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

Yiyang Yang[1], Keith Dufault-Thompson[1], Rafaela Salgado Fontenele[1], Xiaofang Jiang[1,*]

[1]National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[*]Corresponding author: xiaofang.jiang@nih.gov

Abstract

Insertions in the SARS-CoV-2 genome have the potential to drive viral evolution, but the source of the insertions is often unknown. Recent proposals have suggested that human RNAs could be a source of some insertions, but the small size of many insertions makes this difficult to confirm. Through an analysis of available direct RNA sequencing data from SARS-CoV-2 infected cells, we show that viral-host chimeric RNAs are formed through what are likely stochastic RNA-dependent RNA polymerase template switching events. Through an analysis of the publicly available GISAID SARS-CoV-2 genome collection, we then identified two genomic insertions in circulating SARS-CoV-2 variants that are identical to regions of the human 18S and 28S rRNAs. These results provide direct evidence of the formation of viral-host chimeric sequences and the integration of host genetic material into the SARS-CoV-2 genome, highlighting the potential importance of host-derived insertions in viral evolution.

Introduction

During the COVID-19 pandemic, insertions have been frequently acquired in the SARS-CoV-2 lineages[1]. Insertions have been associated with several globally circulating lineages, including the insertion of one amino acid at position 146 of the S protein (ins146N) of the variant of interest Mu (B.1.621)[2], insertions at the recurrent insertion site 214 of the NTD region on the S protein that occurred in the lineages B.1.214.2 (ins214TDR) and A.2.5 (ins214AAG)[1], and the insertion ins214EPE in the recently-emerged variant of concern Omicron[3]. Although there is insufficient evidence to show the direct impact these insertions have on viral spread and interference on immune response, the fact that variants carrying those insertions have circulated for long periods suggests that these insertions might be advantageous for the transmission. Results from a long-term *in vitro* experiment where SARS-CoV-2 was co-incubated with highly neutralizing antibodies have also shown that an 11 amino acid insertion (ins248KTRNKSTSRRE) at the NTD N5 loop of the S protein was able to drive antibody escape suggesting a potential role of insertions in enhancing infectivity and virulence[4]. Taken together, insertions have the potential to increase genetic diversity in SARS-CoV-2 and contribute to the continued evolution of the virus.

Previous research has shown that most small insertions in the SARS-CoV-2 genome likely originated from template sliding, local duplication, or template switching between viruses[5]. Longer insertions (equal or larger than nine nucleotides) have been detected in multiple coronavirus genomes, including in variants of concern like the Omicron variant, but their origin remains unknown. Host genetic material has been suggested as a possible source for these insertions[3,6]. Venkatakrishnan et al. suggested that the unique insertion (ins214EPE) in the Omicron variant could have originated from the human common cold virus HCoV-229E or the human genome based on BLAST search[3], and the human genome has been speculated to be the source of multiple other small insertions[6]. However, given that these insertion sequences are typically short, sequence comparisons tend to be less informative, and false-positive matches have a high chance of occurring. Additionally, coronavirus replication occurs in modified endoplasmic reticulum-derived double-membrane vesicles, providing a physical barrier between viral and host genetic material[7], and coronavirus replication complexes are known to contain enzymes with proofreading activity[8], both of which likely play roles in limiting the formation of host-virus chimeric sequences.

Human-derived insertions in the SARS-CoV-2 genome would likely be generated through RdRp-driven template switching events between SARS-CoV-2 and host mRNA. However, this mechanism, although possible in theory, has never been documented to our knowledge. Chimeric reads between SARS-CoV-2 RNA and human RNA have been detected but were interpreted as a signal of SARS-CoV-2 integration into the human genome[9]. However, this study is controversial and other studies suggested that the chimeric reads were likely to be template switching artifacts mediated by reverse transcriptase or PCR during library preparation[10–13]. One possible explanation that was largely omitted in these studies is that the SARS-CoV-2-host chimeric RNA could be generated by RdRp-driven template switching.

Here, to investigate the possible existence of SARS-CoV-2-host chimeric RNA, we take advantage of the publically available Nanopore direct RNA sequencing data of SARS-CoV-2. Direct RNA-seq sequences the individual polyadenylated RNAs directly mitigating the possible formation of chimeric reads during library preparation or amplification. We first identified SARS-CoV-2-host chimeric RNA from direct RNA-seq data and showed that RdRp-driven template switching between SARS-CoV-2 and host mRNA occurs, but it is infrequent and stochastic. We also found that highly expressed host genes and structural RNA genes have a higher chance to be observed in chimeric RNA reads. We then systematically analyzed the SARS-CoV-2genomes deposited in the GISAID database[14], resulting in the identification of two insertions in functional SARS-CoV-2 genomes that likely originated from the host 18S and 28S rRNAs.

Results

Host-virus mRNA chimera are rare but do exist

We first analyzed direct RNA-seq data from SARS-CoV-2 infected cell lines to identify sequences formed from chimeric host-viral RNAs. The direct RNA-seq data were quality filtered and mapped to both the host and SARS-CoV-2 transcriptomes to identify potential chimeric sequences. Out of the 30 samples that were analyzed, host-viral chimeric reads were detected in 16 of the samples with an average of 0.027% (standard deviation 0.045%) of the reads mapped to SARS-CoV-2 being chimeric (Supplementary Table 1). Chimeric reads were typically rare, making up 0.207% of one sample, but less than 0.06% of the other 15 samples, and these rates may be an overestimation due to the cell lines used compared to what would be observed in *in vivo* conditions.

We then analyzed the chimeric reads to identify trends in how the viral and host RNA sequences were joined. All the viral-derived sequences in chimeric reads were annotated as positive-sense RNA and a majority (92.24%) of the reads contained host-derived positive-sense sequences. Upon further examination, the few host reads that were identified as being negative-sense were largely long non-coding RNAs that were present in the raw reads as the negative-sense sequences making it likely that they were mis-annotated rather than actually being derived from negative-sense RNA. These results suggest that the host-viral chimeric sequences are not the result of the integration of the viral genetic material into the host genome, which would have resulted in a nearly equal mix of positive and negative sense viral sequences[9]. Most likely, these host-viral chimeric sequences were created from positive-to-positive-strand template switching events[15,16].

