Virology Journal

RESEARCH Open Access

Intragenomic rearrangements involving 5′-untranslated region segments in SARS-CoV-2, other betacoronaviruses, and alphacoronaviruses

Roberto Patarca and William A. Haseltine*

Abstract

Background Variation of the betacoronavirus SARS-CoV-2 has been the bane of COVID-19 control. Documented variation includes point mutations, deletions, insertions, and recombination among closely or distantly related coronaviruses. Here, we describe yet another aspect of genome variation by beta- and alphacoronaviruses that was first documented in an infectious isolate of the betacoronavirus SARS-CoV-2, obtained from 3 patients in Hong Kong that had a 5′-untranslated region segment at the end of the *ORF6* gene that in its new location translated into an ORF6 protein with a predicted modified carboxyl terminus. While comparing the amino acid sequences of translated ORF8 genes in the GenBank database, we found a subsegment of the same 5′-UTR-derived amino acid sequence modifying the distal end of ORF8 of an isolate from the United States and decided to carry out a systematic search.

Methods Using the nucleotide and in the case of SARS-CoV-2 also the translated amino acid sequence in three reading frames of the genomic termini of coronaviruses as query sequences, we searched for 5′-UTR sequences in regions other than the 5′-UTR in SARS-CoV-2 and reference strains of alpha-, beta-, gamma-, and delta-coronaviruses.

Results We here report numerous genomic insertions of 5′-untranslated region sequences into coding regions of SARS-CoV-2, other betacoronaviruses, and alphacoronaviruses, but not delta- or gammacoronaviruses. To our knowledge this is the first systematic description of such insertions. In many cases, these insertions would change viral protein sequences and further foster genomic flexibility and viral adaptability through insertion of transcription regulatory sequences in novel positions within the genome. Among human *Embecorivus* betacoronaviruses, for instance, from 65% to all of the surveyed sequences in publicly available databases contain inserted 5′-UTR sequences.

Conclusion The intragenomic rearrangements involving 5′-untranslated region sequences described here, which in several cases affect highly conserved genes with a low propensity for recombination, may underlie the generation of variants homotypic with those of concern or interest and with potentially differing pathogenic profiles. Intragenomic rearrangements thus add to our appreciation of how variants of SARS-CoV-2 and other beta- and alphacoronaviruses may arise.

Keywords Alphacoronaviruses, Betacoronaviruses, Variants, Intragenomic rearrangements

*Correspondence:
William A. Haseltine
william.haseltine@accessh.org
ACCESS Health International, 384 West Lane, Ridgefield, CT 06877, USA

Background

Coronaviruses (CoVs) are positive, singe stranded RNA viruses of the order *Nidovirales*, family *Coronaviridae*, subfamily *Orthocoronavirinae*, with four genera, namely alpha [α], beta [β], gamma [γ] and delta [δ], which have been further subdivided into 25 subgenera, including five for β-CoVs: *Sarbecovirus*, *Merbecovirus*, *Embecovirus*, *Nobecovirus* and *Hibecovirus* [1], and fifteen for α-CoVs: *Luchacovirus*, *Decacovirus*, *Nyctacovirus*, *Minunacovirus*, *Pedacovirus*, *Colacovirus*, *Myotacovirus*, *Duvinacovirus*, *Setracovirus*, *Rhinacovirus*, *Tegacovirus*, *Minacovirus*, *Sunacovirus*, *Soracovirus*, and *Amalacovirus* [2]. Seven CoVs infect humans; two of the α-genus (the *Duvinacovirus* hCoVs 229E and the *Setracovirus* NL63) and five of the β-genus: the *Sarbecoviruses* severe acute respiratory syndrome (SARS)-CoVs 1 and 2, the latter responsible for a pandemic since 2019 [3–6]; the *Merbecovirus* Middle East respiratory syndrome (MERS) CoV; and the *Embecoviruses* hCoV-OC43 and -HKU1. Human CoVs have a zoonotic origin, with bats as key reservoir [7] and possibly other hosts [8, 9]. Bat β-CoVs related to human CoVs belong to the *Sarbecovirus*, *Nobecovirus*, and *Hibecovirus* subgenera [10–12].

Coronaviruses display substantial genomic plasticity and resilience [13, 14] via recombination, point mutations, deletions, and insertions, which are reported to drive variant emergence, host range, gene expression, transmissibility, immune escape, and virulence [15–20]. The use of an RNA-dependent-RNA polymerase (RdRp)-driven template switching mechanism for transcription and control of structural and accessory gene expression in CoVs [20] has been reported to account for the high frequency of recombination [13, 18, 21–27].

In template switching, a leader transcription regulatory sequence (TRS-L; ACGAAC core in β-CoVs) [28] in the 5′-untranslated region (UTR) interacts with homologous TRS-body (B) elements upstream of viral genes in the last third of the genome (illustrated for SARS-CoV-2 in Additional file 1: Fig. S1A) [29, 30]. Template switching renders the neighborhood of TRS-Bs, especially that for the spike gene, a recombination hotspot during viral transcription [3, 16, 21, 22, 24, 27, 31–34].

Viral subgenomic messenger RNAs contain a 5′-leader sequence that spans from the terminal 5′-cap (m^7G) structure to the TRS-L and harbors three conserved stem-loop (SL1-3) regulatory elements of gene expression and replication (Additional file 1: Fig. S1B) [35–37]. The TRS-L core sequence and the secondary structure of the leader sequence are conserved within but not among coronavirus genera (Rfam database: http://rfam.xfam.org/covid-19).

The entire 5′-leader nucleotide sequence of SARS-CoV-2, and beyond up to almost SL5 can be translated into a peptide sequence (Additional file 1: Fig. S1B), and although there is no evidence for the functionality of any open reading frame within the UTRs [36, 38], the 5′-leader sequence could be translated after most of it (nucleotides 8–80, including SL1-3 and TRS-L) is duplicated and translocated to the distal end of the accessory ORF6 gene of a SARS-CoV-2 variant with deleted ORFs 7a, 7b and 8 isolated from 3 patients in Hong Kong [39]. We also found that a shorter portion of the 5′-leader sequence (nucleotides 50–75) is duplicated and translocated to the end of the accessory ORF8 gene of a USA variant (accession number: QUP34336) that could be translated into a modified ORF-8 protein, which prompted us to conduct a systematic analysis.

In the present study, using 5′-leader nucleotide sequences and amino acid sequences translated in the three reading frames as queries to search public databases, we document the presence of intragenomic rearrangements involving segments of the 5′-leader sequence in geographically and temporally diverse isolates of SARS-CoV-2. The intragenomic rearrangements could modify the carboxyl-termini of the ORF8 (also in *Rhinolophus* bat *Sarbecovirus* β-CoVs) and ORF7b proteins; the serine-arginine-rich region of the nucleocapsid protein, generating the well characterized R203K/G204R paired mutation; and two sites of the NiRAN domain of the RdRp (nsp12).

Beyond SARS-CoV-2, we found similar rearrangements of 5′-UTR leader sequence segments including the TRS-L in all subgenera of β-CoVs except for *Hibecovirus* (possibly secondary to the availability of only 3 sequences in GenBank). These rearrangements are in the intergenic region between ORFs 3 and 4a, and at the distal end of ORF4b of the *Merbecovirus* MERS-CoV; intergenic regions in the *Embecoviruses* hCoV-OC43 (between S and Ns5) and hCoV-HKU-1 (between S and NS4); and in the distal end that encodes the Y1 cytoplasmic tail domain of nsp3 of *Nobecoviruses* of African *Rousettus* and *Eidolon* bats. We also found intragenomic rearrangements in α-CoVs in nsp2 (*Luchacovirus* subgenus), nucleocapsid (*Nyctacovirus* subgenus), and ORF5b or ORF4b (*Decacovirus* subgenus). No rearrangements involving 5′-UTR sequences were detected for the β-CoV SARS-CoV-1; the other 12 subgenera of α-CoVs including hCoV-229E and hCoV-NL63 infecting humans; or δ (*Andecovirus*, *Buldecovirus*, and *Herdecovirus* subgenera) and γ CoVs (*Brangacovirus*, *Cegacovirus*, and *Igacovirus* subgenera) for which wild birds are the main reservoir [12, 40].

The present study highlights an intragenomic source of variation involving duplication, inversion (in two α-CoVs subgenera) and translocation of 5′-UTR sequences to the body of the genome with potential implications on gene expression and immune escape of α- and β-CoVs in humans and bats causing mild-to moderate or severe disease in endemic, epidemic, and pandemic settings. Genome-wide annotations had revealed 1516 nucleotide-level variations at various positions throughout the entire SARS-CoV-2 genome [41] and a recent study documented outspread variations of each of the six accessory proteins across six continents of all complete SARS-CoV-2 proteomes which was suggested to reflect effects on SARS-CoV-2 pathogenicity [42]. However, the function and even expression of some of these accessory proteins remains a matter of debate due to inconsistencies derived from the use of bioinformatics predictions, and studies in different cell types and not in in vivo infection settings. The intragenomic rearrangements involving 5′-UTR sequences described here, which in several cases affect highly conserved genes with a low propensity for recombination, may underlie the generation of variants homotypic with those of concern or interest and with potentially differing pathogenic profiles.

Methods

Detection of 5′-UTR sequences in SARS-CoV-2 and SARS-CoV-related viruses in GenBank

To assess the presence of 5′-UTR sequence insertions in the body of the genome, we used 5- to 10-amino acid stretches from the three reading frames of the translated 5′-UTR nucleotide sequence of SARS-CoV-2 (Wuhan reference, NC_045512) as query sequences to search the GenBank® database using the Basic Alignment Search Tool (BLAST)P® (Protein BLAST: search protein databases using a protein query (nih.gov); [43]) for SARS-CoV-2 and SARS-CoV-related viral proteins encoding similar stretches. All nonredundant translated CDS + PDB + SwissProt + PRF excluding environmental samples from WGS projects were searched specifying severe acute respiratory syndrome coronavirus 2 as organism.

Using the accession number listed in PubMed (SARS-CoV-2 Resources—NCBI (nih.gov)) for the viral protein sequence, we obtained the respective nucleotide sequence and translated it using the insilico (DNA to protein translation (ehu.es) [44] and Expasy (ExPASy—Translate tool [45]) tools to determine by manual inspection and the BLASTN program [46] if the nucleotide sequences encoding said stretches were identical to those in the 5′-UTR nucleotide sequence of SARS-CoV-2 or SARS-CoV-related viruses.

Using nucleotide sequences instead of translated amino acid sequences from the 5′-UTR in the three reading frames as query sequences was unproductive to detect insertions in SARS-CoV-2 because of the large number of SARS-CoV-2 sequences in the GenBank database and the limit of 5000 results in the BLAST algorithm settings which yielded solely 5′-UTR sequences.

Detection and validation of 5′-UTR sequences in regions other than the 5′-UTR of SARS-CoV-2 and SARS-CoV-related viruses in other databases

To detect isolates with similar insertions whose sequences had not been included in GenBank, we then searched the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu™ database of SARS-CoV-2 sequences (GISAID—Initiative; [47–49]) using as queries the nucleotide sequences of the insertions plus adjoining 20 nucleotides on either side from the viral genomes. This approach is limited by the fact that maximum number of search results in GISAID is 30. Information on location and timing of isolate collection was obtained from the GenBank and GISAID databases.

Detection of 5′-UTR sequences in regions other than the 5′-UTR in coronaviruses other than SARS-CoV-2 and SARS-CoV-related viruses

We used the Rfam database (http://rfam.xfam.org/covid-19) with the curated Stockholm files containing UTR sequences, alignments and consensus RNA secondary structures of major genera of *Coronoviridae*; the representative RefSeq sequences for each genus obtained from the International Committee on Taxonomy of Viruses (ICTV) taxonomy Coronaviridae Study Group [2]); the reference sequences in the GenBank database; and listings in publications involving phylogenetic analyses of alpha-, delta-, and gamma-coronaviruses from NCBI Taxonomy [34, 50] to derive the 5′-UTRs of various CoVs.

We then utilized the 5′-UTR segments as query sequences to search for insertions in their respective genomes (nucleotide collection [nr/nt]; expect threshold: 0.05; mismatch scores: 2, − 3; gap costs: linear). The GSAID database does not include sequences of CoVs other than SARS-CoV-2 and therefore could not be used for this analysis.

If the intragenomic rearrangement detected using the 5′-UTR sequences involved a coding region, we translated the 5′-UTR insertion and adjacent segments using the insilico (DNA to protein translation (ehu.es) [44] and Expasy tools [45].

Localization and sorting of intragenomic rearrangements

In terms of the locations of the insertions in the body of the genomes, the boundaries of nonstructural, structural, and accessory open reading frames were determined based on GenBank annotation and from manual inspection of multiple alignments and sequence similarities.

Sorting and collection of further information on viral isolates with intragenomic rearrangements

In the results presented, we excluded matches to entries corresponding to the 5′-leader sequences in mRNAs from full viruses or defective interfering RNA particles, as well as protein sequences with > 80% unknown amino acids (represented by the letter X) in GenBank. The Supplementary section includes the accession numbers and collection site and date, and in some cases the SARS-CoV-2 lineages, for the isolates with intragenomic rearrangements involving 5′-UTR sequences.

Detection of possible intragenomic rearrangements involving 3′-UTR sequences

We also searched for intragenomic rearrangements involving 3′-UTR sequences using the same approaches and datasets described for the 5′-UTR ones.

Visualization of RNA secondary structures in segments with intragenomic rearrangements

RNA secondary structures of the 5′-UTR sequence insertion and adjacent sequences of the intragenomic rearrangement were visualized using forna, a force directed graph layout (ViennaRNA Web services; [51]), and the optimal secondary structures and their minimal free energies were determined using the RNAfold webserver [52, 53].

Results

Using the approaches described in the Methods section, we conducted a systematic analysis of SARS-CoV-2 and other coronaviruses and detected insertions involving 5′-UTR sequences at various locations in β- and α-CoVs, as described below by subgenus.

Intragenomic rearrangements at the distal end of ORF8 and ORF7b (Sarbecoviruses)

We found a U.S. isolate of SARS-CoV-2 in which a segment encompassing nucleotides 50–75 of the 5′-UTR was duplicated and translocated to the end of the accessory *ORF8* gene giving rise to a predicted ORF8 protein with modified carboxyl-terminus encoded by the translocated 5′-UTR sequences. Figure 1 summarizes the results of our systematic search which revealed 240 similar insertions of various lengths of the same 5′-UTR sequence at various points in a stretch of 7 amino acids ($_{115}$RVVLDFI$_{121}$) of the carboxyl-terminal sequence of the predicted ORF8 protein. As depicted in Additional file 1: legend to Fig. S1, these internal rearrangements were detected in temporally and geographically diverse isolates, collected from March 2020 to December 2021 in 38 USA states, Bahrain, China, Kenya, and Pakistan, which is not exhaustive of what exists. All translocated 5′-UTR nucleotide sequence segments include TRS-L with variable extents of SL3 and SL2, that could affect expression of the nucleocapsid gene located immediately after the *ORF8* gene [19], and all insertions alter the carboxyl-termini of predicted ORF8 proteins. The analysis also revealed that the insertions in some isolates had further changes involving point mutations, deletions, and insertions. Moreover, as shown in Fig. 2A, a similar 5′-UTR sequence insertion at the distal end of *ORF8* is seen in five *Sarbecovirus* β-CoVs from what is considered the animal reservoir for SARS-CoV-2, the *Rhinolophus* (horseshoe) bats residing in Indochina and Southwest China [54] all the way to England [55].

Figure 1C depicts the predicted secondary structure of ORF8 RNA without and with (ORF8x) the 5′-UTR sequence insertion. Both structures have similar predicted minimum free energy. The insertion involves the TRS-B sequence located in the intergenic region between *ORF8* and *N* genes and is preceded by a uridine-adenosine (U/A)-rich region including a sequence similar to the torovirus attenuation sequence [56], which like TRS-B, might cause the RNA-dependent RNA polymerase to pause, thereby facilitating the intragenomic rearrangement as it is theorized to do during subgenomic RNA synthesis. Additional file 1: Fig. S2A shows the predicted RNA structures of the most common ORF8x variants with similar predicted minimum free energy, while Additional file 1: Fig. S2B shows an alternative RNA structure involving the interaction between the TRS-B in the

(See figure on next page.)

