EMBO Molecular Medicine

Rev.EMM-2021-14062 SARS-CoV-2 Worldwide Replication Drives Rapid Rise and Selection of Mutations:

Stefanie Weber, Christina Ramirez, Barbara Weiser, Harold Burger, and Walter Doerfler
DOI: 10.15252/emmm.202114062

Corresponding author: Walter Doerfler (walter.doerfler@t-online.de)

Review Timeline:

Submission Date:	2nd Feb 21
Editorial Decision:	24th Mar 21
Revision Received:	13th Apr 21
Editorial Decision:	14th Apr 21
Revision Received:	18th Apr 21
Accepted:	20th Apr 21

Editor: Zeljko Durdevic

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. Depending on transfer agreements, referee reports obtained elsewhere may or may not be included in this compilation. Referee reports are anonymous unless the Referee chooses to sign their reports.)

1st Editorial Decision 24th Mar 2021

24th Mar 2021

Dear Prof. Doerfler,

Thank you for the submission of your manuscript to EMBO Molecular Medicine, and please accept my apologies for the delay in getting back to you. We have received feedback from two of the three reviewers who agreed to evaluate your manuscript. Should referee #3 provide a report, we will send it to you, with the understanding that we will not ask for an additional revision. As you will see from the reports below, both referees find the study interesting and important. However, referee #2 also raises important criticism that I would like you to address in a major revision of the current manuscript.

EMBO Molecular Medicine encourages a single round of revision only and therefore, acceptance or rejection of the manuscript will depend on the completeness of your responses included in the next, final version of the manuscript. For this reason, and to save you from any frustrations in the end, I would strongly advise against returning an incomplete revision.

We would welcome the submission of a revised version within three months for further consideration. However, we realize that the current situation is exceptional on the account of the COVID-19/SARS-CoV-2 pandemic. Please let us know if you require longer to complete the revision.

I look forward to receiving your revised manuscript.

Yours sincerely,

Zeljko Durdevic

***** Reviewer's comments *****

Referee #1 (Remarks for Author):

The important manuscript by Weber, Ramirez, Weiser, Burger and Doerfler builds substantially on their initial publication in September 2020 focused on tracking SARS-CoV2 variants arising in 10 different countries representing 5 continents as their genomes become fully sequenced, right up to date through January 20, 2021 and deposited in the Global Initiative of Sharing All Influenza Data - GISAID. In this report, they analyzed over 383,000 complete viral genomes deposited in the global database from 10 countries that are carrying out considerable sequencing of isolates from patients.

Starting from a baseline of the original SARS-CoV2 isolate from Wuhan, China, a major objective of this investigations was to track the emergence of amino acid sequence variants throughout the viral genome and look for identities, similarities and differences in various populations around the world with goals of distinguishing independent emergence and positive selection of certain mutations as well as transmission of genetically identifiable lineages by travelers between the

nations, near and far. Some of the original "wild type" sequences were observed to "dropout" in selected viral populations over time, indicative of insufficient fitness relative to rapidly emerging variants. Some of these are especially worrisome from a public health and containment perspective because they can negatively impact PCR and other screening tests, while other "escape mutations" can diminish the neutralizing ability of the vaccines that are now becoming available, notably the Oxford/AstraZeneca vaccine against the rapidly emerging South African variant.

