growth rate is not significant.



5. I also wonder if the selection estimates are in part reflecting the presence of transient deleterious mutations, that by definition increase in frequency toward the present? I note that the dn/ds values are very close to the neutral expectation. I therefore wonder if it is worth comparing dn/ds on tips versus internal branches phylogenetic trees?

The dN/dS analysis presented considers only internal branches, so deleterious mutations should be diminished and should contribute relatively little to dN. We calculate accumulation of dN/dS, so the increasing ratio in S1 indicates that viruses are becoming increasingly diverged from the root and that this divergence is enriched in nonsynonymous change. We have updated Figure S5 to present the exact same analysis for H3N2, H1N1pdm and OC43, beginning in 2009 when H1N1pdm emerged. This shows that H1N1pdm HA1 also had an elevated dN/dS ratio in years directly after its emergence, indicating that this may be a common phenomenon in the receptor-binding subunit of a virus after switching hosts. However, in H1N1pdm, dN/dS reached a maximum of 0.72, which emphasizes how remarkable the dN/dS of 1.8 that we observe in SARS-CoV-2 S1 is.

We have updated the Results and Methods text to clarify how the dN/dS calculations are done. We have also added panel A to Figure 2, which diagrams the branches on which divergence accumulation is computed to further explain the methodology.

6. Figure 2. Are the 95% confidence intervals really that narrow?

We have double checked the 95% confidence intervals and they are correct. To clarify what is being plotted in Figure 2: for each time window (spanning 2 months of time) accumulation of divergence is calculated on all internal branches within the window. After dividing nonsynonymous divergence by synonymous divergence for each gene, this results in a per-gene dN/dS for every internal branch within the time window. The mean number of branches in a time window is 978 (range: 121, 2155). The mean dN/dS of all branches in a time window is plotted. The 95% CI is calculated according to the equation: mean +/- t(std_dev - sample_size), where t is the t-value corresponding to 95% confidence level.

7. I'm not sure I agree with this statement "clades that happened to accumulate more S1 substitutions should have, on average, higher fitness (and hence faster growth in frequency) than clades that have accumulated fewer S1 substitutions". I believe that the gamma VOC has more S mutations than delta, but the latter is clearly of far higher fitness. I strongly believe that fitness depends on the specific mutation, rather than the total number.

This is a good point: the reviewer is correct that different mutations have different fitness effects, and even that these fitness effects depend on genetic context. However, the observation is that lineages are accumulating many S1 substitutions, and so our hypothesis is that *on average* clades with more S1 substitutions have higher fitness. This general phenomenon of S1 substitutions correlating with fitness indicates that at least some of these mutations are adaptive, it does not mean all of them are, nor does it mean S1 substitution count is an absolute predictor of evolutionary success.

We have inserted a sentence into the relevant Results paragraph to emphasize that this correlation is on average, and that different mutations have different fitness effects.

> **“We hypothesize that adaptive evolution is driving the high rate of S1 nonsynonymous substitutions relative to S1 synonymous substitutions and RdRp nonsynonymous substitutions. And, though each S1 substitution will have a different effect on fitness, this observation suggests that this class of mutations is, on average, under positive selection. If this is the case, we would expect a correlation between S1 substitutions and a clade’s evolutionary success: clades that happened to accumulate more S1 substitutions should have, on average, higher fitness (and hence faster growth in frequency) than clades that have accumulated fewer S1 substitutions.”**

8. The clade success measure is interesting, but I wonder if it is adversely impacted by demography. This was the problem with the old McDonald-Kreitman test if I recall - it assumed a constant population. How does population growth impact this measure? I understand that it would impact the results in a major way because it uses a gene-by-gene comparisons, but wonder how robust this measure is in growing populations, or worse when comparing populations characterised by very different demographic structures? Finally, given clades are nested by definition which clades were used to define success?

Although McDonald-Kreitman tests are substantially more robust to demographic influences compared to tests like Tajima's D that focus on distribution of polymorphism, we were also concerned about the potential for demographic influences in our analysis. That said, we believe that the observation of r = -0.17 (p=0.93) for synonymous sites at S1 should be convincing that demographic influences are not creating false positive results. Synonymous mutations fall across the tree and polymorphism, divergence and their association with growing clades should be driven by demography.