Fig. 1 Modified carboxyl-termini of ORF8 protein predicted to be encoded by 5′-UTR sequence insertions in SARS-CoV-2. **A**. The largest 5′-UTR segment that was duplicated and translocated as an insertion to the carboxyl terminus of ORF8 is shown at the nucleotide and amino acid levels (latter underlined). All translocated 5′-UTR nucleotide sequence segments include TRS-L (dark blue box) with variable extents of SL3 (blue) and SL2 (red). Examples are shown, and corresponding similar sequences in GenBank as of February 20, 2022, are listed in the Additional file 1: legend to Fig. S1. The C-terminus of ORF8 in the Wuhan reference strain is depicted using orange letters with mutations in ochre; the asterisks over the C-terminus sequence designate residues contributing to the covalent dimer interface (Arg115, Asp119, Phe120, Iso121; [80]). The 5′-UTR insertions are shown as underlined letters in black with mutations, deletions, and insertions within them highlighted in green. **B**. Secondary structures of ORF8 RNA in reference strain and in that with longest 5′-UTR sequence intragenomic rearrangement. Nucleotide and amino acid sequences of the carboxyl termini of ORF8 from Wuhan reference (NC_0445512) and from isolate QUP34336 (USA/Minnesota, 2021-04-05) with the longest 5′-UTR sequence. ORF8 protein from the latter has modified carboxyl terminus and is therefore designated ORF8x. Amino acid sequence from the reference strain is highlighted in yellow while that encoded by the duplicated and translocated 5′-UTR segment is highlighted in green. Stop codon of ORF8 protein is depicted with a red asterisk, and initiation codon of N is in green letters. TRS-B core sequence and complementary TRS-Bs in *ORF8* and in *ORF8x* are highlighted in blue; and the uridine/adenosine tracks, including the torovirus-like attenuation sequence [56] are highlighted in fuchsia. 5′-UTR nucleotide and predicted translated amino acid sequences in ORF8X are in bold letters and highlighted in green

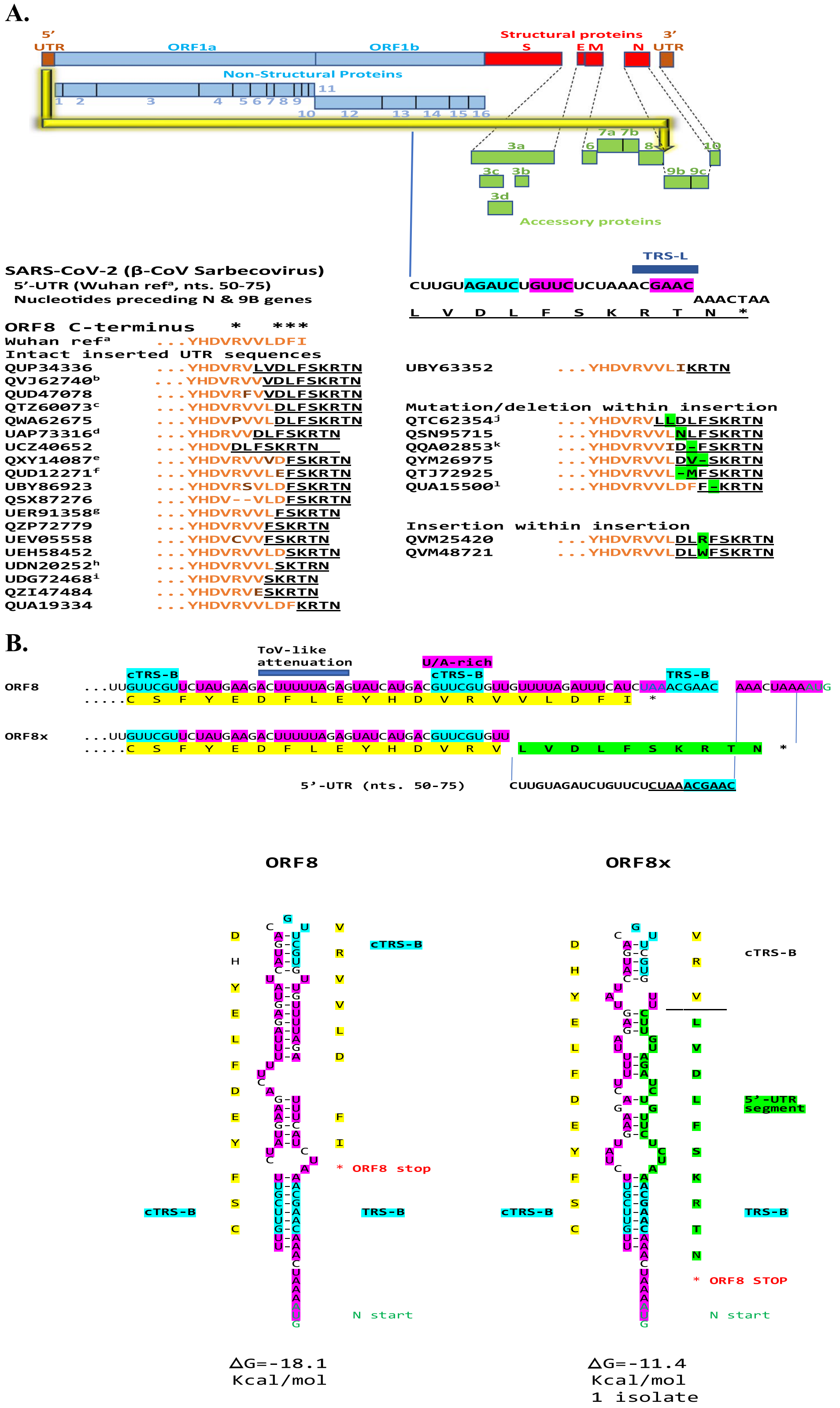

A.
5' UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3' UTR
Non-Structural Proteins
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
3a 3c 3b 3d 6 7a 7b 8 9b 9c 10
Accessory proteins
SARS-CoV-2 (β-CoV Sarbecovirus)
5'-UTR (Wuhan ref, nts. 50-75)
Nucleotides preceding N & 9B genes
TRS-L
CUUGUAGAUCUGUUCUCUAAACGAAC
AAACTAA
L V D L F S K R T N *
ORF8 C-terminus * ***
Wuhan ref ...YHDVRVVLDFI
Intact inserted UTR sequences
QUP34336 ...YHDVRVLVDLFSKRTN
QVJ62740 ...YHDVRVVVDLFSKRTN
QUD47078 ...YHDVRFVVDLFSKRTN
QTZ60073 ...YHDVRVVLDLFSKRTN
QWA62675 ...YHDVPVVLDLFSKRTN
UAP73316 ...YHDRVVDLFSKRTN
UCZ40652 ...YHDVDLFSKRTN
QXY14087 ...YHDVRVVVDFSKRTN
QUD12271 ...YHDVRVVLEFSKRTN
UBY86923 ...YHDVRSVLDFSKRTN
QSX87276 ...YHDV--VLDFSKRTN
UER91358 ...YHDVRVVLFSKRTN
QZP72779 ...YHDVRVVFSKRTN
UEV05558 ...YHDVCVVFSKRTN
UEH58452 ...YHDVRVVLDSKRTN
UDN20252 ...YHDVRVVLSKTRN
UDG72468 ...YHDVRVVSKRTN
QZI47484 ...YHDVRVESKRTN
QUA19334 ...YHDVRVVLDFKRTN
UBY63352 ...YHDVRVVLIKRTN
Mutation/deletion within insertion
QTC62354 ...YHDVRVLLDLFSKRTN
QSN95715 ...YHDVRVVLNLFSKRTN
QQA02853 ...YHDVRVVID-FSKRTN
QYM26975 ...YHDVRVVLDV-SKRTN
QTJ72925 ...YHDVRVVL-MFSKRTN
QUA15500 ...YHDVRVVLDFF-KRTN
Insertion within insertion
QVM25420 ...YHDVRVVLDLRFSKRTN
QVM48721 ...YHDVRVVLDLWFSKRTN
B.
ToV-like attenuation
U/A-rich
cTRS-B
TRS-B
ORF8 ...UUGUUCGUUCUAUGAAGACUUUUUAGAGUAUCAUGACGUUCGUGUUGUUUUAGAUUUCAUCUAAACGAAC AAACUAAAUG
.....C S F Y E D F L E Y H D V R V V L D F I *
ORF8x ...UUGUUCGUUCUAUGAAGACUUUUUAGAGUAUCAUGACGUUCGUGUU
.....C S F Y E D F L E Y H D V R V L V D L F S K R T N *
5'-UTR (nts. 50-75) CUUGUAGAUCUGUUCUCUAAACGAAC
ORF8
ORF8x
cTRS-B
TRS-B
ORF8 stop
N start
5'-UTR segment
ORF8 STOP
ΔG=-18.1 Kcal/mol
ΔG=-11.4 Kcal/mol 1 isolate

Fig. 1 (See legend on previous page.)

intergenic region with a second and closer complementary TRS-B, yielding a similar predicted minimum free energy.

A shorter segment of the SARS-CoV-2 5′-UTR leader sequence (nts. 57–95, including TRS-L and SL3) than that described for *ORF8* insertions was also duplicated and translocated to the end of *ORF7*b in two SARS-CoV-2 isolates (Fig. 2B), one with a truncated ORF7b and the other with a truncated *ORF8*, which may have favored the internal rearrangements. Figure 2C shows the predicted secondary structures of the region with the intragenomic rearrangement, and as that of ORF8, involves the intergenic TRS-B sequence which is preceded and followed by a U/A-rich region, in this case also incorporating an HIV-like attenuation sequence (AAA UUU; [57]. Figure 2C also shows a region of similarity between *ORF8* and *ORF7b* which precedes the intragenomic rearrangement.

Intragenomic rearrangements at the end of the segment encoding the serine-arginine-rich region of the N protein (SARS-CoV-2)

In terms of structural proteins of SARS-CoV-2, we found a similar segment of the 5′-UTR corresponding to the leader sequence (nucleotides 56–76 of the Wuhan reference strain [NC_045512], including TRS-L, SL3 and part of SL2, and encoding the 7-amino acid sequence DLFSKRT) within the *N* gene at the end of the nucleotide segment encoding the serine-arginine region, as exemplified by isolate QTO33828 (USA/Texas, Fig. 3A). The 5′-UTR segment changes 5 of 7 positions, including R203K/G204R, which are known to be frequent co-occurring mutations in the N protein; however, the rest of the N protein sequences are well conserved with only 1 or 2 amino acid differences in the isolates identified. In another set of SARS-CoV-2 isolates, as exemplified by isolate EPI-ISL_3434731 (Brazil/Espirito Santo) in Fig. 3A, the same 5′-UTR sequence is present in *N* but without the predicted translated leucine (L) residue and the phenylalanine (F) changed to serine (S), more closely approaching the Wuhan reference strain sequence.

The predicted RNA structures of *N* without and with (*Nx*) the 5′-UTR sequence insertions are shown in Fig. 3B, with that for Nx being less stable with almost half the minimum free energy. Although there is no TRS-B in the N region where the intragenomic rearrangement was found, there is an inverted TRS-B that can pair with a complementary inverted TRS-B, both surrounded by U/A-rich regions which could facilitate the intragenomic rearrangement.

In total, 37 SARS-CoV-2 isolates had 5′-UTR sequences in their *N* gene, in contrast to ~ 336,000 isolates with either R203K or G204K as per NCBI Virus (mutations in SARS-CoV-2 SRA data); most were isolates of the variant of concern gamma GR/501Yv3 (P1) lineage (first detected in Brazil and Japan) from Brazil, Chile, and Peru, but also alpha (B.1.1.7; first detected in Great Britain) from USA and Canada (Additional file 1: legend to Fig. S3). The R203K/G204R co-mutation has been associated with B.1.1.7 (alpha) lineage emergence, which along with variants with the co-mutation including the P1 (gamma) lineage [58], possess a replication advantage over the preceding lineages and show increased nucleocapsid phosphorylation, infectivity, replication, virulence, fitness, and pathogenesis as documented in a hamster model, human cells, and COVID-19 patients including an analysis of association between COVID-19 severity and sample frequency of R203K/G204R co-mutations [59–61]. The intragenomic rearrangement in N might be one rare way for SARS-CoV-2 to acquire the R203/G204K co-mutation.

Intragenomic rearrangements in the region encoding the Nidovirus RNA-dependent RNA polymerase associated nucleotidyl transferase (NiRAN) domain (SARS-CoV-2)

Another example of intragenomic rearrangement is the presence of the translated sequence (DLFSK) of a shorter segment of 5′-UTR sequence (nts. 56–70 in Wuhan reference strain, including parts of SL2 and SL3 but not TRS-L) at amino acids 36–40 of the NiRAN domain of the viral RdRp (nsp12) in isolates QVL75820 (EPI_ISL_1209225, USA/Seattle, 2021-03-28; lineage: B.1.2 [Pango v.3.1.20 2022-02-02]) and EPI_ISL_1524008 (USA/Washington, 2021-03-28; VOC Alpha GRY (B.1.1.7 + Q.*) first detected in the UK) and at amino acids 146–150 in isolates UFT72204 (EPI_ISL_6912949, USA/Colorado, 2021-10-27; VOC Delta GK [B.1.617.2 + AY.*] first detected in India), EPI_ISL_1384819 (India/Maharashtra, 2021-02-12; lineage:

(See figure on next page.)

Fig. 2 **A**. Modified carboxyl-termini of ORF8 predicted to be encoded by an insertion of a 5′-UTR segment in SARS-related β-coronaviruses of *Rhinolophus* bats from China. For SARS-related bat β-CoVs (BatSARSCoV Rf1/2004 and Bat CoV 273/2005 are subgroup 2b; [7]), all inserted terminal sequences were the same. The nucleotide sequence of the inserted 5′-UTR segment differed from that of SARS-CoV-2 by two nucleotides: a C to U change (underlined) which translates into an amino acid change (serine [S] to phenylalanine [F]), and a U to A (underlined) which introduces a stop codon. **B**. Modified carboxyl termini of ORF7b protein predicted to be encoded by an insertion of a 5′-UTR segment in SARS-CoV-2. The two isolates with predicted modified ORF-7 proteins are QXH28554 (USA/Alabama, 2021/04/14), and QSV08409 (USA/California; 2021/02/26); the latter has a truncated *ORF7b* and the former a truncated *ORF8*. Color codes and abbreviations are as in Fig. 1. **C**. Secondary structures of *ORF7b* and *ORF7bx* RNAs. Color scheme is as in Fig. 1B. An HIV-like attenuation sequence [57] is also highlighted

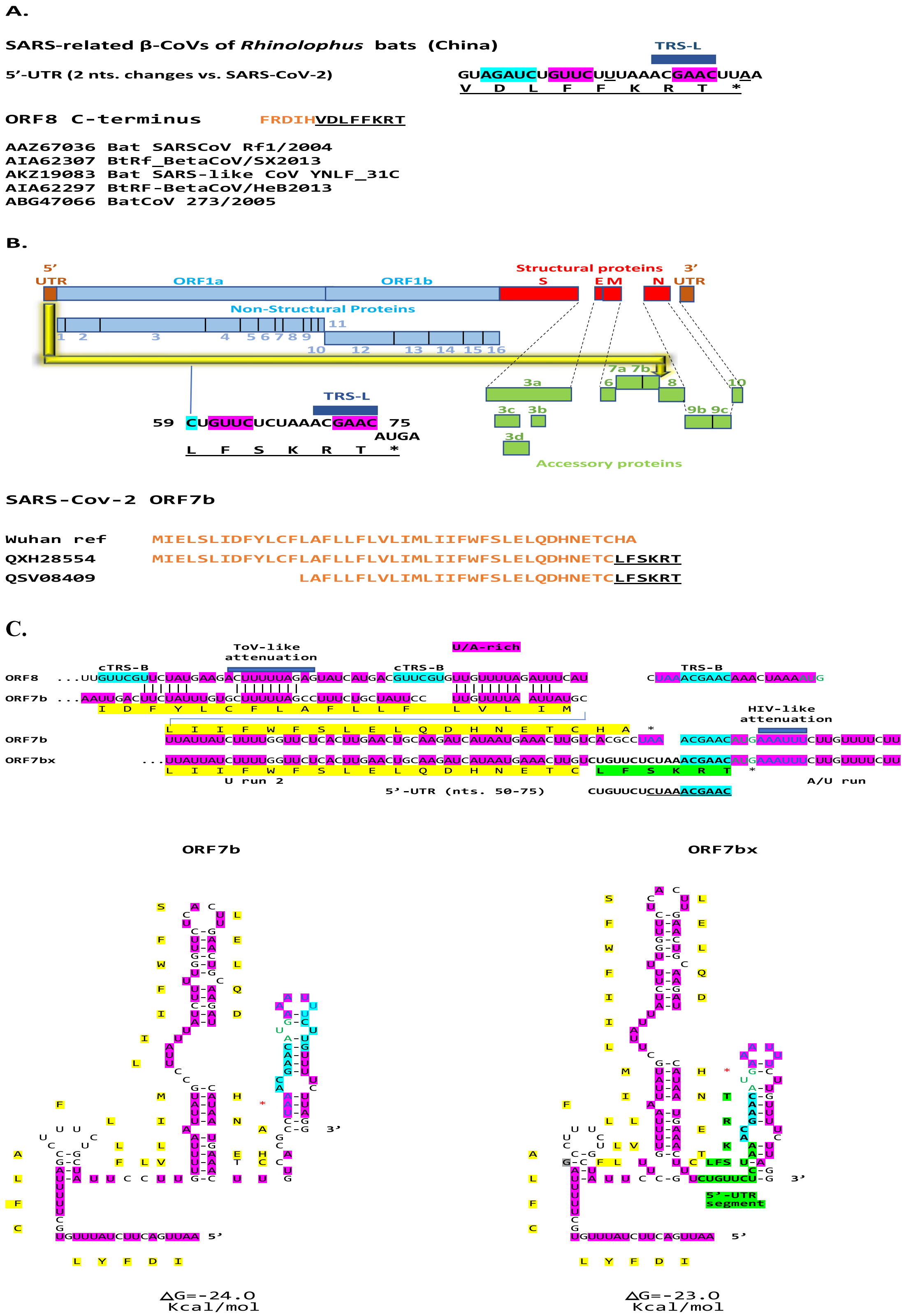


Fig. 2 (See legend on previous page.)

