Intragenomic rearrangements of SARS-CoV-2 and other β-coronaviruses

Roberto Patarca, MD, PhD, and William A. Haseltine, PhD

ACCESS Health International, 384 West Lane, Ridgefield, Connecticut 06877, USA

Corresponding author: William A. Haseltine, PhD; e-mail: william.haseltine@accessh.org

Abstract

The continuation of the SARS-CoV-2 pandemic depends on the generation of new viral variants. 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 β-coronaviruses, including SARS-CoV-2. Specifically, we report numerous genomic insertions of 5'-untranslated region sequences into coding regions of SARS-CoV-2 and other β-coronaviruses. To our knowledge this is the first systematic description of such insertions. In many cases, these insertions 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 β-coronaviruses, for instance, from one-third to one-half of surveyed sequences in publicly available databases contain 5'-UTR-derived inserted sequences. In limited instances, there is mounting evidence that these insertions alter the fundamental biological properties of mutant viruses. Intragenomic rearrangements add to our appreciation of how SARS-CoV-2 variants may arise.

Introduction

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], and five subgenera of β-CoVs: Sarbeco-, Merbeco-, Embeco-, Nobeco- and Hibecovirus (Weiss and Navas-Martin 2005). Seven CoVs infect humans; two of the α-genus (hCoVs 229E & 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 (Pollett et al. 2021; Jackson et al. 2021; VanInsberghe et al. 2021; Turkahia et al. 2021); 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 (Menachery et al. 2015) and possibly intermediate hosts (Fan et al. 2019; Pickering et al. 2021; Reusken et al. 2013; Song et al. 2005). Bat β-CoVs related to human CoVs belong to the Sarbeco-, Nobeco-, and Hibecovirus subgenera (Latinne et al. 2020; Wong et al. 2019; Woo et al. 2007).

Coronaviruses display substantial genomic plasticity and resilience (Amoutzias et al. 2022; Andersen et al. 2020) via recombination, point mutations, deletions, and insertions, which are reported to drive variant emergence, host range, gene expression, transmissibility, immune escape, and virulence (Decaro et al. 2009; Goldstein et al. 2021; Gussow et al., 2020; Simon-Loriere et al. 2011; Throne et al. 2022). 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 (Sawicki et al. 2007) has been reported to account for the high frequency of recombination (Amoutzias et al. 2022; Bobay et al. 2020; Boni et al. 2020; Forni et al. 2017, 2020; Lau et al. 2018; Makino et al. 1986; Simon-Loriere et al. 2011; Su et al. 2016; Yang et al. 2021).

In template switching, a leader transcription regulatory sequence (TRS-L; ACGAAC core in β-CoVs) (Wang et al. 2021) 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 Figure 1) (Bentley et al. 2013; Sawicki et al. 2007; Sola et al. 2015; Van Marle et al. 1995). Template switching renders the neighborhood of TRS-Bs, especially that for the spike gene, a recombination hotspot during viral transcription (Bobay et al. 2020; Boni et al. 2020; Forni et al. 2020; Graham et al. 2010, 2018; Goldstein et al. 2021; Lytras et al. 2022; Nikolaidis et al. 2021; Pollett et al. 2021; Yang et al. 2021).

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 (Figure 1) (Madhugiri et al. 2018; Miao et al. 2021; Zhang et al. 1994). 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 (Figure 2), and although there is no evidence for the functionality of any open reading frame within the UTRs (Chen et al., 2010; Miao et al., 2021), the 5'-leader sequence is 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 (Tse et al. 2021). We (Patarca and Haseltine 2021) also reported 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 USA variant generating a modified ORF-8 protein.

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 modify the carboxyl-termini of ORF-8 (also in *Rhinolophus* bat Sarbecovirus β-CoVs) and ORF7b; 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 carboxyl-terminus 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 Y1 cytoplasmic tail domain of nsp3 of Nobecoviruses of African *Rousettus* and *Eidolon* bats. No rearrangements involving 5'-UTR sequences were detected for the β-CoV SARS-CoV-1 and the α-CoVs hCoV-229E and hCoV-NL63 infecting humans, or for other α-CoVs or CoVs of the γ and δ subgenera.

The present study highlights an intragenomic source of variation involving duplication and translocation of 5'-UTR sequences to the body of the genome with implications on gene expression and immune escape of β-CoVs in humans and bats causing mild-to moderate or severe disease in endemic, epidemic and pandemic settings. Genome-wide annotations had revealed 1,516 nucleotide-level variations at different positions throughout the entire SARS-CoV-2 genome (Islam et al. 2020) 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 (Hassan et al. 2022). 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 differing pathogenic profiles.

Results

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

Intragenomic rearrangements alter the carboxyl termini of ORF8 and ORF7b (Sarbecoviruses)

We had reported on 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 an ORF8 protein with modified carboxyl-terminus encoded by the translocated 5'-UTR sequences (Patarca and Haseltine, 2021). Figure 3 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 ORF8 carboxyl-terminal sequence. As depicted in Supplementary Figure 3 legend, these internal rearrangements were detected in geographically and temporally 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 (Thorne et al. 2022), and all insertions alter the carboxyl-terminus of ORF8. The analysis also revealed that the insertions in some isolates had further changes involving point mutations, deletions, and insertions. Moreover, as shown in Figure 4A, a similar 5'-UTR-derived insertion at the carboxyl-terminus 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 (Temmam et al. 2021) all the way to England (Crook et al. 2021).

Crystal structure of ORF8 of SARS-CoV-2 revealed a ~60-residue core similar to that of SARS-CoV-2 ORF7a (from which ORF8 has been postulated to originate by non-homologous recombination) (Neches et al. 2021) with the addition of two dimerization interfaces, one covalent and the other noncovalent, unique to SARS-CoV-2 ORF8 (Flores et al. 2021). In the C-terminus of ORF8 that is altered by 5'-UTR-derived insertions (i.e., $_{115}$RVVLDFI$_{121}$), R115, D119, F120, and I121 contribute to the covalent dimer interface (marked with asterisks in Figure 3) with R115 and D119 forming salt bridges that flank a central hydrophobic core in which V117 interacts with its symmetry-related counterpart (Flores et al. 2021).

How the C-terminal insertions and changes therein affect the dimerization of ORF8 remains to be determined and described functions for ORF8 remain a matter of debate (Redondo et al. 2021). However, the changes caused by insertions may contribute to immune evasion by SARS-CoV-2 by affecting the interactions of ORF8 as a glycoprotein homodimer with intracellular transport signaling, leading to down-regulation of MHC-I by selective targeting for lysosomal degradation via autophagy (Zhang et al. 2021) and/or extracellular signaling (Matsuoka et al. 2022) involving interferon-I signaling (Li et al. 2020), mitogen-activated protein kinases growth pathways (Valcarcel et al. 2021), the tumor growth factor-β1 signaling cascade (Stukalov et al., 2021) and interleukin-17 signaling promoting inflammation and contributing to the COVID-19-associated cytokine storm (Lin et al., 2021).

The carboxyl-terminal region 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, and ORF8 accounts for 10% of CD8+ T cell reactivity in COVID-19 recovered subjects (Gordon et al. 2020; Griffoni et al. 2020). Another possible effect of the insertions stems from the fact that anti-ORF8 antibodies are detected in both symptomatic and asymptomatic patients early during infection by SARS-CoV-2 (Hachim et al. 2020; Wang et al. 2020) and diagnostic assays for SARS-CoV-2 infection that target only accessory genes or proteins such as ORF8 may be affected (Tse et al. 2021).

A shorter segment of the SARS-CoV-2 5'-UTR leader sequence (nts. 57 to 95, including TRS-L and SL3) than that described for ORF8 insertions was also duplicated and translocated to the end of ORF7b in two SARS-CoV-2 isolates (Figure 4B), one with a truncated ORF7b and the other with a truncated ORF8, which may have favored the internal rearrangements. The function of the SARS-CoV-2 ORF7b remains to be determined and has been suggested to mediate tumor necrosis factor-α-induced apoptosis based on cell culture data (Yang R et al. 2021) and theoretically in the dysfunction of olfactory receptors by triggering autoimmunity (Khavison et al. 2020).

Intragenomic rearrangements alter 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 to 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 its SR region, as exemplified by isolate QTO33828 (USA/Texas, Figure 5). The 5'-UTR-derived 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 Figure 5, the same 5'-UTR-derived sequence is present in N but without the leucine (L) residue and the phenylalanine (F) changed to serine (S), more closely approaching the Wuhan reference strain sequence.

In total, 37 SARS-CoV-2 isolates had 5'-UTR-derived sequences in their N gene; 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.1.17; first detected in Great Britain) from USA and Canada (Supplementary Figure 5 legend). 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 (Franco-Muñoz et al. 2020), 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 an analysis of association between COVID-19 severity and sample frequency of R203K/G204R co-mutations (Johnson et al. 2021; Wu et al. 2021).

The nucleocapsid is the most abundant protein in CoVs, interacts with membrane protein (He et al 2004; Lu et al. 2021), self-associates to provide for efficient viral assembly (Yao et al. 2020), binds viral RNA (McBride et al. 2014) 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 (Lo et al. 2019). 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 (Carlson et al. 2020; Lu et al. 2021).

The phosphorylation-rich stretch encompassing amino acid residues 180 to 210 (SR region) in which the 5'-UTR-derived sequences were found, 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 (reviewed in Carlson et al. 2020). Serine 202 (numbering of reference Wuhan strain), which is phosphorylated by GSK-3, is conserved in the 5'-UTR-derived sequence next to the R203K/G204R co-mutation, as is threonine 205, which is phosphorylated by PKA (Kemp et al. 1977; Kennelly et al. 1991). 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 (Surjit et al. 2005; Tugaeva et al. 2021; Tung and Limtung 2021).

The N gene displays rapid and high expression, high sequence conservation, and a low propensity for recombination (Dutta et al. 2020; Jaroszewski et al. 2021; Nikolaidis et al. 2021). 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 SR region being a strong immunodominant B-cell epitope (Oliveira et al. 2020) as highlighted in Figure 5.

Intragenomic rearrangements alter 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 (nucleotides 56 to 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: 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 (Figure 6). The latter strains have only one amino acid change outside of the insertions relative to the Wuhan reference strain. A subsegment of 5'-UTR (nucleotides 62 to 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 Supplementary Figure 6 legend) and exemplified by isolate UHP90975 [USA/Wisconsin, 2021-12-13] in Figure 6. 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-derived segments. 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) (Hartenian et al. 2020; Lauber et al. 2013), are highly conserved and have a low propensity for recombination among CoVs (Nikolaidis et al. 2021). 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 (Gorbalenya et al. 2020). However, as in other examples here of conserved genes, it is involved in intragenomic rearrangements of 5'-UTR-derived sequences.