To the reviewer's question about clade nesting: Figure1A shows all clades, and Figure1B shows all clades within 6 weeks of the end date of the analysis (2021-05-15), which is the time period we are calculating growth rates for. It is true that clades are nested and this phylogenetic nonindependence makes it tricky to decide what level of nesting to include in analyses like this. To deal with this, we opted to use all clades and create a null expectation based on this decision. The null expectation uses the exact tree topology as the empirical phylogeny (including the same growth rates associated with each clade) but randomizes the positions of all observed mutations across this phylogeny. Correlation coefficients are then computed from 1000 randomized phylogenies in the exact same way as is done for the empirical data (using all nested clades). In this way, we form an expectation of the correlation coefficient r that would be seen if mutations and growth rate are not actually correlated, given the observed phylogeny, growth rates, and number of mutations.

9. Small style thing: I think it is bad practice to write the results in the last paragraph of the Introduction. That's what the Abstract is for. Indeed, the Abstract in this paper is vague and should be improved.

We have changed the last paragraph of the introduction exclude results:

> **"With this method, we aim to present a rigorous quantification of the evolutionary process during this time and to show that the observed success of variant viruses is a result of adaptive, not neutral, evolution. We conduct these analyses across the SARS-CoV-2 genome to identify foci of adaptive evolution. We complement these results with analyses of d_N/d_S accumulation, evolutionary dynamics, and convergent evolution to provide evidence that genetic changes are contributing to viral fitness and identify genomic regions that are responsible. "**

We have revised the Abstract to better summarize analyses conducted altered the Abstract to read:

> **"Given the importance of variant SARS-CoV-2 viruses with altered receptor-binding or antigenic phenotypes, we sought to quantify the degree to which adaptive evolution is driving accumulation of mutations in the SARS-CoV-2 genome. Here we assessed adaptive evolution across genes in the SARS-CoV-2 genome by correlating clade growth with mutation accumulation as well as by comparing rates of nonsynonymous to synonymous divergence, clustering of mutations across the SARS-CoV-2 phylogeny and degree of convergent evolution of individual mutations. We find that spike S1 is the focus of adaptive evolution, but also identify positively-selected mutations in other genes that are sculpting the evolutionary trajectory of SARS-CoV-2. Adaptive changes in S1 accumulated rapidly, resulting in a remarkably high ratio of nonsynonymous to synonymous divergence that is 2.5X greater than that observed in HA1 at the beginning of the 2009 H1N1 pandemic."**

Some smaller points:

A little more care is needed with discussions of positive selection here. Yes dN/dS can find positively selected change, but we expect to see elevated dN too in the case of diversifying selection.

dN/dS>1 could indeed be driven by an excess of nonsynonymous divergence or a depletion of synonymous divergence. The assumption in the literature is that synonymous change is neutral, and thus, should not be selected against. Therefore, different dN/dS ratios are assumed to reflect selection driving nonsynonymous change. In our case, we observe a plethora of nonsynonymous change in S1 (Figure 1a), which fits well with dN/dS>1 being driven by an excess of nonsynonymous divergence. This excess of nonsynonymous change above neutral expectation is a result of natural selection (typically referred to as positive selection).

The population genetic definitions here are unfortunately not used completely consistently in the literature, but our understanding is that dN/dS > 1 requires *positive selection* in the sense of nonsynoymous mutations being propelled to increase in frequency by natural selection. This positive selection may be *directional* in the sense that 501Y may be a systematically "better" residue for SARS-CoV-2 in humans than 501N, or it may be *diversifying* in the sense that 484A or 484Q are both "better" because they are different than 484E and as population immunity changes, selection pressure on 484 may shift along side. However, although we appreciate the reviewer's point that dN/dS>1 can be driven by a variety of sources, we do not think this discussion is central to the conclusions of the paper.

First para of results: this is written as if we expect neutral evolution. We don't in a major surface antigen! Neutrality is rather the null hypothesis in the statistical test of selection.

At the beginning of the pandemic, it was fairly widely touted that coronaviruses don't evolve antigenically, and that the observed mutations were neutral, not adaptive. The first paragraph is written with that assumption in mind. Indeed, we observe little evidence for adaptive evolution spike protein of seasonal coronavirus NL63 (Kistler and Bedford, 2021, eLife) and measles surface proteins are famously antigenically stable and show consistent purifying selection.

