nature research

Peer Review File

Manuscript Title: Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein

Editorial Notes:

Redactions – Mention of other journals

This document only contains reviewer comments, rebuttal and decision letters for versions considered at *Nature*. Mentions of the other journal have been redacted.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Tegally et al. found new SARS-CoC-2 variants in South Africa, and described their epidemiology, evolution and mutants. This provided very important information. I suggest to make minor revision as follow.

1. In this manuscript, other three SARS-CoV-2 lineages (B.1.1.54, B.1.1.56 and C.1) have been circulating in South Africa. Please describe them and do comprehensive comparison with 501Y.V2. How many mutational sites in these three lineages and where are they located?
2. In Figure 2B, mutations characterizing the cluster in branches were not clear, especially for S: L18F, please indicate them with arrows.
3. These lineages have D614G mutation or not?
4. Of the eight mutations of 501Y.V2, five showed as positive selection, K417 showed no evidence of positive selection. What about the other two, R246I and A701V?
5. The word "Genotypes" had been used in several sentences from line 188 to 192 or other parts in text. Why are there some many genotypes such as "2589 South African genotypes" in line 189? The "genotype" should be strain. Please make clear what's different meaning of Lineage, Cluster, Genotype and strain in whole text.
6. In Methods, there are "structure modelling" part, but no result was showed in "Results" part.

Referee #2 (Remarks to the Author):

The authors present important data on a newly discovered SARS-CoV-2 lineage in South Africa associated with increased transmission. The article is clearly written and presented, and the analyses are sound (a few minor points below). While experimental data is still needed to help test the hypothesis that 501Y.V2 is more transmissible, other experimental data with only the N501Y mutation suggests that at least 1 mutation in the lineage is better adapted to human infection. Because of the phylogenetic data presented here and the experimental data presented elsewhere, I think that this is a very important study to be made available quickly.

Major concern:

1) My only concern is that the authors need to present an argument for why 501Y.V2 did not become dominant via (1) founders effects and (2) surveillance biases. The tMRCA for the lineage and subsequent detection was during a period of relatively low transmission: conditions that can lead to random selection leading to lineage replacement. It is also a common concern about both this variant and B.1.1.7 in the UK, and therefore should be preemptively addressed. In addition, while the sampling distribution looks impressive (especially from my vantage point in the US), the authors would be wise to include more information about how samples were selected for sequencing, and how this represents an unbiased assessment of SARS-CoV-2 lineages in SA. More importantly for the last point, it would be important to note if the sample selection process changed over time to ensure that the perceived replacement is not just a reflection in a shift of location/population representation.

Minor comments:

1) Genome assembly: I haven't used Genome Detective before, so not sure if this step is included, but the authors should indicate if the primer sequences were removed from the reads prior to generating the consensus genome. Leaving primer sequences can mask variations that occur within this region.

2) Genome assembly: The authors should indicate their specs for calling a consensus nucleotide at each site – i.e. read depth and % in agreement with the consensus.

3) Phylogeography: the authors indicate that the lat/long was attributed to the health facility where sample collection occurred. The authors should state some additional justification for this, such as the general catchment area in case some facilities cover a large geographical area. This doesn't need to be precise for each facility, but something to convey that the facility is a reasonable proxy for the individual's residence would be helpful.

4) Lineage name: That authors' state: "we have assigned it the name 501Y.V2 until PANGOLIN reflects the lineage in a future update". Isn't 501Y.V2 the same as B.170 in PANGOLIN? If so, the authors should use this nomenclature to avoid confusion.

Referee #3 (Remarks to the Author):

Tegally et al reported the detection of a new SARS-CoV-2 variant with multiple spike mutations in South Africa, and characterization of its genomic epidemiology and transmission. The finding has high importance and timely public health significance. The study was well conducted - I do not have major comments but the following questions that hopefully could be clarified by authors.

Result, Line 282: It would be clearer by introducing which clade (NextStrain/Pangolin/GISAID) this 501Y.V2 monophyletic lineage emerges from. And what are the non-501Y.V2 viruses closely related to this V2 variant lineage? I saw in Fig 2B, there are some South African sequences (red circles) near the tree base. Are there any other non-SA non-501Y.V2 sequences falling closer that might suggest an importation of 501Y.V2 from foreign countries? Or otherwise does it suggest the imported non-501Y.V2 virus lineage into SA and evolved locally to give rise to the 501Y.V2? Which SA region were those red circle sequences at the Fig 2B tree base were sampled from? Did these geographic and epidemiological backgrounds of those patients (those red circles in Fig 2B, as well as those early cases of 501Y.V2 variants) give some clues into where and how this 501Y.V2 possibly evolved. E.g. Any of these patients are immunosuppressed or co-infected with HIV etc?

Result, Line 300: It indicates that the 501Y.V2 lineage exhibited hypermutations. It is obvious from the comparison of total mutations between various SA lineages; yet the total amount could be affected by the lineage duration. May be better compared in the substitution rates,

subs/site/month? Is this 501Y.V2 lineage evolving significantly faster than the other prior lineages in the world?

Could the authors estimate the gain in transmissibility by this new 501Y.V2 variant, using the epidemiologic model or phylodynamics model. This will provide further and more quantitative evidence supporting the increased transmissibility/fitness of this new variants.

Line 330-331, I think another effort is to understand where and how this mutant evolved and emerged in SA. This is less urgent than dealing with the current outbreak wave and global dissemination, but important to monitor and prevent the next emergence, if there is any geography-specific factors that drive the virus evolution.

