Putative host-derived insertions in the genomes of circulating SARS-CoV-2 variants

Yiyan Yang, Keith Dufault-Thompson, Rafaela Fontenele, and Xiaofang Jiang

Corresponding Author(s): Xiaofang Jiang, National Library of Medicine, National Institutes of Health

Review Timeline:

| Event                          | Date          |
|-------------------------------|---------------|
| Submission Date               | February 23, 2022 |
| Editorial Decision            | April 6, 2022  |
| Revision Received             | April 23, 2022 |
| Accepted                      | May 2, 2022    |

Editor: Sergio Baranzini

Reviewer(s): Disclosure of reviewer identity is with reference to reviewer comments included in decision letter(s). The following individuals involved in review of your submission have agreed to reveal their identity: Jan Postberg (Reviewer #2)

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. The original formatting of letters and referee reports may not be reflected in this compilation.)

DOI: https://doi.org/10.1128/msystems.00179-22
Dear Dr. Xiaofang Jiang:

Thank you for submitting your manuscript to mSystems. We have completed our review and I am pleased to inform you that, in principle, we expect to accept it for publication in mSystems. However, acceptance will not be final until you have adequately addressed the reviewer comments.

Thank you for the privilege of reviewing your work. Below you will find instructions from the mSystems editorial office and comments generated during the review.

Preparing Revision Guidelines
To submit your modified manuscript, log onto the eJP submission site at https://msystems.msubmit.net/cgi-bin/main.plex. Go to Author Tasks and click the appropriate manuscript title to begin the revision process. The information that you entered when you first submitted the paper will be displayed. Please update the information as necessary. Here are a few examples of required updates that authors must address:

• Point-by-point responses to the issues raised by the reviewers in a file named "Response to Reviewers," NOT IN YOUR COVER LETTER.
• Upload a compare copy of the manuscript (without figures) as a "Marked-Up Manuscript" file.
• Each figure must be uploaded as a separate file, and any multipanel figures must be assembled into one file.
• Manuscript: A .DOC version of the revised manuscript
• Figures: Editable, high-resolution, individual figure files are required at revision, TIFF or EPS files are preferred

ASM policy requires that data be available to the public upon online posting of the article, so please verify all links to sequence records, if present, and make sure that each number retrieves the full record of the data. If a new accession number is not linked or a link is broken, provide production staff with the correct URL for the record. If the accession numbers for new data are not publicly accessible before the expected online posting of the article, publication of your article may be delayed; please contact the ASM production staff immediately with the expected release date.

For complete guidelines on revision requirements, please see the journal Submission and Review Process requirements at https://journals.asm.org/journal/mSystems/submission-review-process. Submission of a paper that does not conform to mSystems guidelines will delay acceptance of your manuscript.

Corresponding authors may join or renew ASM membership to obtain discounts on publication fees. Need to upgrade your membership level? Please contact Customer Service at Service@asmusa.org.

Thank you for submitting your paper to mSystems.

The ASM Journals program strives for constant improvement in our submission and publication process. Please tell us how we can improve your experience by taking this quick Author Survey.

Sincerely,

Sergio Baranzini

Editor, mSystems

Journals Department
American Society for Microbiology
1752 N St., NW
Washington, DC 20036
E-mail: peerreview@asmusa.org
Reviewer comments:

Reviewer #1 (Comments for the Author):

Yang et al. reported an interesting finding of identifying two genomic insertions of human 18S and 28S rRNA fragments into circulating SARS-CoV-2 RNAs by computationally analyzing available Nanopore long reads and SARS-CoV-2 genomes. The authors proposed that these chimeric reads are generated by viral RdRP through stochastic template switching. They showed that the junction sites tend to be in unpaired state; the host-derived chimeric fragments tend to be from structural RNAs or highly expressed genes; the 18S and 28S fragments in the chimera tend to be in the outside of their structures and potentially accessible to RdRP. However, several major points below should be addressed.