Even in 2020, we (the authors) expected some adaptive evolution in spike, but weren't sure whether this would be a little or a lot. However, we think it's fair to treat neutrality as the *null hypothesis* as is standard practice in population genetics. Null hypothesis is not the prior expectation.

The comments that these convergent changes have "giving rise to successful viral clades each time" are not borne out by the data. Some of these mutations, discussed above, have happened but not reliably produced a successful strain. This is an interesting question, which I would like to see discussed more. It is certainly true that N501Y, E484K and del69/70 seem important -especially in combination - but each alone is not sufficient.

We have removed this statement from the text. We agree that it would be very interesting to assess how the success of a variant is sculpted by a variety of factors like epistatic interactions, the genome of other competing variants, and societal mitigation procedures, in addition to the individual mutations that it possesses. However, that is beyond the scope of this paper.

The observations about nsp6 and ORF1a:3675-3677del are really striking.

Agreed. It's also interesting that a very similar deletion is present in Omicron.

First line of discussion - while I agree that adaptive evolution is really important and of great interest. It is not clear that detecting it will directly (or even indirectly) help to "curb the transmission of infectious disease".

Whether or not a virus evolves antigenically impacts how vaccines will be produced. We have updated this paragraph to try to be clearer about this point:

> **"Detecting adaptive evolution is both highly interesting from a basic scientific perspective as we seek to understand how and when this type of evolution occurs, and highly relevant from a public health perspective as we strive to curb**

the transmission of infectious diseases. As widespread SARS-CoV-2 circulation continues, our best defense is through vaccination. The SARS-CoV-2 vaccines showed high efficacy in clinical trials, but we must be proactive to ensure their continued effectiveness. Vaccines against viruses that undergo antigenic drift, like influenza, must be continually updated to match circulating variants. Therefore, the propensity of SARS-CoV-2 to evolve adaptively in spike S1 (the location of most neutralizing antibody epitopes) has important bearing as to whether the SARS-CoV-2 vaccine will also need to be regularly updated."

There seems to be an inconsistency in the paper between a method to detect 'subtle changes' over the short term and the fact that these changes are anything but subtle. A dN/dS of ~2 is, as stated, very high (it might be worth thinking about fair comparisons here. I like the choice of the emergence of H1N1 in 2009 but would like to see it discussed more.

We have added Figure S5, which calculates dN/dS for H1N1pdm in the same manner as is done for SARS-CoV-2 in the main figures. We have added the following text to the Results to emphasize how high the S1 dN/dS ratio is compared to H1N1pdm dN/dS at the beginning of that pandemic:

"However, dN/dS within S1 increases over time, with an apparent inflection point in mid-2020, and the dN/dS ratio exceeding 1 in late-2020 and 2021 with the most recent time point measured at 1.80. As a point of comparison, we used the same methodology to compute dN/dS for influenza H1N1pdm following its emergence in humans in 2009, and found that dN/dS in HA1 subunit peaked at 0.72 roughly a year after the beginning of that pandemic (Figure S5)."

We have also revised the paragraph in the Discussion relating to this:

"An initially high rate of protein-coding changes is consistent with the idea that, soon after a spillover event, there are many evolutionarily-accessible mutations that are advantageous in the new host environment. This was observed in the influenza H1N1 pandemic virus (H1N1pdm). Su et al [24] report that H1N1pdm had elevated genome-wide dN/dS rates for 2 years following its emergence in 2009, and evolution during this period is thought to largely have been adaptation to a new host, including increased transmission in humans. From 2011 onward, the adaptive evolution of H1N1pdm has been dominated by antigenic changes [24]. In agreement with these results, we observe that the accumulation of dN/dS in the HA1 subunit of H1N1pdm peaks around a year after the virus' emergence at 0.72, roughly twice the mean dN/dS ratio between 2009 and 2021 (Figure S5). It is possible that SARS-CoV-2 is following a similar trajectory of adaptive evolution, with initially high dN/dS due to host adaptation to be followed by sustained antigenic drift."

The color scheme in the figures needs attention. It is hard to distinguish some of the most important variants (alpha in particular). Do the major contributions of A.23.1, B.1.318 and B.1.1.519 reflect corrections for undersampled regions? I was very confused

We have changed the color scheme in Figure 1 and supplements that use colors to identify different clades. The color scheme now has a color for each of the prominent VOCs Alpha, Beta, Delta and Gamma and groups all other VOIs into another color.