The NiRAN domain of nsp12 is involved in the NMPylation of nsp9 (Slanina et al. 2021) during the formation of the replication-transcription complex (interface regions [Yan et al. 2021] are shown with yellow bars and key residues therein with ochre letters in Figure 6). The 5'-UTR-derived 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 (Dwivedy et al. 2021), is involved in protein-primed initiation of RNA synthesis (Lehmann et al. 2015) and catalyzes the formation of the cap core structure (GpppA; contact regions with GDP [Yan et al. 2021] indicated with blue boxes and key residues therein in ochre in Figure 6) (Park et al. 2022). The 5'-UTR-derived sequence at the proximal site in nsp12 NiRAn domain is close to the first contact region with GDP.

Intragenomic rearrangements in Merbecovirus, Embecovirus, and Nobecovirus subgenera of β-CoVs

As shown in Figure 7, 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 p3 and p4a in isolate MG923473 (Burkina Faso, 2015) and at the carboxyl-terminal end of 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 replaced by two amino acids (QL). The Q residue is encoded by a cytosine present in the reference sequence (indicated in orange color in Fig. 8) and two adenosines incorporated by the 5'-UTR-drived sequence.

p4a, a double stranded RNA-binding protein, as well as p4b and p5 of MERS-CoV are type-I IFN antagonists (Liu et al. 2014; Matthews et al. 2014; Niemeyer et al. 2913; Siu et al. 2014). 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 (Niemeyer et al. 2013; Siu et al. 2014). 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 (Yang et al. 2013). Therefore, the intragenomic rearrangements found in MERS-CoV may facilitate immune evasion by bringing regulatory sequences to the intergenomic regions preceding the 4a and 5 genes and facilitating their expression.

Out of 239 isolates of the β-CoV Embevovirus hCoV-OC43 in GenBank, 89 (~37%) had 5'-UTR-leader derived sequences (largest spanning nucleotides 34-78 of the hCoV-OC43 reference strain KJ958218) between the spike (S) and Nsp5a genes (Figure 8). The insertions did not affect the protein sequences of either S or Nsp5a; nucleotide changes relative to the 5'-UTR sequence are underlined in Figure 8. The hCoV-OC43 5'-UTR sequence inserted is identical to that of bovine coronavirus (BCoV) 5'-UTR except for one nucleotide (underlined adenosine [A] is a guanosine [G] in BCoV), which is consistent with a most probable bovine or swine coronavirus origin for hCoV-OC43 (Vijgen et al. 2005). The 5'-UTR-derived insertion sequence is also present in a molecularly characterized cloned hCoV-OC43 S protein gene (Mounir et al. 1993).

hCoV-OC43 ns5a, as well as ns2a, M, or N protein 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 (Beidas et al. 2018a). Like SARS-CoVs and MERS-CoV, hCoV-OC43 can downregulate the transcription of genes critical for the activation of different antiviral signaling pathways (Beidas et al., 2018b), and the intragenomic rearrangements described in the intergenic region preceding hCov-OC43 ns5a may facilitate immune evasion as was mentioned above for other immunomodulatory accessory proteins.

The β-CoV Embecovirus hCoV-HKU1 is a sister taxon to murine hepatitis virus and rat sialodacyoadenitis virus (Corman et al. 2018). Out of 51 HKU-1 isolates in GenBank, a 5'-UTR sequence including TRS-L, SL3 and most of SL2 (nucleotides 43-75 in hCoV-HKU-1 references NC_006577 and AY597011) is present in 29 isolates (~51%) between the S and Ns4 genes (Figure 9A).

The Spike (S) gene encodes a structural protein that binds to the host receptors and determines cell tropism as well as the host range. As mentioned in the Introduction, 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 (Nikolaidis 2021). In the cases described above for hCoV-OC43 and hCoV-HKU-1, 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.

An intragenomic rearrangement involving a 5'-UTR sequence (nucleotides 1-55) to 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 database: AWV67046), CMR705-P13 (MG693172, protein database: 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 Figure 9B. 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.

Nsp3 protein, the largest protein encoded by coronaviruses encompasses up to 16 modular domains. The N-terminal cytosolic domains include a mono-ADP-ribosylhydrolase (Alhammad et al. 2021), a papain-like protease (Lei et al. 2018), and a scaffold region that participates in replication-transcription complex assembly (Imbert et al., 2008). 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 nsp3 segment encoded by the 5'-UTR-derived 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 (Angelini et al. 2013; Hagemeijer et al. 2014). Although there are structural data on the CoV-Y domain (Pustovalova et al. 2021), its function is unknown as is that of the Y1 domain. Therefore, although the nsp3 sequence is well conserved among bat Nobecoviruses, the significance of the nsp3 segment encoded by the 5'-UTR-derived sequence, which might be involved in double vesicle membrane formation, remains to be determined.

Intragenomic rearrangements were not detected in some β-coronaviruses or in α-, γ-, and δ-CoVs

Using 5' UTRs from reference isolates (in parentheses) as query sequences, no 5'-UTR insertions were detected in the genome bodies of other coronaviruses infecting humans including the Sarbecovirus β-CoV SARS-CoV-1 (NC_004718) and the human α-CoVs hCoV-229E (MW532103 and KU291448) and hCoV-NL63 (NC_005831). In addition, no insertions were found in: α-CoVs subgenus Tegacovirus feline CoV and feline infectious peritonitis virus (FECV and FIPV; NC_002306), and transmissible gastroenteritis virus (TGEV; DQ811788), subgenus Rhinacovirus severe acute diarrhea syndrome CoV (MK651076), subgenus Pedavovirus porcine epidemic diarrhea virus (MK841495); and subgenus Tegacovirus transmissible gastroenteritis CoV (TGEV; DQ811788); β-CoVs subgenus Embecovirus murine hepatitis virus (MHV; NC_048217; AF208067), rat CoV Parker (NC_006213), rabbit CoV (JN874562), and bovine CoV (BCoV, U00735 and NC_003045; in this case except for sequences in related CoVs like hCoV-OC43 in Figure 8), and subgenus Hibecovirus Bat-Hp-betacoronavirus/ZHeijang 2013 (KF636752 and NC_025217) and Zaria bat CoV strain ZBCoV (HQ166910); δ-CoVs (all subgenus Buldecovirus) Porcine deltacoronavirus (USA/Ohio444/2014, KR265862; MN942260); Common-moorhen CoV HKU21 (NC_016996); Night-heron CoV HKU19 (NC_016994); Munia CoV HKU13-3514 (NC_011550); Bulbul CoV HKU11-934 (NC_011547); White-eye CoV HKU16 (NC_016991); Wigeon CoV HKU20 (NC_016995); Sparrow CoV HKU17 strain HKU17-6124 (JQ065045); and Thrush CoV HKU12-600 (FJ376621); and γ-CoVs subgenus Igavirus Infectious avian bronchitis virus (IABV; NC_001452; AY319651) and Turkey CoV (NC_010800), and subgenus Cegacovirus Beluga Whale CoV SW1 (subgenus Cegacovirus; NC_010646) and Bottlenose dolphin CoV HKU22 isolate CF090327 (KF793825).

Intragenomic rearrangements involving 3'-UTR sequences were not detected

The directionality of translocation appears to be strictly in the 5' to 3' direction as further underscored by the absence of 3'-UTR-derived insertions in any of the viruses analyzed here. We had documented insertion of segments derived from the 3'-end of the nucleocapsid gene and/or the beginning of the

ORF10 gene to the end of the 3'-UTR of two CoVs from *Rhinolophus* bats, exemplifying again translocation in the 5' to 3' direction (Patarca and Haseltine, 2020).

Discussion

We here describe intragenomic rearrangements involving 5'-UTR-derived sequences and the coding section of the genome. Figure 10 summarizes the locations of insertions (yellow arrows) 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. We had previously reported (Patarca and Haseltine 2022) on the presence of conserved complementary sequences (CCSs) in the 5'- and 3'-UTRs potentially involved in circularization of the genome during subgenomic RNA synthesis. As shown in Figure 10, the 5'-UTR-derived sequences involved in intragenomic rearrangements in SARS-CoV-2 shown here 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.

Most of the 5'-UTR sequences duplicated and translocated include TRS-L. Introduction of a new TRS-L and adjoining 5'-UTR sequences to an intra- or intergenic region by the intragenomic rearrangements described here may facilitate template switching during subgenomic messenger RNA synthesis by 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 of the genome, thereby optimizing free minimum 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 (Bentley et al. 2013) 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 (Nomburg et al. 2020).

The structural genes control genome dissemination (Lauber et al. 2013) 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 (Cui et al. 2021; Fang et al. 2021). Gaining insight into the effect of the amino acid changes introduced by the 5'-UTR-derived 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, that is generally lethal (Rottier et al. 2005).

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 frequent 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 also 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 β-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. Nonetheless, for SARS-CoV-2, the frequency of variants with full-length insertions appears low relative to those with subsegments or other mutations relative 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 (28 in California; 20 in Michigan; 18 in Florida; 17 in Minnesota; 15 in Maryland; and 13 in Pennsylvania, among the most representative in the case of mutations affecting the carboxyl terminus of ORF8) would be consistent with human-to-human transmission. Moreover, Turakhia et al. (2020), 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 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 Supplemental section), 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 various subgenera in at least two hosts (humans and bats). Moreover, it is unlikely that the insertion in the nucleocapsid gene which encodes for a common co-mutation of adjacent sites that has been shown experimentally to have functional significance reflects an artifactual event.

Intragenomic rearrangements might be more common in isolates with large deletions, as exemplified by those involving the ORF6 (Tse et al. 2021), ORF7b and ORF8, which in all cases affect the C-termini of the encoded proteins. The length of the insertion does not notably affect that of the protein 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.

Variation driven by internal rearrangements is distinct from the non-homologous recombination events proposed as origins of Sarbecovirus/Hibecovirus/Nobecovirus β-CoV ORF3a by gene duplication followed by rapid divergence from M (Nikolaidis et al. 2021; Ouzounis et al. 2020) or of SARS-CoV-2

ORF8 from ORF7a (Neches et al. 2021). 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.

Materials and methods

To assess the presence of 5’-UTR-derived insertions in the body of the genome, we used 5- to 10-amino acid stretches from the 3 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 BLASTP® (Protein BLAST: search protein databases using a protein query (nih.gov); Altschul et al. 1997) 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) [Bikandy et al. 2004] and Expasy (ExPASy - Translate tool [Duvaud et al. 2021]) tools to determine by manual inspection and the BLASTN program 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.

To detect isolates with similar insertions whose sequences had not been included in GenBank, we then searched the GISAID EpiFlu™ database of SARS-CoV-2 sequences (GISAID - Initiative; Elbe et al. 2017; Khare et al. 2021; Shu et al. 2017) 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.