Do the authors have intra-host sequencing data to present? How stable are these 501Y.V2 mutations?

S501.V2, 501Y.V2 and S501Y.V2 are used in the text. Better have consistent naming.

Figure 2A. Would be helpful to indicate where the major SA lineages C1, B.1.1.56, B.1.1.54 are.

Line 83: ", becoming the dominant … Provinces within weeks."

Tommy Lam

Referee #4 (Remarks to the Author):

The authors describe a new variant of SARS-CoV-2 that they first detected in South Africa, which is reported to have spread rapidly throughout the country and has now appeared internationally. In the following review I restrict my comments to the genome sequencing and primary bioinformatic analysis of the sequence data, as this is my area of expertise.

The authors use the ARTIC v3 protocol on the Illumina MiSeq for whole genome sequencing, which is an amplicon-based approach and is the predominant method of sequencing SARS-CoV-2. The authors use "Genome Detective', a tool developed by a number of the authors in previous publications, to perform the primary analysis of this sequencing data. Genome Detective appears to use a de novo assembly approach (through the SPAdes/metaSPAdes software) which may not be an appropriate choice for accurate analysis of amplicon-based sequencing data. De novo assemblers typically assume a uniform distribution of reads across the genome, and this assumption is violated by amplicon-based data. The authors used a subsequent polishing step based on mpileup/bcftools but few details are provided (see additional comments section below) which makes this step difficult to assess. In the previous publications describing Genome Detective its accuracy is assessed only in the context of detecting whether a virus is present or absent, not the base-level accuracy of the assembly.

Further, the previous papers describing Genome Detective do not discuss removal of amplicon primer sequences, which is a critical step to avoid false calls of reference sequence. Two of the important variants discussed in the manuscript, the 9bp deletion at 22286-22294 and the K417N mutation, are in ARTIC V3 primer sequences. The K417N mutation is listed as "fixed" in supplementary figure S8 but it is called as reference in a few samples, which may be due to primer trimming issues. Please comment on whether amplicon primer trimming is implemented in the analysis pipeline.

I was unable to find the raw sequence reads for the samples discussed in the paper (see below)

but inspecting the reads for sample of this variant (England/ALDP-C497BF/2020) indicates that ARTIC V3 amplicon 74 is dropped out, likely as a result of the 9bp deletion in the left primer of this amplicon. I note that some of the sequences posted by the authors to GISAID have coverage in the amplicon 74 region. Please comment on whether amplicon 74 is dropped out in the authors' data and whether this issue contributes to the inability to resolve the 9bp deletion noted in the manuscript.

The key mutations discussed in the manuscript (K417N, E484K and N501Y), and others, are replicated in externally generated data from COG-UK and Switzerland so I am confident they are real mutations, not artifacts. I anticipate however there will be considerable interest in using the South African genomes to track and reconstruct the spread of this variant throughout the world, so I feel it is justified to request the authors expend additional effort to ensure these genomes are of the highest quality. The status of which mutations are fixed or not is also relevant to the proposed models of how this variant originated (discussion lines 372-375) so deserves this extra attention. I suggest the authors re-analyze their data using the de facto standard approach for ARTIC sequencing on Illumina instruments, which is mapping the reads to the reference genome using bwa mem, followed by various trimming steps (both sequencing adapters and primers) and variant/consensus calling with ivar. COG-UK has an easy-to-use pipeline (https://github.com/connor-lab/ncov2019-artic-nf) that implements these steps for ARTIC data.

Additional comments

The manuscript states that the raw reads have been uploaded to NCBI but I was unable to find them. Please add the accessions to supplementary table 1.

Please state which reference genome is used and double check the coordinates of the 9bp deletion. The manuscript states the deletion is at 22286-22294 but when I map the consensus sequences to MN908947.3 the deleted bases are 22281-22289 (inclusive). The coordinates of the other mutations appear to match reference MN908947.3.

The following parts of the methods require additional detail:

-"aligning mapped reads to the references and filtering out low-quality mutations using bcftools 1.7-2 mpileup method"

What aligner was used? What is considered "low-quality"? Were any coverage thresholds applied?

-"We filtered out 99 South African genotypes due to low coverage, and a further 16 due to poor sequence quality"

Please describe the threshold for "low coverage" and how "poor sequence quality" was determined

Author Rebuttals to Initial Comments:

Referee #1

Comment 1: In this manuscript, other three SARS-CoV-2 lineages (B.1.1.54, B.1.1.56 and C.1) have been circulating in South Africa. Please describe them and do comprehensive

comparison with 501Y.V2. How many mutational sites in these three lineages and where are they located?