Viral-host chimeric read formation is likely a stochastic process

The chimeric reads were then analyzed to determine if there were any patterns in the composition of the sequences and in which positions relative to the references they were formed. Both viral to host and host to viral chimeric sequences were detected in the direct RNA-seq data, but the chimeric reads did not show a preference for either organization (Supplementary Table 1). Both types of sequences were seen in approximately the same frequency, with viral to host reads making up 58% of the chimeric sequences and host to viral reads making up 42%. This lack of strong preference may indicate that host RNA can be readily recognized by viral RdRp, but other factors

like the exclusion of host RNA by the formation of the double-membrane vesicles might prevent the formation of chimeric RNAs. When examining the positions of the junctions on the viral RNA sequences, we found there was a bias toward the junction sites being located in the dense coding region near the three prime end of the sequence, with fewer junctions being identified in the ORF1ab genes, the largest region of the genome (Fig.1). This is likely due to the ORF1ab region not being retained in the canonical SARS-CoV-2 subgenomic RNAs resulting in fewer viral RNAs being synthesized with these regions that could form chimeric RNAs[17]. It suggests that the process by which chimeric sequences are formed is likely stochastic, depending on the availability of template RNA molecules.

Previous studies have also found that indel formation and template switching events preferentially occur in the loops and stems formed in the RNA secondary structure[5,18]. To assess if the location of the junctions in the chimeric reads is associated with the viral RNA secondary structure, a permutation test was used to investigate if the junction sites were commonly located in stems (positions that form base-pairs) or non-stem regions (non-base-paired positions) in the viral RNA. The results of this test showed a significant (P-value equals 0.005) preference for the formation of junctions in non-base-paired regions of the RNA secondary structure (Fig.1), suggesting that these chimera forming template switching events are more likely to happen at the non-base-paired segments of the RNA. We speculate that these non-base-paired regions of the SARS-CoV-2 RNA may be more susceptible to stochastic template-switching events due to their more "open" configurations where the viral RdRp could easily attach or detach from these regions.

An examination of the types of human gene sequences found in the chimeric sequences revealed an enrichment of non-coding RNAs and highly expressed genes. We found that a disproportionate number of non-coding RNAs, mainly long non-coding RNAs (lncRNAs), were forming parts of the chimeric reads compared to their abundance in the human genomes. These non-coding RNA chimeric sequences made up 9.2% and 19% of the chimeric reads detected in the Caco and Calu cell lines, respectively, while non-coding sequences made up only 4% of the genes annotated in the human genome. This enrichment of non-coding RNA chimeric sequences was tested using Fisher's exact test confirming that the trend was significant (Caco cells: odds ratio=2.3, P-value=0.037; Calu cells: odds ratio=5.6, P-value=5.06e-6). When analyzed in the context of the expression level of the host genes in each sample, we also observed an enrichment for highly expressed genes forming parts of the chimeric sequences (Fig.2). This enrichment was confirmed through the Mann-Whitney U tests showing that the trend was significant in the two human cell lines (P-value < 2.2e-16 for both) and the *Chlorocebus sabaeus* (green monkey) cell line (P-value < 2.2e-16). These results appear to highlight two groups of sequences that are forming chimeric RNAs, structural RNAs like lncRNAs, which may be susceptible due to their secondary structures, or highly expressed genes, which would have more RNA molecules present for template-switching events to occur with. This suggests that the formation of chimeras is largely stochastic, with factors like the abundance of RNAs playing a large role, but that certain RNA molecules may be more susceptible to these events due to their structure.

Systematic search for host-derived insertions in SARS-CoV-2 genomes

We performed a survey of the GISAID SARS-CoV-2 genomes to identify insertions with potential host origins. Insertions were detected based on alignments and comparison to the Wuhan-Hu-1/2019 reference genome. Only insertions greater than or equal to 21 nucleotides long and that were found outside of the 5' and 3' untranslated regions were considered in subsequent analyses (Supplementary Table 2). Of the 36 insertions that were found, 17 of them were found in multiple SARS-CoV-2 genomes but were not monophyletic. Upon further examination, the genomes containing these insertions tended to be sequenced by the same labs around the same times making it likely that these detected insertions are due to library preparation or sequencing errors rather than the result of multiple independent insertion events in different viral lineages. Of the 19 other insertions, 16 of them were only detected in a single genome, and while many of these had plausible hits to human genes, it is difficult to assess if these are true insertions or library preparation or sequencing artifacts due to their limited presence.

The three remaining insertions were from monophyletic virus variants and were further examined to determine if they had plausible homologous sequences in the human genome. Two of the insertions were found to be identical to conserved segments of the 28S and 18S rRNAs and were analyzed further. The remaining insertion was 21 nucleotides long and was found in 6 SARS-CoV-2 genomes of the Alpha B.1.1.7 lineage. These genomes were collected in early March of 2021 from England, United Kingdom by two laboratories, and sequenced at the same location using the same sequencing platform. The raw reads were available for two of the genomes and were examined directly, providing confirmation that the insertion was present and likely not an artifact. Unfortunately, no plausible source for this insertion was able to be identified and it was not analyzed further.

28S rRNA-derived insertion in SARS-CoV-2 genomes

We detected a 27-nucleotide long insertion in five SARS-CoV-2 genomes (Supplementary Table 3 and Fig.3a) at position 7120 of the reference genome (China/Wuhan-Hu-1/2019). By performing the Blast search for this insertion against the human transcripts (Release 109), an exact match (E-value: 2e-06) of this insertion was found in the nucleotide sequences of 28S ribosomal RNA (Fig.3b). We observed an extra three overlapping bases in the pairwise alignment of SARS-CoV-2 variants containing the insertion and the human 28S rRNA sequence, extending the length of identity nucleotide bases from 27 nucleotides to 30 nucleotides. The identical region was located at positions 4969-4998 of the human 28S rRNA (based on the structure of PDB 5AJ0 Chain A2) and makes up part of the highly conserved loop 94 stem of domain 7 of the rRNA molecule according to the Gorski et al.'s segmentation of human 28S rRNA[19] (Fig.3b).