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 (2020 release; https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/223/coronaviridae-figures); and the reference sequences in the GenBank database to derive the 5’-UTRs of various coronaviruses and utilized them 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 coronaviruses other than SARS-CoV-2 and therefore could not be used for this analysis. Using nucleotide sequences instead of translated amino acid sequences from the 5’-UTR in the 3 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.

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. 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. We also tested the 3'-UTR sequences using the same approaches described for the 5'-UTR ones. 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.

Data availability

The data underlying this article are available in GenBank (pubmed.ncbi.nlm.nih.gov) and GISAID at gisaid.org and all accession numbers are provided in the text and in Supplementary material.

References

Alhammad YMO, Kashipathy MM, Roy A, Gagné JP, McDonald P, Gao P, Nonfoux L, Battaile KP, Johnson DK, Holmstrom ED, Poirier GG, Lovell S, Fehr AR. The SARS-CoV-2 Conserved Macrodomain Is a Mono-ADP-Ribosylhydrolase. *J Virol*. 2021;95(3):e01969-20. doi: 10.1128/JVI.01969-20.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389-3402.

Amoutzias GD, Nikolaidis M, Tryfonopoulou E, Chlichlia K, Markoulatos P, Oliver SG. The remarkable evolutionary plasticity of coronaviruses by mmutation and recombination: Insights for the COVID-19 pandemic and the future evolutionary paths of SARS-CoV-2. *Viruses*. 2022;14:78. https://doi.org/10.3390/v14010078.

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat. Med*. 2020;26(4):450-452.

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): e00524-13.

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):30-35. doi: 10.1159/000490566.

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(8):2065-2072. doi: 10.1007/s00705-018-3832-8.

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; 87:2128-2136.

Bikandi, J., San Millán, R., Rementeria, A., and Garaizar, J. *In silico* analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction. *Bioinformatics*. 2004;20:798-799. DOI: 10.1093/bioinformatics/btg491

Bobay L-M, O'Donnell AC, Ochman H. Recombination events are concentrated in the sspike protein region of betacoronaviruses. *PLoS Genet*. 2020;16:e1009272.

Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, Rambaut A, Robertosn DL. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol*. 2020;5:1408-1417.

Carlson CR, Asfaha JB, Ghent CM, Howard CJ, Hartooni N, Safari M, Frankel AD, Morgan DO. Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. *Mol Cell*. 2020 Dec 17;80(6):1092-1103.e4. doi: 10.1016/j.molcel.2020.11.025.

Chen SC, Olsthoorn RCL. Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology*. 2010;401(1):29-41.

Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and sources of endemic human coronaviruses. Adv *Virus Res*. 2018;100:163-188.

Crook JM, Murphy I, Carter DP, Pullan ST, Carroll M, Vipond R, Cunningham AA, Bell D. Metagenomic identification of a new sarbecovirus from horseshoe bats in Europe. *Sci Rep*. 2021;11:14723.

Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol*. 2019;17:181-192.

Decaro N, Mari V, Campolo M, Lorusso A, Camero M, Elia G, Martella V, Cardioli P, Enjuanes L, Buonavoglia C. Recombinant canine coronaviruses related to transmissible gastroenteritis virus of Swine and circulating in dogs. *J Virol*. 2009;83(3): 1532-1537.

Dudas G, Rambaut A. MERS CoV recombination: Implications about the reservoir and potential for adaptation. *Virus Evol*. 2016;2: vsv023.

Dutta NK, Mazumdar K, Gordy JT. The nucleocapsid protein of SARS–CoV-2: a target for vaccine development. *J Virol*. 2020;94(13): e00647-20.

Duvaud S, Gabella C, Lisacek F, Stockinger H, Ioannidis V, Durinx C. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its user. *Nucleic Acids Research*. 2021. doi: 10.1093/nar/gks225.

Dwivedy A, Mariadasse R, Ahmad M, Chakraborty S, Kar D, Tiwari S, Bhattacharyya S, Sonar S, Mani S, Tailor P, Majumdar T, Jeyakanthan J, Biswal BK. 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. doi: 10.1371/journal.pcbi.

Elbe, S. and Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*. 2017;1:33-46. doi:10.1002/gch2.1018 PMCID: 31565258

Fang P, Fang L, Zhang H, Xia S, Xiao S. Functions of coronavirus accessory proteins: Overview of the state of the art. *Viruses*. 2021;13:1139.

Flower TG, Buffalo CZ, Hooy RM, Allaire M, Ren X, Hurley JH. Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proc Natl Acad Sci USA*. 2021;118(2):e2021785118. doi: 10.1073/pnas.2021785118.

Forni D, Cagliani R, Clerici M, Sironi M. Molecular evolution of human coronavirus genomes. *Trends Microbiol*. 2017;25:35-48.

Forni D, Cagliani R, Sironi M. Recombination and positive selection differentially shaped the diversity of betacoronavirus subgenera. *Viruses*. 2020;12:1313.

Franco-Muñoz C, Álvarez-Díaz DA, Laiton-Donato K, Wiesner M, Escandón P, Usme-Ciro JA, Franco-Sierra ND, Flórez-Sánchez AC, Gómez-Rangel S, Rodríguez-Calderon LD, Barbosa-Ramirez J, Ospitia-Baez E, Walteros DM, Ospina-Martinez ML, Mercado-Reyes M. Substitutions in Spike and Nucleocapsid proteins of SARS-CoV-2 circulating in South America. *Infect Genet Evol*. 2020;85:104557. doi: 10.1016/j.meegid.2020.104557.

Goldstein SA, BrownJ, Pedersen BS, Quinlan AR, Elde NC. Extensive recombination-driven coronavirus diversification expands the pool of potential pandemic pathogens. *bioRxiv : the preprint server for biology*, 2021.02.03.429646. https://doi.org/10.1101/2021.02.03.429646

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagman BL, Lauber C, Leontovich AM, Neuman BW, et al. The species Severe Acute Respiratory Syndrome-related coronavirus: Classifying 2019-NCoV and naming it SARS-CoV-2. *Nat Microbiol*. 2020;5:536-544.

Gordon DE, Hiatt J, Bouhaddou M, Rezelj VV, Ulferts S, Braberg H, Jureka AS, Obernier K, Guo JZ, Batra J, Kaake RM. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science*. 2020;370(6521):eabe9403.

Graham RL, Baric RS. Recombination, reservoirs, and the modular spike. Mechanisms of coronavirus cross-species transmission. *J Virol*. 2010;84:3134-3146.

Graham RL, Deing DJ, Deming ME, Yount BL, Baric RS. Evaluation of a recombination-resistant coronavirus as a broadly applicable, rapidly implementable vaccine platform. *Commun Biol*. 2018; 1(1): 1-10.

Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, Rawlings SA, Sutherland A, Premkumar L, Jadi RS, Marrama D, de Silva AM, Frazier A, Carlin AF, Greenbaum JA, Peters B, Krammer F, Smith DM, Crotty S, Sette A. 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. doi: 10.1016/j.cell.2020.05.015

Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences of the United States of America*. 2020; 117(26):15193–15199. https://doi.org/10.1073/pnas.2008176117

Hachim A, Kavian N, Cohen CA, Chin AW, Chu DK, Mok CK, Tsang OT, Yeung YC, Perera RA, Poon LL, Peiris JS. ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection. *Nature immunology*. 2020 Oct;21(10):1293-301.

Hagemeijer MC, Monastyrska I, Griffith J, van der Sluijs P, Voortman J, en Henegouwen PM, Vonk AM, Rottier PJ, Reggiori F, De Haan CA. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology*. 2014;458:125-135.

Hartenian F, Nandakumar D, Lari A, Ly M, Tucker JM, Glausinger BA. The molecular virology of coronaviruses. *J Biol Chem*. 2020;295:12910-12934.

Hassan SS, Choudhury PP, Dayhoff GW 2nd, Aljabali AAA, Uhal BD, Lundstrom K, Rezaei N, Pizzol D, Adadi P, Lal A, Soares A, Mohamed Abd El-Aziz T, Brufsky AM, Azad GK, Sherchan SP, Baetas-da-Cruz W, Takayama K, Serrano-Aroca Ã, Chauhan G, Palu G, Mishra YK, Barh D, Santana Silva RJ, Andrade BS, Azevedo V, Góes-Neto A, Bazan NG, Redwan EM, Tambuwala M, Uversky VN. The importance of accessory protein variants in the pathogenicity of SARS-CoV-2. *Arch Biochem Biophys*. 2022;717:109124. doi: 10.1016/j.abb.2022.109124.

He R, Leeson A, Ballantine M, Andonov A, Baker L, Dobie F, Li Y, Bastien N, Feldmann H, Strocher U, Theriault S, Cutts T, Cao J, Booth TF, Plummer FA, Tyler S, Li X. Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res*. 2004;105(2):121-125. doi: 10.1016/j.virusres.2004.05.002.

Imbert I, Snijder EJ, Dimitrova M, Guillemot JC, Lécine P, Canard B. The SARS-Coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. *Virus research*. 2008; 133(2):136-148.

Islam MR, Hoque MN, Rahman MS, Alam ASMRU, Akther M, Puspo JA, Akter S, Sultana M, Crandall KA, Hossain MA. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Sci Rep*. 2020;10(1):14004. doi: 10.1038/s41598-020-70812-6.

Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, Haldenby S, Hill V, Lucaci A, McCrone JT, Nicholls SM. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*. 2021 Sep 30;184(20):5179-5188.

Jaroszewski L, Iyer M, Alisoltani A, Sedova M, Godzik A. The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. *PLoS computational biology*. 2021; 17(7):e1009147.

Johnson BA, Zhou Y, Lokugamage KG, Vu MN, Bopp N, Crocquet-Valdes PA, Schindewolf C, Liu Y, Scharton D, Plante JA, Xie X, Aguilar P, Weaver SC, Shi PY, Walker DH, Routh AL, Plante KS, Menachery VD. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *bioRxiv* [Preprint]. 2021 Oct 15:2021. doi: 10.1101/2021.10.14.464390.

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):4888-4894.

Kennelly PJ, Krebs EG. Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J Biol Chem*. 1991;266(24):15555-15558.

Khavinson V, Terekhov A, Kormilets D, Maryanovich A. Homology between SARS CoV-2 and human proteins. *Sci Rep*. 2021;11:17199. doi: 10.1038/s41598-021-96233-7.

Khare, S., et al. GISAID's Role in Pandemic Response. *China CDC Weekly*. 2021;3(49): 1049-1051. doi: 10.46234/ccdcw2021.255 PMCID: 8668406

Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, Li H, Chmura AA, Field HE, Zambrana-Torrelio C, Epstein JH, Li B. Origin and cross-species transmission of bat coronaviruses in China. *Nature Communications*. 2020 Aug 25;11(1):1-5.

Lau SKP, Wong EYM, Tsang CC, Ahmed SS, Au-Yeung RKH, Yuen K-Y, Wernery U, Woo PCY. 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: e00265-18.