> ***Response:*** *We described the other major SA lineages in a previous manuscript deposited as a preprint in medRxiv (reference 10) and forthcoming in [redacted]. The three lineages mentioned (B.1.1.54, B.1.1.56, and C.1) were the main lineages detected during the first wave of the pandemic, representing just under half of all the samples sequenced in the country. Comparison with these lineages in this report served to contrast the widespread co-circulation of major lineages without phenotypic variance with the rapid expansion of the 501Y.V2 variant and to emphasize the replacement of these major lineages with the 501Y.V2. To answer the question, in the manuscript the C.1 lineage was detected in five South African provinces (North-West, Gauteng, Free State, Limpopo and KwaZulu-Natal). Additional sequencing since then has picked up this lineage in the Western Cape as well. Lineage B.1.1.54 was detected in three provinces (North-West, Gauteng and KwaZulu-Natal). In the above-mentioned manuscript the B.1.1.56 lineage was detected only in the province of KwaZulu-Natal. However, subsequent sampling and sequencing has also detected this variant in the Free State and Gauteng provinces. We have updated Fig 2A to make it clearer that these lineages cluster separately from the 501Y.V2 variant (Fig 2A) and to make clear that while these three lineages accounted for a significant proportion of infections in South Africa during the first wave, none of them were a precursor to the 501Y.V2 variant.*
>
> *We have added the following description to the results section (Phylogenetic and phylogeographic analysis): "The 501Y.V2 cluster was phylogenetically distinct from three main lineages circulating widely (>42% of samples sequenced before October 2020) in South Africa during the first wave (B.1.1.54, B.1.1.56, and C.1) (Fig 2A). These three lineages had been circulating in the provinces of KZN, WC, Gauteng, Free State, Limpopo, and North West). By mid-November, the 501Y.V2 lineage had superseded B.1.1.54, B.1.1.56 and C.1, and rapidly became the dominant lineage in samples from EC, KZN and WC (Fig. 2C, Suppl Fig. S6,S7)."*
>
> *To the mutational profile section of the results, we have also added: "The B.1.1.54, B.1.1.56 and C.1 lineages only contained the one non-synonymous spike mutation (D614G)."*

Comment 2: In Figure 2B, mutations characterizing the cluster in branches were not clear, especially for S: L18F, please indicate them with arrows.

> ***Response:*** *We thank the reviewer for suggesting this. We have now added lines and node points to indicate more clearly which mutations characterize the cluster branches on the tree in 2B.*

Comment 3: These lineages have D614G mutation or not?

> ***Response:*** *Yes, all the lineages that we describe in this study, including the new 501Y.V2 variant, contain the D614G mutation, as specified in the Mutational profile section of the Results: "At the point of first sampling on 15 October this lineage had,*

in addition to D614G, five other non-synonymous mutations in the spike protein…". It is also shown on Fig 3A (mutational map of the 501Y.V2 spike protein), where, annotated in black, this is shown as a non-lineage-defining mutation of the 501Y.V2.

Comment 4: Of the eight mutations of 501Y.V2, five showed as positive selection, K417 showed no evidence of positive selection. What about the other two, R246I and A701V?

Response: *We have now indicated that up until 14 November (i.e. before the addition of the 501Y.V2 sequences to the GISAID database) there was no significant evidence of positive selection at codons 417, 246 and 701.*

The last line of the selection analysis section now reads: "Up until 14 November 2020 there was no statistical evidence of positive selection at codons 417, 246 and 701".

Comment 5: The word "Genotypes" had been used in several sentences from line 188 to 192 or other parts in text. Why are there some many genotypes such as "2589 South African genotypes" in line 189? The "genotype" should be strain. Please make clear what's different meaning of Lineage, Cluster, Genotype and strain in whole text.

Response: *We agree with the reviewer that using genotype in this part of the text could be confusing. Here we are simply referring to individual genomes of SARS-CoV-2. For consistency of terminology, we have replaced all "genotypes" from line 188 to 192 to "genomes". While we refrain from using the term "strain" until we have demonstrated that it is clearly associated with a different phenotype , we agree about clarifying the two most referred to terminologies in the text. We have added the following lines to the method section (Under heading "Lineage classification")*

"A lineage is a linear chain of viruses in a phylogenetic tree showing connection from the ancestor to the last descendant. Variant refers to a genetically distinct virus with different mutations to other viruses."

Comment 6: In Methods, there are "structure modelling" part, but no result was showed in "Results" part.

Response: *We thank the reviewer for pointing this out. We have addressed this concern by adding a summary of the structural modelling in the results section (Mutational profile) as follows: "Structural modelling of the spike trimer with these mutations reveals that three of the spike mutations are at key residues in the RBD (N501Y, E484K and K417N), three are in the N-terminal domain (L18F, D80A and D215G) and one is in loop 2 (A701V) (Fig 3D). The 3-amino acid deletion (242-244) also lies on the NTD. Two of the RBD sites in particular (417 and 484) also appeared to be key regions for binding of neutralising antibodies (Suppl Fig S11)."*

Referee #2

Major comment 1: My only concern is that the authors need to present an argument for why 501Y.V2 did not become dominant via (1) founders effects and (2) surveillance biases. The

tMRCA for the lineage and subsequent detection was during a period of relatively low transmission: conditions that can lead to random selection leading to lineage replacement. It is also a common concern about both this variant and B.1.1.7 in the UK, and therefore should be preemptively addressed. In addition, while the sampling distribution looks impressive (especially from my vantage point in the US), the authors would be wise to include more information about how samples were selected for sequencing, and how this represents an unbiased assessment of SARS-CoV-2 lineages in SA. More importantly for the last point, it would be important to note if the sample selection process changed over time to ensure that the perceived replacement is not just a reflection in a shift of location/population representation.

Response: *Thank you for this important comment - on reflection we recognize we should have preemptively addressed these points in the manuscript. Whilst we can't completely exclude a founder effect, we believe the evidence strongly points to 501Y.V2 having a selective advantage. Actually, although the epidemic in the Eastern Cape was contracting in mid-July to mid-August (the estimated tMRCA), this was not really a period of low transmission. Incidence was still above 20 per 100 000 per week in this period and the test positivity rate remained above 10%, suggesting moderate-to-high levels of transmission. As there were many different lineages circulating, the rapid expansion of 501Y.V2 and almost complete displacement of other lineages in multiple regions strongly suggests a selective advantage for this variant.*

As mentioned in the Results section, we did purposefully intensify surveillance in the Eastern Cape, and specifically Nelson Mandela Bay, in response to the rapid resurgence in infections in October-November 2020. However, both before and after the detection of 501Y.V2 in KwaZulu-Natal and Western Cape, and following the detection of 501Y.V2 in Eastern Cape, our genomic surveillance involved regular sequencing of a random selection of residual samples from routine diagnostic services. In Suppl Fig S1 we show that 501Y.V2 was detected in samples from 197 different health facilities in multiple districts across the three provinces. Therefore we are confident that whilst our sequencing coverage is relatively low, the sequences are representative of the circulating viruses in these provinces.