B.1.540 [Pango v.3.1.20 2022-02-02]) and EPI_ISL_1703925 (India/Maharashtra, 2021-02-07; B.1.540 lineage), respectively (Fig. 4A). The latter strains have only one amino acid change outside of the insertions relative to the Wuhan reference strain. A subsegment of 5′-UTR (nts. 62–70) translated as FSK is present at the more proximal site (amino acids 38–40) in 230 isolates isolated from diverse populations at various times (listed in Additional file 1: legend to Fig. S4) and exemplified by isolate UHP90975 [USA/Wisconsin, 2021-12-13] in Fig. 4A. Isolate QZM71485 (USA/New York, 2021-08-05) exemplifies isolates with the FSK sequence at the more distal site (amino acids 148–150). Examples of the most common single amino acid changes in overlapping segments of other isolates are listed as comparators, and they have similar or lower frequency than those of the 5′-UTR segments (summarized in Additional file 1: Table S1). However, the Wuhan reference strain sequence corresponding to the areas with 5′-UTR sequences is the most abundant among SARS-CoV-2 isolates.

Genes encoding components of the replication-transcription complex, such as the RdRp (nsp12) [62, 63], are highly conserved and have a low propensity for recombination among CoVs [34]. The nsp12 NiRAN domain is one of the five replicative peptides that are common to all Nidovirales and used for species demarcation because it is not involved in cross-species homologous recombination [64]. However, as in other examples here of conserved genes, it is involved in intragenomic rearrangements of 5′-UTR sequences. Figure 4B shows the predicted structure RNA structures for the proximal site of intragenomic rearrangement in nsp12 and nsp12x (with 5′-UTR sequence). As in the case of the example in the intragenomic rearrangement in N there is no TRS-B at the site of intragenomic rearrangement, which is however preceded by a sequence similar to the torovirus-like attenuation sequence within a U/A-rich region which may facilitate pausing of the RdRp.

The NiRAN domain of nsp12 is involved in the NMPylation of nsp9 [65] during the formation of the replication-transcription complex (interface regions [66] are shown with yellow bars and key residues therein with ochre letters in Fig. 4). The 5′-UTR sequence at the proximal site in the nsp12 NiRAN domain overlaps with one of the interface regions with nsp9 but does not affect key interface residues or alter the charge distribution of amino acid side chains in the overlap region. The nsp12 NiRAN domain also exhibits a kinase/phosphotransferase like activity [67], is involved in protein-primed initiation of RNA synthesis [68] and catalyzes the formation of the cap core structure (GpppA; contact regions with GDP [66] indicated with blue boxes and key residues therein in ochre in Fig. 4A) [69]. The 5′-UTR sequence at the proximal site in nsp12 NiRAn domain is close to the first contact region with GDP.

Intragenomic rearrangements in β-CoVs of Merbecovirus, Embecovirus, and nob$ecovirus subgenera

Merbecovirus

As shown in Fig. 5A, a segment of the 5′-UTR of the β-CoV *Merbecovirus* MERS-CoV including TRS-L and part of the second of the two stem-loops is present in the intergenic region between the genes encoding p3 and p4a in isolate MG923473 (Burkina Faso, 2015) and at the distal end of the gene encoding p4b in isolate MK564475 (Ethiopia, 2017). In the latter case, the last 4 amino acids (HPGF) of p4b in the reference MERS-CoV sequence (NC_019843) are predicted to be replaced by two amino acids (QL). The Q residue is encoded by a cytosine present in the reference sequence (indicated in orange in Fig. 6A) and two adenosines incorporated by the 5′-UTR sequence. Figure 5B depicts the predicted RNA secondary structures without and with the insertion corresponding to the intragenomic rearrangement between the genes encoding p3 and p4b. The structures have similar predicted minimum free energy, and the rearrangement involves the intergenic TRS-B sequence which is preceded and succeeded by torovirus-like attenuation sequences. It is unknown whether these sequences, which function as attenuation sequences in other viruses, are functional or simply secondary to the fact that AU-rich sequences are frequent in coronavirus genomes.

Embecoviruses

The β-CoV *Embecovirus* hCoV-HKU1 is a sister taxon to murine hepatitis virus and rat sialodacyoadenitis virus [70]. Out of 48 HKU-1 isolates in GenBank, a 5′-UTR sequence including TRS-L, SL3 and most of SL2 (nucleotides 42–74 in hCoV-HKU-1 references NC_006577 and AY597011) is present in 31 isolates (65%) between the *S* and *Ns4* genes (Fig. 6A). The hCoV-HKU-1 NS4 protein is structurally similar to the hCoV-OC43 ns5a protein whose function is detailed in the Discussion section.

All 245 full genome isolates of the β-CoV *Embevovirus* hCoV-OC43 in GenBank had 5′-UTR-leader sequences (largest spanning nucleotides 35–67 of the

(See figure on next page.)

Fig. 3 A. Insertion of 5′-UTR segment into the nucleotide segment encoding the serine-arginine-rich region of the nucleocapsid (N) in SARS-CoV-2. The R203K and G204R amino acid substitutions (blue arrows) which are commonly present concomitantly are encoded in this case by the insertion of a 5′-TR segment into the serine-arginine (SR)-rich region of the N protein at the end of a strong immunodominant B-cell epitope (purple box; [105]). Examples of isolates (a. and b.) with 5′-UTR sequences are provided in the figure and a full listing is provided in the Additional file 1: legend to Fig. S3. **B**. Secondary structures of *N* and *Nx* RNAs. Color scheme is as in preceding figures

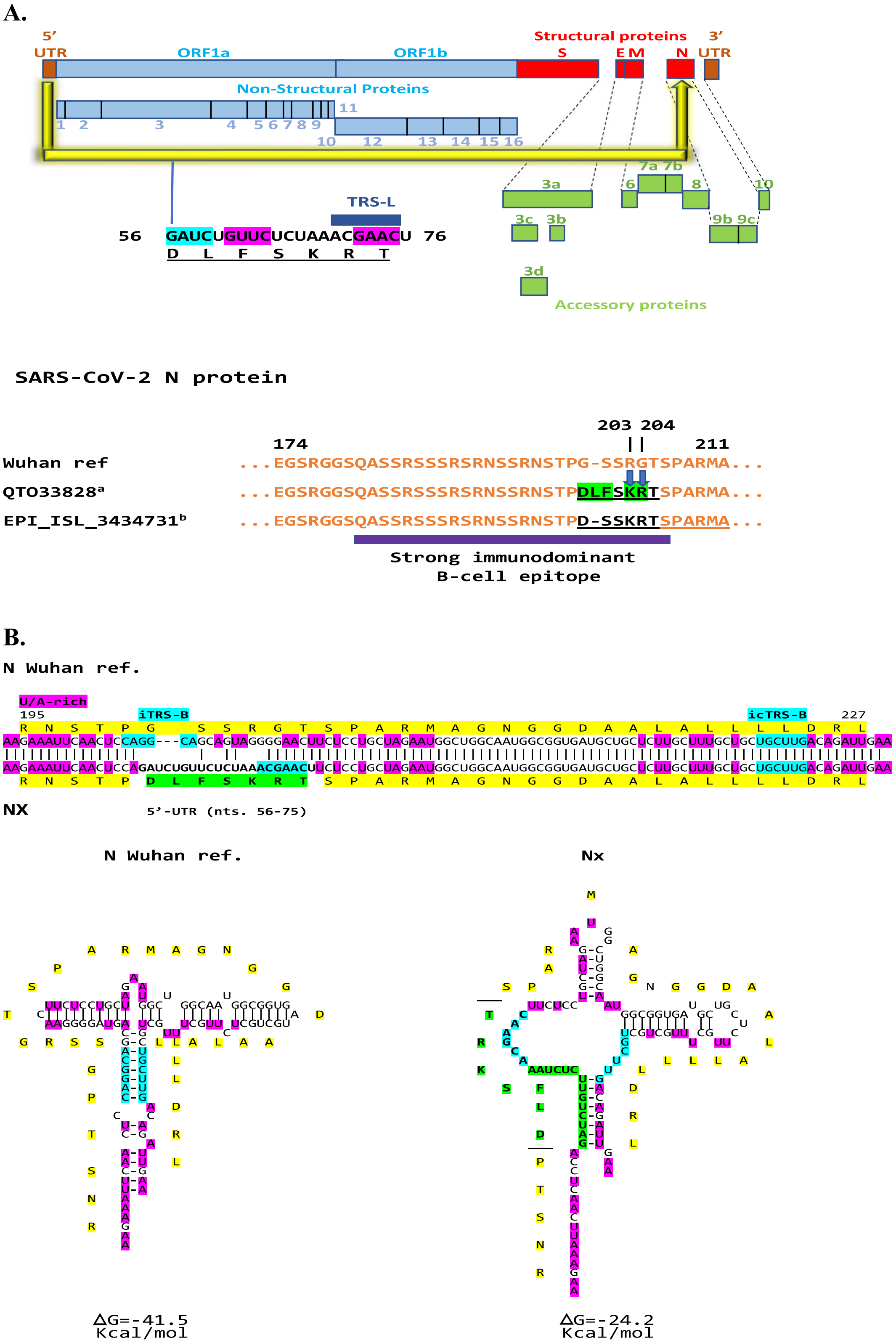

A.
5' UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3' UTR
Non-Structural Proteins
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
TRS-L
56 GAUCUGUUCUCUAAACGAACU 76
D L F S K R T
3a 3c 3b 3d 6 7a 7b 8 9b 9c 10
Accessory proteins
SARS-CoV-2 N protein
174 203 204 211
Wuhan ref ...EGSRGGSQASSRSSSRSRNSSRNSTPG-SSRGTSPARMA...
QT033828ª ...EGSRGGSQASSRSSSRSRNSSRNSTPDLFSKRTSPARMA...
EPI_ISL_3434731ᵇ ...EGSRGGSQASSRSSSRSRNSSRNSTPD-SSKRTSPARMA...
Strong immunodominant B-cell epitope
B.
N Wuhan ref.
U/A-rich
195 iTRS-B icTRS-B 227
R N S T P G S S R G T S P A R M A G N G G D A A L A L L L L D R L
AAGAAAUUCAACUCCAGG---CAGCAGUAGGGGAACUUCUCCUGCUAGAAUGGCUGGCAAUGGCGGUGAUGCUGCUCUUGCUUUGCUGCUGCUUGACAGAUUGAA
AAGAAAUUCAACUCCAGAUCUGUUCUCUAAACGAACUUCUCCUGCUAGAAUGGCUGGCAAUGGCGGUGAUGCUGCUCUUGCUUUGCUGCUGCUUGACAGAUUGAA
R N S T P D L F S K R T S P A R M A G N G G D A A L A L L L L D R L
NX 5'-UTR (nts. 56-75)
N Wuhan ref.
Nx
ΔG=-41.5 Kcal/mol
ΔG=-24.2 Kcal/mol

Fig. 3 (See legend on previous page.)

hCoV-OC43 reference strain KJ958218) between the spike (*S*) and *Nsp5a* genes (Fig. 6B). The insertions did not affect the protein sequences of either S or Nsp5a. The hCoV-OC43 5′-UTR sequence inserted is identical to that of bovine coronavirus (BCoV) 5′-UTR except for one nucleotide (an adenosine substituted by a guanosine in BCoV) up to the TRS-B, and sequences of varying length after the TRS-B show similarities to BCoV 5′-UTR, which is consistent with a most probable bovine or swine coronavirus origin for hCoV-OC43 [71]. The 5′-UTR sequence insertion sequence is also present in a molecularly characterized cloned hCoV-OC43 *S* gene [72].

Nobecoviruses

An intragenomic rearrangement involving a 5′-UTR sequence (nucleotides 1–55) to distal end encoding the C-terminal cytoplasmic Y1 domain of nsp3 (nucleotides 6837–6891; amino acids 2188–2205), is seen in the β-CoV subgenus *Nobecovirus* of African bats, namely isolates MIZ240 (OK067321) and MIZ178 (OK067320) from *Rousettus madagascariensis* bats and isolates CMR900 (MG693169; protein: AWV67046), CMR705-P13 (MG693172, protein: AWV67070), and unclassified (NC_048212) from *Eidolon helvum* bats (Cameroon). Using the translated nucleotide sequence as query, the following additional isolates were detected: *Eidolon helvum* (Cameroon) isolates CMR704-P12 (YP_009824989 and YP_009824988), and CMR891-892 (AWV67062). The 5′-UTR sequence involved in this intragenomic rearrangement does not include the TRS-L and includes a stem-loop structure highlighted in grey in Fig. 7A. The position of the translated sequence of the 5′-UTR identical sequence is amino acids 2188–2205, which corresponds to amino acids 1567–1584 in SARS-CoV-2 nsp3. Figure 7B depicts the predicted secondary structures of *nsp3* and *nsp3x* RNAs with the intragenomic rearrangement. Both structures have similar predicted minimum free energy. Although there are no TRS-B sequences present in this region the rearrangement takes place adjacent to an inverted complementary TRS-B within a U/A-rich region.

Intragenomic rearrangement in nsp2 of rodent α-CoVs subgenus Luchacovirus

As shown in Fig. 8A, a segment of the 5′-UTR (nts. 1–119) of the *Luchacovirus* AcCoV-JC34 (KX964649; isolated in China, 2011-10 from the rodent *Apodemus chevieri*) was duplicated, inverted, and translocated to the genomic region encoding the nonstructural protein nsp2 (nts. 1679-1760). The latter sequence in *nsp2* differs by only one nucleotide from that in the 5′-UTR (99% similarity), and it is also present with varying degrees of similarity in other rodents. Two examples are shown in Fig. 8A for isolates from rat (Lucheng Rn rat CoV isolate Ruian 83; MT820626; isolated from *Rattus norvegicus* in China, 2014, 76% similar), and mouse (Fievel mouse CoV strain FiCoV/UMN2020 (OK655840; isolated from *Mus musculus* in USA, 2018, 59% similar). Other isolates (listed by rodent of origin) with intragenomic rearrangements in *nsp2* with nucleotide sequences up to 75% similar to the 5′-UTR sequences include: *Apodemus chevrieri* (MT820625, China, 2015, 93% similar); *Apodemus agrarium* (MZ328302, China, 2016, 93% similar); *Eothenomys miletus* (MT820627, China, 2014, 81% similar); *Eothynomys melanogaster* (KY370054, China, 2015-12, 79% similar); *Myodes rufocanus* (KY370045, China, 2014-08, 79% similar); *Rattus losea* (KY370050, China, 2015-05, 78% similar); *Rattus norvegicus* (MK163627, United Kingdom, 2014-06-23, 78% similar; NC_032730, China, 2013; MT549854, China, 2016-12, 76% similar; MW802582, China, 2017-03-07, 76% similar); and *Brylmys bowersi* (MZ328301, China, 2016, 77% similar).

There appears to be a temporal gradient with the most similar sequence (99%) in isolate KX964649 (China, 2011-10) to the least similar (59%) in isolate OK655840 (USA, 2018). The temporal gradient of decreasing similarity holds within rodents from the same genus, which would suggest that the translocated sequence is the oldest and the rest reflect more recent mutations. This is consistent with a possible common ancestor for all rodent α-CoVs sampled so far, with phylogenetic analyses suggesting relatively frequent host-jumping among the different rodent species [50]. The minimum free energy of the predicted RNA secondary structures of the intragenomic rearrangement and adjacent sequences increases from the most similar to the least similar to the 5′-UTR

(See figure on next page.)