Lauber C, Goeman JJ, Parquet MDC, Thi Nga P, Snijder EJ, Morita K, Gorbalenya AE. The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathogen*. 2013;9:e1003500.

Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GM, Ruben M, Overkleeft HS, van Veelen PA, Samborskiy DV, Kravchenko AA, Leontovich AM, Sidorov IA, Snijder EJ, Posthuma CC, Gorbalenya AE. 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):8416-8434. doi: 10.1093/nar/gkv838.

Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral research*. 2018;149:58-74.

Li JY, Liao CH, Wang Q, Tan YJ, Luo R, Qiu Y, Ge XY. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus research*. 2020;286:198074.

Lin X, Fu B, Yin S, Li Z, Liu H, Zhang H, Xing N, Wang Y, Xue W, Xiong Y, Zhang S, Zhao Q, Xu S, Zhang J, Wang P, Nian W, Wang X, Wu H. ORF8 contributes to cytokine storm during SARS-CoV-2 infection by activating IL-17 pathway. *iScience*. 2021;24(4):102293. doi: 10.1016/j.isci.2021.102293.

Liu DX, Fung TS, Chong KK, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res*, 2014;109:97-109. doi: 10.1016/j.antiviral.2014.06.013.

Lo C-Y, Tsai T-L, Lin C-N, Lin C-H, Wu H-Y. 2019. 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;286:3222-3239.

Lu S, Ye Q, Singh D, Cao Y, Diedrich JK, Yates JR 3rd, Villa E, Cleveland DW, Corbett KD. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat Commun*. 2021;12(1):502. doi: 10.1038/s41467-020-20768-y.

Lytras S, Hughes J, Martin D, de Klerk A, Lourens R, Kosakovsky Pond SL, Xia W, Jiang X, Robertson DL. Exploring the natural origins of SARS-CoV-2 in the light of recombination. *Genome Biol Evol*. 2022;evac018. doi: 10.1093/gbe/evac018

Madhugiri R, Karl N, Petersen D, Lamkiewicz K, Fricke M, Wend U, Scheuer R, Marz M, Ziebuhr J. Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions. *Virology*. 2018;517:44-55. https://doi.org/10.1016/j.virol.2017.11.025

Makino S, Keck JG, Stohlman SA, Lai MM. High-frequency RNA recombination of murine coronaviruses. *J Virol.* 1986;57:729-737.

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. *The Journal of general virology*. 2014;95(Pt 4):874.

McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. *Viruses*. 2014;6(8):2991-3018. doi: 10.3390/v6082991.

Menachery, V. D., Yount, B. L., Jr, Debbink, K., Agnihothram, S., Gralinski, L. E., Plante, J. A., Graham, R. L., Scobey, T., Ge, X. Y., Donaldson, E. F., Randell, S. H., Lanzavecchia, A., Marasco, W. A., Shi, Z. L., & Baric, R. S. (2015). A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nature medicine*. 2015;21(12):1508–1513. https://doi.org/10.1038/nm.3985

Miao Z, Tidu A, Eriani G, Martin F. Secondary structure of the SARS-CoV-2 5'-UTR. *RNA Biology*. 2021;18(4):447-456.

Mounir S, Talbot PJ. Molecular characterization of the S protein gene of human coronavirus OC43. *J Gen Virol*. 1993;74:1981-1987. **https://doi.org/10.1099/0022-1317-74-9-1981**

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):e03014-20.

Niemeyer D, Zillinger T, Muth D, Zielecki F, Horvath G, Suliman T, Barchet W, Weber F, Drosten C, Müller MA. Middle East respiratory syndrome coronavirus accessory protein 4a is a type I interferon antagonist. *Journal of virology*. 2013;87(22):12489-12495.

Nikolaidis M, Markoulatos P, van de Peer Y, Oliver SG, Amoutzias GD. The neighborhood of the spike gene is a hotspot for modular intertypic homologous and non-homologous recombination in coronavirus genomes. *Mol Biol Evol*. 2021;msab 292.

Nomburg J, Meyerson M, De Caprio JA. Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. *Genome Medicine*. 2020;12:108.

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. doi: 10.3389/fimmu.2020.587615.

Ouzounis C. A. A recent origin of Orf3a from M protein across the coronavirus lineage arising by sharp divergence. *Computational and structural biotechnology journal*. 2020;18:4093–4102. https://doi.org/10.1016/j.csbj.2020.11.047

Park GJ, Osinski A, Hernandez G, et al. The mechanism of RNA capping by SARS-CoV-2. bioRxiv 2022.02.07.479471; doi: https://doi.org/10.1101/2022.02.07.479471

Patarca R, Haseltine WA. Structural flexibility of the SARS-CoV-2 genome relevant to variation, replication, pathogenicity and immune evasion. *bioRxiv*. [Preprint] 2021.12.20.473542; doi: https://doi.org/10.1101/2021.12.20.473542

Patarca R, Haseltine WA. Circularization via complementary sequences in the 5' and 3' termini may facilitate replication of SARS coronaviruses. *Authorea*. [Preprint] January 04, 2022. doi: 10.22541/au.164132044.46753705/v1

Pickering B, Lung O, Maguire F, Kruckiewicz P, Kotwa JD, Buchanan T, Gagnier M, Guthrie JL, Jardine CM, Marchand-Austin A, Massé A, McClinchey H, Nirmalarajah K, Aftanas P, Blais-Savoie J, Chee H-Y, Chien E, Yim W, Goolia M, Sudermna M, Pinette M, Smith G, Sullivan D, Rudar J, Adey E, Nebroski M, Côté M, Laroche G, McGeer AJ, Nituch L, Mubareka S, Bowman J. Highly divergent white-tailed deer SARS-CoV-2 with potential deer-to-human transmission. *bioRxiv* [Preprint] 2022.02.22.481551. doi: https://doi.org/10.1101/2022.02.22.481551

Pollett S, Conte MA, Sanborn M, Jarman RG, Lidl GM, Modjarrad K, Maljkovic Berry I. A comparative recombination of analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic. *Sci Rep*. 2021;11:17365.

Pustovalova Y, Gorbatyuk O, Li Y, Hao B, Hoch JC. 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. doi: 10.1007/s12104-021-10059-y.

Redondo N, Zaldívar-López S, Garrido JJ, and Montoya M. SARS-CoV-2 Accessory Proteins in Viral Pathogenesis: Knowns and Unknowns. *Front. Immunol*. 2021;12:708264

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;79:14122-14130.

Sawicki SG, Sawicki DL, Siddell SG. A contemporary view of coronavirus transcription. *J Virol*. 2007;81:20-29.

Shang J, Han N, Chen Z, Peng Y, Li L, Zhou H, Ji C, Meng J, Jiang T, Wu A. Compositional diversity and evolutionary pattern of coronavirus accessory proteins. *Brief. Bioinform.* 2021; 221267-1268.

Shu, Y. and McCauley, J. GISAID: from vision to reality. *EuroSurveillance* 2017;22(13) doi:10.2807/1560-7917.ES.2017.22.13.30494 PMCID: PMC5388101

Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? *Nat Rev Microbiol*. 2011;9(8):617-626.

Siu KL, Yeung ML, Kok KH, Yuen KS, Kew C, Lui PY, Chan CP, Tse H, Woo PC, Yuen KY, Jin DY. 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. *Journal of virology*. 2014; 88(9): 4866-4876.

Slanina H, Madhugiri R, Bylapudi G, Schultheiß K, Karl N, Gulyaeva A, Gorbalenya AE, Linne U, Ziebuhr J. 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. doi: 10.1073/pnas.2022310118.

Song H-D, Tu C-C, Zhang G-W, Wang S-Y, Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol*. 2016;24:490-502.

Stukalov A, Girault V, Grass V, Karayel O, Bergant V, Urban C, Haas DA, Huang Y, Oubraham L, Wang A, Hamad MS, Piras A, Hansen FM, Tanzer MC, Paron I, Zinzula L, Engleitner T, Reinecke M, Lavacca TM, Ehmann R, Wölfel R, Jores J, Kuster B, Protzer U, Rad R, Ziebuhr J, Thiel V, Scaturro P, Mann M, Pichlmair A. Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature*. 2021 Jun;594(7862):246-252. doi: 10.1038/s41586-021-03493-4.

Surjit M, Kumar R, Mishra RN, Reddy MK, Chow VT, Lal SK. 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):11476-11486. doi: 10.1128/JVI.79.17.11476-11486.2005.

Temmam S, Vongphayloth K, Baquero Salazar E, et al. Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. https://doi.org/10.21203/rs.3.rs-871965/v1 (2020).

Thorne LG, Bouhaddou M, Reuschl AK, Zuliani-Alvarez L, Polacco B, Pelin A, Batra J, Whelan M, Hosmillo M, Fossati A, Ragazzini R, Jungreis I, Ummadi M, Rojc A, Turner J, Bischof ML, Obernier K, Braberg H, Soucheray M, Richards A, et al. Evolution of enhanced innate immune evasion by SARS-CoV-2. *Nature*. 2022;602(7897), 487–495. https://doi.org/10.1038/s41586-021-04352-y

Tse H, Lung DC, Wong SC, Ip KF, Wu TC, To KK, Kok KH, Yuen KY, Choi GK. Emergence of a Severe Acute Respiratory Syndrome Coronavirus 2 Virus variant with novel genomic architecture in Hong Kong. *Clin Infect Dis*. 2021 Nov 2;73(9):1696-1699. doi: 10.1093/cid/ciab198.

Tugaeva KV, Hawkins DEDP, Smith JLR, Bayfield OW, Ker DS, Sysoev AA, Klychnikov OI, Antson AA, Sluchanko NN. The Mechanism of SARS-CoV-2 Nucleocapsid Protein Recognition by the Human 14-3-3 Proteins. *J Mol Biol*. 2021 Apr 16;433(8):166875. doi: 10.1016/j.jmb.2021.166875.

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;532:134-138.

Turakhia, Y., De Maio, N., Thornlow, B., Gozashti, L., Lanfear, R., Walker, C. R., Hinrichs, A. S., Fernandes, J. D., Borges, R., Slodkowicz, G., Weilguny, L., Haussler, D., Goldman, N., & Corbett-Detig, R. (2020). Stability of SARS-CoV-2 phylogenies. *PLoS genetics*. *16*(11); e1009175. https://doi.org/10.1371/journal.pgen.1009175

Turakhia Y, Thornlow B, Hinrichs AS, McBroome J, Ayala N, Ye C, De Maio N, Haussler D, Lanfear R, Corbett-Detig R. Pandemic-Scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. *bioRxiv*. 2021 Jan 1.

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. doi: 10.3389/fgene.2021.693227.