We have added the following text to the discussion section: "We detected this new variant through intensified genomic surveillance in response to a rapid resurgence of cases in the Eastern Cape Province. However, both before and after the detection of 501Y.V2, our genomic surveillance involved regular sequencing of a random selection of residual samples from routine diagnostic services. We show that 501Y.V2 was detected in samples from 197 different health facilities in multiple districts across four provinces. Therefore we are confident that whilst our sequencing coverage is relatively low, the sequences are representative of the circulating viruses in these provinces. Although the epidemic in the Eastern Cape was contracting in mid-July to mid-August (the estimated tMRCA), this was not really a period of low transmission. Incidence was still above 20 per 100 000 per week at this time and the test positivity rate remained above 10%, suggesting moderate-to-high levels of transmission. As there were many different lineages circulating at this time, the rapid expansion of 501Y.V2 and almost complete displacement of other lineages in multiple regions strongly suggests a selective advantage for this variant."

Minor comment 1: Genome assembly: I haven't used Genome Detective before, so not sure if this step is included, but the authors should indicate if the primer sequences were removed from the reads prior to generating the consensus genome. Leaving primer sequences can mask variations that occur within this region.

> ***Response:*** *The Genome Detective pipeline by default includes adapter trimming and base quality trimming but not primer trimming. As the reviewer suggests, this workflow could potentially result in false negative calls because of primers, potentially causing the workflow to miss the mutation (such as K417N pointed out by Reviewer 4). To address this valid concern, we investigated this potential problem by processing the read data also using the ARTIC Illumina pipeline [git revision 9ac3119a87], as suggested further by Reviewer 4. Calls were highly consistent regarding the lineage key mutations, including the missing 22813G>T mutation (which is the nucleotide change underlying K417N) in these particular samples.*
>
> *We described this in the manuscript:*
>
> *"In some samples, the K417N mutation was covered by the sequencing but not called. To avoid an assembly concern, these samples were also analyzed using the ARTIC Illumina pipeline [connor-lab/ncov2019-artic-nf, git revision 9ac3119a87]. Results between the two pipelines were highly consistent with respect to the lineage defining mutations, but also consistent with respect to the missing 22813G>T (K417N) mutation in these samples despite being considered covered by both pipelines (Supplementary Table S1)."*

Minor comment 2: Genome assembly: The authors should indicate their specs for calling a consensus nucleotide at each site – i.e. read depth and % in agreement with the consensus.

> ***Response:*** *We agree with the reviewer this is an important detail. We have added the following sentence to the methods section under Whole genome sequence and genome assembly: "To call the consensus sequence, GATK HaplotypeCaller is used with default settings, followed by GATK VariantFiltration to select only variants with a variant confidence normalized by unfiltered depth of variant samples of at least 10 (QualByDepth >= 10)."*

Minor comment 3: Phylogeography: the authors indicate that the lat/long was attributed to the health facility where sample collection occurred. The authors should state some additional justification for this, such as the general catchment area in case some facilities cover a large geographical area. This doesn't need to be precise for each facility, but something to convey that the facility is a reasonable proxy for the individual's residence would be helpful.

> ***Response:*** *Many thanks for this sensible suggestion. We don't collect individual geolocators for the routine submissions in the genomic surveillance, so the health facility where the diagnostic sample was collected is the best geolocator we have for the phylogeographic analysis. We believe this serves as a reasonable proxy, given that two-thirds of the population in South Africa lives less than 2km from their nearest*

health facility (ref. McLaren Z et al. Distance decay and persistent health care disparities in South Africa. BMC Health Services Research 2014).

To the Methods section (Phylogeographic analysis), we have added the following: "Given that we don't have access to residential geolocators within the genomic surveillance, the location of the health facility serves as a reasonable proxy, especially as two-thirds of the population live within 2km of their nearest health facility."

Minor comment 4: Lineage name: That authors' state: "we have assigned it the name 501Y.V2 until PANGOLIN reflects the lineage in a future update". Isn't 501Y.V2 the same as B.170 in PANGOLIN? If so, the authors should use this nomenclature to avoid confusion.

Response: *We agree it can be confusing. The 501Y.V2 is actually equivalent to PANGO lineage B.1.351. We use the official name given in South Africa to this variant (501Y.V2) throughout this manuscript to match other reports documenting the properties of this variant. To eliminate any naming confusion, we have updated the information in the methods section (Under heading "Lineage classification"): "For the new variant identified in South Africa in this study, we have assigned it the name 501Y.V2; the corresponding PANGO lineage classification is B.1.351 (lineages version 2021-01-06)."*

Referee #3

Comment 1: Result, Line 282: It would be clearer by introducing which clade (NextStrain/Pangolin/GISAID) this 501Y.V2 monophyletic lineage emerges from. And what are the non-501Y.V2 viruses closely related to this V2 variant lineage? I saw in Fig 2B, there are some South African sequences (red circles) near the tree base. Are there any other non-SA non-501Y.V2 sequences falling closer that might suggest an importation of 501Y.V2 from foreign countries? Or otherwise does it suggest the imported non-501Y.V2 virus lineage into SA and evolved locally to give rise to the 501Y.V2? Which SA region were those red circle sequences at the Fig 2B tree base were sampled from? Did these geographic and epidemiological backgrounds of those patients (those red circles in Fig 2B, as well as those early cases of 501Y.V2 variants) give some clues into where and how this 501Y.V2 possibly evolved. E.g. Any of these patients are immunosuppressed or co-infected with HIV etc?