This investigation is most timely, expansive and inclusive, as well as exceptionally granular, as presented in a large number of Tables and associated data files. By limiting the data set to viral genomes fully sequenced, and by emphasizing reproducible sites of variation to deemphasize the inevitable one-off random sequencing errors, the authors carefully curated the data sets to focus on evolutionary patterns of changing viral fitness over the full year of the global pandemic. The major sites of change observed were most often in the Spike gene but also significantly in a number of genes associated with viral replication and packaging, all of which are conceptually consistent with successful attachment of infection virions to target cells, rapid and robust amplification, efficient packaging, and short infection cycle turn-around. Fairly often, the variant replacement often rose to 50% upwards to over 90%, relative to the dominant sequence just months previously, so the selection can be quite rapid. Interestingly, some significant differences in mutant sites were seen continent to continent, implying ethnic and socio-economic factors were at play in the selection for SARS-CoV2 fitness. The authors took particular note of the types of mutation, especially those that resulted in non-synonymous changes in the chemical nature of the replacement amino acid. These results will inspire and guide protein biophysical chemists and enzymologists to investigate the molecular changes and their mechanistic consequences. A most notable consistency was the very disproportionate preference for cytidine to uracil (thymidine) transitions across viral genes and human populations. This points to deamination of cytosine, which has been linked to the APOBEC protein, an RNA editing factor. Intriguingly, the SARS-CoV2 cell surface receptor ACE interacts with APOBEC. The details of these suggestively linked phenomena await future mechanistic research. Nonetheless, it may be that this corona virus co-ops a cellular antiviral defense to establish an environment where the virus can potentially benefit from occasional mutagenesis, noting that SARS-CoV2 RNA replication has its own proof-reading capabilities so is not as likely as most RNA viruses to generate variants in the course of genomic replication.

The manuscript is especially clearly written. I note only a single confusing typographical problem in line 2 of the top of page 17, apparently due to the insertion of a series of country names. I am not convinced of the relevance or value of the two parenthetic notes added during submission.

The comprehensive and rather easy to navigate data in this report will be a most valuable resource for virologists, epidemiologists and global public health agencies and practitioners as well as commercial providers of screening tests or vaccines who continue to monitor and react to the emergence of variants. This manuscript should be accepted for rapid publication with enthusiasm. Notably, I strongly recommend that the authors be accorded the rather unusual opportunity to update their database now and over the coming months should they choose to do so. They are to be congratulated on a Herculean task of consolidating such a vast database into a functional format.

Referee #2 (Comments on Novelty/Model System for Author):

This is an exhaustive characterization of the emergence of SARS-CoV-2 mutants worldwide, using standard bioinformatic pipelines and an open database of SARS-CoV-2 sequences. It is technically well done, and may have high medical impact. The use of pipelines and GISAID is not novel, though the spatiotemporal evaluation of mutations in this exhaustive manner is novel.

Referee #2 (Remarks for Author):

This is an exhaustive evaluation of the emergence of SARS-COV-2 mutants over time, from about 10 predominant mutants to May 2020, followed by an increase in the number of new mutants from June 2020 to December 2020. The work in general is a straightforward characterization of the emergence of these mutations in various countries, and should be useful as a reference for these mutations.

In a manuscript this exhaustive, certain statements are made that require further support or modification.

(1) The authors postulate APOBEC mRNA editing enzymes should be active in the emergence of these mutations, given the high number of C to U transitions. The authors take this a step further, and postulate that ACE2 binding of the spike has something to do with this process. While an interesting speculation, there is nothing provided in the manuscript by literature or possible mechanism linking APOBEC to ACE2. This needs further explanation or modification.

(2) The authors do state briefly that synonymous mutation may have effects outside of protein modification, such as changes in mRNA secondary structure. Given the number of synonymous mutations found, this explanation should be expanded, and possibly include codon choice and translation efficiency as one possible explanation for differences in viral transmissibility and pathogenesis.

(3) A separate section on limitations of this study should be included, especially for the non-specialist in this field. In particular, selection bias in sequencing should be emphasized. This is important in that viral sequences may have been provided to GISAID from patients with the most severe disease, which may give an inaccurate picture of viral mutations in less severely affected patients.

(4) The choice of 2% cut off for mutation prevalence to report is as the authors claim arbitrary. This choice should be further justified.

(5) There is a large literature on viral evolution when viruses enter novel hosts, starting with the analysis of the MYXV virus in rabbits in Australia in the 1950s. Examination of this literature, with reference to it in the manuscript, would potentially be helpful in explaining the emergence of mutations in SARS CoV 2 in humans.

(6) Page 16 at the top: "High current incidence of Covid-19 is paralleled by high numbers of new mutations and variants, although this relationship was not observed in Brazil or Russia." Can the authors provide a possible explanation?