Due to the high level of sequence conservation of 28S rRNA, asserting the origin of the insertion-related 30 nucleotide sequences is impossible based on sequence identity alone. In the human genome (GRCh38 release 105), three 28S rRNA gene copies in chromosome 21 and one copy in chromosome 12 contain the exact 30 nucleotide sequences. When we searched the 30 nucleotide

sequences in the LSU rRNA database downloaded from SILVA[20], 98 organisms were found to contain the sequences. The last common ancestor of these 98 organisms is bony vertebrates (*Euteleostomi*). Given the fact that the insertion emerged from the SARS-CoV-2 variant circulating in humans, the originating organism of the 28S rRNA-derived insertion is most likely humans.

The nine amino acid insertion is located at position 1467 of the ectodomain (3Ecto) in the Nsp3 protein, the only domain of this protein located on the lumenal side of the endoplasmic reticulum (Fig.3c). Nsp3 along with Nsp4 and Nsp6 have been shown to be involved in the formation of double membrane vesicles in coronavirus infected cells[21,22]. The 3Ecto domain is specifically involved in the recruitment of Nsp4 and has been shown to be an essential component of Nsp3 for correct double-membrane vesicle formation[21]. At this point, it is unclear if this insertion would have had an effect on viral fitness, but given its location in the 3Ecto domain, it is possible that the insertion could have an effect on the interactions between Nsp3 and other proteins and on the membrane rearrangement process.

The five SARS-CoV-2 genomes with the 28S rRNA-derived insertion belonged to a monophyletic clade within the AY.103 group of the delta lineage[23] (Fig.3a). The AY.103 variant was first detected worldwide on January 1st, 2021 and in the USA on January 2nd, 2021. The clade containing the 28S rRNA-derived insertion is defined by five nucleotide mutations (T7900C, A10420T, C18646T, C25721T, and C29668T). By September 2021, AY.103 had become the most common delta lineage in the United States and has continued to be responsible for a significant fraction of cases until the recent emergence of the Omicron variant[24]. The five genomes containing the 28S rRNA-derived insertion were collected between October 9th and November 10th in 2021 from the states of Washington, Idaho, Massachusetts, and California, indicating that these variants were likely being transmitted over this timeframe, but the extent to which it was being spread seems to be low as Idaho was the only state where multiple genomes were collected from and no genomes containing the insertion have been reported since.

The five genomes containing the 28S rRNA-derived insertions were collected by different laboratories and were sequenced on different sequencing platforms, making it extremely unlikely that laboratory error is responsible for the presence of the insertions. Three other variants with genomes fall within the same clade as the genomes with the insertion but did not have any detected insertion sequence in their genomes. Unfortunately, the raw data from the genome sequencing of these viruses was not available and could not be checked to confirm if the insertion was actually missing or was absent in the assembly. Based on the limited spread of the viruses containing the 28S rRNA-derived insertion, it is likely that the insertion might not confer phenotypic advantages or is possibly disadvantageous to the virus. Nonetheless, our data show that AY.103 lineages containing this insertion were viable and were transmitted for a short period of time.

18S rRNA-derived insertion in SARS-CoV-2 genomes

A 24-nucleotide insertion was detected in two genomes at position 27494 in the genome of the reference genome (China/Wuhan-Hu-1/2019) (Supplementary Table 3). A sequence search against

human transcripts (Release 109) was performed using BLAST[25], resulting in the identification of an exact match to a 24 nucleotide stretch (E-value: 2e-5) of the 18S rRNA sequence. When aligned to the full 18S rRNA sequence, it was found that the identical region extended one additional nucleotide outside of the insertion region, bringing the identical stretch to 25 nucleotides (Fig.4a). The insertion was identical to a highly conserved region of the 18S rRNA (at positions 399-423 in 18S rRNA), consisting of a portion of the helix 12 of the 5' domain[26,27]). In the human genome alone there are five copies of the 18S rRNA gene on chromosome 21 that contain identical matches for this 25 nucleotide sequence. When compared to the LSU rRNA SILVA database[20], identical sequences were found in the 18S sequences of 2609 organisms, which had a common ancestor of animals with bilateral symmetry (*Bilateria*) indicating that this is a highly conserved region of the rRNA sequence across many animals. Considering that the viral samples were circulating in human populations, it is highly likely that the insertion was derived from human 18S rRNA.

The insertion is in the SARS-CoV-2 ORF7a protein, encoding an eight amino acid sequence that is located between the proline and cysteine at positions 34 and 35 in the reference protein sequence (Fig.4b). The cysteine at position 35 is known to form a disulfide bond with a cysteine at position 67 and is thought to help stabilize the beta-sheet structure[28,29] and the possible functions of the proline at position 34 are not known. The ORF7a protein has been shown to contain an immunoglobulin-like ectodomain between residues 16 and 96 on the protein which is thought to have a role of binding to human immune cells and modulating immune response[28–30]. Given the proximity of the insert to the disulfide bond forming cysteine at position 34 and the size of the insert it is possible that this insert would have an effect on the overall structure and immunoregulatory functions of ORF7a, but without additional evidence, the effect of this insertion on the fitness of the virus remains unknown.

The two genomes containing the 18S rRNA insertion were from the same clade in the Alpha B.1.1.7 SARS-CoV-2 lineage, which was first identified in England, United Kingdom in mid-December of 2020 (Fig. 4c). This variant was designated as a variant of concern due to its transmissibility and large number of mutations and quickly became the dominant variant in England while spreading to other countries[31]. The genomes containing the 18S rRNA-derived insertion, along with the other four genes in the same clade, were collected in April and May of 2021 in Oregon, United States. The genomes from the variants containing the insertion were collected and sequenced by different labs using different sequencing platforms, making it unlikely that the insertion was a sequencing or library preparation artifact. We did not detect the insertion in any of the other four genomes from this clade, indicating that either they do not have the insertion, they have it but it was not detected, or that the insertion was only acquired in a sub-clade within this group. After May of 2021, no new genomes containing this insertion were collected, indicating that the period during which these lineages were circulating may have been brief. While these viral variants seem to be viable and transmitted for a short period of time, the insertion likely does not confer a significant advantage or may be disadvantageous for the virus resulting in its limited spread.