VanInsberghe D, Neish AS, Lowen AC, Koelle K. Recombinant SARS-CoV-2 genomes are currently circulating at low levels. *bioRxiv*. 2021 Jan 1:2020-08.

Van Marle G, Luytjes W, Van der Most RG, et al. Regulation of coronavirus mRNA transcription. *J Virol*. 1995; 69(12):7851-7856.

Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, Vandamme A-M, Van Ranst M. Complete genomic sequence of human coronavirus OC43: Molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J Virol*. 2005;79:1595-1604.

Wang X, Lam JY, Wong WM, Yuen CK, Cai JP, Au SW, Chan JF, To KKW, Kok KH, Yuen KY. Accurate Diagnosis of COVID-19 by a Novel Immunogenic Secreted SARS-CoV-2 orf8 Protein. *mBio*. 2020 Oct 20;11(5):e02431-20. doi: 10.1128/mBio.02431-20.

Wang D, Jiang A, Feng J, et al. The SARS-CoV-2 subgenome landscape and its novel regulatory features. Molecular *Cell*. 2021;81:2135-2147.

Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiology and molecular biology reviews*. 2005;69(4):635-664.

Wong AC, Li X, Lau SK, Woo PC. Global epidemiology of bat coronaviruses. *Viruses*. 2019;11(2):174.

Woo PC, Lau SK, Huang Y, Yuen KY. Coronavirus diversity, phylogeny and interspecies jumping. *Experimental Biology and medicine*. 2009;234(10):1117-27.

Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, Tang W, Xiao Y, Liu G, Luo H, Zhu W, Lin X, Meng G, Zhu Z. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host & Microbe*. 2021;29:1788-1801.

Yan L, Ge J, Zheng L, Zhang Y, Gao Y, Wang T, Huang Y, Yang Y, Gao S, Li M, Liu Z, Wang H, Li Y, Chen Y, Guddat LW, Wang Q, Rao Z, Lou Z. 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. doi: 10.1016/j.cell.2020.11.016.

Yang R, Zhao Q, Rao J, Zeng F, Yuan S, Ji M, Sun X, Li J, Yang J, Cui J, Jin Z, Liu L, Liu Z. SARS-CoV-2 Accessory Protein ORF7b Mediates Tumor Necrosis Factor-α-Induced Apoptosis in Cells. *Front Microbiol*. 2021;12:654709. doi: 10.3389/fmicb.2021.654709.

Yang Y, Zhang L, Geng H, Deng Y, Huang B, Guo Y, Zhao Z, Tan W. 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):951-961.

Yang Y, Yan W, Hall AB, Jiang X. Characterizing transcriptional regulatory sequences in coronaviruses and their role in recombination. *Mol Biol Evol*. 2021;38:1241-1248.

Yao H, Song Y, Chen Y, Wu N, Xu J, Sun C, Zhang J, Weng T, Zhang Z, Wu Z, Cheng L, Shi D, Lu X, Lei J, Crispin M, Shi Y, Li L, Li S. Molecular Architecture of the SARS-CoV-2 Virus. *Cell*. 2020;183(3):730-738.e13. doi: 10.1016/j.cell.2020.09.018.

Zavadil J, Bitzer M, Liang D, Yang YC, Massimi A, Kneitz S, Piek E, Böttinger EP. Genetic programs of epithelial cell plasticity directed by transforming growth factor-β. *Proceedings of the National Academy of Sciences*. 2001;98(12):6686-6691.

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):4738-4746.

Zhang Y, Chen Y, Li Y, Huang F, Luo B, Yuan Y, Xia B, Ma X, Yang T, Yu F, Liu J, Liu B, Song Z, Chen J, Yan S, Wu L, Pan T, Zhang X, Li R, Huang W, He X, Xiao F, Zhang J, Zhang H. The ORF8 protein of SARS-CoV-2 mediates immune evasion through down-regulating MHC-I. *Proc Natl Acad Sci USA*. 2021;118(23):e2024202118. doi: 10.1073/pnas.2024202118.

Figure legends

Figure 1.

Discontinuous synthesis of SARS-CoV-2 negative strand subgenomic RNA. For the synthesis of subgenomic RNA, the leader transcription regulatory sequence (TRS-L, blue box) within the 5' leader sequence interacts with homologous TRSs in the body (TRS-B) of the genome that precede structural (red boxes) and accessory (green boxes) genes. Overlapping genes for ORF3a (namely, ORF3b-d) and N (namely, ORF9b, c) that would be translated from the sgRNAs shown for ORF3a and N, respectively, are shown at the bottom of the figure.

Figure 2.

A. Secondary structure of the 5'-UTR and 5'-leader sequence. The secondary structure of the 5'-UTR is shown as presented in Miao et al. (2021). The 5' leader sequence extends from the cap structure (m^7G) to the TRS-L and encompasses stem-loops (SL)1-3, which have been associated with viral replication and gene expression.

B. Translated sequence of the 5'-leader sequence and beyond until the stop codon before SL5. An open reading frame spans most of the 5'-UTR (Wuhan reference strain, NC_045512 shown) except for SL5, where ORF1ab starts. The segment of the open reading frame that is translocated as a whole or partially in SARS-CoV-2 variants is underlined. At the nucleotide level, the segment includes the TRS-L, and at the amino acid level, the translated 5'-leader sequence and beyond includes a predicted upstream open reading frame (uORF, grey box) which has not been shown to be functional and whose predicted initiation codon is highlighted (underlined methionine [M] in gray). The underlined sequence has been shown to be duplicated and translocated in place of an 882-nucleotide deletion within the coding portion of the viral genome of a SARS-CoV-2 variant isolated from 3 cases in Hong Kong with absent ORF7a, ORF7b, and ORF8 and a C-terminally modified ORF6 product (Tse et al., 2021).

Figure 3.

Modified carboxyl-termini of ORF8 encoded by an insertion of a 5'-UTR segment in SARS-CoV-2. 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 January 20, 2022, are listed below. 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; Flower et al. 2021). The 5'-UTR insertions are shown as underlined letters in black with mutations, deletions, and insertions within them highlighted in green.

Figure 4.

A. Modified carboxyl-termini of ORF8 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; Menachery et al. 2015), 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. Color codes and abbreviations are as in Figures 1 and 2.

B. Modified carboxyl termini of ORF7b encoded by an insertion of a 5'-UTR segment in SARS-CoV-2. The two isolates with 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 Figures 1 and 2.

Figure 5.

Insertion of 5'-UTR segment into the serine-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 SR-rich region of the N protein at the end of a strong immunodominant B-cell epitope (purple box; Oliveira et al. 2020).

Figure 6.

Insertions of a 5'-UTR sequence into two sites within 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-derived sequences at the proximal and distal sites are provided in the figure and a full listing is provided in the Supplement Figure 6 legend. Variants with single amino acid changes relative to the Wuhan reference strain in the segment corresponding to the insertion are also listed in the Supplement Figure 6 legend. 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.

Figure 7.

Intragenomic rearrangement with 5'-UTR sequences present in the intergenic regions between p3 and 4a as well as between 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 p3 and p4a in isolate MG923473 (Burkina Faso, 2015) and between p4b and p5 affecting the 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-drived sequence.

Figure 8.

Presence in in the intergenic region between the S and Ns5a genes hCoV-OC43 (β-CoV Embevovirus) of sequences of various lengths of the same 5'-UTR region.

The hCoV-OC43 5'-UTR sequence inserted is identical to that of bovine coronavirus (BCoV) 5'-UTR except for one nucleotide (underlined adenosine [A] is a guanosine [G] in BCoV). Changes between intergenic region and 5’-UTR are also underlined. All variants detected are listed in the Supplement Figure 8 legend.

Figure 9.

A. Presence of hCoV-HKU-1 (β-CoV Embecovirus) of 5’-UTR-derived sequence in the intergenic region between the spike (S) and the Ns4 genes
B. Presence of 5’-UTR sequence in the Bat β-CoV Nobecovirus nsp3 gene

All variants detected are listed in the Supplement Figure 9 legend.

Figure 10.

The 5’-UTR nucleotide segment that is translocated to viral genes partially overlaps a predicted sequence potentially involved in circularization of genome during viral replication.
The top of the figure summarizes the locations of insertions (yellow arrows) in nonstructural, structural, and accessory genes of SARS-CoV-2. The previously reported conserved complementary sequences (CCSs) in the 5’- and 3’ UTRs potentially involved in circularization of the genome during subgenomic RNA synthesis (Patarca and Haseltine 2022) are then shown. The insertion sequences 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.

Figure 1.