(7) Page 17: The link between ACE2 binding and APOBEC enzyme activation is unsupported. Additionally, the authors fail to note that SARS-CoV-2 may have multiple accessory factors for entry, such as CD147, NRP-1, and DC-SIGN among others. How do the mutations affect binding to

these accessory factors, and do they have a link to APOBEC enzymes? This section is highly speculative, and need better literature support for its assertions.

(8) Page 17: This paragraph at the bottom, discussing the impact of variants on vaccine efficacy, while interesting, is cursory for what is turning out to be a very complicated topic. Perhaps it may be best to reduce or eliminate the part of the paragraph beginning with "The emergence..." since this is speculation.

(9) The note added during submission at the top of page 18 references a very speculative report, and I am not sure it belongs in this manuscript, which should focus on the categorization of mutations through time.

(10) I would eliminate the last paragraph of the manuscript "Conceptual bridge.." in that while interesting, it does not add much for the general reader, and the analogies are a bit confusing.

Reviewers' Reports and our Point-by-Point Response (Response in yellow)

Referee #1 (Remarks for Author)

The important manuscript by Weber, Ramirez, Weiser, Burger and Doerfler builds substantially on their initial publication in September 2020 focused on tracking SARS-CoV2 variants arising in 10 different countries representing 5 continents as their genomes become fully sequenced, right up to date through January 20, 2021 and deposited in the Global Initiative of Sharing All Influenza Data - GISAID. In this report, they analyzed over 383,000 complete viral genomes deposited in the global database from 10 countries that are carrying out considerable sequencing of isolates from patients.

Starting from a baseline of the original SARS-CoV2 isolate from Wuhan, China, a major objective of this investigations was to track the emergence of amino acid sequence variants throughout the viral genome and look for identities, similarities and differences in various populations around the world with goals of distinguishing independent emergence and positive selection of certain mutations as well as transmission of genetically identifiable lineages by travelers between the nations, near and far. Some of the original "wild type" sequences were observed to "dropout" in selected viral populations over time, indicative of insufficient fitness relative to rapidly emerging variants. Some of these are especially worrisome from a public health and containment perspective because they can negatively impact PCR and other screening tests, while other "escape mutations" can diminish the neutralizing ability of the vaccines that are now becoming available, notably the Oxford/AstraZeneca vaccine against the rapidly emerging South African variant.

This investigation is most timely, expansive and inclusive, as well as exceptionally granular, as presented in a large number of Tables and associated data files. By limiting the data set to viral genomes fully sequenced, and by emphasizing reproducible sites of variation to deemphasize the inevitable one-off random sequencing errors, the authors carefully curated the data sets to focus on evolutionary patterns of changing viral fitness over the full year of the global pandemic. The major sites of change observed were most often in the Spike gene but also significantly in a number of genes associated with viral replication and packaging, all of which are conceptually consistent with successful attachment of infection virions to target cells, rapid and robust amplification, efficient packaging, and short infection cycle turn-around. Fairly often, the variant replacement often rose to 50% upwards to over 90%, relative to the dominant sequence just months previously, so the selection can be quite rapid. Interestingly, some significant differences in mutant sites were seen continent to continent, implying ethnic and socio-economic factors were at play in the selection for SARS-CoV2 fitness. The authors took particular note of the types of mutation, especially those that resulted in non-synonymous changes in the chemical nature of the replacement amino acid. These results will inspire and guide protein biophysical chemists and enzymologists to investigate the molecular changes and their mechanistic consequences. A most notable consistency was the very disproportionate preference for cytidine to uracil (thymidine) transitions across viral genes and human populations. This points to deamination of cytosine, which has been linked to the APOBEC protein, an RNA editing factor. Intriguingly, the SARS-CoV2 cell surface receptor ACE interacts with APOBEC. The details of these suggestively linked phenomena await future mechanistic research. Nonetheless, it may be that this corona virus co-ops a cellular antiviral defense to establish an environment where the virus can potentially benefit from occasional mutagenesis, noting that SARS-CoV2 RNA replication has its own proof-reading capabilities so is not as likely as most RNA viruses to generate variants in the course of genomic replication.