Response: *We thank the reviewer for this comment. We agree it can be confusing particularly with so many different classification systems/schemes. With regards to NextStrain clades system, the 501Y.V2 variant emerged from the 20C cluster that dominated the early epidemic in the United States. Since the release of the sequences publicly on GISAID, NextStrain has moved to naming the 501Y.V2 clade as 20H/501.V2. This is clearly visible in the latest updated NextStrain build for this manuscript (https://nextstrain.org/groups/ngs-sa/COVID19-ZA-2021.01.18) as well as on the latest global NextStrain build (https://nextstrain.org/sars-cov-2/).*

With regards to the PANGO lineage assignment scheme, the 501Y.V2 variant emerged from B.1 viral isolates and is now classified as PANGO lineage B.1.351. We

have decided to keep the 501Y.V2 classification as the PANGOLIN and NextStrain schemes are dynamic.

To address the reviewer's comment with regards to the basal sequences and the origin of the 501Y.V2 variant, we can confirm that there are seven South African sequences that cluster basal to the 501Y.V2 (Now only shown in Fig2A basal to the yellow cluster), with their geographical range involving the provinces of the Eastern Cape (oldest isolate), Western Cape, Gauteng and KwaZulu-Natal. To the best of our understanding, this variant must have emerged from an introduction of a 20C or B.1 isolate from the United States early on in the epidemic. Since this introduction the virus diversified and spread throughout the country before the emergence of 501Y.V2. Given that the source of samples in which we first detected 501Y.V2 was Nelson Mandela Bay in the Eastern Cape, at the time of a resurgence of infections that occurred prior to other areas, we hypothesize that the 501Y.V2 variant most likely emerged in this region. Our phylogeography analysis supports our hypothesis as we strongly infer the origin on the 501Y.V2 cluster to be Nelson Mandela Bay, South Africa. However, given relatively sparse sampling in Eastern Cape prior to the resurgence in cases in October 2020, we can't exclude the possibility that 501Y.V2 emerged elsewhere and was then amplified in Nelson Mandela Bay. To this effect, we added the following lines to the Phylogenetic and phylogeographic analysis section: "Seven South African sequences that appear to be basal to the 501Y.V2 cluster (Fig 2A) were sampled in the provinces of the Eastern Cape (oldest isolate), Western Cape, Gauteng and KwaZulu-Natal between late June and early September. While these do not have any of the defining mutations of the 501Y.V2 variant, they form part of the B.1.351 lineage. This suggests that the precursor to the new variant had been circulating in the country and that 501Y.V2 expanded from within the country rather than as a result of an imported infection."

With regards to reviewers question about clinical details of individual cases, unfortunately we don't have access to that information. For the routine genomic surveillance, we only collect limited metadata (sex, age, and health facility) and don't have access to clinical data.

Comment 2: Result, Line 300: It indicates that the 501Y.V2 lineage exhibited hypermutations. It is obvious from the comparison of total mutations between various SA lineages; yet the total amount could be affected by the lineage duration. May be better compared in the substitution rates, subs/site/month? Is this 501Y.V2 lineage evolving significantly faster than the other prior lineages in the world?

Response: *We agree fully with the reviewer that a higher number of accumulating nucleotide mutations could be a factor of lineage duration. However, if we look closely, the number of non-synonymous (amino acid) changes in 501Y.V2 is also significantly higher, particularly in the spike protein, while the numbers for the three other lineages remain fairly consistent. We also still detect these three lineages in November and December, albeit at low frequency, which lessens the chances of 501Y.V2 mutation accumulation being just a function of time.*

To fully address the reviewer's concerns, we have estimated and reported the mean evolutionary rates (from root-to-tip regression plots) of the 501Y.V2 lineage compared to the three other lineages (Suppl Fig S2).

We added the following to the mutation profile section of the results: "B.1.1.54, B.1.1.56 and C.1 only contained one non-synonymous change on the spike protein (D614G) despite following the expected temporal accumulation of mutations and therefore did not show any concerning mutation pattern like the 501Y.V2. An estimate of the evolutionary rates indicates that substitutions on the 501Y.V2 lineage are happening at 1.917E-3 nucleotide changes/site/year, compared to 5.344E-4, 4.251E-4 and 9.781E-4 respectively for B.1.1.54, B.1.1.56 and C.1 (Suppl Fig S2)."

Comment 3: Could the authors estimate the gain in transmissibility by this new 501Y.V2 variant, using the epidemiologic model or phylodynamics model. This will provide further and more quantitative evidence supporting the increased transmissibility/fitness of this new variants.

Response: *Since submission of the manuscript, preliminary estimates of the relative transmissibility of 501Y.V2 have been generated by the Centre for Mathematical Modelling of Infectious Diseases at the London School of Hygiene & Tropical Medicine using their well-established epidemiological model. We have now added information about these estimates.*

Comment 4: Line 330-331, I think another effort is to understand where and how this mutant evolved and emerged in SA. This is less urgent than dealing with the current outbreak wave and global dissemination, but important to monitor and prevent the next emergence, if there is any geography-specific factors that drive the virus evolution.