Fig. 4 A. Insertions of a 5′-UTR sequence into two sites within the nucleotide segment encoding the *Nidovirus* RdRp associated nucleotidyl transferase (NiRAN) domain of the RNA-dependent RNA polymerase (nsp12) of SARS-CoV-2. Examples of isolates with 5′-UTR sequences at the proximal and distal sites are provided in the figure and a full listing is provided in the Additional file 1: legend to Fig. S4, as is a listing of variants with single amino acid changes relative to the Wuhan reference strain in the segment corresponding to the insertion. The Wuhan reference strain sequence corresponding to the insertion areas is the most abundant among SARS-CoV-2 isolates. The nsp12-nsp9 interface regions are shown with yellow bars and key residues therein with ochre letters, while the contact regions with GDP are indicated with blue boxes and key residues therein in ochre. **B**. Secondary structures for RNAs in the proximal sites in *nsp12* and *nsp12x*. Color scheme is as in previous figures. The site for -1 frameshifting is also highlighted

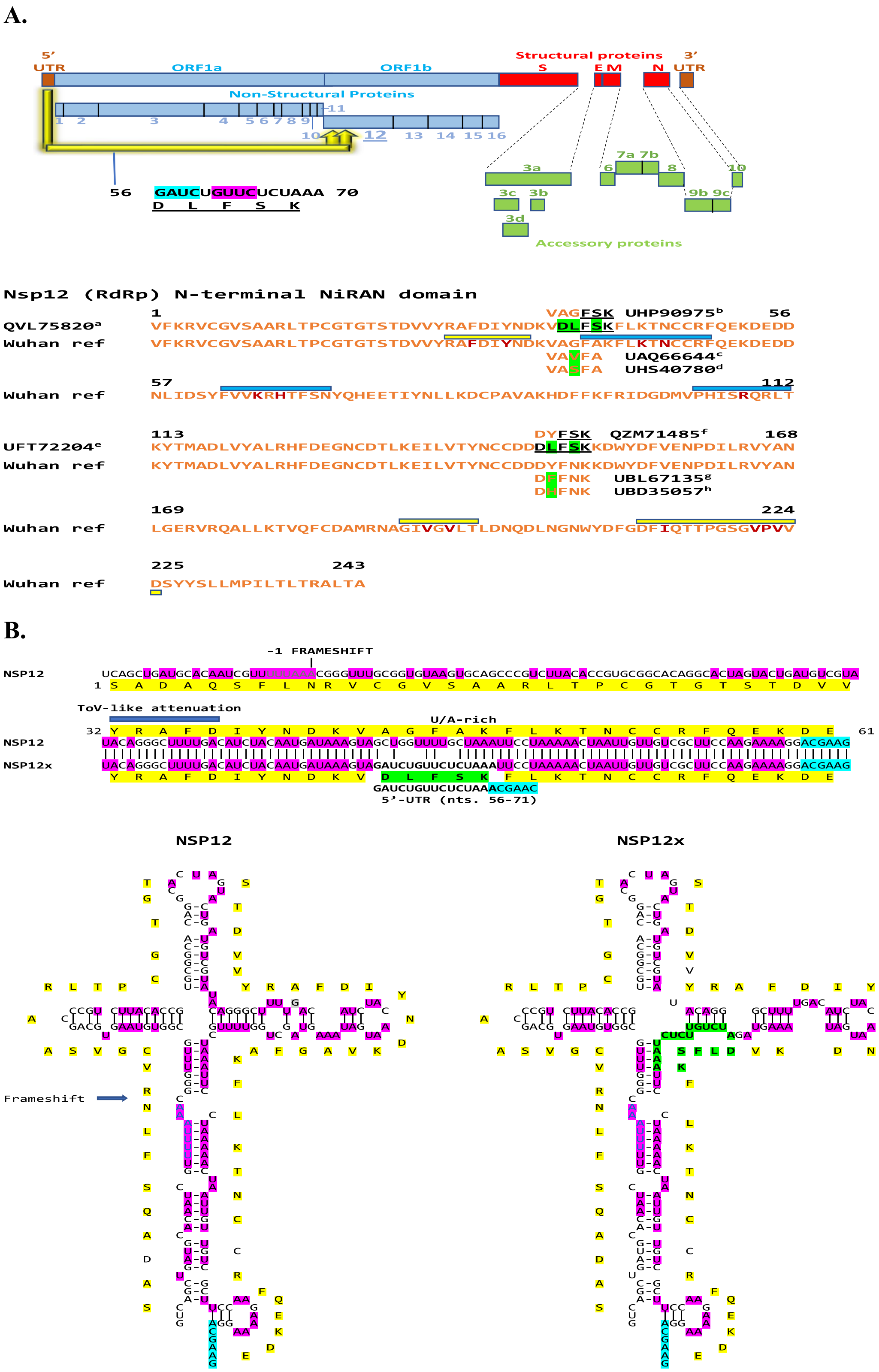

A.
5' UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3' UTR
Non-Structural Proteins
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
56 GAUCUGUUCUCUAAA 70
D L F S K
3a 3b 3c 3d 6 7a 7b 8 9b 9c 10
Accessory proteins
Nsp12 (RdRp) N-terminal NiRAN domain
1
VAGFSK UHP90975b 56
QVL75820a VFKRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVDLFSKFLKTNCCRFQEKDEDD
Wuhan ref VFKRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVAGFAKFLKTNCCRFQEKDEDD
VAVFA UAQ66644c
VAAFA UHS40780d
57 112
Wuhan ref NLIDSYFVVKRHTFSNYQHEETIYNLLKDCPAVAKHDFFKFRIDGDMVPHISRQRLT
113 168
DYFSK QZM71485f
UFT72204e KYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDLFSKKDWYDFVENPDILRVYAN
Wuhan ref KYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDYFNKKDWYDFVENPDILRVYAN
DYFNK UBL67135g
DHFNK UBD35057h
169 224
Wuhan ref LGERVRQALLKTVQFCDAMRNAGIVGVLTLDNQDLNGNWYDFGDFIQTTPGSGVPVV
225 243
Wuhan ref DSYYSLLMPILTLTRALTA
B.
-1 FRAMESHIFT
NSP12 UCAGCUGAUGCACAAUCGUUUUUAAACGGGUUUGCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUA
1 S A D A Q S F L N R V C G V S A A R L T P C G T G T S T D V V
ToV-like attenuation
U/A-rich
32 Y R A F D I Y N D K V A G F A K F L K T N C C R F Q E K D E 61
NSP12 UACAGGGCUUUUGACAUCUACAAUGAUAAAGUAGCUGGUUUUGCUAAAUUCCUAAAAACUAAUUGUUGUCGCUUCCAAGAAAAGGACGAAG
NSP12x UACAGGGCUUUUGACAUCUACAAUGAUAAAGUAGAUCUGUUCUCUAAAUUCCUAAAAACUAAUUGUUGUCGCUUCCAAGAAAAGGACGAAG
Y R A F D I Y N D K V D L F S K F L K T N C C R F Q E K D E
GAUCUGUUCUCUAAAACGAAC
5'-UTR (nts. 56-71)
NSP12
NSP12x
Frameshift

Fig. 4 (See legend on previous page.)

insertion (Additional file 1: Fig. S3). The function of the region of intragenomic rearrangement in nsp2 remains to be determined and it does not overlap with that contributing to inflammation via NF-κB activation in the α-CoV porcine transmissible gastroenteritis virus [73].

Intragenomic rearrangements in N of bat α-CoVs subgenus Nyctacovirus

As shown in Fig. 8B, in this intragenomic rearrangement in bat α-CoVs subgenus *Nyctacovirus*, a 115-nucleotide-long segment of the 5′-UTR is duplicated, inverted (negative-sense strand) and translocated to the proximal end of the nucleocapsid gene thereby encoding the predicted first 38 amino acids of the amino-terminus of N. Other variants share the sequence with lesser similarity to the 5′-UTR sequence. There is a TRS-B sequence (AACUAA) at the beginning of the insertion, and the negative strand 5′-UTR sequence also has a AACUAA sequence, which may have mediated the intragenomic rearrangement.

Intragenomic rearrangements in ORF5b/4b of bat α-CoVs subgenus Decacovirus

Orb5b and ORF4b proteins are 53% similar (including conservative substitutions) between bat alphacoronaviruses subgenus *Decacovirus* shown in Fig. 8C. In both sets of viruses, *ORF5b* or *ORF4b* overlap the beginning of the membrane (*M*) gene, i.e., there is no intergenic region between them and *M*. However, there is a TRS-B sequence (AACUAA) within the 3′ end of *ORF5b* and *ORF4b* where the intragenomic rearrangement occurs. Viruses with similar intragenomic rearrangements in *ORF5b* include: *Rhinolophus* bat coronavirus HKU32 strain TLC28A (MK720946), *Rhinolophus* bat coronavirus HKU32 strain TLC26A (MK720945; Hong Kong, 08-06-2005), Alphacoronavirus sp. strain bat/Yunnan/HcYN26/2020 (MZ081384; *Hipposideros cineraceus*; China, 07-29-2020), Alphacoronavirus sp. strain bat/Yunnan/RsYN12/2019 (MZ081386; *Rhinolophus sinicus*; China, 10-22-2019), Alphacoronavirus sp. strain bat/Yunnan/MmYN16/2020 (MZ081385; *Myotis muricola*; China, 04-18-2020), Alphacoronavirus sp. strain bat/Yunnan/RmYN21/2020 (MZ081387; *Rinolophus malayanus*; China, 06-03-2020). Viruses with similar intragenomic rearrangements in *ORF4b* include: Bat coronavirus isolate BtCoV/Rh/YN2012_Rs4259 (MG916903; China, 04-17-2013), Bat coronavirus isolate BtCoV/Rh/YN2012_Rs4125 (MG916902; China, 09-16-2012). The functional significance of this intragenomic rearrangement remains to be determined.

Intragenomic rearrangements of 5′-UTR sequences were not detected in some β-or α-, or in any γ- and δ-CoVs, and no intragenomic rearrangements of 3′-UTR sequences were detected in any coronavirus

A listing of the other coronaviruses analyzed beyond the ones found to have intragenomic rearrangements is provided at the end of the Additional file 1. The directionality of potential translocation appears to be in the 5′–3′ direction as further underscored by the absence of 3′-UTR sequence insertions in any of the viruses analyzed.

Discussion

We here describe intragenomic rearrangements involving 5′-UTR sequences and the coding section of the genome of beta- and alphacoronaviruses. Additional file 1: Fig. S4A summarizes the locations of insertions in accessory, structural, and nonstructural genes of SARS-CoV-2, which for at least the accessory and structural genes appear to involve and/or affect the template switching mechanism by creating new regions of homology for interaction with TRS-L. The presence of conserved complementary sequences (CCSs) in the 5′- and 3′-UTRs potentially involved in circularization of the genome during subgenomic RNA synthesis has been reported [74]. As shown in Additional file 1: Fig. S4B, the 5′-UTR sequences involved in intragenomic rearrangements in SARS-CoV-2 shown in the present work usually include the TRS-L and span approximately half of the 5′ CCS, thus potentially facilitating circularization of the genome from locations closer to the 3′-UTR. The 5′-UTR sequences involved in intragenomic rearrangements may also facilitate other long-distance RNA-RNA interactions contributing to the complex coronavirus transcription process [75].

Most of the 5′-UTR sequences duplicated and translocated include TRS-L. Extending the homology region of interaction between the TRS-L in the 5′-leader and the TRS-L introduced in a particular area of the body

(See figure on next page.)

Fig. 5 **A**. Intragenomic rearrangement with 5′-UTR sequences present in the intergenic regions between genes encoding p3 and 4a as well as between those encoding p4b and p5 of the *Merbecovirus* Middle East respiratory syndrome (MERS)-CoV. A segment of the 5′-UTR of the MERS-CoV including TRS-L and part of the second of the two stem-loops is present in the intergenic region between genes encoding p3 and p4a in isolate MG923473 (Burkina Faso, 2015) and between those encoding p4b and p5 affecting the predicted carboxyl-terminal end of ORF4b in isolate MK564475 (Ethiopia, 2017). In the latter case, the last 4 amino acids (HPGF) of ORF4b in the reference MERS-CoV sequence (NC_019843) are replaced by two amino acids (QL). The Q residue is encoded by a cytosine present in the reference sequence (indicated in orange color) and two adenosines incorporated by the 5′-UTR sequence. **B**. RNA secondary structures of the intergenic region between genes encoding p3 and p4a in MERS-CoV without and with intragenomic rearrangement. Color scheme is as in previous figures

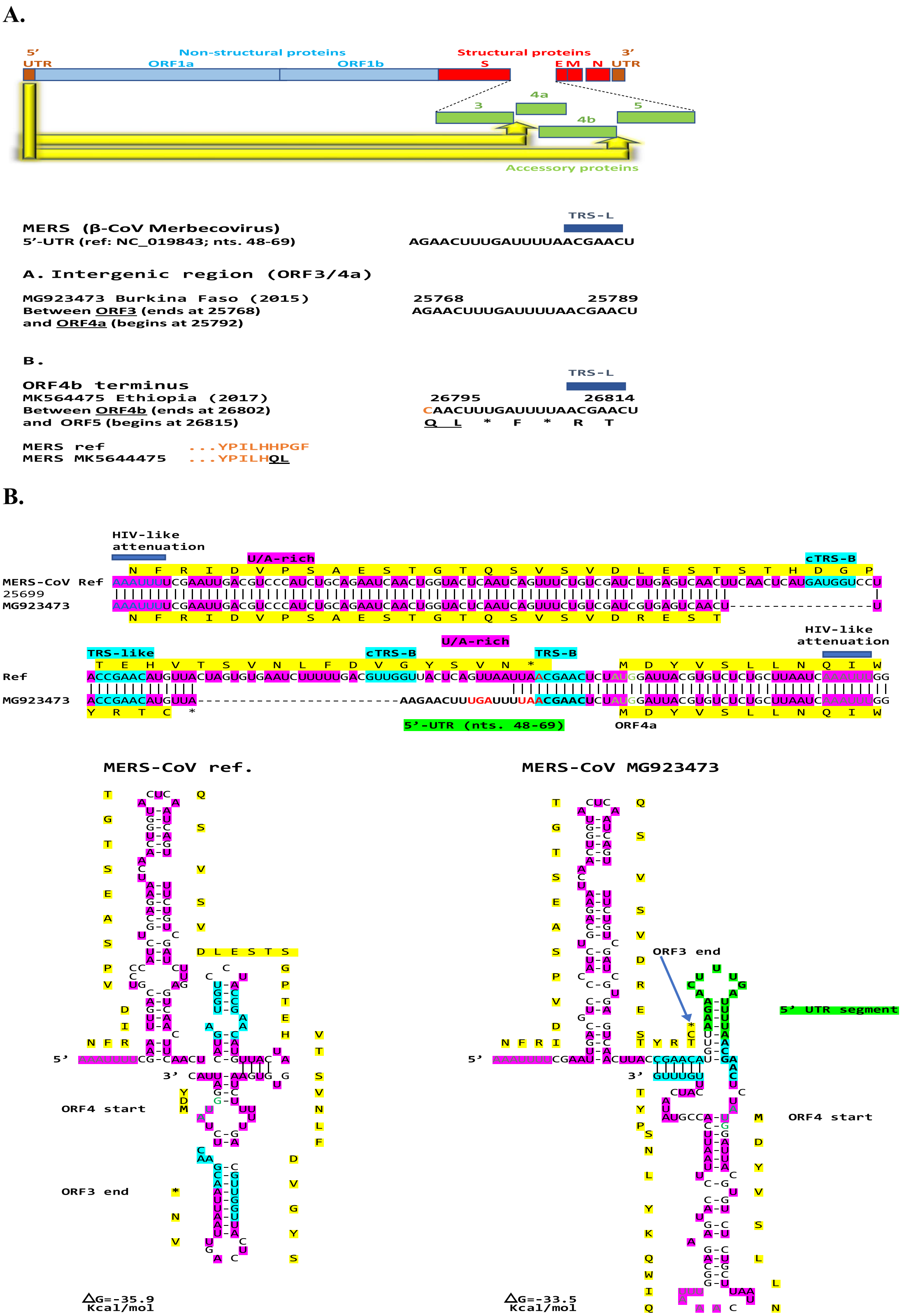


Fig. 5 (See legend on previous page.)