The manuscript is especially clearly written. I note only a single confusing typographical problem in line 2 of the

top of page 17, apparently due to the insertion of a series of country names. I am not convinced of the relevance or value of the two parenthetic notes added during submission.

Response:

1. As will also be explained in response to items (1) and (7) raised by referee #2, the remark on a hypothetical association between the ACE-2 receptor and the deaminase functions of APOBEC has been deleted. Text passages on pages 3 (Synopsis) and 17 (second paragraph) have been completely rewritten. Now, we just mention that cellular deaminases, possibly of the APOBEC type, might be responsible for the frequent C to T transitions
2. On page 18, the "Note added during submission" has been removed. Similarly, the typographical problem now on page 17, first paragraph, lines 3 to 4 from the bottom – ~~UK, South Africa, Brazil,~~ ... – that had been added inadvertently - has been amended. We have updated the data bases in Tables 1 and 2 as of March 31, 2021. In the interest of not delaying publication, we decided to forego the referee's offer to update additional data bases. We thank you for the opportunity. Hence Tables 1 and 2 have been revised. In contrast Tables 3 to 12B have not been altered and will be resubmitted in the original (February 1) format.
3. The updated numbers of variant incidence have also been added to the description of mutants arising in individual countries. Please see pages 7 (UK), 8 (South Africa), 9 (Brazil), 10 (US), 11 (India), 12 (Russia and France), 14 (Spain and Germany), 15 (China, only a comment).

The comprehensive and rather easy to navigate data in this report will be a most valuable resource for virologists, epidemiologists and global public health agencies and practitioners as well as commercial providers of screening tests or vaccines who continue to monitor and react to the emergence of variants. This manuscript should be accepted for rapid publication with enthusiasm. Notably, I strongly recommend that the authors be accorded the rather unusual opportunity to update their database now and over the coming months should they choose to do so. They are to be congratulated on a Herculean task of consolidating such a vast database into a functional format.

Referee #2 (Comments on Novelty/Model System for Author):

This is an exhaustive characterization of the emergence of SARS-CoV-2 mutants worldwide, using standard bioinformatic pipelines and an open database of SARS-CoV-2 sequences. It is technically well done, and may have high medical impact. The use of pipelines and GISAID is not novel, though the spatiotemporal evaluation of mutations in this exhaustive manner is novel.

Referee #2 (Remarks for Author):

This is an exhaustive evaluation of the emergence of SARS-COV-2 mutants over time, from about 10 predominant mutants to May 2020, followed by an increase in the number of new mutants from June 2020 to December 2020. The work in general is a straightforward characterization of the emergence of these mutations in various countries, and should be useful as a reference for these mutations.

In a manuscript this exhaustive, certain statements are made that require further support or modification.

(1) The authors postulate APOBEC mRNA editing enzymes should be active in the emergence of these

mutations, given the high number of C to U transitions. The authors take this a step further, and postulate that ACE2 binding of the spike has something to do with this process. While an interesting speculation, there is nothing provided in the manuscript by literature or possible mechanism linking APOBEC to ACE2. This needs further explanation or modification.

Response:

Items (1) and (7). As mentioned above (item 1 of referee #1), the remark on a hypothetical association between the ACE-2 receptor and the deaminase functions of APOBEC has been deleted. Text passages on pages 3 (Synopsis) and 17 (second paragraph) have been completely rewritten. Now, we just mention that cellular deaminases, possibly of the APOBEC type, might be responsible for the frequent C to T transitions.

(2) The authors do state briefly that synonymous mutation may have effects outside of protein modification, such as changes in mRNA secondary structure. Given the number of synonymous mutations found, this explanation should be expanded, and possibly include codon choice and translation efficiency as one possible explanation for differences in viral transmissibility and pathogenesis.

Response:

As recommended by the reviewer, we have expanded this discussion on page 13/14.