Discussion

Insertions in the SARS-CoV-2 genome can be introduced through multiple mechanisms and have the potential to give rise to new variants with enhanced infectivity, pathogenicity, and antibody escape[4,5], but the source of these insertions is often difficult to determine and has been hotly debated[3,6]. Leveraging available direct RNA sequencing data and an analysis of SARS-CoV-2 genomes, we have found evidence of the formation of viral-host chimeric RNA sequences and described two novel human-derived genomic insertions present in circulating variants of SARS-CoV-2.

Through our screening of direct RNA-seq data from SARS-CoV-2 infected cell lines we found that viral-host chimeric RNAs were rare but were present in approximately half of the samples analyzed. The chimeric reads all contained positive-sense viral RNA sequences indicating that these chimeric sequences are not the result of the integration of the viral genetic material into the host genome, which would have resulted in a nearly equal mix of positive and negative sense viral sequences[9]. This process does appear to be stochastic in nature though, with no preference for starting with host or viral sequences during chimera formation and a higher frequency of chimeras being formed with highly expressed genes in the cells. The regions in the RNA where these template switching events occur appears to be influenced by the secondary structure of the viral RNA, possibly due to certain structures being more susceptible to template switching events similar to what has been reported in previous studies[5,18].

The formation of host-viral chimeric mRNAs or subgenomic RNAs could be transient events, not having a long-term impact on viral fitness, but the possibility of human-derived insertions in the coronavirus genomes could have significant implications considering the role that genomic insertions seem to have in the evolution of new SARS-CoV-2 variants[3,4]. The putative 18S and 28S-derived insertions were identified in circulating variants of the SARS-CoV-2, and while these particular variants did not seem to spread widely, they do provide evidence that human genetic material can be a source of genomic insertions in SARS-CoV-2. Interestingly, rRNAs have been established to be a source of insertions in influenza genomes, in some cases resulting in significantly more pathogenic viral variants[32,33]. It has been speculated that these recombination events often occur with host rRNAs due to their abundance in the cells, the presence of recombination hotspots on rRNA molecules, and the utilization of host rRNAs during viral replication[32]. Similar factors may play a role in the formation of these rRNA-derived insertions in SARS-CoV-2, but the formation of double-membrane vesicles during SARS-CoV-2 would seemingly complicate this process. There may be unintentional capture of host RNAs inside of the double-membrane vesicles during their formation or some crossover of host RNA from the cytosol, but evidence of this is lacking and warrants further investigation.

Overall, our results suggest that viral-host chimeric sequences can be formed, likely through stochastic RdRp template switching events. Furthermore, we have identified two long insertions in SARS-CoV-2 genomes in previously circulating variants which are likely derived from human

ribosomal RNAs. While the source of smaller insertions that are present in many SARS-CoV-2 genomes are still difficult to identify due to their short lengths, these results provide evidence that bolsters the hypothesis that some of them are derived from human genetic material. The mechanisms at work in the formation of these chimeric RNAs and genomic insertions are still unclear but warrant further study considering the potential importance of these processes in viral evolution and the emergence of new variants.

Methods

Identification of host-virus chimeric reads in SARS-CoV-2 direct-RNA seq data

The nanopore direct RNA-seq data from SARS-CoV-2 infected cell lines were downloaded from the NCBI SRA database (Supplementary Table 1). All reads were quality trimmed using NanoFilt v2.8.0[34], to remove the first 50 nucleotides of each read and require an average quality score of at least 10 over the length of the read. The trimmed reads were then mapped using Minimap2 v2.23[35] to the SARS-CoV-2 reference genome (NCBI GenBank accession: NC_045512.2)[36], and either a reference *Chlorocebus sabaeus* transcriptome (ftp://ftp.ensembl.org/pub/release-105/fasta/chlorocebus_sabaeus/) or human transcriptome (ftp://ftp.ensembl.org/pub/release-105/fasta/homo_sapiens/). The mapping files were converted to the Pairwise mApping Format (PAF) using the paftools script that is part of Minimap2[35] Reads that mapped to both the host and SARS-CoV-2 transcriptomes were extracted for analysis as potential chimeric sequences. To avoid including chimeric reads that resulted from technical artifacts such as those caused by misinterpretation of open-pore states by base-calling softwares[10], additional quality filtering was applied to the chimeric reads. The distance between the mapped regions of the virus and the host sequence on the chimeric reads was required to be less than 15 nucleotides, the junction was required to be formed in the middle of the genes (not within the last 50 nucleotides of the first gene sequence, nor the first 50 nucleotides of the second gene sequence), and the quality score within 20 bp of either side of the junction was required to be higher than the 20th percentile quality score for that read.

Analysis of junction positions in relation to viral RNA secondary structure

The RNA secondary structure of the SARS-CoV-2 reference genome was obtained from a previous study[37]. A junction site was considered in the stem if it was flanked on both sides by residues known to be paired. To investigate if junctions tend to happen in non-stem regions, the number of junctions occurring in base-paired positions were calculated and compared with a background distribution for the numbers of junctions located in stems derived from a 1000-time random sampling of the same number of sites along the viral RNA strand.

Analysis of the expression level of host genes observed in chimeric reads

Gene expression profiles for two SARS-CoV-2 infected Caco-2 cell line samples (GSM4477888, GSM4477889), two SARS-CoV-2 infected Calu-3 cell line samples (GSM4477962,

GSM4477963), and three SARS-CoV-2 infected Vero-6 cell line samples (GSM4916368, GSM4916369, GSM4916370) were downloaded from the GEO database. The read counts of each gene were normalized by the total number of reads in each sample and by the gene length (RPKM) to represent the gene expression level. The background gene set was composed of all expressed protein-coding genes in the cell line. To evaluate whether the expression level of the host protein-coding genes in chimeric reads is significantly greater than the expression level of the background gene set, a one-sided Mann-Whitney U test was performed for each sample.