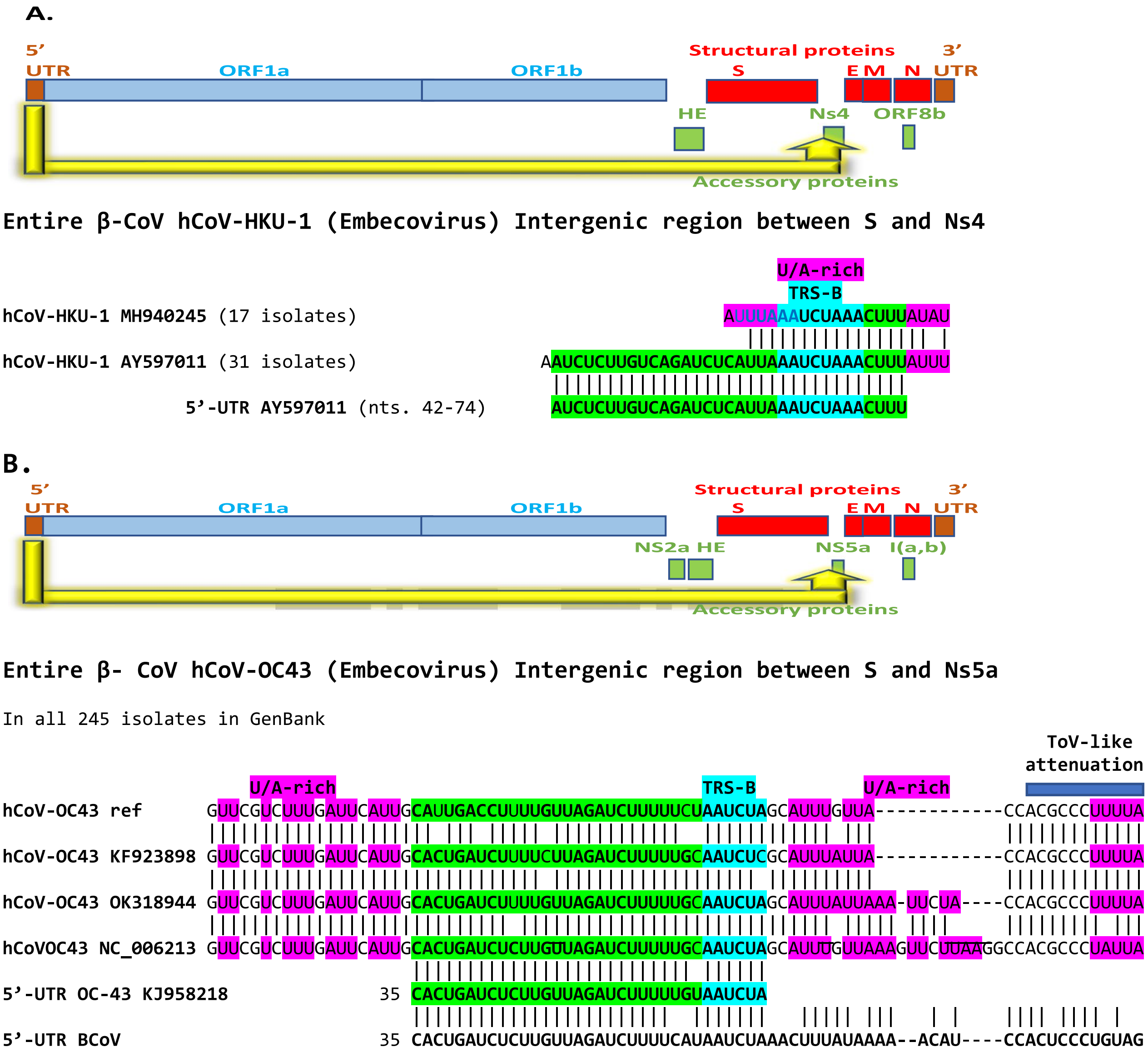

A.
5' UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3' UTR
HE
Ns4
ORF8b
Accessory proteins
Entire β-CoV hCoV-HKU-1 (Embecovirus) Intergenic region between S and Ns4
U/A-rich
TRS-B
hCoV-HKU-1 MH940245 (17 isolates) AUUUAAAUCUAAACUUUAUAU
hCoV-HKU-1 AY597011 (31 isolates) AAUCUCUUGUCAGAUCUCAUUAAAUCUAAACUUUAUUU
5'-UTR AY597011 (nts. 42-74) AUCUCUUGUCAGAUCUCAUUAAAUCUAAACUUU
B.
5' UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3' UTR
NS2a HE
NS5a
I(a,b)
Accessory proteins
Entire β- CoV hCoV-OC43 (Embecovirus) Intergenic region between S and Ns5a
In all 245 isolates in GenBank
ToV-like attenuation
U/A-rich
TRS-B
U/A-rich
hCoV-OC43 ref GUUCGUCUUUGAUUCAUUGCAUUGACCUUUUGUUAGAUCUUUUUCUAAUCUAGCAUUUGUUA------------CCACGCCCUUUUA
hCoV-OC43 KF923898 GUUCGUCUUUGAUUCAUUGCACUGAUCUUUUCUUAGAUCUUUUUGCAAUCUCGCAUUUAUUA------------CCACGCCCUUUUA
hCoV-OC43 OK318944 GUUCGUCUUUGAUUCAUUGCACUGAUCUUUUGUUAGAUCUUUUUGCAAUCUAGCAUUUAUUAAA-UUCUA----CCACGCCCUUUUA
hCoVOC43 NC_006213 GUUCGUCUUUGAUUCAUUGCACUGAUCUCUUGUUAGAUCUUUUUGCAAUCUAGCAUUUGUUAAAGUUCUUAAGGCCACGCCCUAUUA
5'-UTR OC-43 KJ958218 35 CACUGAUCUCUUGUUAGAUCUUUUUGUAAUCUA
5'-UTR BCoV 35 CACUGAUCUCUUGUUAGAUCUUUUCAUAAUCUAAACUUUAUAAAA--ACAU----CCACUCCCUGUAG

Fig. 6 **A**. Presence of hCoV-HKU-1 (β-CoV *Embecovirus*) of 5′-UTR sequence in the intergenic region between the spike (*S*) and the *Ns4* genes. The hCoV-OC43 5′-UTR sequence inserted is identical to that of bovine coronavirus (BCoV) 5′-UTR (shown at the bottom of the figure) except for one nucleotide (an adenosine [A] instead of a guanosine [G] in BCoV). 31 out of 48 variants (65%) in GenBank have this intragenomic rearrangement. **B**. Presence in in the intergenic region between the *S* and *Ns5a* genes hCoV-OC43 (β-CoV *Embecovirus*) of sequences of various lengths of the same 5′-UTR region. As in the case of hCoV-HKU1, the 5′-UTR segment translocated to the intergenic region between *S* and *Ns5a* of hCoV-OC43 variants is similar to that of BCoV. Differences among them (all 245 isolates in GenBank) are distally to the TRS-B and involve various extents of sequences similar to the 5′-UTR of BCoV

of the genome optimizes minimum free energy of the interaction. Such facilitation may favor expression of certain genes over that of others, thereby altering the hierarchy in gene expression. Because insertions are in various locations of viral genes, including some encoding nonstructural proteins, they may propitiate formation of new subgenomic RNAs thereby expanding the repertoire of proteins and even transforming noncanonical subgenomic messenger RNAs, i.e., not associated with TRS homology, to canonical ones. SARS-CoV-2 and other CoVs have been reported to generate noncanonical subgenomic RNAs in abundance, accounting for up to a third of subgenomic messenger RNAs in cell culture models of infection and increasing in proportion over time [76].

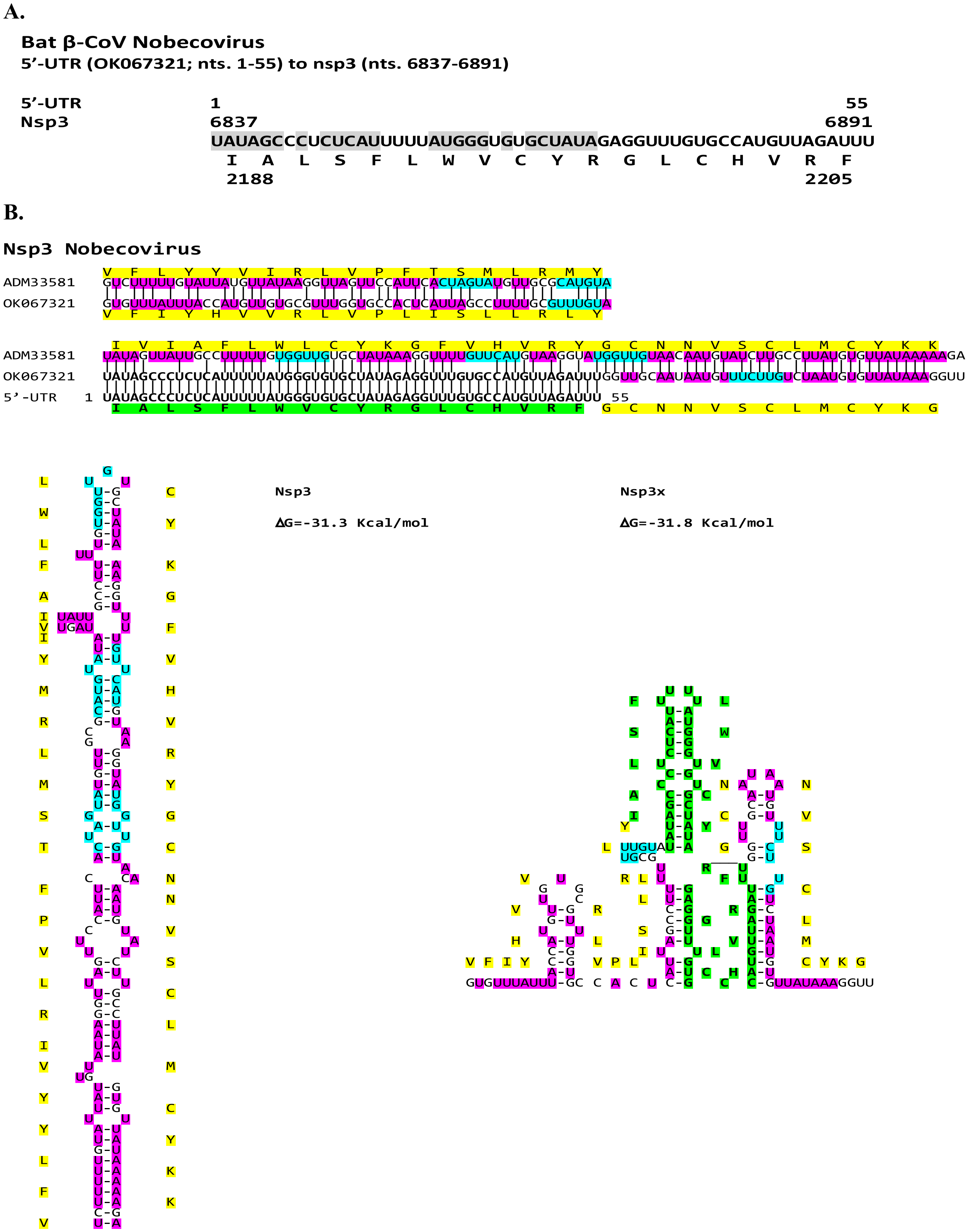

A.
Bat β-CoV Nobecovirus
5'-UTR (OK067321; nts. 1-55) to nsp3 (nts. 6837-6891)
5'-UTR 1 55
Nsp3 6837 6891
UAUAGCCCUCUCAUUUUUAUGGGUGUGCUAUAGAGGUUUGUGCCAUGUUAGAUUU
I A L S F L W V C Y R G L C H V R F
2188 2205
B.
Nsp3 Nobecovirus
V F L Y Y V I R L V P F T S M L R M Y
ADM33581 GUCUUUUUGUAUUAUGUUAUAAGGUUAGUUCCAUUCACUAGUAUGUUGCGCAUGUA
OK067321 GUGUUUAUUUACCAUGUUGUGCGUUUGGUGCCACUCAUUAGCCUUUUGCGUUUGUA
V F I Y H V V R L V P L I S L L R L Y
I V I A F L W L C Y K G F V H V R Y G C N N V S C L M C Y K K
ADM33581 UAUAGUUAUUGCCUUUUUGUGGUUGUGCUAUAAAGGUUUUGUUCAUGUAAGGUAUGGUUGUAACAAUGUAUCUUGCCUUAUGUGUUAUAAAAAGA
OK067321 UAUAGCCCUCUCAUUUUUAUGGGUGUGCUAUAGAGGUUUGUGCCAUGUUAGAUUUGGUUGCAAUAAUGUUUCUUGUCUAAUGUGUUAUAAAGGUU
5'-UTR 1 UAUAGCCCUCUCAUUUUUAUGGGUGUGCUAUAGAGGUUUGUGCCAUGUUAGAUUU 55
I A L S F L W V C Y R G L C H V R F G C N N V S C L M C Y K G
Nsp3
ΔG=-31.3 Kcal/mol
Nsp3x
ΔG=-31.8 Kcal/mol

Fig. 7 **A**. Presence of 5′-UTR sequence in the Bat β-CoV *Nobecovirus nsp3* gene. **B**. Secondary structures of *nsp3* and *nsp3x* RNAs. Color scheme is as in previous figures

The structural genes control genome dissemination [63] while the accessory genes in the same region of the genome may be involved in adaptation to specific hosts, modulation of the interferon signaling pathways, the production of pro-inflammatory cytokines, or the induction of apoptosis [77], among other mechanisms underlying immune evasion and pathogenesis. Gaining insight into the effect of the amino acid changes introduced by the 5′-UTR sequences is likely to shed light into pathogenesis and immune evasion mechanisms. For instance, a few point mutations can have a profound effect as exemplified by the few mutations in the C-terminus of the spike protein that transform the feline CoV associated with mild disease to one, the feline infectious peritonitis virus, which is generally lethal [78].

ORF8 had been postulated to originate from *ORF7a* by non-homologous recombination, and a predicted structure model of the ORF8 protein of SARS-CoV-2 revealed a ~ 60-residue core like that of SARS-CoV-2 ORF7a protein [79] with the addition of two dimerization interfaces, one covalent and the other noncovalent, unique to SARS-CoV-2 ORF8 [80]. In the C-terminus of ORF8 that would be predicted to be altered by 5′-UTR sequence insertions (i.e., $_{115}$RVVLDFI$_{121}$), R115, D119, F120, and I121 contribute to the covalent dimer interface (marked with asterisks in Fig. 1) with R115 and D119 forming salt bridges that flank a central hydrophobic core in which V117 interacts with its symmetry-related counterpart [80].

How the C-terminal insertions and changes therein affect the dimerization of ORF8 protein remains to be determined and described functions for ORF8 protein remain a matter of debate [81]. However, the predicted changes caused by insertions might contribute to immune evasion by SARS-CoV-2 by affecting the interactions of the ORF8 glycoprotein homodimer with intracellular transport signaling, leading to down-regulation of MHC-I by selective targeting for lysosomal degradation via autophagy [82], and/or extracellular signaling involving interferon-I signaling [83], mitogen-activated protein kinases growth pathways [84], the tumor growth factor-β1 signaling cascade [85] and interleukin-17 signaling promoting inflammation and contributing to the COVID-19-associated cytokine storm [86].

The carboxyl-terminal region of the ORF8 protein may include T- and/or B-cell epitopes that may be affected by the variations described. To this end, approximately 5% of CD4+ T cells in most COVID-19 cases are specific for ORF8 protein, and ORF8 protein accounts for 10% of CD8+ T cell reactivity in COVID-19 recovered subjects [87, 88]. Another possible effect of the insertions stems from the fact that anti-ORF8 protein antibodies are detected in both symptomatic and asymptomatic patients early during infection by SARS-CoV-2 [89] and diagnostic assays for SARS-CoV-2 infection that target only accessory genes or proteins such as ORF8 may be affected [39].