1. For the chimeric RNAs, the authors should also identify those short reads spanning the junctions in available RNA-seq data as another layer of evidence supporting the existence of host-virus mRNA chimera. The reads could be shown underneath the chimera and the numbers can be quantified; the larger the counts, the higher the confidence.

2. What's the molecular basis for the observed viral-host chimera? Do they have the L-TRS and B-TRS like RNA-RNA pairings? Potential RNA-RNA interactions around the junctions mediating the template switching should be investigated.

3. For the 28S rRNA-viral chimera in Figure 3, a 1 nt-shifted insertion sequence is also compatible with the data, as the upstream T (red highlight) can be matched with the downstream T, as shown below. The authors should take care of the mapping problem.

Original insert ‘AGCAGAGCAGCTCCCTCGCTGCGATCT’:
T---------------------------ATAC
TAGCAGAGCAGCTCCCTCGCTGCGATCTATAT
TAGCAGAGCAGCTCCCTCGCTGCGATCTATTG

Another compatible insert ‘TAGCAGAGCAGCTCCCTCGCTGCGATC’:
---------------------------TATAC
TAGCAGAGCAGCTCCCTCGCTGCGATCTATAT
TAGCAGAGCAGCTCCCTCGCTGCGATCTATTG

Reviewer #2 (Comments for the Author):

Yang et al. provide an interesting and relevant manuscript entitled 'Putative host-derived insertions in the genomes of circulating SARS-CoV-2 variants'. Therein they carefully and convincingly conclude that host-derived rRNA sequences could be source of SARS-CoV-2 genome insertions. Importantly, since they postulate a stochastic template switching mechanism rRNA insertion does not seem to be a specific event, but in this study rRNAs serve as plausible bioindicators. From my perspective, few queries remain, which (in parts optionally) should be addressed in a revised manuscript (minor revisions).

1. Can the current emergence of deltacron/deltamicron lineages be taken as further evidence for template switching, which can happen, when one host is infected by 2 variants simultaneously, and its importance for virus evolution?

2. Line 39 (Introduction): Sentence is slightly misleading. "frequently acquired" does not properly describe the content of the single citation. Therein, multiple insertion in a restricted N-terminal Spikes region are described.

3. Line 48: “might be advantageous” - please consider adding: "or neutral".

4. Line 93: separate SARS-CoV-2 genomes

5. Please consider my general thought, if appropriate: Theoretically, it should be possible to detect insertions of putative zoonotic origin host species, I assume the putative rRNA sequences are much too short to distinguish between mammals. Chapter lines 174 to 199: No plausible source for insertion could be identified. Was your search restricted to humans, or did you also cross-check putative earlier host species? Can data-mining in genomes of other coronaviruses support your data/my thought, regardless whether these data would come from human pathogenic coronaviruses or other hosts’ pathogenic coronaviruses?
Response to Reviewers

Dear Dr. Sergio Baranzini,

Thank you for giving us the opportunity to submit a revised draft of our manuscript titled "Putative host-derived insertions in the genomes of circulating SARS-CoV-2 variants" to *mSystems*. We appreciate the time and effort that you and the reviewers have dedicated to providing your valuable feedback on the manuscript. We have been able to incorporate changes to reflect most of the suggestions provided by the reviewers. We mapped the chimeric reads with short reads from the same samples to add another layer of evidence supporting their existence. We found an error in our previous code and have fixed it and updated the data. All the conclusions remain the same. All changes have been marked in red within the marked-up manuscript.

Here is a point-by-point response to the reviewers’ comments.

**Reviewer comments:**

Reviewer #1 (Comments for the Author):

Yang et al. reported an interesting finding of identifying two genomic insertions of human 18S and 28S rRNA fragments into circulating SARS-CoV-2 RNAs by computationally analyzing available Nanopore long reads and SARS-CoV-2 genomes. The authors proposed that these chimeric reads are generated by viral RdRP through stochastic template switching. They showed that the junction sites tend to be in unpaired state; the host-derived chimeric fragments tend to be from structural RNAs or highly expressed genes; the 18S and 28S fragments in the chimera tend to be in the outside of their structures and potentially accessible to RdRP. However, several major points below should be addressed.