Response: *We do agree this is really important, but the reviewer is correct that for now this has taken a back seat whilst we prioritize controlling transmission of 501Y.V2 and characterizing the phenotype of 501Y.V2. We have established a national consortium of scientists to address key questions around the detection of 501Y.V2 – one of the questions that has been prioritized is understanding how this emerged in the Eastern Cape.*

Comment 5: Do the authors have intra-host sequencing data to present? How stable are these 501Y.V2 mutations?

Response: *We thank the reviewer for this very constructive comment. We have now analyzed intra-host sequencing data for three of our sequencing runs at each of the nucleotide mutation sites that we call lineage-defining for the 501Y.V2. We investigated the possibility of any minority alleles and the stability of the called mutant alleles at these sites and concluded that the 501Y.V2 mutations are all relatively stable. We have described this in the methods section, as follows, and added a supplementary figure (Suppl Fig S12) to illustrate the results.*

Methods section under Whole genome sequencing and assembly: "LoFreq was used to detect minor viral variants to study the intra-host heterogeneity of viral variants

(quasi-species) (Suppl Fig S12). Variants were called with at minimum coverage of 10% and conservative false discovery rate (FDR) p-value of 0.1. LoFreq models sequencing error rate and implements a Poisson distribution to probe the statistical significance of nucleotide variants at each position filtering out all variants falling below the p-value threshold."

Suppl Fig S12 legend: "Showing allele proportions at each 501Y.V2 lineage defining mutations sites, with the black line and dots showing the mutant allele proportion and the grey line and dots showing the reference allele proportion in individual samples in three sequencing runs."

Comment 6: S501.V2, 501Y.V2 and S501Y.V2 are used in the text. Better have consistent naming.

Response: *We thank the reviewer for pointing this out to us. We have carefully edited the manuscript and removed the S501.V2 and S501Y.V2 instances. For consistency we kept the naming of 501Y.V2.*

Comment 7: Figure 2A. Would be helpful to indicate where the major SA lineages C1, B.1.1.56, B.1.1.54 are.

Response: *We have updated Fig 2A to make it clearer that these other lineages cluster separately from the 501Y.V2 variant in the phylogenetic tree (Fig 2A), and to make clear that while these three lineages accounted for a significant proportion of infections in South Africa during the first wave, none of them were a precursor to the 501Y.V2 variant.*

Comment 8: Line 83: ", becoming the dominant … Provinces within weeks."

Response: *Thank you, this was edited as suggested and now reads: "becoming the dominant lineage in the Eastern Cape and Western Cape Provinces within weeks."*

Referee #4

Comment 1: The authors use the ARTIC v3 protocol on the Illumina MiSeq for whole genome sequencing, which is an amplicon-based approach and is the predominant method of sequencing SARS-CoV-2. The authors use "Genome Detective', a tool developed by a number of the authors in previous publications, to perform the primary analysis of this sequencing data. Genome Detective appears to use a de novo assembly approach (through the SPAdes/metaSPAdes software) which may not be an appropriate choice for accurate analysis of amplicon-based sequencing data. De novo assemblers typically assume a uniform distribution of reads across the genome, and this assumption is violated by amplicon-based data. The authors used a subsequent polishing step based on mpileup/bcftools but few details are provided (see additional comments section below) which makes this step difficult to assess. In the previous publications describing Genome Detective its accuracy is assessed only in the context of detecting whether a virus is present or absent, not the base-level accuracy of the assembly.

Response: *We agree with the reviewer. Since February 2020, Genome Detective has been continuously improved to make the software also usable for the accurate calling of variants, so that it is useful in particular for SARS-CoV-2 sequencing data and epidemiological applications, for read sets generated using a large diversity of protocols, but including Illumina and Nanopore datasets generated using the ARTIC lab protocol. Most of the changes suggested by the reviewer have already been implemented in the software. We have updated the manuscript to clarify this.*

"To accurately call mutations and short indels for SARS-CoV-2, Genome Detective software was updated with an additional assembly step after the de novo assembly and strain identification. When the de novo assembly indicates a nucleotide similarity higher than 95% to the reference strain, a new assembly is made by read mapping against the reference. In this process, for strains satisfying this criterion, reads are mapped using minimap2 [1] against the reference rather than the de novo consensus sequence, and subsequently final mutations and indels are called using GATK HaplotypeCaller [2], with low quality variants (with QD < 10) filtered using GATK VariantFiltration [2]."

[1] Heng L. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018; 34(18): 3094–3100

[2] McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20(9):1297-303

Comment 2: Further, the previous papers describing Genome Detective do not discuss removal of amplicon primer sequences, which is a critical step to avoid false calls of reference sequence. Two of the important variants discussed in the manuscript, the 9bp deletion at 22286-22294 and the K417N mutation, are in ARTIC V3 primer sequences. The K417N mutation is listed as "fixed" in supplementary figure S8 but it is called as reference in a few samples, which may be due to primer trimming issues. Please comment on whether amplicon primer trimming is implemented in the analysis pipeline.