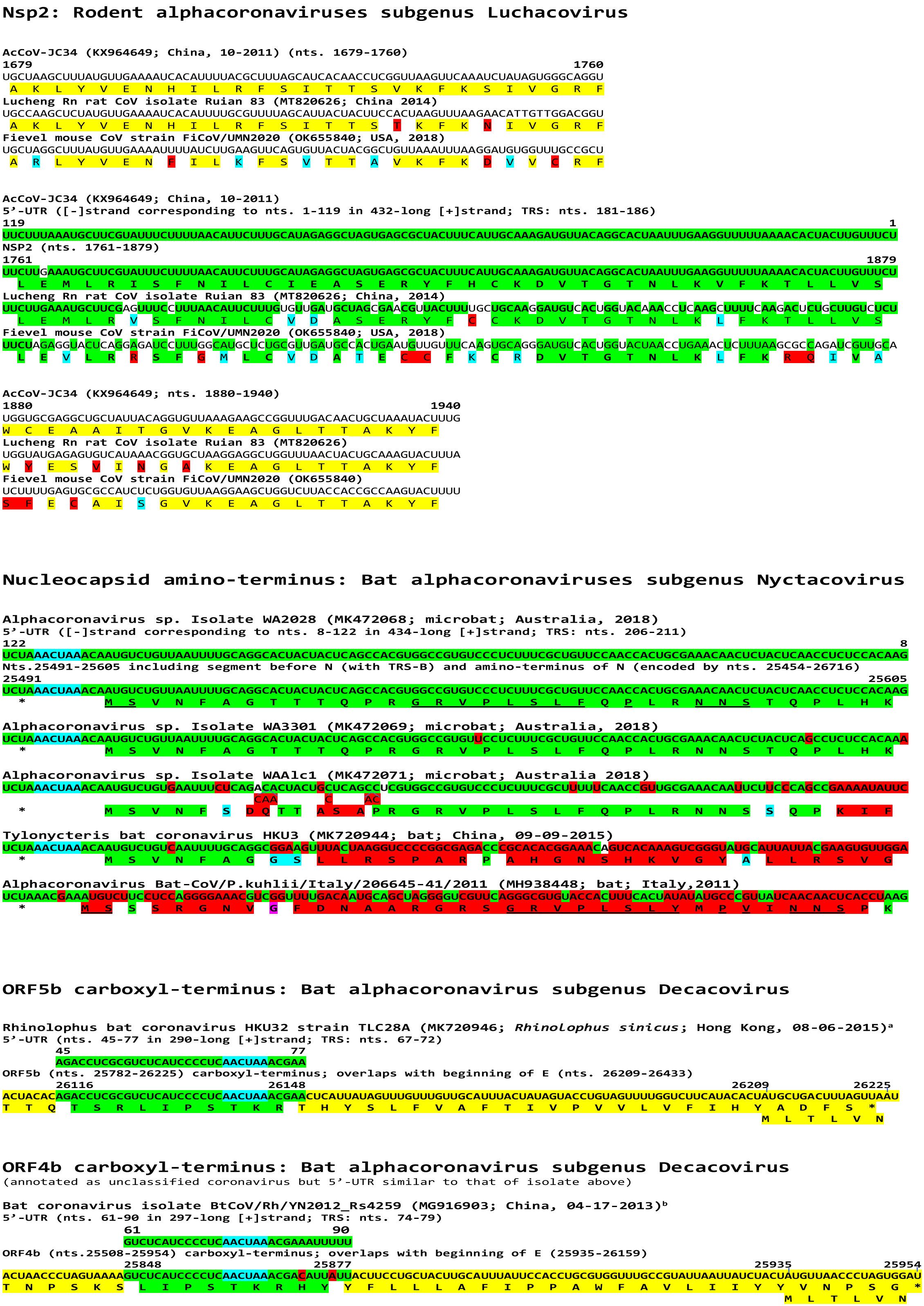
In terms of the potential consequences of intragenomic rearrangements involving *ORF7b* of SARS-CoV-2, the function of the SARS-CoV-2 ORF7b protein remains to be determined and has been suggested to mediate tumor necrosis factor-α-induced apoptosis based on cell culture data [90] and theoretically the dysfunction of olfactory receptors by triggering autoimmunity [91].

We also found intragenomic rearrangements in the nucleocapsid gene of SARS-CoV-2 and bat α-CoVs subgenus *Nyctacovirus*. The nucleocapsid is the most abundant protein in CoVs, interacts with membrane protein [92, 93], self-associates to provide for efficient viral assembly [94], binds viral RNA [95] and has been involved in circularization of the murine hepatitis virus genome via interaction with 3′- and 5′-UTR sequences which may facilitate template switching during subgenomic RNA synthesis [96]. Phosphorylation transforms N-viral RNA condensates into liquid-like droplets, which may provide a cytoplasmic-like compartment to support the protein's function in viral genome replication [93, 97].

The phosphorylation-rich stretch encompassing amino acid residues 180–210 (SR region) encoded by the nucleotide segment where 5′-UTR sequences were detected in SARS-CoV-2, serves as a key regulatory hub in N protein function within a central disordered linker for dimerization and oligomerization of the N protein, which is phosphorylated early in infection at multiple sites by cytoplasmic kinases [97]. Serine 202 (numbering

(See figure on next page.)

Fig. 8 Intragenomic rearrangement in *nsp2* of rodent alphacoronaviruses subgenus *Luchacovirus*, *N* of bat alphacoronaviruses subgenus *Nyctacovirus* (A) and *ORF5b* or *ORF4b* of bat alphacoronaviruses subgenus *Decacovirus*. Nucleotide and amino acid sequences of intragenomic rearrangements are shown. 5′-UTR sequence (negative strand) is highlighted in green. Conservative amino acid substitutions are highlighted in blue while non-conservative ones are highlighted in red. For the intragenomic rearrangement in *nsp2* of rodent alphacoronaviruses subgenus *Luchacovirus*, two examples of isolates (listed by rodent of origin) with *nsp2* nucleotide sequences up to 75% similar to the 5′-UTR sequences are shown. There appears to be a temporal gradient with the most similar sequence (99%) in isolate KX964649 (China, 2011-10) to the least similar (59%) in isolate OK655840 (USA, 2018). The temporal gradient holds within animals from the same genus, which would suggest that the translocated sequence is the oldest and the rest reflect more recent mutations. For the predicted secondary structures of the RNAs corresponding to the intragenomic rearrangement and adjacent sequences, the minimum free energy increases among variants from those with the most to those with the least similar sequence to the 5′-UTR insertion (Additional file 1: Fig. S3). Color scheme is as in previous figures



```
Nsp2: Rodent alphacoronaviruses subgenus Luchacovirus

AcCoV-JC34 (KX964649; China, 10-2011) (nts. 1679-1760)
1679                                                                               1760
UGCUAAGCUUUAUGUUGAAAAUCACAUUUUACGCUUUAGCAUCACAACCUCGGUUAAGUUCAAAUCUAUAGUGGGCAGGU
 A  K  L  Y  V  E  N  H  I  L  R  F  S  I  T  T  S  V  K  F  K  S  I  V  G  R  F
Lucheng Rn rat CoV isolate Ruian 83 (MT820626; China 2014)
UGCCAAGCUCUAUGUUGAAAAUCACAUUUUGCGUUUUAGCAUUACUACUUCCACUAAGUUUAAGAACATTGTTGGACGGU
 A  K  L  Y  V  E  N  H  I  L  R  F  S  I  T  T  S  T  K  F  K  N  I  V  G  R  F
Fievel mouse CoV strain FiCoV/UMN2020 (OK655840; USA, 2018)
UGCUAGGCUUUAUGUUGAAAAUUUUAUCUUGAAGUUCAGUGUUACUACGGCUGUUAAAUUUAAGGAUGUGGUUUGCCGCU
 A  R  L  Y  V  E  N  F  I  L  K  F  S  V  T  T  A  V  K  F  K  D  V  V  C  R  F


AcCoV-JC34 (KX964649; China, 10-2011)
5’-UTR ([-]strand corresponding to nts. 1-119 in 432-long [+]strand; TRS: nts. 181-186)
119                                                                                                                      1
UUCUUUAAAUGCUUCGUAUUUCUUUUAACAUUCUUUGCAUAGAGGCUAGUGAGCGCUACUUUCAUUGCAAAGAUGUUACAGGCACUAAUUUGAAGGUUUUUAAAACACUACUUGUUUCU
NSP2 (nts. 1761-1879)
1761                                                                                                                  1879
UUCUUGAAAUGCUUCGUAUUUCUUUUAACAUUCUUUGCAUAGAGGCUAGUGAGCGCUACUUUCAUUGCAAAGAUGUUACAGGCACUAAUUUGAAGGUUUUUAAAACACUACUUGUUUCU
  L  E  M  L  R  I  S  F  N  I  L  C  I  E  A  S  E  R  Y  F  H  C  K  D  V  T  G  T  N  L  K  V  F  K  T  L  L  V  S
Lucheng Rn rat CoV isolate Ruian 83 (MT820626; China, 2014)
UUCUUGAAAUGCUUCGAGUUUCCUUUAACAUUCUUUGUGUUGAUGCUAGCGAACGUUACUUUUGCUGCAAGGAUGUCACUGGUACAAACCUCAAGCUUUUCAAGACUCUGCUUGUCUCU
  L  E  M  L  R  V  S  F  N  I  L  C  V  D  A  S  E  R  Y  F  C  C  K  D  V  T  G  T  N  L  K  L  F  K  T  L  L  V  S
Fievel mouse CoV strain FiCoV/UMN2020 (OK655840; USA, 2018)
UUCUAGAGGUACUCAGGAGAUCCUUUGGCAUGCUCUGCGUUGAUGCCACUGAAUGUUGUUUCAAGUGCAGGGAUGUCACUGGUACUAACCUGAAACUCUUUAAGCGCCAGAUCGUUGCA
  L  E  V  L  R  R  S  F  G  M  L  C  V  D  A  T  E  C  C  F  K  C  R  D  V  T  G  T  N  L  K  L  F  K  R  Q  I  V  A


AcCoV-JC34 (KX964649; nts. 1880-1940)
1880                                                       1940
UGGUGCGAGGCUGCUAUUACAGGUGUUAAAGAAGCCGGUUUGACAACUGCUAAAUACUUUG
W  C  E  A  A  I  T  G  V  K  E  A  G  L  T  T  A  K  Y  F
Lucheng Rn rat CoV isolate Ruian 83 (MT820626)
UGGUAUGAGAGUGUCAUAAACGGUGCUAAGGAGGCUGGUUUAACUACUGCAAAGUACUUUA
W  Y  E  S  V  I  N  G  A  K  E  A  G  L  T  T  A  K  Y  F
Fievel mouse CoV strain FiCoV/UMN2020 (OK655840)
UCUUUUGAGUGCGCCAUCUCUGGUGUUAAGGAAGCUGGUCUUACCACCGCCAAGUACUUUU
S  F  E  C  A  I  S  G  V  K  E  A  G  L  T  T  A  K  Y  F



Nucleocapsid amino-terminus: Bat alphacoronaviruses subgenus Nyctacovirus

Alphacoronavirus sp. Isolate WA2028 (MK472068; microbat; Australia, 2018)
5’-UTR ([-]strand corresponding to nts. 8-122 in 434-long [+]strand; TRS: nts. 206-211)
122                                                                                                                  8
UCUAAACUAAACAAUGUCUGUUAAUUUUGCAGGCACUACUACUCAGCCACGUGGCCGUGUCCCUCUUUCGCUGUUCCAACCACUGCGAAACAACUCUACUCAACCUCUCCACAAG
Nts.25491-25605 including segment before N (with TRS-B) and amino-terminus of N (encoded by nts. 25454-26716)
25491                                                                                                            25605
UCUAAACUAAACAAUGUCUGUUAAUUUUGCAGGCACUACUACUCAGCCACGUGGCCGUGUCCCUCUUUCGCUGUUCCAACCACUGCGAAACAACUCUACUCAACCUCUCCACAAG
  *         M  S  V  N  F  A  G  T  T  T  Q  P  R  G  R  V  P  L  S  L  F  Q  P  L  R  N  N  S  T  Q  P  L  H  K

Alphacoronavirus sp. Isolate WA3301 (MK472069; microbat; Australia, 2018)
UCUAAACUAAACAAUGUCUGUUAAUUUUGCAGGCACUACUACUCAGCCACGUGGCCGUGUUCCUCUUUCGCUGUUCCAACCACUGCGAAACAACUCUACUCAGCCUCUCCACAAA
  *         M  S  V  N  F  A  G  T  T  T  Q  P  R  G  R  V  P  L  S  L  F  Q  P  L  R  N  N  S  T  Q  P  L  H  K

Alphacoronavirus sp. Isolate WAAlc1 (MK472071; microbat; Australia 2018)
UCUAAACUAAACAAUGUCUGUGAAUUUCUCAGACACUACUGCUCAGCCUCGUGGCCGUGUCCCUCUUUCGCUUUUUCAACCGUUGCGAAACAAUUCUUCCCAGCCGAAAAUAUUC
                                 CAA      C    AC
  *         M  S  V  N  F  S  D Q T T  A S  A P R  G  R  V  P  L  S  L  F  Q  P  L  R  N  N  S  S  Q  P  K  I  F

Tylonycteris bat coronavirus HKU3 (MK720944; bat; China, 09-09-2015)
UCUAAACUAAACAAUGUCUGUCAAUUUUGCAGGCGGAAGUUUACUAAGGUCCCCGGCGAGACCCGCACACGGAAACAGUCACAAAGUCGGGUAUGCAUUAUUACGAAGUGUUGGA
  *         M  S  V  N  F  A  G  G  S  L  L  R  S  P  A  R  P  A  H  G  N  S  H  K  V  G  Y  A  L  L  R  S  V  G

Alphacoronavirus Bat-CoV/P.kuhlii/Italy/206645-41/2011 (MH938448; bat; Italy,2011)
UCUAAACGAAAUGUCUUCCUCCAGGGGAAACGUCGGUUUGACAAUGCAGCUAGGGUCGUUCAGGGCGUGUACCACUUUCACUAUAUAUGCCCGUUAUCAACAACUCACCUAAG
  *       M  S  S  S  R  G  N  V  G  F  D  N  A  A  R  G  R  S  G  R  V  P  L  S  L  Y  M  P  V  I  N  N  S  P  K



ORF5b carboxyl-terminus: Bat alphacoronavirus subgenus Decacovirus

Rhinolophus bat coronavirus HKU32 strain TLC28A (MK720946; Rhinolophus sinicus; Hong Kong, 08-06-2015)a
5’-UTR (nts. 45-77 in 290-long [+]strand; TRS: nts. 67-72)
       45                              77
       AGACCUCGCGUCUCAUCCCCUCAACUAAACGAA
ORF5b (nts. 25782-26225) carboxyl-terminus; overlaps with beginning of E (nts. 26209-26433)
       26116                           26148                                                                 26209          26225
ACUACACAGACCUCGCGUCUCAUCCCCUCAACUAAACGAACUCAUUAUAGUUUGUUUGUUGCAUUUACUAUAGUACCUGUAGUUUUGGUCUUCAUACACUAUGCUGACUUUAGUUAAU
T  T  Q  T  S  R  L  I  P  S  T  K  R  T  H  Y  S  L  F  V  A  F  T  I  V  P  V  V  L  V  F  I  H  Y  A  D  F  S  *
                                                                                                    M  L  T  L  V  N

ORF4b carboxyl-terminus: Bat alphacoronavirus subgenus Decacovirus
(annotated as unclassified coronavirus but 5’-UTR similar to that of isolate above)

Bat coronavirus isolate BtCoV/Rh/YN2012_Rs4259 (MG916903; China, 04-17-2013)b
5’-UTR (nts. 61-90 in 297-long [+]strand; TRS: nts. 74-79)
                61                          90
                GUCUCAUCCCCUCAACUAAACGAAAUUUUU
ORF4b (nts.25508-25954) carboxyl-terminus; overlaps with beginning of E (25935-26159)
                25848                       25877                                                        25935           25954
ACUAACCCUAGUAAAAGUCUCAUCCCCUCAACUAAACGACAUUAUUACUUCCUGCUACUUGCAUUUAUUCCACCUGCGUGGUUUGCCGUAUUAAUUAUCUACUAUGUUAACCCUAGUGGAU
T  N  P  S  K  S  L  I  P  S  T  K  R  H  Y  Y  F  L  L  L  A  F  I  P  P  A  W  F  A  V  L  I  I  Y  Y  V  N  P  S  G  *
                                                                                                         M  L  T  L  V  N
```

Fig. 8 (See legend on previous page.)

of reference Wuhan strain), which is phosphorylated by GSK-3, is conserved in the predicted translated 5′-UTR sequence next to the R203K/G204R co-mutation, as is threonine 205, which is phosphorylated by PKA [98, 99]. R203 and G204 mutations affect the phosphorylation of serines 202 and 206 in turn affecting binding to protein 14-3-3 and replication, transcription, and packaging of the SARS-CoV-2 genome [100–102].