**R1.1 For the chimeric RNAs, the authors should also identify those short reads spanning the junctions in available RNA-seq data as another layer of evidence supporting the existence of host-virus mRNA chimera. The reads could be shown underneath the chimera and the numbers can be quantified; the larger the counts, the higher the confidence.**

Thank you for your suggestion. We managed to find six paired-end sequencing datasets on five samples with corresponding direct RNA-sequencing data. By mapping short reads onto the chimeric reads, read pairs spanning the junction regions of chimeric reads were counted. We have also added a new Supplementary Table 2 including the numbers of read pairs supporting the chimeric reads.

We have revised the Results (lines 110-115) and Methods (lines 413-420) to expand upon our description of these results.

In Results: "Additionally, chimeric reads detected in five samples were further investigated using paired-end sequencing short reads from the same samples (Supplementary Table 2). Approximately 1.4% (5 out of 357) of chimeric reads were supported by at least five read pairs spanning the junctions. This finding implies that a small fraction of the host-viral chimeric mRNA molecules could function as templates for RNA replication."
In Methods: "We collected paired-end sequencing data on five samples with corresponding direct RNA-sequencing data. The short reads were first preprocessed with fastp v0.23.1 (44) and then mapped to the chimeric reads from the same samples by using Minimap2 v2.23 (42) with options ‘-ax sr -w 5’ to tolerate the high error rate of the Nanopore direct RNA-sequencing reads (45). Read pairs spanning the junctions were detected and counted with a custom script. The numbers of read pairs supporting the chimeric reads are provided in Supplementary Table 2."

R1.2 What's the molecular basis for the observed viral-host chimera? Do they have the L-TRS and B-TRS like RNA-RNA pairings? Potential RNA-RNA interactions around the junctions mediating the template switching should be investigated.

We investigated multiple aspects of the regions around the junctions to address this question. We didn’t find the full SARS-CoV-2 TRS core sequence (i.e., ACGAAC) in any of the junction regions. We aligned the host transcriptome sequence and the viral genome sequence mapped by each chimeric read’s junction region to check if there were any matching stretches of ribonucleotides. Although long stretches of identical sequences were found in some cases, a majority of them have no or short common sequences (please see figure below). Identical sequences at least 6-nt long in the junction regions were observed in only 4.97% of human-viral chimeric reads and 3.51% of monkey-viral chimeric reads. No conserved motifs were found in the junction regions. These numbers are rough estimates due to 1) the difficulties in identifying the exact chimeric junctions due to the high error rate of Nanopore sequencing, 2) different alignments can be accounted for the sequences around the junction, 3) the lack of observed conserved motifs (unlike the core sequences) to support the base-pairings, 4) no observed consistent pattern given the limited number of chimeric reads we identified. Overall, the evidence shown here is inconclusive, and we are not confident to determine the specific molecular basis for the observed viral-host chimeric reads. It is also possible that different mechanisms might exist to generate the chimeric reads.

In addition, we also added sentences on the lines 354-359 in Discussion: "The accurate determination of the exact junction boundaries and potential base-pairings were hindered by the high error rate of 14% in direct RNA-sequencing data and the limited number of host-viral chimeras detected in this study. The exact molecular basis for the viral-host chimera remains unclear and future investigation with larger sets of error-corrected direct RNA-seq data of SARS-CoV-2 could be beneficial to address this question."
The distribution of identical sequence lengths in (A) human-viral chimeric reads and (B) monkey-viral chimeric reads. (C) An example showing the longest match (12 nt) between the human region and viral region of a chimeric read.

R1.3 For the 28S rRNA-viral chimera in Figure 3, a 1 nt-shifted insertion sequence is also compatible with the data, as the upstream T (red highlight) can be matched with the downstream T, as shown below. The authors should take care of the mapping problem.