Identification of insertions in SARS-CoV-2 genomes

The SARS-CoV-2 genomes available at GISAID (https://www.gisaid.org/) on 2021-12-17 were downloaded for analysis (n=6,163,073). The sequences were then processed by NextClade CLI v1.7.0[38] which generated a multiple sequence alignment against the reference genome (Wuhan-Hu-1/2019) and provided a list of single nucleotide polymorphisms, insertions, and deletions associated with each genome sequence. Only sequences that passed all quality controls and were assessed as "good" applied by NextClade were used for further analysis (n=5,226,229).

Monophyletic test

To check if the insertions of interest formed a monophyletic group, all genomes that contained the same insertion were analyzed using UShER: Ultrafast Sample placement on Existing tRee v.0.5.1[39] against a phylogenetic tree with available genomes (n=6,257,569) from GISAID, GenBank, COG-UK and CNCB generated by sarscov2phylo pipeline v.13-11-20[40]. The sequences are placed within an updated global subsampled SARS-CoV-2 phylogenetic tree and local subtrees are computed to show more sequences with the same context of the ones being analyzed.

Author Statements

Author contributions: YY was involved in the execution of the analyses, interpretation of the results, and writing and revision of the manuscript. KD was involved in the interpretation of the results and writing and revision of the manuscript. RF was involved in the execution of the analyses and writing of the manuscript. XJ was involved in the conceptualization, planning, interpretation of the results, and revision of the manuscript. All authors read and approved the final manuscript.

Funding information: All authors are supported by the Intramural Research Program of the NIH, National Library of Medicine.

Conflicts of interest: The authors declare that there are no conflicts of interest.

Acknowledgments: This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). We gratefully acknowledge the researchers from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based (Supplementary Table 4).

Data and materials availability: The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

References

1. Gerdol, M., Dishnica, K. & Giorgetti, A. Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein. doi:10.1101/2021.04.17.440288.
2. Laiton-Donato, K. *et al.* Characterization of the emerging B.1.621 variant of interest of SARS-CoV-2. *Infect. Genet. Evol.* **95**, 105038 (2021).
3. Venkatakrishnan, A. J. *et al.* Omicron variant of SARS-CoV-2 harbors a unique insertion mutation of putative viral or human genomic origin. (2021) doi:10.31219/osf.io/f7txy.
4. Andreano, E. *et al.* SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
5. Garushyants, S. K., Rogozin, I. B. & Koonin, E. V. Template switching and duplications in SARS-CoV-2 genomes give rise to insertion variants that merit monitoring. *Commun Biol* **4**, 1343 (2021).
6. thomasppeacock, Kristian_Andersen & profbillg. Putative host origins of RNA insertions in SARS-CoV-2 genomes. https://virological.org/t/putative-host-origins-of-rna-insertions-in-sars-cov-2-genomes/761 (2021).
7. Knoops, K. *et al.* SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. *PLoS Biol.* **6**, e226 (2008).
8. Robson, F. *et al.* Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. *Mol. Cell* **80**, 1136–1138 (2020).
9. Zhang, L. *et al.* Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
10. Parry, R., Gifford, R. J., Lytras, S., Ray, S. C. & Coin, L. J. M. No evidence of SARS-CoV-2 reverse transcription and integration as the origin of chimeric transcripts in patient tissues. *Proceedings of the National Academy of Sciences of the United States of America* vol. 118 (2021).
11. Zhang, L. *et al.* Response to Parry et al.: Strong evidence for genomic integration of SARS-CoV-2 sequences and expression in patient tissues. *Proceedings of the National Academy of Sciences of the United States of America* vol. 118 (2021).
12. Briggs, E. *et al.* Assessment of potential SARS-CoV-2 virus integration into human genome reveals no significant impact on RT-qPCR COVID-19 testing. *Proceedings of the National Academy of Sciences of the United States of America* vol. 118 (2021).
13. Smits, N. *et al.* No evidence of human genome integration of SARS-CoV-2 found by long-read DNA sequencing. *Cell Rep.* **36**, 109530 (2021).
14. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).

15. Wang, D. *et al.* The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Mol. Cell* **81**, 2135–2147.e5 (2021).
16. Wu, H.-Y. & Brian, D. A. 5'-proximal hot spot for an inducible positive-to-negative-strand template switch by coronavirus RNA-dependent RNA polymerase. *J. Virol.* **81**, 3206–3215 (2007).
17. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914–921.e10 (2020).
18. Chrisman, B. S. *et al.* Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Min.* **14**, 20 (2021).
19. Gorski, J. L., Gonzalez, I. L. & Schmickel, R. D. The secondary structure of human 28S rRNA: the structure and evolution of a mosaic rRNA gene. *J. Mol. Evol.* **24**, (1987).
20. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
21. Hagemeijer, M. C. *et al.* Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology* **458-459**, 125–135 (2014).
22. Lei, J., Kusov, Y. & Hilgenfeld, R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res.* **149**, 58–74 (2018).
23. Khare, S. *et al.* GISAID's Role in Pandemic Response. *CCDCW* **3**, 1049–1051 (2021).
24. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
25. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
26. Granneman, S., Petfalski, E., Swiatkowska, A. & Tollervey, D. Cracking pre-40S ribosomal subunit structure by systematic analyses of RNA-protein cross-linking. *EMBO J.* **29**, 2026–2036 (2010).
27. Gopanenko, A. V., Malygin, A. A. & Karpova, G. G. Exploring human 40S ribosomal proteins binding to the 18S rRNA fragment containing major 3'-terminal domain. *Biochim. Biophys. Acta* **1854**, 101–109 (2015).
28. Zhou, Z. *et al.* Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14 monocytes. *iScience* **24**, 102187 (2021).
29. Cao, Z. *et al.* Ubiquitination of SARS-CoV-2 ORF7a promotes antagonism of interferon response. *Cell. Mol. Immunol.* **18**, 746–748 (2021).
30. Su, C.-M., Wang, L. & Yoo, D. Activation of NF-κB and induction of proinflammatory cytokine expressions mediated by ORF7a protein of SARS-CoV-2. *Sci. Rep.* **11**, 13464 (2021).
31. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
32. Gultyaev, A. P., Spronken, M. I., Funk, M., Fouchier, R. A. M. & Richard, M. Insertions of codons encoding basic amino acids in H7 hemagglutinins of influenza A viruses occur by recombination with RNA at hotspots near snoRNA binding sites. *RNA* **27**, 123–132 (2021).