(3) A separate section on limitations of this study should be included, especially for the non-specialist in this field. In particular, selection bias in sequencing should be emphasized. This is important in that viral sequences may have been provided to GISAID from patients with the most severe disease, which may give an inaccurate picture of viral mutations in less severely affected patients.

Response:

Following up on this recommendation, a new section has been added on page 19: (v) "***Limitations of this study"*** in which a number of items, including the ones suggested by referee #2, have been discussed. Factors that might have led to selection bias in the choice of analyzed sequences have also been discussed on page 5, last paragraph.

(4) The choice of 2% cut off for mutation prevalence to report is as the authors claim arbitrary. This choice should be further justified.

Response:

We have added an explanation on page 6, third paragraph:

Of course, it can be argued that a cut off for the registration of mutants at 2% incidence is arbitrary. However, we cannot predict with certainty which mutations at low incidence of occurrence at present will become more predominant in the future during rapid worldwide viral replication in the current pandemic. A feasible strategy will be to install mutant watch programs and remain on the alert for the rise of new mutations. This strategy can be implemented only by highly efficient SARS-CoV-2 RNA sequencing strategies that will have to be instituted as widely as possible and

without delay. Moreover, 2% was also chosen for publication practicality reasons as well; due to the table lengths becoming very large when using 1% as a cutoff. We thought that 2% was the right balance in being able to flag rare mutations that may become predominant in the future in several countries with readability for interested researchers.

(5) There is a large literature on viral evolution when viruses enter novel hosts, starting with the analysis of the MYXV virus in rabbits in Australia in the 1950s. Examination of this literature, with reference to it in the manuscript, would potentially be helpful in explaining the emergence of mutations in SARS CoV 2 in humans.

Response:

Viral evolution is definitely a most interesting topic. In the context of this study, we can consider it, however, only to a limited extent. Accordingly, we have added paragraph 2 on page 16 plus two new references (MacLean et al, 2021) and [WHO-2019-nCoV-FAQ-Virus_origin-2020.1-eng.pdf (122.6K].

(6) Page 16 at the top: "High current incidence of Covid-19 is paralleled by high numbers of new mutations and variants, although this relationship was not observed in Brazil or Russia." Can the authors provide a possible explanation?

Response:

We surmise that the relevant data on this point have not been available. A short remark has been added on page 16, first paragraph.

(7) Page 17: The link between ACE2 binding and APOBEC enzyme activation is unsupported. Additionally, the authors fail to note that SARS-CoV-2 may have multiple accessory factors for entry, such as CD147, NRP-1, and DC-SIGN among others. How do the mutations affect binding to these accessory factors, and do they have a link to APOBEC enzymes? This section is highly speculative, and need better literature support for its assertions.

Response: Please, see response to item (1)

(8) Page 17: This paragraph at the bottom, discussing the impact of variants on vaccine efficacy, while interesting, is cursory for what is turning out to be a very complicated topic. Perhaps it may be best to reduce or eliminate the part of the paragraph beginning with "The emergence..." since this is speculation.

Response:

We want to keep and consider this discussion, now on page 18 under „iv [***Will the constant selection of new mutants impinge upon the success of therapeutic or vaccination strategies?].*** In the context of present reports on the spread of constantly evolving new variants like B.1.1.7, B.1.135, P1, B.1.429 + B.1.427, B.1.525E and possibly others (see also the updated versions of Tables 1 and 2) this part of the manuscript is of high actual interest and describes the current state of the Covid-19 pandemic realistically. New variants of concern and mutations with yet un-identified pathogenicity might surface as SARS-CoV-2 continues to replicate actively in many parts of the world. We therefore, with due respect, should like to keep this section. Thank you.

(9) The note added during submission at the top of page 18 references a very speculative report, and I am not sure it belongs in this manuscript, which should focus on the categorization of mutations through time.

Response:

The note added during submission now on page 18 together with the reference were removed from the manuscript. ~~Domingo and Perales [J. Virol. 2021, doi: 10.1128/JVI.02437-20~~

(10) I would eliminate the last paragraph of the manuscript "Conceptual bridge.." in that while interesting, it does not add much for the general reader, and the analogies are a bit confusing.