Original insert 'AGCAGAGCAGCTCCCTCGCTGCATCT':
T---------------------------ATAC
TAGCAGAGCAGCTCCCTCGCTGCATCTATAT
TAGCAGAGCAGCTCCCTCGCTGCATCTATTG

Another compatible insert 'TAGCAGAGCAGCTCCCTCGCTGCATCT':
---------------------------TATAC
TAGCAGAGCAGCTCCCTCGCTGCATCTATAT
TAGCAGAGCAGCTCCCTCGCTGCATCTATTG

There are five possible alignments for the inserted sequence (please see figure below). We chose to show the alignment that results in the insertion inserted after the 3rd position of 2285th codon in ORF1ab and with the longest length hit to the 28S rRNA sequence. We have added a description of this to the Figure 3
legend "There are five possible alignments for mapping this insertion to the reference. Only the alignment with the sequence inserted after the 3rd position of 2285th codon in ORF1ab is shown."

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| B |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| C |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| E |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

Five possible alignments for the 28S rRNA-derived insert. The match to 28S rRNA is highlighted in red. Codons are colored in blue and yellow in turn. The amino acids encoded by the nucleotides covering the insert are listed below. The alignment (D) is chosen and shown in the manuscript.

Reviewer #2 (Comments for the Author):

Yang et al. provide an interesting and relevant manuscript entitled 'Putative host-derived insertions in the genomes of circulating SARS-CoV-2 variants'. Therein they carefully and convincingly conclude that host-derived rRNA sequences could be source of SARS-CoV-2 genome insertions. Importantly, since they postulate a stochastic template switching mechanism rRNA insertion does not seem to be a specific event, but in this study rRNAs serve as plausible bioindicators. From my perspective, few queries remain, which (in parts optionally) should be addressed in a revised manuscript (minor revisions).

R2.1 Can the current emergence of deltacron/deltamicron lineages be taken as further evidence for template switching, which can happen, when one host is infected by 2 variants simultaneously, and its importance for virus evolution?
Yes. We think the current emergence of the deltacron variant can be taken as further evidence for template switching in coronavirus. An early review suggested that template switching is a major factor contributing to recombination in RNA viruses, including the Coronaviridae (Simon-Loriere and Holmes, 2011). Recently, researchers have found evidence for Delta and Omicron variant co-infections and recombination and implied the recombination may be mediated by template switching (Bolze et al., 2022). We have cited these two studies to highlight the importance of template switching to virus evolution and modified the sentence on lines 76-79 in the manuscript: "While template switching events between coronaviruses are common (10-13) and likely contribute to the emergence of SARS-CoV-2 lineages including the deltacron variant (14), template switching events between coronaviruses and host RNAs are rarely documented (15, 16)."

**R2.2 Line 39 (Introduction): Sentence is slightly misleading. "frequently acquired" does not properly describe the content of the single citation. Therein, multiple insertion in a restricted N-terminal Spikes region are described.**

We added additional citations to this sentence to provide other examples of the acquisition of insertions in SARS-CoV-2 lineages.

**R2.3 Line 48: "might be advantageous" - please consider adding: "or neutral".**

We have modified the sentence as "... the fact that variants carrying those insertions have circulated for long periods suggests that they might be advantageous or neutral for the transmission."

**R2.4 Line 93: separate SARS-CoV-2 genomes**

We have corrected this typo.

**R2.5 Please consider my general thought, if appropriate: Theoretically, it should be possible to detect insertions of putative zoonotic origin host species, I assume the putative rRNA sequences are much too short to distinguish between mammals. Chapter lines 174 to 199: No plausible source for insertion could be identified. Was your search restricted to humans, or did you also cross-check putative earlier host species? Can data-mining in genomes of other coronaviruses support your data/my thought, regardless whether these data would come from human pathogenic coronaviruses or other hosts' pathogenic coronaviruses?**

Unfortunately, the putative rRNA sequences are too short to distinguish between different mammals. In fact, the 28S-derived insertion and 18S-derived insertion could be found, respectively, in the LSU rRNA and SSU rRNA of multiple mammal species (lines 251-253 and 297-299 in the main text).