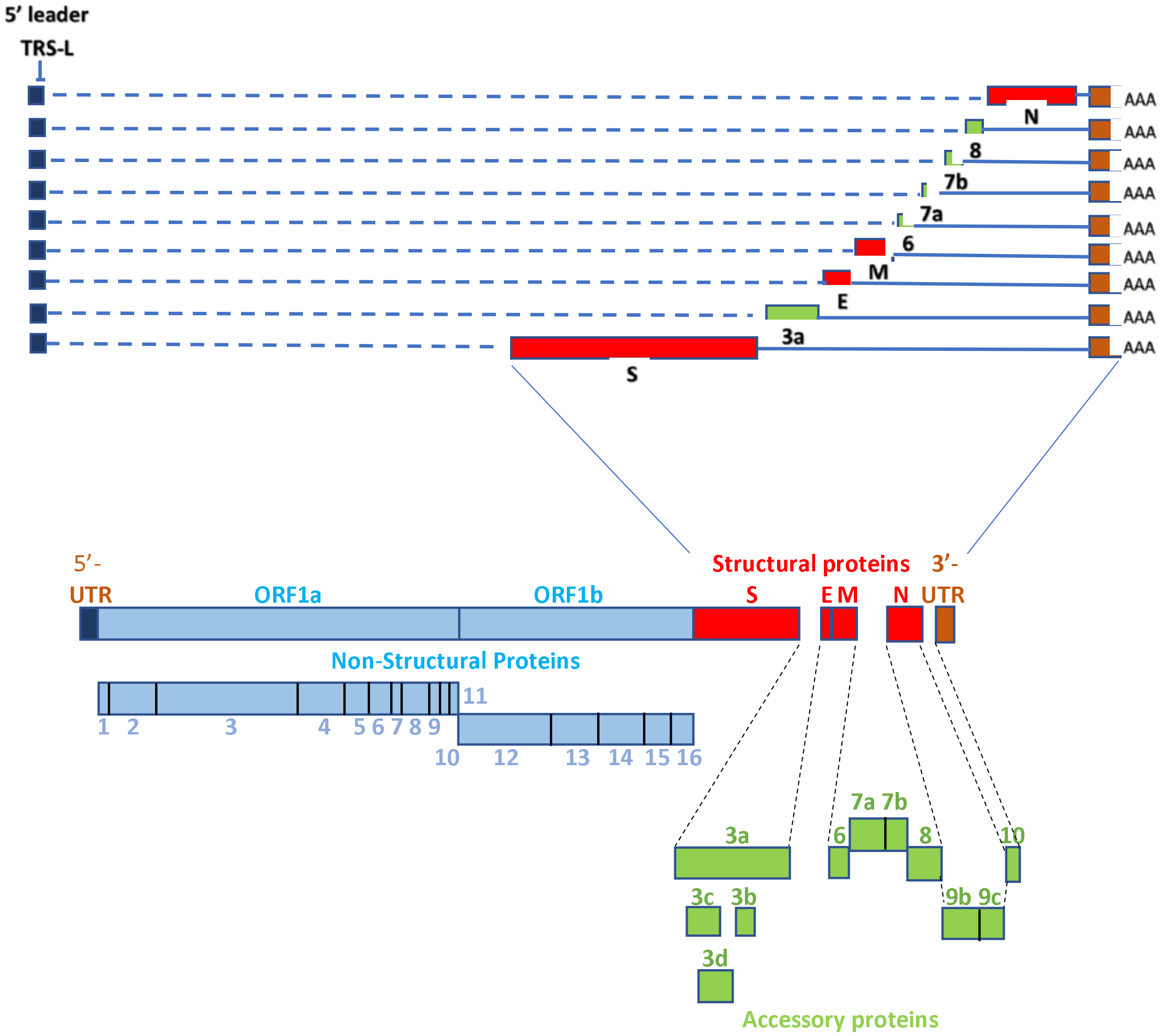

5' leader
TRS-L
N
8
7b
7a
6
M
E
3a
S
AAA
5'-
UTR
ORF1a
ORF1b
Structural proteins
3'-
S
E M
N
UTR
Non-Structural Proteins
11
1 2 3 4 5 6 7 8 9
10 12 13 14 15 16
7a 7b
3a
6
8
10
3c 3b
9b 9c
3d
Accessory proteins

Figure 2.

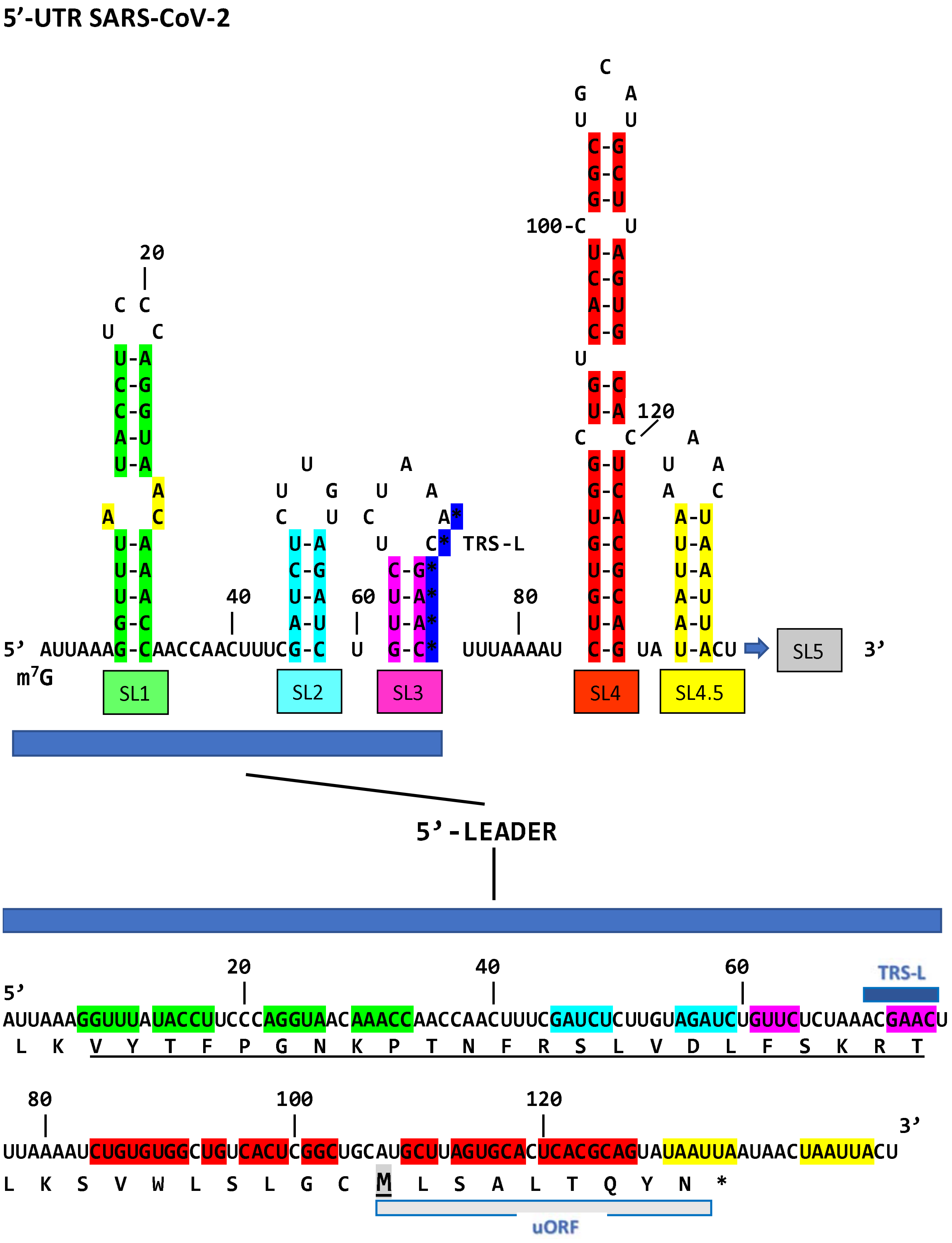
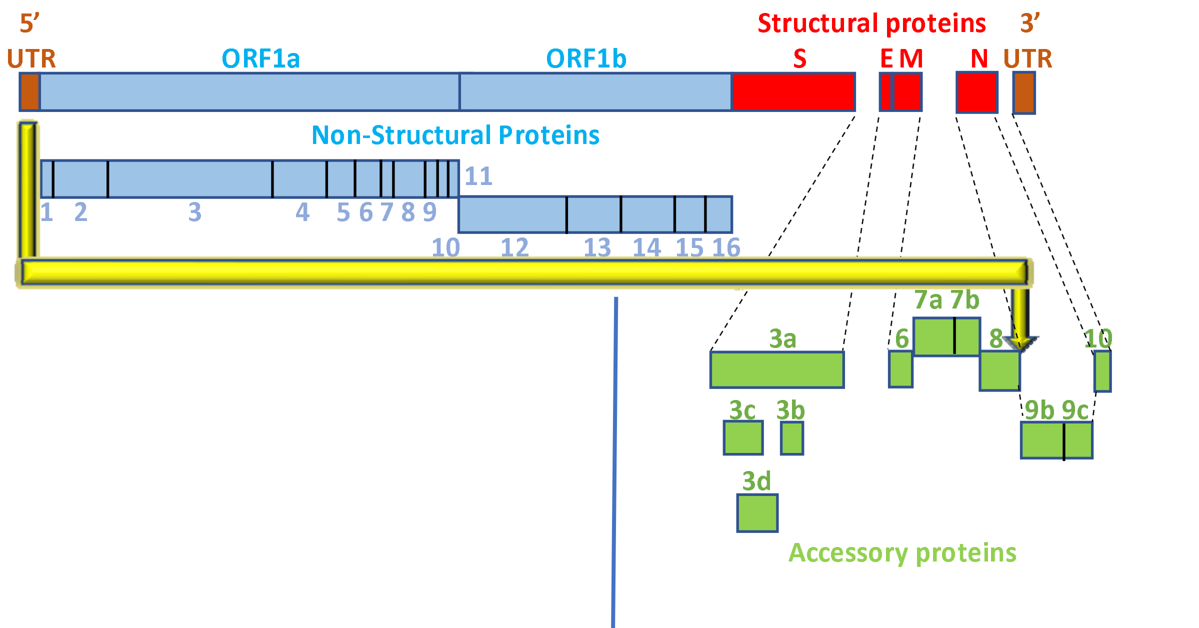

5'-UTR SARS-CoV-2
20
40
60
80
100
120
TRS-L
5' AUUAAAG-CAACCAACUUUCG-C U UUUAAAAU C-G UA U-ACU
m^7G
3'
SL1
SL2
SL3
SL4
SL4.5
SL5
5'-LEADER
5'
AUUAAAGGUUUAUACCUUCCCAGGUAACAAACCAACCAACUUUCGAUCUCUUGUAGAUCUGUUCUCUAAACGAACU
L K V Y T F P G N K P T N F R S L V D L F S K R T
TRS-L
80
100
120
3'
UUAAAAUCUGUGUGGCUGUCACUCGGCUGCAUGCUUAGUGCACUCACGCAGUAUAAUUAAUAACUAAUUACU
L K S V W L S L G C M L S A L T Q Y N *
uORF

Figure 3.



SARS-CoV-2 (β-CoV Sarbecovirus)

5'-UTR (Wuhan ref[a], 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[a]	...YHDVRVVLDFI

Intact inserted UTR sequences

QUP34336	...YHDVRVLVDLFSKRTN
QVJ62740[b]	...YHDVRVVVDLFSKRTN
QUD47078	...YHDVRFVVDLFSKRTN
QTZ60073[c]	...YHDVRVVLDLFSKRTN
QWA62675	...YHDVPVVLDLFSKRTN
UAP73316[d]	...YHDRVVDLFSKRTN
UCZ40652	...YHDVDLFSKRTN
QXY14087[e]	...YHDVRVVVDFSKRTN
QUD12271[f]	...YHDVRVVLEFSKRTN
UBY86923	...YHDVRSVLDFSKRTN
QSX87276	...YHDV--VLDFSKRTN
UER91358[g]	...YHDVRVVLFSKRTN
QZP72779	...YHDVRVVFSKRTN
UEV05558	...YHDVCVVFSKRTN
UEH58452	...YHDVRVVLDSKRTN
UDN20252[h]	...YHDVRVVLSKTRN
UDG72468[i]	...YHDVRVVSKRTN
QZI47484	...YHDVRVESKRTN
QUA19334	...YHDVRVVLDFKRTN

UBY63352	...YHDVRVVLIKRTN

Mutation/deletion within insertion

QTC62354[j]	...YHDVRVLLDLFSKRTN
QSN95715	...YHDVRVVLNLFSKRTN
QQA02853[k]	...YHDVRVVID-FSKRTN
QYM26975	...YHDVRVVLDV-SKRTN
QTJ72925	...YHDVRVVL-MFSKRTN
QUA15500[l]	...YHDVRVVLDFF-KRTN

Insertion within insertion

QVM25420	...YHDVRVVLDLRFSKRTN
QVM48721	...YHDVRVVLDLWFSKRTN

Figure 4.