Response:

~~B H, Zhong E, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2021;371~~ Indeed, an interesting paper. However, we agree that the section "Conceptual Bridge" and the accompanying reference now on page 19 would not add much and has been removed from the manuscript.

In addition:

In accordance with journal style, the Materials and Methods section has now been moved to pages 19 to 20 (top), including the comment on Data Availability. There are no changes in this section.

Minor numerical corrections had to be made in Table 13 (highlighted yellow). All calculations have now been based on the results presented in the A Tables. Minor discrepancies, now corrected, arose because we had based some of the calculations on the data in the B Tables.

A "four bullet points" synopsis (page 3, also submitted separately) of the main findings of our study have been incorporated into the manuscript as well as the "Paper explained" statements (page 21).

Comments on items 1) to 15) from pages 2/3 in the Decision Letter

1) Revised manuscript has been formatted .docx. All changes have been highlighted yellow.
2) Quality i n TIFF has been included in the submission.
3) This letter includes both the reviewers' reports and our point-by-point responses (in yellow) that have been intercalated with the referees' comments.
4) A completed author checklist will be provided as a separate file.
5) Walter Doerfler's ORCID ID is: orcid.org/0000-0002-9971-0138
6) Datasets have been deposited as described on page 20 under **Data availability**
7) Not required.
8) Data citations in reference list have been converted to journal style.
9) There is no Supplementary Information for this manuscript.
10) "Paper explained" section can be found on page 21 of the manuscript.
11) Not relevant.

12) Author contributions have been described on page 20 of the manuscript.
13) Conflict of Interests statement is now on page 21 of the manuscript.
14) A Synopsis has been added on page 3 and has also been submitted separately. Moreover, Figure 1 has been added as a TIFF file which could be used as **visual abstract.**
15) We understand the journal to provide that part.

Finally, we wish to thank the referees and you for your constructive comments. We understand that the revised manuscript will now be acceptable for publication.

Best wishes and regards,

Walter Doerfler – for all authors
Professor *emeritus*, Molecular Genetics, Cologne, Guest Professor, Virology, FAU Erlangen

1st Revision - Editorial Decision 14th Apr 2021

14th Apr 2021

Dear Prof. Doerfler,

Thank you for the submission of your revised manuscript to EMBO Molecular Medicine. I am pleased to inform you that we will be able to accept your manuscript pending the following final amendments:

1) Tables: Tables should be uploaded as one file per table. Where there are several parts that don't fit onto one page, they could be combined as tabs in an excel table. Main tables are typeset and can only be displayed in black and white. Please remove the red colour from some of the tables and replace it with bolded font.
2) In the main manuscript file, please do the following:
- Correct/answer the track changes suggested by our data editors by working from the attached/uploaded document.
- Please shorten keywords to max. 5.
- Remove text highlight colour.
- Add callouts for Figure 1 and Table 2.
3) Synopsis: Every published paper now includes a 'Synopsis' to further enhance discoverability. Synopses are displayed on the journal webpage and are freely accessible to all readers. They include separate synopsis image and synopsis text.
- Synopsis image: Please provide a striking image or visual abstract as a high-resolution jpeg file 550 px-wide x (250-400)-px high to illustrate your article.
4) For more information: There is space at the end of each article to list relevant web links for further consultation by our readers. Could you identify some relevant ones and provide such information as well? Some examples are patient associations, relevant databases, OMIM/proteins/genes links, author's websites, etc...
5) Press release: Please inform us as soon as possible and latest at the time of submission of the revised manuscript if you plan a press release for your article so that our publisher could coordinate publication accordingly.
6) Please be aware that we use a unique publishing workflow for COVID-19 papers: a non-typeset PDF of the accepted manuscript is published as "Just Accepted" on our website. With respect to a possible press release, we have the option to not post the "Just Accepted" version if you prefer to wait with the press release for the typeset version. Please let us know whether you agree to publication of a "Just accepted" version or you prefer to wait for the typeset version.
7) As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://embomolmed.embopress.org/content/2/9/329), EMBO Molecular Medicine will publish online a Review Process File (RPF) to accompany accepted manuscripts. This file will be published in conjunction with your paper and will include the anonymous referee reports, your point-by-point response and all pertinent correspondence relating to the manuscript. Let us know whether you agree with the publication of the RPF and as here, if you want to remove or not any figures from it prior to publication. Please note that the Authors checklist will be published at the end of the RPF.
8) Please provide a point-by-point letter INCLUDING my comments as well as the reviewer's reports and your detailed responses (as Word file).