The *N* gene displays rapid and high expression, high sequence conservation, and a low propensity for recombination [34, 103, 104]. However, it can show variation driven by internal rearrangement which does not affect the length of the protein. The N protein is highly immunogenic, and its amino acid sequence is largely conserved, with the serine-arginine (SR) region being a strong immunodominant B-cell epitope [105] as highlighted in Fig. 3A.

The functional significance of the intragenomic rearrangement in *N* of bat α-CoVs subgenus *Nyctacovirus* remains to be determined. Although in infectious bronchitis virus, the amino terminal domain of N protein has been shown to interact with nucleotide sequences in the 3′-UTR which is relevant to viral RNA packaging, the amino acids that are critical for such interaction are more distally located in the amino terminus (amino acids 76 or 94) [106, 107] than those encoded by the intragenomic rearrangement in this case.

The intragenomic rearrangements found in MERS-CoV may modulate immune evasion by bringing regulatory sequences to the intergenomic regions preceding the *4a* and *5* genes and modulating their expression. p4a, a double stranded RNA-binding protein, as well as p4b and p5 of MERS-CoV are type-I IFN antagonists [108–111]. p4a prevents dsRNA formed during viral replication from binding to the cellular dsRNA-binding protein PACT and activating the cellular dsRNA sensors RIG-I and MDA5 [110, 111]. p4a is the strongest in counteracting the antiviral effects of IFN via inhibition of both its production and Interferon-Stimulated Response Element (ISRE) promoter element signaling pathways [112]. The latter findings were obtained in cell cultures and studies in an in vivo infection are warranted. To this end, a more recent study associated p4b with inflammatory pathology and suppression of autophagy in murine lungs thereby highlighting the complex interplay of proteins during virus replication under in vivo physiological conditions [113].

Like SARS-CoVs and MERS-CoV, hCoV-OC43 can downregulate the transcription of genes critical for the activation of different antiviral signaling pathways [114], and the intragenomic rearrangements described in the intergenic region preceding hCov-OC43 *ns5a* may modulate immune evasion. To this end, hCoV-OC43 ns5a, as well as ns2a, M, or N proteins significantly reduced the transcriptional activity of ISRE, IFN-β promoter, and NF-κB-RE following challenge of human embryonic kidney 293 (HEK-293) cells with Sendai virus, IFN-α or tumor necrosis factor-α [115].

In hCoV-HKU-1 and hCoV-OC43, intragenomic rearrangements involved the intergenic region at the end of the *S* gene highlighting a potential source of regulatory sequences that may affect expression of adjoining genes. The Spike (*S*) gene encodes a structural protein that binds to the host receptors and determines cell tropism as well as the host range. The neighborhood of the spike gene, particularly the region before the S gene, is a hotspot for modular intertypic homologous and non-homologous recombination in coronavirus genomes [34].

Although the nsp3 protein sequence is well conserved among bat *Nobecoviruses*, the significance of the nsp3 segment encoded by the 5′-UTR sequence, which might affect double vesicle membrane formation, remains to be determined. Nsp3 protein, the largest protein encoded by CoVs encompasses up to 16 modular domains. The N-terminal cytosolic domains include a mono-ADP-ribosylhydrolase, a papain-like protease [116], and a scaffold region that participates in replication-transcription complex assembly [117]. After the latter domains, there are two transmembrane domains (TM1 and TM2) with an endoplasmic reticulum luminal loop (Ecto3) between them, and two cytosolic domains (Y1 and CoV-Y) following TM2. The predicted nsp3 segment encoded by the 5′-UTR sequence falls in the cytosolic domain Y1. Nsp3C anchors nsp3 to the endoplasmic reticulum membrane and induces membrane rearrangement leading to double membrane vesicle formation via a yet unknown molecular mechanism [118, 119]. Although there are structural data on the CoV-Y domain [120], its function is unknown as is that of the Y1 domain.

The discontinuous RNA synthesis of the polymerase machinery of coronaviruses along with the use of canonical and noncanonical TRS-L and TRS-B pairing may enhance the occurrence of insertions (via intragenomic rearrangements or other means) and deletions, which can remain uncorrected by the proofreading activity of nsp14 exoribonuclease [121]. Most insertion and deletions likely negatively affect viral fitness [122] and duplication of TRS sequences in coronaviruses led to attenuation [123] and when affecting essential genes frequently to viral genetic instability [124]. However, a small number of insertions/deletions emerge and spread in viral populations, suggesting a positive effect on fitness and adaptive evolution [125–131]. Thus, analyzing these insertion/deletions may reveal evolutionary trends and provide new insight into the surprising variability and rapidly spreading capability that SARS-CoV-2 has shown since its emergence. One usual target of deletions is the accessory ORFs in the distal third of the genome, because they do not appear to participate in viral replication but can allow the virus to evade host defenses. Variants with these deletions occur naturally in SARS-CoV-2 and spread without apparently affecting virus infectivity.

Some of the intragenomic rearrangements described here in *ORF8* and *ORF7a* and one previously in *ORF6* occurred in viruses with deletions that removed or truncated ORFs, such as the deletion in the B.1.36.27 lineage from Hong Kong which lacks *ORFs 7a*, *7b*, and *8* and has the last 12 nucleotides of the *ORF6* replaced by ~ 60 nucleotides from the 5′-UTR [39]. An 872-nucleotide deletion described in the AY.4 lineage (Delta variant) from Southern Poland also eliminated *ORFs 7a*, *7b* and *8* [132], as did a 872-nucleotide deletion documented in late 2021 in Uruguay in a different Delta lineage (AY.20), with viruses without the deletion coexisting with wild-type AY.20 and AY.43 strains [128, 129].

Two large and phylogenetically unrelated deletions (392 and 227 nucleotides long) fused *ORF7a* with downstream ORFs [133]. One, a 392-nucleotide deletion, lacked *ORF7b* and created a new ORF including *ORF7a* and *ORF8*, while the other, a 227-nucleotide deletion, resulted in a new ORF by combining the proximal end of *ORF7a* with *ORF7b*. These deletions have become extinct or appear as sporadic or unique variants [39, 133]. On the other hand, a 382-nucleotide deletion that removes most of the ORF8 was a circulating form hypothesized to lead to an attenuated phenotype of SARS-CoV-2 [130, 131].

Intragenomic rearrangements in isolates with large deletions, as exemplified by those involving *ORF6* [39], *ORF7b* and *ORF8* of SARS-CoV-2, in all cases thus far affect the carboxyl-termini of the predicted encoded proteins. The length of the insertions does not notably affect that of the predicted proteins in isolates without major genomic deletions. For 5′-UTR segments within viral genes, such as the examples shown in *N*, *nsp12* and *nsp3*, or intergenic regions, the length of the protein or intergenic region appears not to be affected.

Intragenomic rearrangements are yet another example of the tremendous genomic flexibility of coronaviruses which underlies changes in transmissibility, immune escape and/or virulence documented during the SARS-CoV-2 pandemic.

Limitations

The intragenomic rearrangements involving 5′-UTR sequences were detected in all subgenera of β-coronaviruses infecting humans (i.e., *Sarbecovirus*, *Embecovirus*, and *Merbecovirus*) and in the *Nobecovirus* but not the *Hibecovirus* subgenera of CoVs infecting bats. There were only 3 *Hibecovirus* genomes in the database, which may account for the lack of detection of internal rearrangements in this subgenus most closely related to *Sarbecoviruses*. In this respect, the most diverse detection of rearrangements in SARS-CoV-2 may reflect the bias generated by the presence in GenBank of SARS-CoV-2 isolates in up to 5 orders of magnitude greater number than any other CoV. However, the relative paucity of α-, γ-, or δ-CoV sequences available also applies to those of β-CoVs other than SARS-CoV-2 for which 5′-UTR rearrangements were found in notable proportions. Moreover, the present analysis included CoVs involved in large outbreaks such as the swine enteric CoVs of the α and δ genera and avian infectious bronchitis virus of the γ genus that have been studied over decades with hundreds of isolates characterized without apparent evidence for intragenomic rearrangements. The apparent absence of internal rearrangements in the latter viruses bodes well for the specificity of the findings described here for 4 of 5 subgenera of β-CoVs and 3 of 12 subgenera of α-CoVs.

Many sequences in the databases have incomplete 5′-UTRs rendering it difficult to comprehensively analyze them and to calculate more reliable proportions of variations. There are also partial genome and protein sequences, and we excluded sequences with undetermined amino acids. Nonetheless, for SARS-CoV-2, the frequency of variants with full-length insertions appears low relative to those with subsegments or other mutations in comparison to the reference strain in the same insertion area. One could posit that for hCoV-OC43 and hCoV-HKU-1, the apparently much higher frequency of intragenomic rearrangements involving 5′-UTR sequences might be driven by characterization of a greater number of isolates during epidemics with rearrangements possibly providing transmissibility, immune evasion and/or virulence advantages.

A limitation of the methods used for detecting these isolates is that they may not be viable, i.e., they may be associated with molecular diagnostic detection of virus but not necessarily culture conversion, or may represent artifacts of sequencing; however, their prevalence with redundancy in various locations and processing laboratories would be consistent with human-to-human transmission. Moreover, Turakhia et al. [134], among others, have pointed out that systematic errors associated with lab-or protocol-specific practices affect some sequences in the repositories, which are predominantly or exclusively from single labs, co-localize with commonly used primer binding sites and are more likely to affect the protein-coding sequences than other similarly recurrent mutations. Although we cannot rule out that such systematic errors as well as wrong short reads alignment may underlie some if not all the rearrangements detected, the possibility is rendered less likely by the geographic and temporal diversity of the isolates with each intragenomic rearrangement (as underscored by the data in the Additional file 1: legends to Figures and Table), their presence in diverse variants of concern, as well as the occurrence of rearrangements in sequences from before the pandemic era and among diverse viruses of two genera and various subgenera in at least three hosts (humans, bats, and rodents). Moreover, it is unlikely that the insertion in the nucleocapsid gene of SARS-CoV-2 which encodes for a common co-mutation of adjacent sites that has been shown experimentally to have functional significance reflects an artifactual event. Finally, when using peptides as query sequences for SARS-CoV-2 we verified that the nucleotide sequences encoding the detected peptides were identical to 5′-UTR sequences. However, we cannot rule out that the sequences detected in intragenomic rearrangements may have arisen from host cell genomes or other sources.

Conclusions

We here describe intragenomic rearrangements involving 5′-UTR sequences and the coding section of the genome of beta and alphacoronaviruses. Variation driven by internal rearrangements is distinct from the non-homologous recombination events proposed as origins of *Sarbecoviru s*/*Hibecovirus*/*Nobecovirus* β-CoV *ORF3a* by gene duplication followed by rapid divergence from M [34, 135] or of SARS-CoV-2 *ORF8* from *ORF7a* [79]. The mechanisms underlying intragenomic rearrangements warrant further study. Understanding the variation that they introduce also is of relevance in the design of prophylactic and therapeutic interventions for all coronaviruses, including a pan-betacoronavirus vaccine.

Abbreviations

CoV	Coronavirus
MERS	Middle east respiratory syndrome
N	Nucleocapsid
NIRAN	*Nidovirus* RNA-dependent RNA polymerase associated nucleotidyl transferase
Nts	Nucleotides
ORF	Open reading frame
SARS-CoV-2	Severe acute respiratory syndrome-coronavirus-2
SL	Stem-loop
SR	Serine-arginine
TRS	Transcription regulatory sequence
U/A	Uridine/Adenosine
UTR	Untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12985-023-01998-0.

Additional file 1. Supplementary figures and tables.

Acknowledgements

Not applicable.

Author contribution

Both authors contributed to all aspects of this work and manuscript preparation. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its Additional file.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 8 May 2022 Accepted: 21 February 2023
Published online: 25 February 2023

References

1. Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. Microbiol Mol Biol Rev. 2005;69(4):635–64.
2. ICTV Coronaviridae Study Group. International Committee on Taxonomy of Viruses (ICTV). 2021. Available from: https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/223/coronaviridae-figures.
3. Pollett S, Conte MA, Sanborn M, Jarman RG, Lidl GM, et al. A comparative recombination of analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic. Sci Rep. 2021;11:17365.
4. Jackson B, Boni MF, Bull MJ, et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell. 2021;184(20):5179–88.
5. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England). 2020;395(10223):497–506.
6. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727–33.
7. Menachery VD, Yount BL Jr, Debbink K, et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med. 2015;21(12):1508–13.
8. Caserta LC, Martins M, Butt SL, et al. White-tailed deer (*Odocoileus virginianus*) may serve as a wildlife reservoir for nearly extinct SARS-CoV-2 variants of concern. Proc Natl Acad Sci USA. 2023;120(6): e2215067120.
9. Song H-D, Tu C-C, Zhang G-W, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 2016;24:490–502.
10. Latinne A, Hu B, Olival KJ, et al. Origin and cross-species transmission of bat coronaviruses in China. Nat Commun. 2020;11(1):1–5.
11. Wong AC, Li X, Lau SK, Woo PC. Global epidemiology of bat coronaviruses. Viruses. 2019;11(2):174.
12. Woo PC, Lau SK, Huang Y, Yuen KY. Coronavirus diversity, phylogeny and interspecies jumping. Exp Biol Med. 2009;234(10):1117–27.
13. Amoutzias GD, Nikolaidis M, Tryfonopoulou E, et al. The remarkable evolutionary plasticity of coronaviruses by mutation and recombination: insights for the COVID-19 pandemic and the future evolutionary paths of SARS-CoV-2. Viruses. 2022;14:78.