A.

SARS-related β-CoVs of *Rhinolophus* bats (China)

TRS-L

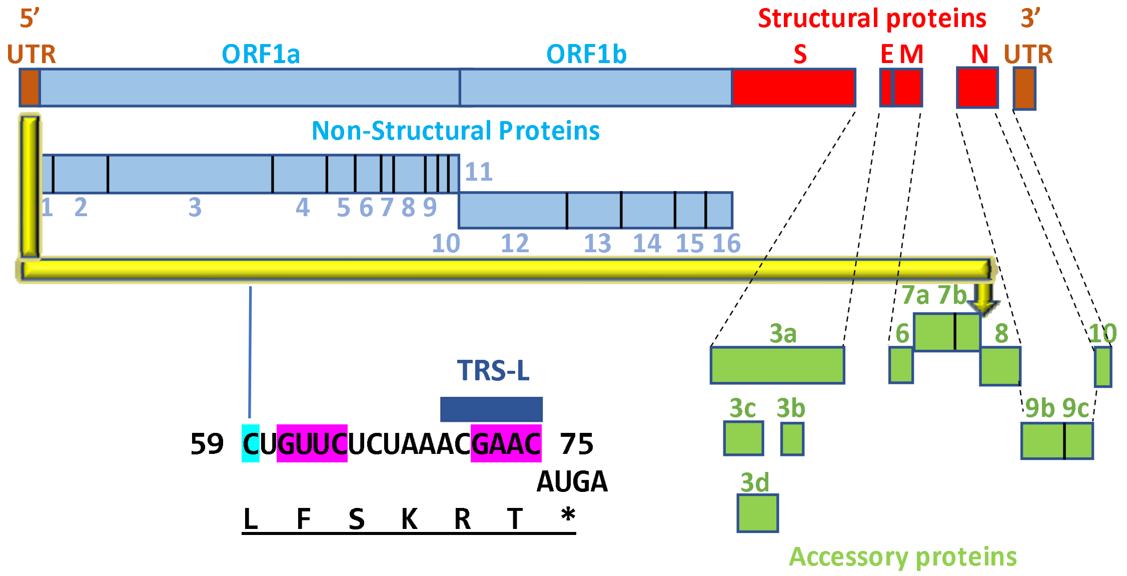
5'-UTR (2 nts. changes vs. SARS-CoV-2) GUAGAUCUGUUCUUUAAACGAACUUAA

V D L F F K R T *

ORF8 C-terminus FRDIHVDLFFKRT

AAZ67036 Bat SARSCoV Rf1/2004
AIA62307 BtRf_BetaCoV/SX2013
AKZ19083 Bat SARS-like CoV YNLF_31C
AIA62297 BtRF-BetaCoV/HeB2013
ABG47066 BatCoV 273/2005

B.



SARS-Cov-2 ORF7b

Wuhan ref	MIELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHNETCHA
QXH28554	MIELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHNETCLFSKRT
QSV08409	LAFLLFLVLIMLIIFWFSLELQDHNETCLFSKRT

Figure 5.

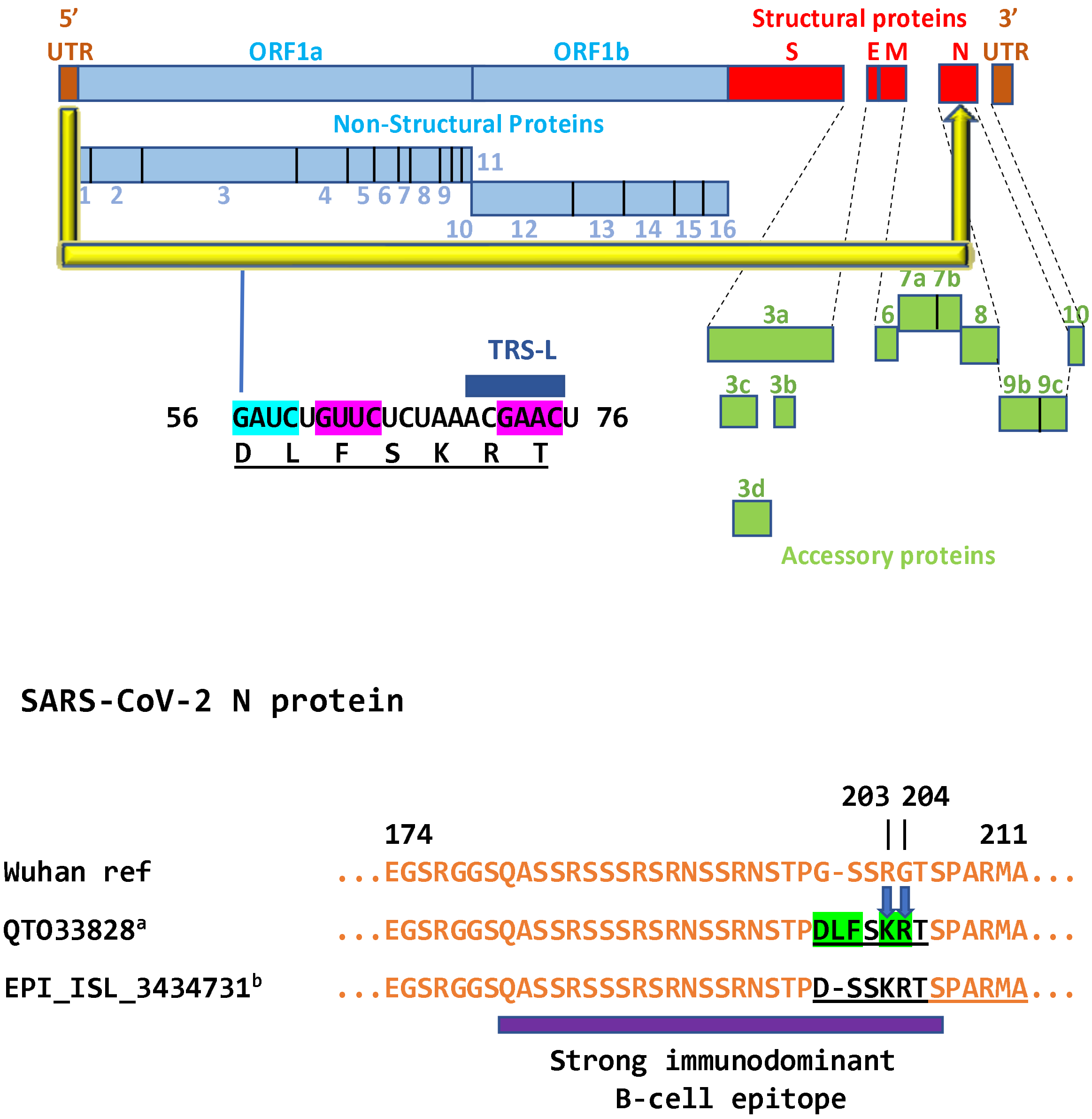
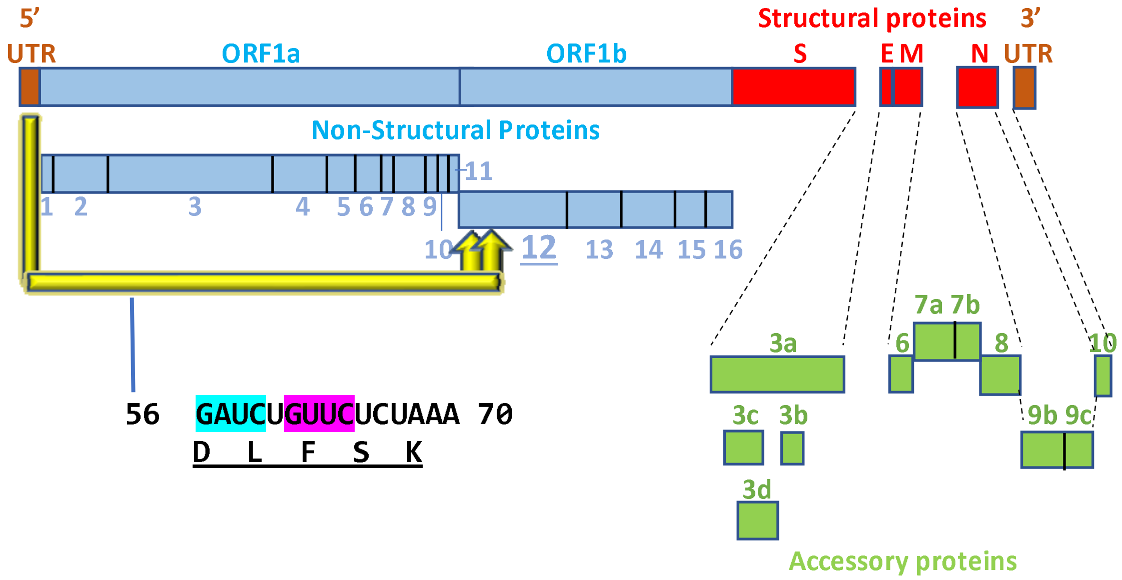

5' UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3' UTR
Non-Structural Proteins
1 2 3 4 5 6 7 8 9
11
10 12 13 14 15 16
7a 7b
3a
6
8
10
TRS-L
56 GAUCUGUUCUCUAAACGAACU 76
D L F S K R T
3c 3b
9b 9c
3d
Accessory proteins
SARS-CoV-2 N protein
203 204
174
211
Wuhan ref ...EGSRGGSQASSRSSSRSRNSSRNSTPG-SSRGTSPARMA...
QTO33828a ...EGSRGGSQASSRSSSRSRNSSRNSTPDLFSKRTSPARMA...
EPI_ISL_3434731b ...EGSRGGSQASSRSSSRSRNSSRNSTPD-SSKRTSPARMA...
Strong immunodominant B-cell epitope

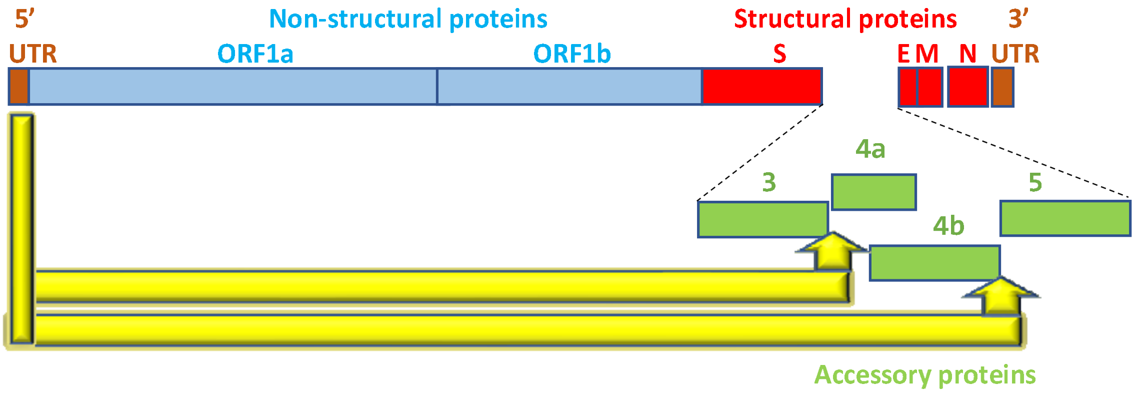
Figure 6.