33. Khatchikian, D., Orlich, M. & Rott, R. Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. *Nature* **340**, 156–157 (1989).
34. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
35. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
36. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
37. Huston, N. C. *et al.* Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell* **81**, 584–598.e5 (2021).
38. Aksamentov, I., Roemer, C., Hodcroft, E. B. & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773 (2021).
39. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
40. Lanfear, R. A global phylogeny of SARS-CoV-2 sequences from GISAID. *ZENODO* (2020) doi:10.5281/zenodo.3958883.

Supplementary Information

Supplementary Table 1: Direct RNA-seq data analysis. Metadata associated with all 30 of the analyzed directed RNA-seq samples is provided along with the number of reads mapped to the SARS-CoV-2 transcriptome and the count and frequency of chimeric reads in each sample. The counts of chimeric reads in the host to virus and virus to host orientation are listed for each sample. The references DOI for each sample is also listed.

Supplementary Table 2: Long insertions identified in GISAID that are not derived from SARS-CoV-2. The location on the SARS-CoV-2 reference genome, insertion sequence, insertion length, and what gene they are located in are provided for each of the 36 detected insertions. SARS-CoV-2 genomes with the insertion, whether those genomes were monophyletic, and short descriptions of putative matches are also provided for each insertion.

Supplementary Table 3: Information on genomes related to the three verified insertions. Metadata associated with each of the SARS-CoV-2 genomes with putative insertions that were analyzed including the variant types, collection dates, collecting labs, and sequencing methods are provided.

Supplementary Table 4: GISAID acknowledgement table.

Figures

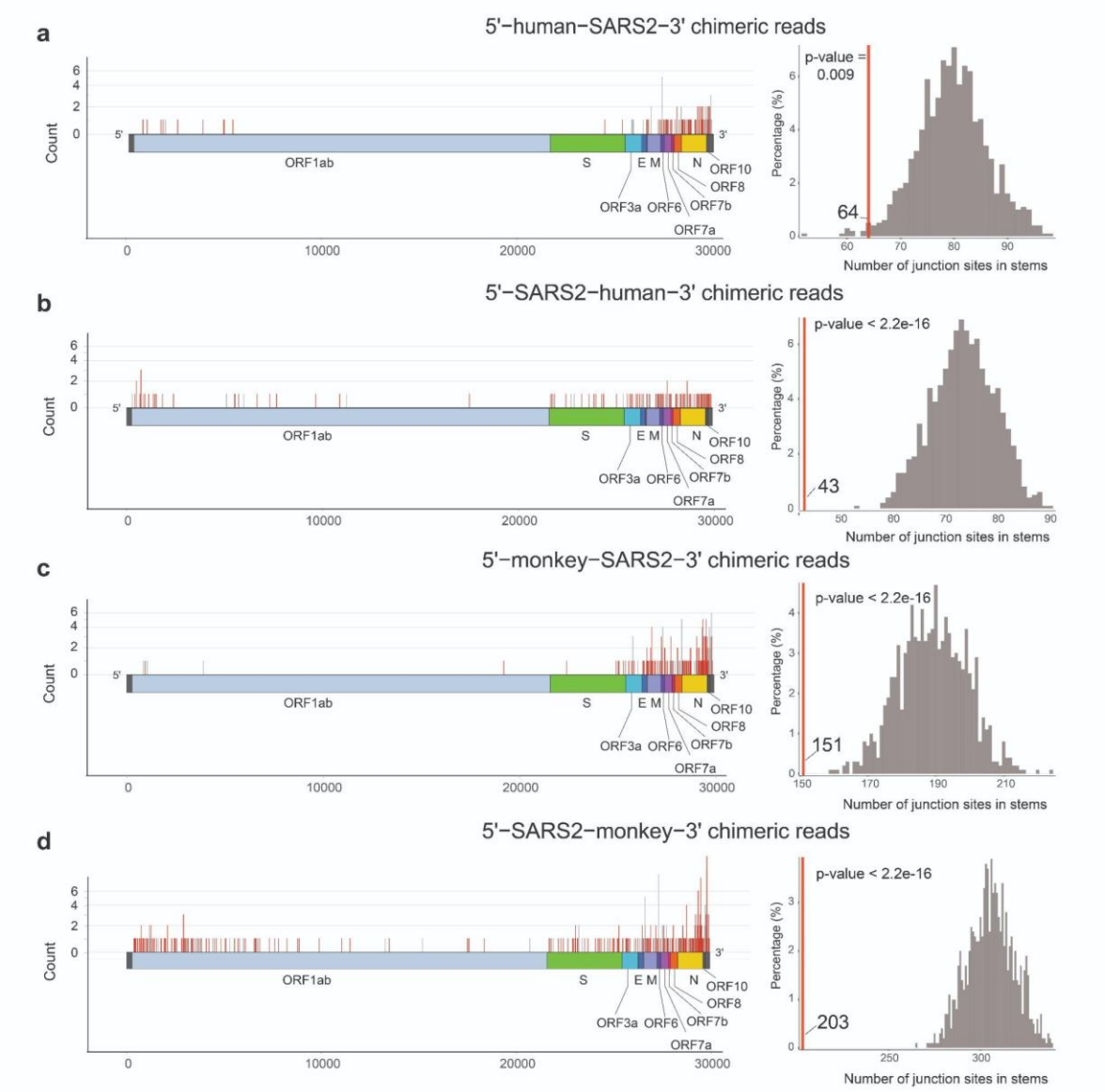

Fig.1: Locations of the chimeric read junction sites and permutation tests for the number of junction sites in stems. Diagrams show how frequently junction sites occur at each position on the SARS-CoV-2 genome for **(a)** 5'-human-SARS2-3', **(b)** 5'-SARS2-human-3', **(c)** 5'-monkey-SARS2-3', and **(d)** 5'-SARS2-monkey-3' chimeric reads. Positions are colored based on the secondary structure of the SARS-CoV-2 RNA, with red lines indicating that the position is in the non-stem region, while gray indicates that the position is located in the stem region. Histograms following each diagram show the corresponding results of permutation tests used to test if the junction sites of chimeric reads are within base-paired regions of the viral RNA. Each test consists of 1000 permutations and the actual frequency of junction sites occurring in the stem regions is marked with a vertical red line.