14. Andersen KG, Rambaut A, Lipkin WI, et al. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):450–2.
15. Decaro N, Mari V, Campolo M, et al. Recombinant canine coronaviruses related to transmissible gastroenteritis virus of Swine and circulating in dogs. J Virol. 2009;83(3):1532–7.
16. Goldstein SA, Brown J, Pedersen BS, Quinlan AR, Elde NC. Extensive recombination-driven coronavirus diversification expands the pool of potential pandemic pathogens. Genome Biol Evol. 2022;14(12):evac161.
17. Gussow AB, Auslander N, Faure G, et al. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Nat Acad Sci USA. 2020;117(26):15193.
18. Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? Nat Rev Microbiol. 2011;9(8):617–26.
19. Thorne LG, Bouhaddou M, Reuschl AK, et al. Evolution of enhanced innate immune evasion by SARS-CoV-2. Nature. 2022;602(7897):487–95.
20. Sawicki SG, Sawicki DL, Siddell SG. A contemporary view of coronavirus transcription. J Virol. 2007;81:20–9.
21. Bobay L-M, O'Donnell AC, Ochman H. Recombination events are concentrated in the spike protein region of betacoronaviruses. PLoS Genet. 2020;16: e1009272.
22. Boni MF, Lemey P, Jiang X, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 2020;5:1408–17.
23. Forni D, Cagliani R, Clerici M, Sironi M. Molecular evolution of human coronavirus genomes. Trends Microbiol. 2017;25:35–48.
24. Forni D, Cagliani R, Sironi M. Recombination and positive selection differentially shaped the diversity of betacoronavirus subgenera. Viruses. 2020;12:1313.
25. Lau SKP, Wong EYM, Tsang CC, et al. Discovery and sequence analysis of four deltacoronaviruses from birds in the Middle East reveal interspecies jumping with recombination as a potential mechanism for avian-to-avian and avian-to-mammalian transmission. J Virol. 2018;92:e00265-e318.
26. Makino S, Keck JG, Stohlman SA, Lai MM. High-frequency RNA recombination of murine coronaviruses. J Virol. 1986;57:729–37.
27. Yang Y, Yan W, Hall AB, Jiang X. Characterizing transcriptional regulatory sequences in coronaviruses and their role in recombination. Mol Biol Evol. 2021;38:1241–8.
28. Wang D, Jiang A, Feng J, et al. The SARS-CoV-2 subgenome landscape and its novel regulatory features. Mol Cell. 2021;81:2135–47.
29. Bentley K, Keep SM, Armesto M, Britton P. Identification of a noncanonically transcribed subgenomic mrna of infectious bronchitis virus and other gammacoronaviruses. J Virol. 2013;87:2128–36.
30. Van Marle G, Luytjes W, Van der Most RG, et al. Regulation of coronavirus mRNA transcription. J Virol. 1995;69(12):7851–6.
31. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike. Mechanisms of coronavirus cross-species transmission. J Virol. 2010;84:3134–46.
32. Graham RL, Deing DJ, Deming ME, et al. Evaluation of a recombination-resistant coronavirus as a broadly applicable, rapidly implementable vaccine platform. Commun Biol. 2018;1(1):1–10.
33. Lytras S, Hughes J, Martin D, et al. Exploring the natural origins of SARS-CoV-2 in the light of recombination. Genome Biol Evol. 2022;5:evac018.
34. Nikolaidis M, Markoulatos P, van de Peer Y, et al. The neighborhood of the spike gene is a hotspot for modular intertypic homologous and non-homologous recombination in coronavirus genomes. Mol Biol Evol. 2022. https://doi.org/10.1093/molbev/msab292.
35. Madhugiri R, Karl N, Petersen D, et al. Structural and functional conservation of cis-acting RNA elements in coronavirus 5′-terminal genome regions. Virology. 2018;517:44–55.
36. Miao Z, Tidu A, Eriani G, Martin F. Secondary structure of the SARS-CoV-2 5′-UTR. RNA Biol. 2021;18(4):447–56.
37. Zhang X, Liao C-L, Lai M. Coronavirus leader RNA regulates and initiates subgenomic mRNA transcription both in trans and in cis. J Virol. 1994;8(8):4738–46.
38. Chen SC, Olsthoorn RCL. Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. Virology. 2010;401(1):29–41.
39. Tse H, Lung DC, Wong SC, et al. Emergence of a severe acute respiratory syndrome coronavirus 2 virus variant with novel genomic architecture in Hong Kong. Clin Infect Dis. 2021;73(9):1696–9.
40. Wille M, Holmes EC. Wild birds as reservoirs for diverse and abundant gamma- and deltacoronaviruses. FEMS Microbiol Rev. 2020;44(5):631–44.
41. Islam MR, Hoque MN, Rahman MS, et al. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci Rep. 2020;10(1):14004.
42. Hassan SS, Choudhury PP, Dayhoff GW 2nd, et al. The importance of accessory protein variants in the pathogenicity of SARS-CoV-2. Arch Biochem Biophys. 2022;717: 109124.
43. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
44. Bikandi J, San Millán R, Rementeria A, Garaizar J. In silico analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction. Bioinformatics. 2004;20:798–9.
45. Duvaud S, Gabella C, Lisacek F, et al. Expasy, the swiss bioinformatics resource portal, as designed by its user. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gks225.
46. Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36:W5-9. https://doi.org/10.1093/nar/gkn201.
47. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Chall. 2017;1:33–46.
48. Khare S, et al. GISAID's role in pandemic response. China CDC Weekly. 2021;3(49):1049–51.
49. Shu Y, McCauley J. GISAID: from vision to reality. EuroSurveillance. 2017. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
50. Tsoleridis T, Chappell JG, Onianwa O, et al. Shared common ancestry of rodent alphacoronaviruses sampled globally. Viruses. 2019;11(2):125.
51. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015;31:3377–9.
52. Gruber AR, Lorenz R, Bernhart SH, et al. The vienna RNA websuite. Nucleic Acids Res. 2008;36:W70–4. https://doi.org/10.1093/nar/gkn188.
53. Lorenz R, Bernhart SH, Höner zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26.
54. Temmam S, Vongphayloth K, Baquero Salazar E, et al. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature. 2022;604(7905):330–6.
55. Crook JM, Murphy I, Carter DP, et al. Metagenomic identification of a new sarbecovirus from horseshoe bats in Europe. Sci Rep. 2021;11:14723.
56. Ujike M, Taguchi F. Recent progress in torovirus molecular biology. Viruses. 2021;13(3):435.
57. Harrison GP, Mayo MS, Hunter E, Lever AM. Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5′ and 3′ of the catalytic site. Nucleic Acids Res. 1998;26(14):3433–42.
58. Franco-Muñoz C, Álvarez-Díaz DA, Laiton-Donato K, et al. Substitutions in spike and nucleocapsid proteins of SARS-CoV-2 circulating in South America. Infect Genet Evol. 2020;85: 104557.
59. Johnson BA, Zhou Y, Lokugamage KG, et al. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. PLoS Pathog. 2022;18(6): e1010627.
60. Mourier T, Shuaib M, Hala S, et al. SARS-CoV-2 genomes from Saudi Arabia implicate nucleocapsid mutations in host response and increased viral load. Nat Commun. 2022;13:601.
61. Wu H, Xing N, Meng K, et al. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe. 2021;29:1788–801.
62. Hartenian F, Nandakumar D, Lari A, et al. The molecular virology of coronaviruses. J Biol Chem. 2020;295:12910–34.
63. Lauber C, Goeman JJ, Parquet MDC, et al. The footprint of genome architecture in the largest genome expansion in RNA viruses. PLoS Pathogen. 2013;9: e1003500.
64. Gorbalenya AE, Baker SC, Baric RS, et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-NCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5:536–44.

65. Slanina H, Madhugiri R, Bylapudi G, et al. Coronavirus replication-transcription complex: vital and selective NMPylation of a conserved site in nsp9 by the NiRAN-RdRp subunit. Proc Natl Acad Sci USA. 2021;118(6): e2022310118.
66. Yan L, Ge J, Zheng L, Zhang Y, et al. Cryo-EM structure of an extended SARS-CoV-2 replication and transcription complex reveals an intermediate state in cap synthesis. Cell. 2021;184(1):184-193.e10.
67. Dwivedy A, Mariadasse R, Ahmad M, et al. Characterization of the NiRAN domain from RNA-dependent RNA polymerase provides insights into a potential therapeutic target against SARS-CoV-2. PLoS Comput Biol. 2021;17(9): e1009384.
68. Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, et al. Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. Nucleic Acids Res. 2015;43(17):8416–34.
69. Park GJ, Osinski A, Hernandez G, et al. The mechanism of RNA capping by SARS-CoV-2. Nature. 2022;609(7928):793–800.
70. Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and sources of endemic human coronaviruses. Adv Virus Res. 2018;100:163–88.
71. Vijgen L, Keyaerts E, Moës E, et al. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J Virol. 2005;79:1595–604.
72. Mounir S, Talbot PJ. Molecular characterization of the S protein gene of human coronavirus OC43. J Gen Virol. 1993;74:1981–7.
73. Wang L, Qiao X, Zhang S, et al. Porcine transmissible gastroenteritis virus nonstructural protein 2 contributes to inflammation via NF-κB activation. Virulence. 2018;9(1):1685–98.
74. Ziv O, Gabryelska MM, Lun ATL, et al. COMRADES determines in vivo RNA structures and interactions. Nat Methods. 2018;15(10):785–8.
75. Sola I, Almazán F, Zúñiga S, Enjuanes L. Continuous and discontinuous RNA synthesis in coronaviruses. Ann Rev Virol. 2015;2(1):265–88.
76. Nomburg J, Meyerson M, De Caprio JA. Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. Genome Med. 2020;12:108.
77. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17:181–92.
78. Rottier PJM, Nakamura K, Schellen P, Volders H, Hajema BJ. Acquisition of macrophage tropism during the pathogenesis of feline infectious peritonitis is determined by mutations in the feline coronavirus spike protein. J Virol. 2005;79:14122–30.
79. Neches RY, Kyrpides NC, Ouzounis CA. Atypical divergence of SARS-CoV-2 Orf8 from Orf7a within the coronavirus lineage suggests potential stealthy viral strategies in immune evasion. MBio. 2021;12(1):e03014-e3020.
80. Flower TG, Buffalo CZ, Hooy RM, et al. Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. Proc Natl Acad Sci USA. 2021;118(2): e2021785118.
81. Redondo N, Zaldívar-López S, Garrido JJ, Montoya M. SARS-CoV-2 accessory proteins in viral pathogenesis: knowns and unknowns. Front Immunol. 2021;12: 708264.
82. Zhang Y, Chen Y, Li Y, et al. The ORF8 protein of SARS-CoV-2 mediates immune evasion through down-regulating MHC-**I**. Proc Natl Acad Sci USA. 2021;118(23): e2024202118.
83. Li JY, Liao CH, Wang Q, et al. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. Virus Res. 2020;286: 198074.
84. Valcarcel A, Bensussen A, Álvarez-Buylla ER, Díaz J. Structural analysis of SARS-CoV-2 ORF8 protein: pathogenic and therapeutic implications. Front Genet. 2021;12: 693227.
85. Stukalov A, Girault V, Grass V, et al. Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. Nature. 2021;594(7862):246–52.
86. Lin X, Fu B, Yin S, et al. ORF8 contributes to cytokine storm during SARS-CoV-2 infection by activating IL-17 pathway. iScience. 2021;24(4):102293.
87. Gordon DE, Hiatt J, Bouhaddou M, et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. Science. 2020;370(6521):eabe9403.
88. Grifoni A, Weiskopf D, Ramirez SI, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell. 2020;181(7):1489-1501.e15.
89. Wang X, Lam JY, Wong WM, et al. Accurate diagnosis of COVID-19 by a novel immunogenic secreted SARS-CoV-2 orf8 protein. MBio. 2020;11(5):e02431-e2520.
90. Yang R, Zhao Q, Rao J, et al. SARS-CoV-2 accessory protein ORF7b mediates tumor necrosis factor-α-induced apoptosis in cells. Front Microbiol. 2021;12: 654709.
91. Khavinson V, Terekhov A, Kormilets D, Maryanovich A. Homology between SARS CoV-2 and human proteins. Sci Rep. 2021;11:17199.
92. He R, Leeson A, Ballantine M, et al. Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. Virus Res. 2004;105(2):121–5.
93. Lu S, Ye Q, Singh D, Cao Y, et al. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. Nat Commun. 2021;12(1):502.
94. Yao H, Song Y, Chen Y, et al. Molecular architecture of the SARS-CoV-2 virus. Cell. 2020;183(3):730-738.e13.
95. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014;6(8):2991–3018.
96. Lo C-Y, Tsai T-L, Lin C-N, et al. Interaction of coronavirus nucleocapsid protein with the 5′- and 3′-ends of the coronavirus genome is involved in genome circularization and negative strand RNA synthesis. FEBS J. 2019;2019(286):3222–39.
97. Carlson CR, Asfaha JB, Ghent CM, et al. Phosphoregulation of phase separation by the SARS-CoV-2 N protein suggests a biophysical basis for its dual functions. Mol Cell. 2020;80(6):1092–103.
98. Kemp BE, Graves DJ, Benjamini E, Krebs EG. Role of multiple basic residues in determining the substrate specificity of cyclic AMP-dependent protein kinase. J Biol Chem. 1977;252(14):4888–94.
99. Kennelly PJ, Krebs EG. Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. J Biol Chem. 1991;266(24):15555–8.
100. Surjit M, Kumar R, Mishra RN, et al. The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. J Virol. 2005;79(17):11476–86.
101. Tugaeva KV, Hawkins DEDP, Smith JLR, et al. The mechanism of SARS-CoV-2 nucleocapsid protein recognition by the human 14-3-3 proteins. J Mol Biol. 2021;433(8): 166875.
102. Tung HYL, Limtung P. Mutations in the phosphorylation sites of SARS-CoV-2 encoded nucleocapsid protein and structure model of sequestration by protein 14-3-3. Biochem Biophys Res Comm. 2020;532:134–8.
103. Dutta NK, Mazumdar K, Gordy JT. The nucleocapsid protein of SARS–CoV-2: a target for vaccine development. J Virol. 2020;94(13):e00647-e720.
104. Jaroszewski L, Iyer M, et al. The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. PLoS Comput Biol. 2021;17(7): e1009147.
105. Oliveira SC, de Magalhães MTQ, Homan EJ. Immunoinformatic analysis of SARS-CoV-2 nucleocapsid protein and identification of COVID-19 vaccine targets. Front Immunol. 2020;11: 587615.
106. Tan YW, Fang S, Fan H, Lescar J, Liu DX. Amino acid residues critical for RNA-binding in the N-terminal domain of the nucleocapsid protein are essential determinants for the infectivity of coronavirus in cultured cells. Nucleic Acids Res. 2006;34(17):4816–25.
107. Zhou M, Collisson EW. The amino and carboxyl domains of the infectious bronchitis virus nucleocapsid protein interact with 3′ genomic RNA. Virus Res. 2000;67(1):31–9.
108. Liu DX, Fung TS, Chong KK, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 2014;109:97–109.
109. Matthews KL, Coleman CM, van der Meer Y, Snijder EJ, Frieman MB. The ORF4b-encoded accessory proteins of middle east respiratory syndrome coronavirus and two related bat coronaviruses localize to the nucleus and inhibit innate immune signalling. J Gen Virol. 2014;95(Pt 4):874.
110. Niemeyer D, Zillinger T, Muth D, et al. Middle East respiratory syndrome coronavirus accessory protein 4a is a type I interferon antagonist. J Virol. 2013;87(22):12489–95.

111. Siu KL, Yeung ML, Kok KH, et al. Middle east respiratory syndrome coronavirus 4a protein is a double-stranded RNA-binding protein that suppresses PACT-induced activation of RIG-I and MDA5 in the innate antiviral response. J Virol. 2014;88(9):4866–76.
112. Yang Y, Zhang L, Geng H, et al. The structural and accessory proteins M, ORF 4a, ORF 4b, and ORF 5 of middle east respiratory syndrome coronavirus (MERS-CoV) are potent interferon antagonists. Protein Cell. 2013;4(12):951–61.
113. Bello-Perez M, Hurtado-Tamayo J, Requena-Platek R, et al. MERS-CoV ORF4b is a virulence factor involved in the inflammatory pathology induced in the lungs of mice. PLoS Pathog. 2022;18(9): e1010834.
114. Beidas M, Chehadeh W. Effect of human coronavirus OC43 structural and accessory proteins on the transcriptional activation of antiviral response elements. Intervirology. 2018;61(1):30–5.
115. Beidas M, Chehadeh W. PCR array profiling of antiviral genes in human embryonic kidney cells expressing human coronavirus OC43 structural and accessory proteins. Arch Virol. 2018;163:2065–72.
116. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. Antiviral Res. 2018;149:58–74.
117. Imbert I, Snijder EJ, Dimitrova M, et al. The SARS-coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. Virus Res. 2008;133(2):136–48.
118. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. MBio. 2013;4(4):e00524-e613.
119. Hagemeijer MC, Monastyrska I, Griffith J, et al. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. Virology. 2014;458:125–35.
120. Pustovalova Y, Gorbatyuk O, Li Y, et al. Backbone and Ile, Leu, Val methyl group resonance assignment of CoV-Y domain of SARS-CoV-2 non-structural protein 3. Biomol NMR Assign. 2021;18:1–6.
121. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020;92(4):418–23.
122. Grubaugh ND, Petrone ME, Holmes EC. We shouldn't worry when a virus mutates during disease outbreaks. Nat Microbiol. 2020;5(4):529–30.
123. Ortego J, Sola I, Almazan F, et al. Transmissible gastroenteritis coronavirus gene 7 is not essential but influences in vivo virus replication and virulence. Virology. 2003;308(1):13–22.
124. Pascual-Iglesias A, Sanchez CM, Penzes Z, et al. Recombinant chimeric transmissible gastroenteritis virus (TGEV)—porcine epidemic diarrhea virus (PEDV) virus provides protection against virulent PEDV. Viruses. 2019;11(8):682.
125. Meng B, Kemp SA, Papa G, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Rep. 2021;35(13):109292.
126. Lau SY, Wang P, Mok BW, et al. Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. Emerg Microbes Infect. 2020;9(1):837–42.
127. McCarthy KR, Rennick LJ, Nambulli S, et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science. 2021;371(6534):1139–42.
128. Panzera Y, Ramos N, Calleros L, et al. Transmission cluster of COVID-19 cases from Uruguay: emergence and spreading of a novel SARS-CoV-2 ORF6 deletion. Mem Inst Oswaldo Cruz. 2022;116: e210275.
129. Panzera Y, Cortinas MN, Marandino A, et al. Emergence and spreading of the largest SARS-CoV-2 deletion in the Delta AY.20 lineage from Uruguay. Gene Rep. 2022;29:101703.
130. Su YCF, Anderson DE, Young BE, et al. Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. MBio. 2020;11(4):e01610-e1620.
131. Young BE, Fong SW, Chan YH, et al. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet. 2020;396(10251):603–11.
132. Mazur-Panasiuk N, Rabalski L, Gromowski T, et al. Expansion of a SARS-CoV-2 delta variant with an 872 nt deletion encompassing *ORF7a*, *ORF7b* and *ORF8*, Poland, July to August 2021. Euro Surveill. 2021;26(39):2100902.
133. Addetia A, Xie H, Roychoudhury P, et al. Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. J Clin Virol. 2020;129: 104523.
134. Turakhia Y, De Maio N, Thornlow B, et al. Stability of SARS-CoV-2 phylogenies. PLoS Genet. 2020;16(11): e1009175.
135. Ouzounis CA. A recent origin of Orf3a from M protein across the coronavirus lineage arising by sharp divergence. Comput Struct Biotechnol J. 2020;18:4093–102.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.