We didn’t restrict our search to humans and searched potential long insertions against the NCBI non-redundant nucleotide database and a collection of coronavirus genomes to explore their possible origins. For the 21-nt long insertion of unknown origin, it has an exact match from a moth genome but with a high E-value of 0.3. We further extended the sequence to two 22-nt long sequences by adding one nucleotide upstream or downstream along the viral genome and re-searched them in the database. However, no 100% hits were found in both of them, and the origin of this insertion remains unclear.

To clarify this, we modified lines 208-210 in Results as "Unfortunately, no plausible source for this insertion could be identified using a BLAST search in the NCBI non-redundant nucleotide database and a collection of coronavirus genomes with a cutoff E-value of 1e-2, and it was not analyzed further.", and added on lines 452-455 in Methods "Insertions greater than or equal to 21 nucleotides long and found outside of the 5’ and 3’ untranslated regions of the viral genomes were kept. They were searched in the
NCBI non-redundant nucleotide database and a collection of coronavirus genomes with BLASTN (E-value $\leq 1e-2$) (32) to explore their possible origins.

References:
Simon-Loriere E, Holmes EC. Why do RNA viruses recombine?. Nature Reviews Microbiology. 2011 Aug;9(8):617-26.
Bolze A, White S, Basler T, Dei Rossi A, Roychoudhury P, Greninger AL, Hayashibara K, Wyman D, Kil E, Dai H, Cassens T. Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. medRxiv. 2022 Jan 1.
May 2, 2022

Dr. Xiaofang Jiang
National Library of Medicine, National Institutes of Health
Bethesda

Re: mSystems00179-22R1 (Putative host-derived insertions in the genomes of circulating SARS-CoV-2 variants)

Dear Dr. Xiaofang Jiang:

Your manuscript has been accepted, and I am forwarding it to the ASM Journals Department for publication. For your reference, ASM Journals' address is given below. Before it can be scheduled for publication, your manuscript will be checked by the mSystems production staff to make sure that all elements meet the technical requirements for publication. They will contact you if anything needs to be revised before copyediting and production can begin. Otherwise, you will be notified when your proofs are ready to be viewed.

ASM policy requires that data be available to the public upon online posting of the article, so please verify all links to sequence records, if present, and make sure that each number retrieves the full record of the data. If a new accession number is not linked or a link is broken, provide production staff with the correct URL for the record. If the accession numbers for new data are not publicly accessible before the expected online posting of the article, publication of your article may be delayed; please contact the ASM production staff immediately with the expected release date.

As an open-access publication, mSystems receives no financial support from paid subscriptions and depends on authors' prompt payment of publication fees as soon as their articles are accepted.

Publication Fees:
You will be contacted separately about payment when the proofs are issued; please follow the instructions in that e-mail. Arrangements for payment must be made before your article is published. For a complete list of Publication Fees, including supplemental material costs, please visit our website.

Corresponding authors may join or renew ASM membership to obtain discounts on publication fees. Need to upgrade your membership level? Please contact Customer Service at Service@asmusa.org.

For mSystems research articles, you are welcome to submit a short author video for your recently accepted paper. Videos are normally 1 minute long and are a great opportunity for junior authors to get greater exposure. Importantly, this video will not hold up the publication of your paper, and you can submit it at any time.

Details of the video are:
· Minimum resolution of 1280 x 720
· .mov or .mp4. video format
· Provide video in the highest quality possible, but do not exceed 1080p
· Provide a still/profile picture that is 640 (w) x 720 (h) max
· Provide the script that was used

We recognize that the video files can become quite large, and so to avoid quality loss ASM suggests sending the video file via https://www.wetransfer.com/. When you have a final version of the video and the still ready to share, please send it to mSystems staff at msystems@asmusa.org.

For mSystems research articles, if you would like to submit an image for consideration as the Featured Image for an issue, please contact mSystems staff at msystems@asmusa.org.

Thank you for submitting your paper to mSystems.

Sincerely,

Sergio Baranzini
Editor, mSystems