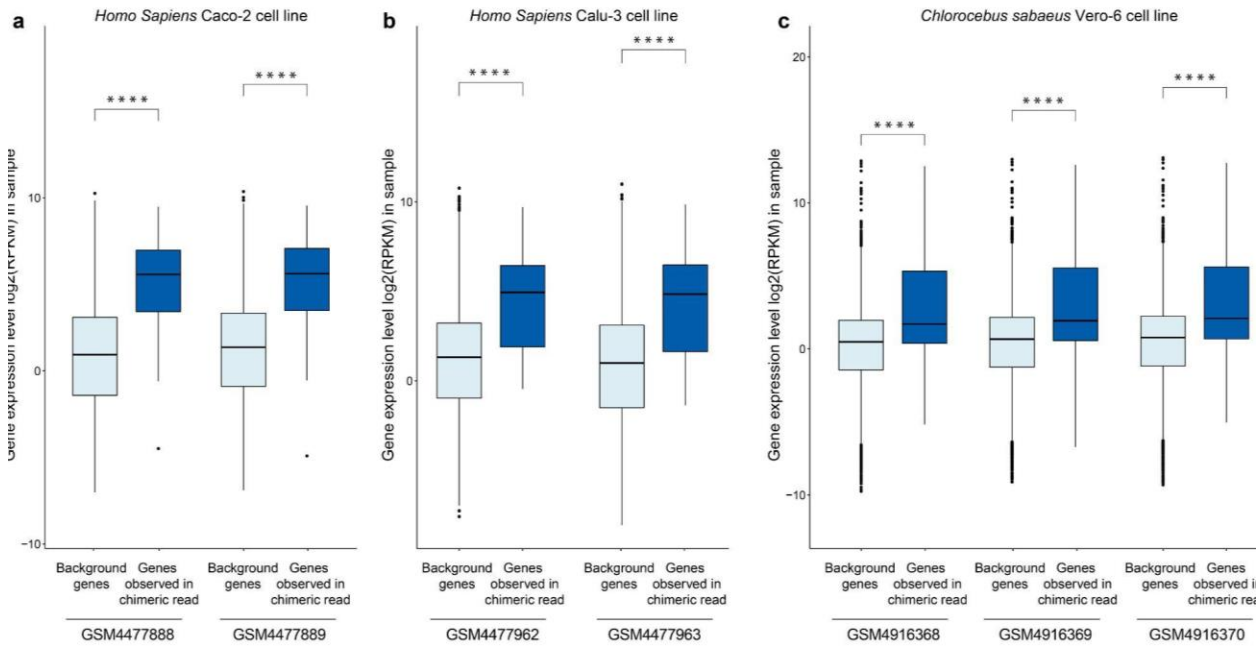

Fig.2: The expression level of host genes observed in chimeric reads. The expression level of host protein-coding genes observed in chimeric reads is significantly higher than the background protein-coding gene expression level based on studies on **(a)** *Homo sapiens* Caco-2 cell line, **(b)** *Homo sapiens* Calu-3 cell line, and **(c)** *Chlorocebus sabaeus* Vero-6 cell line.