Response: *The Genome Detective pipeline by default includes adapter trimming and base quality trimming but not primer trimming. As the reviewer suggests, this workflow could potentially result in false negative calls because of primers, potentially causing the workflow to miss the mutation (in this example, K417N). To address this valid concern, we also processed the read data using the ARTIC Illumina pipeline [git revision 9ac3119a87]. Calls were highly consistent regarding the lineage key mutations, including the missing 22813G>T mutation (which is the nucleotide change underlying K417N) in these particular samples.*

We described this in the manuscript:

"In some samples, the K417N mutation was previously not called. To avoid an assembly concern, these samples were also analyzed using the ARTIC Illumina pipeline [connor-lab/ncov2019-artic-nf, git revision 9ac3119a87]. Results between the two pipelines were highly consistent with respect to the lineage defining mutations, but also consistent with respect to the missing 22813G>T (K417N) mutation in these samples despite being considered covered by both pipelines

(Supplementary Table S1). In addition, we have implemented a Sanger sequencing method that cover the main RBD sites and this was used to confirm the K417N and other mutations (i.e. 484K and 501Y) in sequences that we were not confident about the call from NGS data."

Comment 3: I was unable to find the raw sequence reads for the samples discussed in the paper (see below) but inspecting the reads for sample of this variant (England/ALDP-C497BF/2020) indicates that ARTIC V3 amplicon 74 is dropped out, likely as a result of the 9bp deletion in the left primer of this amplicon. I note that some of the sequences posted by the authors to GISAID have coverage in the amplicon 74 region. Please comment on whether amplicon 74 is dropped out in the authors' data and whether this issue contributes to the inability to resolve the 9bp deletion noted in the manuscript.

Response: *Apologies that the raw sequence reads were not available at the time of your review. The BioProject with the raw reads is now available on SRA (accession PRJNA694014). Yes, we do observe a drop in this amplicon just following the 9bp spike deletion in most of our sequences, but, as the reviewer points out, not all. The inability to resolve the 9bp deletion was a bioinformatics issue due to this region being a high-repeat region. However, this has now been resolved and we confidently call the deletion, which also clearly appears in our BAM files. To ensure that we were not missing any additional mutation in this drop-out region, we have further performed Sanger sequencing covering this region in some future samples and we report no further mutation in this area. However, since this was done with more recent sequences not included in the analysis for this paper, we do not present these results here, but we will be happy to consider otherwise if necessary.*

Comment 4: The key mutations discussed in the manuscript (K417N, E484K and N501Y), and others, are replicated in externally generated data from COG-UK and Switzerland so I am confident they are real mutations, not artifacts. I anticipate however there will be considerable interest in using the South African genomes to track and reconstruct the spread of this variant throughout the world, so I feel it is justified to request the authors expend additional effort to ensure these genomes are of the highest quality. The status of which mutations are fixed or not is also relevant to the proposed models of how this variant originated (discussion lines 372-375) so deserves this extra attention. I suggest the authors re-analyze their data using the de facto standard approach for ARTIC sequencing on Illumina instruments, which is mapping the reads to the reference genome using bwa mem, followed by various trimming steps (both sequencing adapters and primers) and variant/consensus calling with ivar. COG-UK has an easy-to-use pipeline (https://github.com/connor-lab/ncov2019-artic-nf) that implements these steps for ARTIC data.

Response: *We fully agree with the reviewer that the assembled genomes must be of the highest quality and thank the reviewer for suggesting this re-analysis. To ensure the quality, we have verified that the consensus sequences generated by the revised Genome Detective pipeline are indeed consistent with the results from the COG-UK ARTIC Illumina pipeline, particularly at the problematic site of S:417 (Supplementary Table S1).*

Comment 5: The manuscript states that the raw reads have been uploaded to NCBI but I was unable to find them. Please add the accessions to supplementary table 1.

Response: *Apologies that the raw sequence reads were not available at the time of your review. The BioProject is now available (accession PRJNA694014). We have added this accession number to the manuscript.*

Comment 6: Please state which reference genome is used and double check the coordinates of the 9bp deletion. The manuscript states the deletion is at 22286-22294 but when I map the consensus sequences to MN908947.3 the deleted bases are 22281-22289 (inclusive). The coordinates of the other mutations appear to match reference MN908947.3.

Response: *The numbering is according to NC_045512.2 which is equivalent to MN908947.3: we have updated the manuscript accordingly in the methods section: "The reference genome used throughout the assembly process was NC_045512.2 (numbering equivalent to MN908947.3"*

We agree with the reviewer that there is a confusion on the exact location of 22286-22294. In fact, this stems from the presence of a repeat region 'CTTT' which makes it impossible to unambiguously conclude which nucleotides are deleted. There could be 7 possibilities:

CAAA[CTTTACTTG]CTTTACAT -> CAAACTTACAT

CAAAC[TTTACTTGC]TTTACAT -> CAAACTTACAT

CAAACT[TTACTTGCT]TTACAT -> CAAACTTACAT

CAAACTT[TACTTGCTT]TACAT -> CAAACTTACAT

CAAACTTT[ACTTGCTTT]ACAT -> CAAACTTACAT

CAAACTTTA[CTTGCTTTA]CAT -> CAAACTTACAT

CAAACTTTAC[TTGCTTTAC]AT -> CAAACTTACAT

Fortunately, any of those possibilities results in exactly the same nucleotide and amino acid sequences in the variant, where we always end with amino acids 'OTLH', even though the alignments might look different. In fact, the deletion could technically be either in amino acids 241-243 or 242-244, and we cannot be sure which it is, but, again, the resulting protein sequence is the same. This pattern has also been seen in 501Y.V2 sequences from other countries.