Nsp12 (RdRp) N-terminal NiRAN domain

```
                1                                  VAGFSK  UHP90975[b]    56
QVL75820[a]     VFKRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVDLFSKFLKTNCCRFQEKDEDD
Wuhan ref       VFKRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVAGFAKFLKTNCCRFQEKDEDD
                                                   VAVFA   UAQ66644[c]
                                                   VASFA   UHS40780[d]

                57                                                    112
Wuhan ref       NLIDSYFVVKRHTFSNYQHEETIYNLLKDCPAVAKHDFFKFRIDGDMVPHISRQRLT


                113                               DYFSK  QZM71485[f]    168
UFT72204[e]     KYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDLFSKKDWYDFVENPDILRVYAN
Wuhan ref       KYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDYFNKKDWYDFVENPDILRVYAN
                                                  DFFNK   UBL67135[g]
                                                  DHFNK   UBD35057[h]

                169                                                   224
Wuhan ref       LGERVRQALLKTVQFCDAMRNAGIVGVLTLDNQDLNGNWYDFGDFIQTTPGSGVPVV


                225                243
Wuhan ref       DSYYSLLMPILTLTRALTA
```

Figure 7.



TRS-L

MERS (β-CoV Merbecovirus)
5'-UTR (ref: NC_019843; nts. 48-69) AGAACUUUGAUUUUAACGAACU

A. Intergenic region (ORF3/4a)

MG923473 Burkina Faso (2015) 25768 25789
Between ORF3 (ends at 25768) AGAACUUUGAUUUUAACGAACU
and ORF4a (begins at 25792)

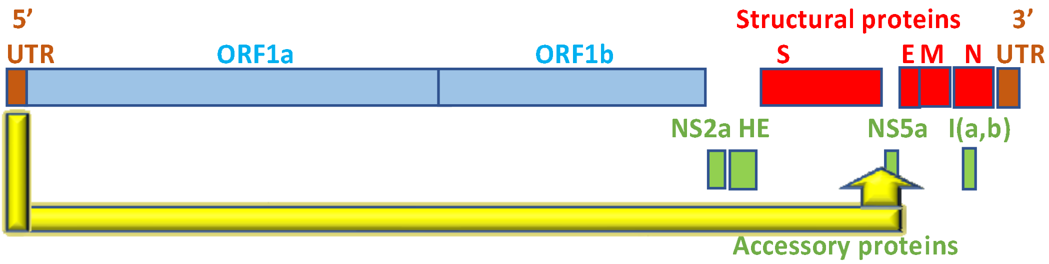
B.

TRS-L

ORF4b terminus
MK564475 Ethiopia (2017) 26795 26814
Between ORF4b (ends at 26802) CAACUUUGAUUUUAACGAACU
and ORF5 (begins at 26815) Q L * F * R T

MERS ref ...YPILHHPGF
MERS MK5644475 ...YPILHQL

Figure 8.



TRS-L

hCoV-OC43 ref (β-CoV Embecovirus)

5'-UTR (KJ958218; nts. 34-78) CACUGAUCUCUUGUUAGAUCUUUUUGUAAUCUAAACUUUAUAAAA

Intergenic region (S/Ns5a)

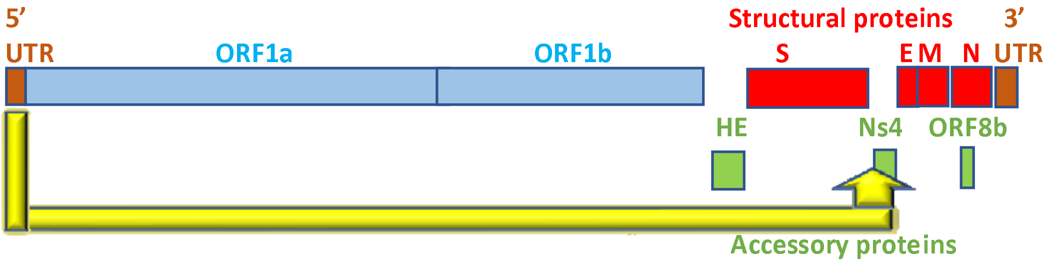
NC_006213 ref ATCC
VR-759 (USA, 2004)[a] 27724 27756
Between S (ends at 27704) CACUGAUCUCUUGUUAGAUCUUUUUGCUAAUCUA
and Ns5a (begins at 27792)

KF923898 (China, 2012)[b]
Between S (ends at 2728) 27748 27788
and Ns5a (begins at 27804) CACUGAUCUUUUGUUAGAUCUUUUUGCUAAUCUAGCAUUUAU

OK318944 (China, 2019)[c]
Between S (ends at 27710) 27730 27774
and Ns5a (begins at 27793) CACUGAUCUUUUGUUAGAUCUUUUUGCUAAUCUAGCAUUUAUUAAA

Figure 9.

A.



TRS-L

hCoV-HKU-1 (β-CoV Embecovirus)

5'-UTR (AY597011; nts. 43-75) **AUCUCUUGUCAGAUCUCAUUAAAUCUAAACUUU**

KY674943 (USA, 2016)[a] **26996** **27028**

Between S (ends at 26994) **AUCUCUUGUCAGAUCUCAUUAAAUCUAAACUUU**

and Ns4 (begins at 27033)

B.

Bat β-CoV Nobecovirus

5'-UTR (OK067321; nts. 1-55) to nsp3 (nts. 6837-6891)

5'-UTR **1** **55**

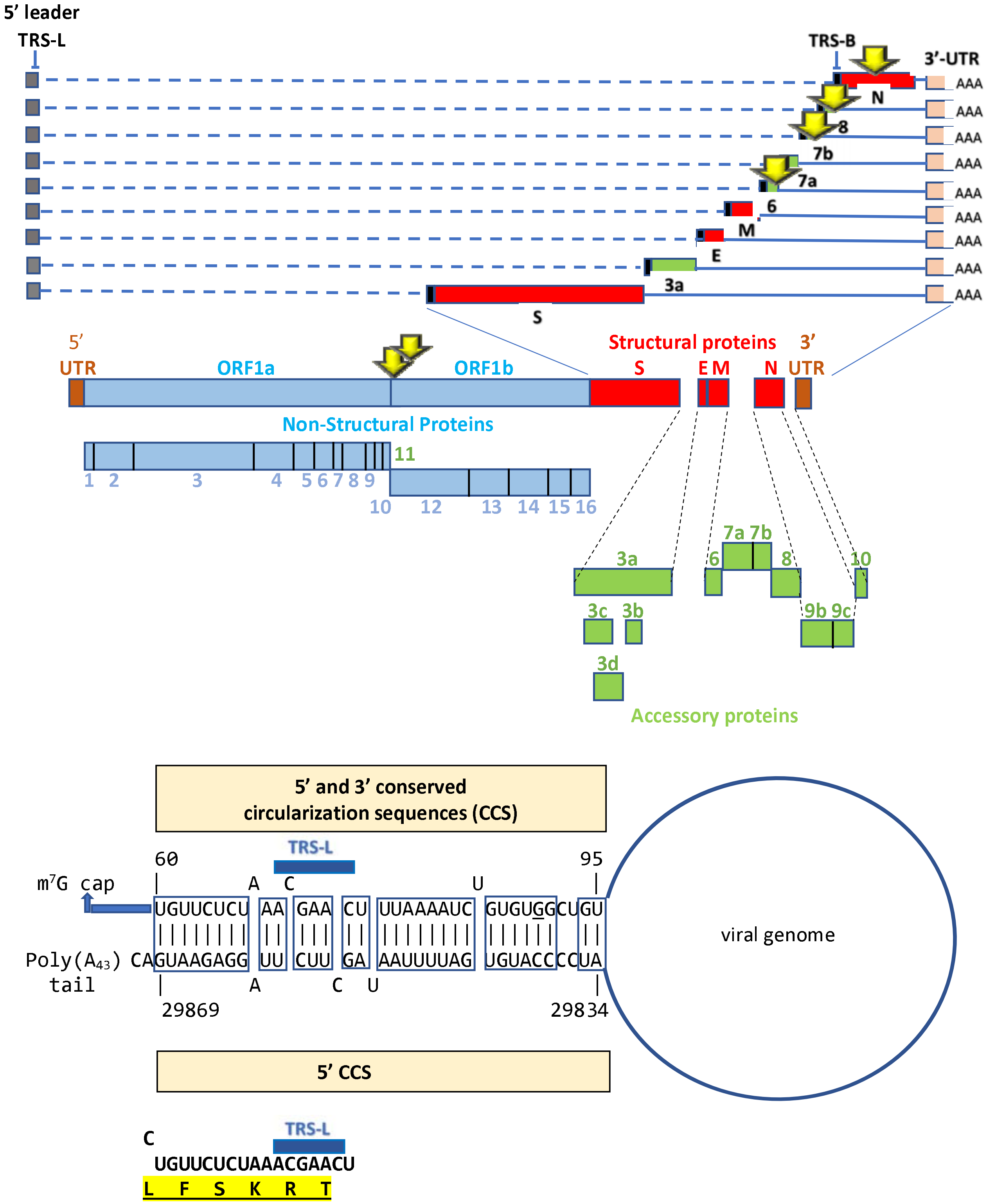
Nsp3 **6837** **6891**

UAUAGCCCUCUCAUUUUUAUGGGUGUGCUAUAGAGGUUUGUGCCAUGUUAGAUUU

I A L S F L W V C Y R G L C H V R F

2188 **2205**

Figure 10.


5’ leader
TRS-L
TRS-B
3’-UTR
AAA
N
8
7b
7a
6
M
E
3a
S
5’ UTR
ORF1a
ORF1b
Structural proteins
S
E M
N
3’ UTR
Non-Structural Proteins
11
1 2 3 4 5 6 7 8 9
10 12 13 14 15 16
3a
7a 7b
6
8
10
3c 3b
9b 9c
3d
Accessory proteins
5’ and 3’ conserved circularization sequences (CCS)
TRS-L
60
95
m⁷G cap
A C U
UGUUCUCU AA GAA CU UUAAAAUC GUGUGGCUGU
Poly(A43) CA GUAAGAGG UU CUU GA AAUUUUAG UGUACCCCUA
tail
A C U
29869
29834
viral genome
5’ CCS
C
TRS-L
UGUUCUCUAAACGAACU
L F S K R T