I look forward to reading a new revised version of your manuscript as soon as possible.

Yours sincerely,

Zeljko Durdevic

The authors performed the requested editorial changes.

2nd Revision - Editorial Decision 20th Apr 2021

We are pleased to inform you that your manuscript is accepted for publication and is now being sent to our publisher to be included in the next available issue of EMBO Molecular Medicine.

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Walter Doerfler
EMBO Molecular Medicine
EMM-2021-14062

Reporting Checklist For Life Sciences Articles (Rev. June 2017)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

A- Figures

1. Data

The data shown in figures should satisfy the following conditions:

- ➔ the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
- ➔ figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
- ➔ graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
- ➔ if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified
- ➔ Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

2. Captions

Each figure caption should contain the following information, for each panel where they are relevant:

- ➔ a specification of the experimental system investigated (eg cell line, species name).
- ➔ the assay(s) and method(s) used to carry out the reported observations and measurements
- ➔ an explicit mention of the biological and chemical entity(ies) that are being measured.
- ➔ an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.

- ➔ the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- ➔ a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
- ➔ a statement of how many times the experiment shown was independently replicated in the laboratory.
- ➔ definitions of statistical methods and measures:
  - common tests, such as t-test (please specify whether paired vs. unpaired), simple χ2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - exact statistical test results, e.g., P values = x but not P values < x;
  - definition of 'center values' as median or average;
  - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

In the pink boxes below, please ensure that the answers to the following questions are reported in the manuscript itself. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable). We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repor

http://grants.nih.gov/grants/olaw/olaw.htm
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm
http://ClinicalTrials.gov
http://www.consort-statement.org
http://www.consort-statement.org/checklists/view/32-consort/66-title

http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumo

http://datadryad.org

http://figshare.com

http://www.ncbi.nlm.nih.gov/gap

http://www.ebi.ac.uk/ega

http://biomodels.net/

http://biomodels.net/miriam/
http://jjj.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?	In this study, 383,000 SARS-CoV-2 complete RNA sequences have been analyzed for the occurrence of mutations in the GISAID database which were deposited by Janurary 20, 2021. For an addional > 1,700 RNA sequences, the amino acid changes due to the mutant status have also been reported.
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	NA
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	Only complete sequences were included in the study.
3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.	NA
For animal studies, include a statement about randomization even if no randomization was used.	NA
4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.	NA
4.b. For animal studies, include a statement about blinding even if no blinding was done	NA
5. For every figure, are statistical tests justified as appropriate?	NA
Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.	NA
Is there an estimate of variation within each group of data?	NA

Is the variance similar between the groups that are being statistically compared?	NA

C- Reagents

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right).	NA
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	NA

* for all hyperlinks, please see the table at the top right of the document

D- Animal Models

8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA
10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.	NA

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	NA
12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.	NA
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	NA
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	NA
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	NA
16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list.	NA
17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines.	NA

F- Data Accessibility

18: Provide a "Data Availability" section at the end of the Materials & Methods, listing the accession codes for data generated in this study and deposited in a public database (e.g. RNA-Seq data: Gene Expression Omnibus GSE39462, Proteomics data: PRIDE PXD000208 etc.) Please refer to our author guidelines for 'Data Deposition'. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Functional genomics data e. Proteomics and molecular interactions	In the manuscript. All sequences are available in GISAID.
19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).	All sequences were downloaded from GISAID according to their terms of use.
20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right).	NA
21. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biomodels (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information.	Listed under Methods.

G- Dual use research of concern

22. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could.	NA