We have added a small description of this in the results section (Mutational profile): "We also observe a deletion of three amino acids at 242-244, seen in samples extracted and generated in various laboratories across the network (Because of a hard-to-align repeat region, the deletion could potential also be in amino acids 241-243 but the resulting sequence of both deletions are exactly the same)."

Comment 7: The following parts of the methods require additional detail: "aligning mapped reads to the references and filtering out low-quality mutations using bcftools 1.7-2 mpileup method". What aligner was used? What is considered "low-quality"? Were any coverage thresholds applied?

> ***Response:*** *We have removed this line in the methods section, as we have replaced the pipeline with the GATK calling as described in Response 1*

Comment 8: The following parts of the methods require additional detail: "We filtered out 99 South African genotypes due to low coverage, and a further 16 due to poor sequence quality". Please describe the threshold for "low coverage" and how "poor sequence quality" was determined

> ***Response:*** *We agree with the reviewer that this needs clarification. Accordingly, we have clarified the selection criteria further in the methods section with the following sentence: "Poor sequence quality was defined as sequences with clustered SNPs and ambiguous bases at >10% of sites, and low coverage genomes were anything with <90% genome coverage against the reference"*

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I am satisfied with the authors' responses and have no further comments.

Referee #2 (Remarks to the Author):

The authors addressed all of my concerns. It is an excellent manuscript.

- Nathan Grubaugh

Referee #3 (Remarks to the Author):

The authors have fully addressed my comments. I have no further questions.

Tommy

Referee #4 (Remarks to the Author):

I thank the authors for their careful consideration of the issues raised in the previous review and I am mostly satisfied with the changes made to the manuscript. I still feel primer trimming is an essential part of an amplicon analysis pipeline so encourage the authors to add it to Genome Detective. Also, it is not appropriate to use GATK to analyze nanopore data as GATK assumes short read input data, which has a very different error model than Oxford Nanopore reads. A

nanopore-specific variant caller should be used instead. Weighed against the public health importance of this work I will leave these comments as strong suggestions to the authors for their future work however, rather than requirements to address immediately.

I appreciate the author's explanation of the 242-244 deletion ambiguity and now agree with them. I will note however that for nucleotide changes the convention is to report the lowest genome coordinate for mutations with multiple possible representations (aka left-aligning indels). Please add a note to the caption in Supp Fig S8 about the representation issues for this deletion. Also I believe that the statement "The site in blue is unresolved for the moment as we detect a non-synonymous mutation in some sequences/reads and a 9-nucleotide long deletion in others." can be removed from the caption as this site is now resolved.

Author Rebuttals to First Revision:

Referee #4 (Remarks to the Author):

Comment 1: I thank the authors for their careful consideration of the issues raised in the previous review and I am mostly satisfied with the changes made to the manuscript. I still feel primer trimming is an essential part of an amplicon analysis pipeline so encourage the authors to add it to Genome Detective.

> ***Response 1:*** *We thank the reviewer for this comment and suggestion which we take very seriously. We are planning to add a step before Genome Detective assembly which will trim primers from our reads using iVar.*

Comment 2: Also, it is not appropriate to use GATK to analyze nanopore data as GATK assumes short read input data, which has a very different error model than Oxford Nanopore reads. A nanopore-specific variant caller should be used instead. Weighed against the public health importance of this work I will leave these comments as strong suggestions to the authors for their future work however, rather than requirements to address immediately.

> ***Response 2:*** *We apologize for the confusion, the nanopore sequence assembly pipeline of Genome Detective does not use GATK. The GATK module was added to the short-read assembly pipeline of Genome Detective only to improve variant calling, as described in the methods section. For nanopore, Genome Detective uses assignment of reads to reference using NCBI blastn and read alignment using AGA aligner with iterative improvement by realignments to consensus and reference sequences. Although most of our data comes from Illumina sequencing, the reviewer is right that some of the data is from Nanopore sequencing, and therefore it is important to make the distinction in assembly methods. As such, we have modified the genome assembly section of the manuscript to now read the following:*
>
> *"For nanopore data, candidate reads are assigned to candidate reference sequences using NCBI blastn with sensitive settings and low gap costs. Candidate reads are then aligned using AGA (Annotated Genome Aligner), after which a draft majority consensus sequence is subsequently called, and iteratively improved by realignment of all reads against the draft consensus sequence and realignment of regions with a putative insert against the reference using global alignment (MAFFT). The resulting consensus sequence is further polished by considering and correcting indels of length*

one or two in homopolymer regions of length 4 or longer that break the open reading frame (likely sequencing errors)"

Comment 3: I appreciate the author's explanation of the 242-244 deletion ambiguity and now agree with them. I will note however that for nucleotide changes the convention is to report the lowest genome coordinate for mutations with multiple possible representations (aka left-aligning indels). Please add a note to the caption in Supp Fig S8 about the representation issues for this deletion.

Response 3: *We thank the reviewer for helping us investigate this deletion region further. We have now added the following note to Supp Fig 8:*

"It is important to note an unresolvable ambiguity in the representation of the exact location of the 22286-22294 nucleotide deletion; because of a hard-to-align repeat region 'CTTT', the deletion could be any 9-nucleotide segment between 22281-22289 and 22286-22294, which means that technically, the deletion could also be in amino acids 241-243, but the resulting amino acid sequence of all the possibilities are exactly the same (OTLH)"

Comment 4: Also I believe that the statement "The site in blue is unresolved for the moment as we detect a non-synonymous mutation in some sequences/reads and a 9-nucleotide long deletion in others." can be removed from the caption as this site is now resolved.

Response 4: *Thank you for catching that this statement was still there, we have now removed it.*