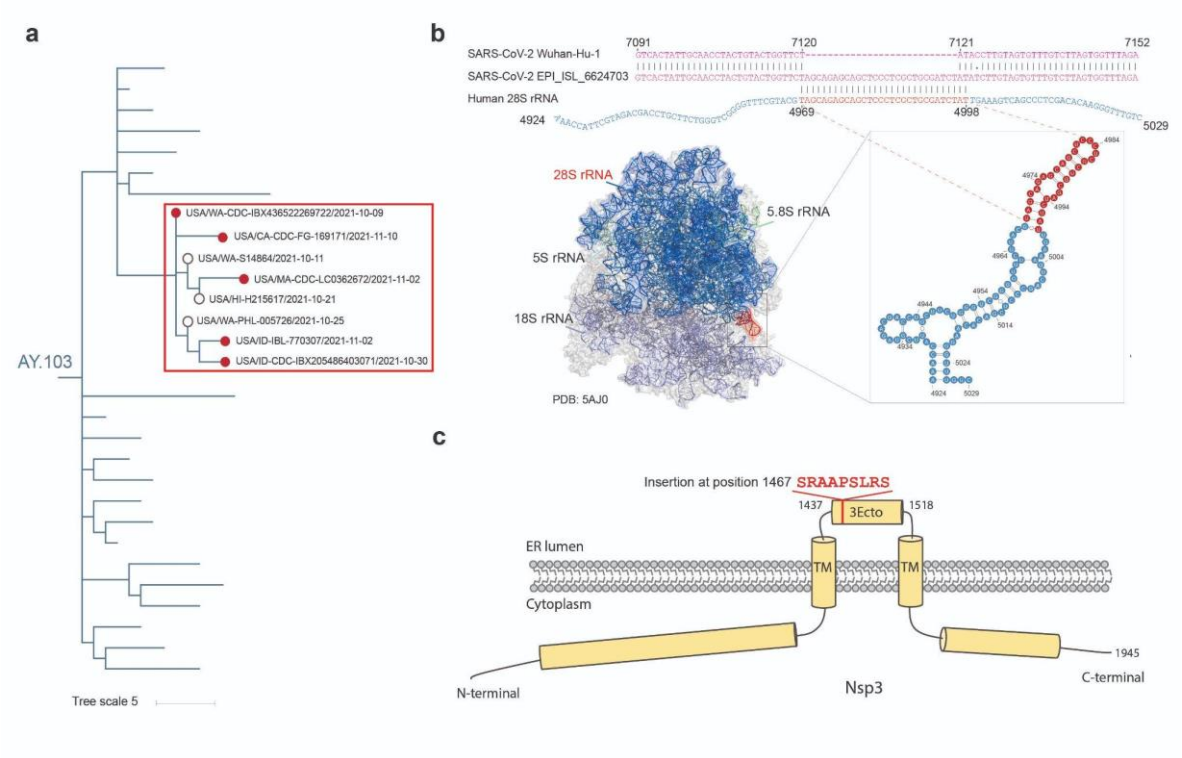

Fig.3: The 28S rRNA-derived insertion in SARS-CoV-2 genomes. (a) The phylogeny tree shows the genomes containing the human 28S-derived insertion. The monophyletic clade where the insertion was detected is highlighted with a red box and the genomes containing the insertion are marked with red circles at the tips. **(b)** The insertion in SARS-CoV-2 genomes potentially originate from the host 28S rRNA shown by the sequence alignment of SARS-CoV-2 reference genome (NCBI accession: NC_045512.2, GISAID accession: China/Wuhan-Hu-1/2019) (pink), USA/CA-CDC-FG-169171/2021 (NCBI accession: OL591909.1, GISAID accession: EPI_ISL_6624703) (pink) and human 28S rRNA (chain A2 of PDB 5AJ0) (blue). The putative insertion origin is colored in red. The numbers listed above and below the alignment indicate the positions of aligned bases in the original sequences. The insertion sequence (red) was mapped to the 28s rRNA (blue) in a human polysome 3D structure (PDB: 5AJ0). A zoom-in view of the RNA secondary structure shows that the insertion is located on the No. 94 stem of domain 7 (position: 4969-4998) 28S rRNA region (highlighted red). **(c)** Diagram shows the position of the human 28S rRNA-derived insertion in the ectodomain (3Ecto) of Nsp3 protein.

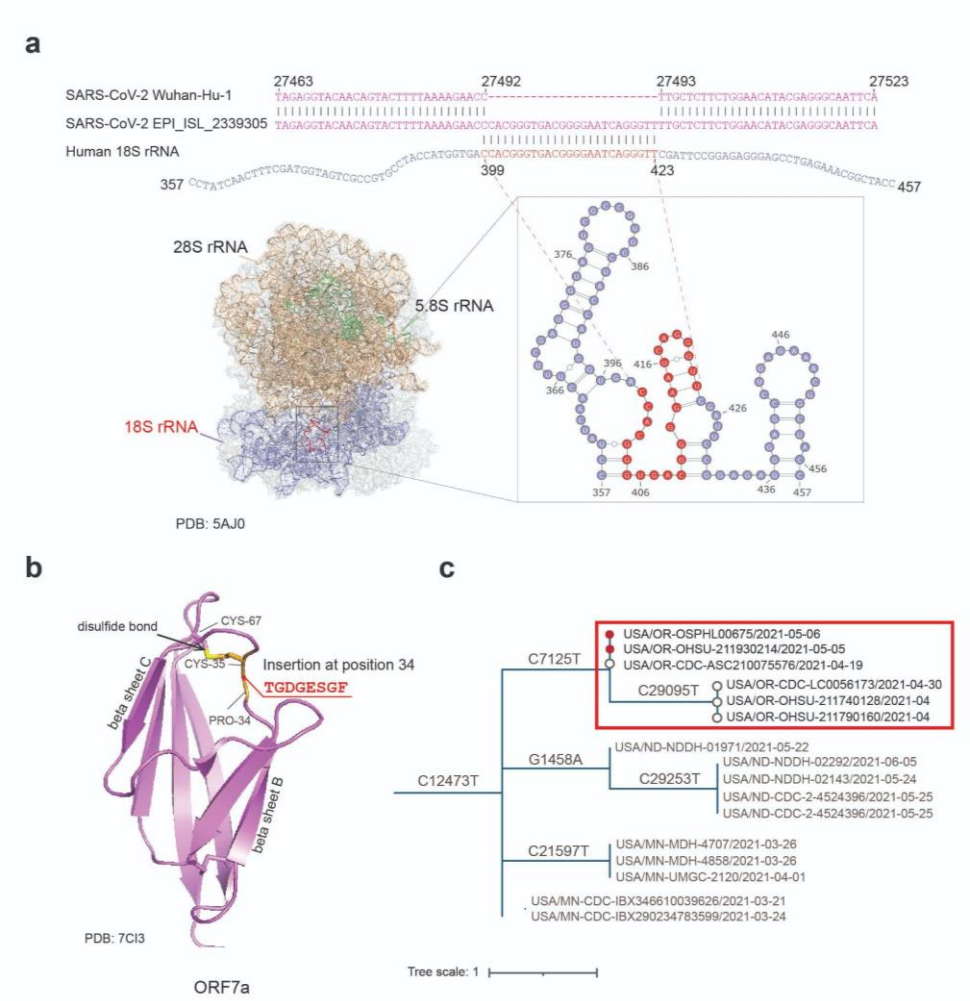

Fig.4: The 18S rRNA-derived insertion in SARS-CoV-2 genomes. (a) The insertion in SARS-CoV-2 genomes potentially originates from the host 18S rRNA shown by the sequence alignment of SARS-CoV-2 reference genome (NCBI accession: NC_045512.2, GISAID accession: China/Wuhan-Hu-1/2019) (pink), USA/OR-OSPHL00675/2021 (GISAID accession: EPI_ISL_2339305) (pink) and human 18S rRNA (purple). The putative insertion origin is colored in red. The numbers listed above and below the alignment indicate the positions of aligned bases in the original sequences. The insertion sequence (red) was mapped to the 18S rRNA (purple) in a human polysome 3D structure (PDB: 5AJ0). A zoom-in view of the RNA secondary structure shows that the insertion covers parts of helices 11 and 12 of the 5' domain of the 18S rRNA. The location of the putative insertion sequence is highlighted red. **(b)** Diagram shows the position of the human 18S-derived insertion on the structure of the SARS-CoV-2 ORF7a protein (PDB: 7CI3). **(c)** The phylogeny tree shows the genomes containing the human 18S-derived insertion. The monophyletic clade where the insertion was detected is highlighted with a red box and the genomes with the insertion are marked with red circles.