Virus Evolution, 2021, 7(1): veab021

doi: 10.1093/ve/veab021
Research Article

Structural phylogenetic analysis reveals lineage-specific RNA repetitive structural motifs in all coronaviruses and associated variations in SARS-CoV-2

Shih-Cheng Chen,[1] René C. L. Olsthoorn,[2] and Chien-Hung Yu[1,*,†]

[1]Department of Biochemistry and Molecular Biology, College of Medicine, National Cheng-Kung University, No.1, University Road, Tainan City 701, Taiwan and [2]Department of Supramolecular Biomaterials Chemistry, Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, Einsteinweg 55, 2333 CC, Leiden,The Netherlands

*Corresponding author: E-mail: chienhung_yu@mail.ncku.edu.tw

†https://orcid.org/0000-0003-0337-4473

Abstract

In many single-stranded (ss) RNA viruses, the *cis-acting* packaging signal that confers selectivity genome packaging usually encompasses short structured RNA repeats. These structural units, termed repetitive structural motifs (RSMs), potentially mediate capsid assembly by specific RNA–protein interactions. However, general knowledge of the conservation and/or the diversity of RSMs in the positive-sense ssRNA coronaviruses (CoVs) is limited. By performing structural phylogenetic analysis, we identified a variety of RSMs in nearly all CoV genomic RNAs, which are exclusively located in the 5′-untranslated regions (UTRs) and/or in the inter-domain regions of poly-protein 1ab coding sequences in a lineage-specific manner. In all alpha- and beta-CoVs, except for *Embecovirus* spp, two to four copies of 5′-gUUYCGUc-3′ RSMs displaying conserved hexa-loop sequences were generally identified in Stem-loop 5 (SL5) located in the 5′-UTRs of genomic RNAs. In *Embecovirus* spp., however, two to eight copies of 5′-agc-3′/guAAu RSMs were found in the coding regions of non-structural protein (NSP) 3 and/or NSP15 in open reading frame (ORF) 1ab. In gamma- and delta-CoVs, other types of RSMs were found in several clustered structural elements in 5′-UTRs and/or ORF1ab. The identification of RSM-encompassing structural elements in all CoVs suggests that these RNA elements play fundamental roles in the life cycle of CoVs. In the recently emerged SARS-CoV-2, beta-CoV-specific RSMs are also found in its SL5, displaying two copies of 5′-gUUUCGUc-3′ motifs. However, multiple sequence alignment reveals that the majority of SARS-CoV-2 possesses a variant RSM harboring SL5b C241U, and intriguingly, several variations in the coding sequences of viral proteins, such as Nsp12 P323L, S protein D614G, and N protein R203K-G204R, are concurrently found with such variant RSM. In conclusion, the comprehensive exploration for RSMs reveals phylogenetic insights into the RNA structural elements in CoVs as a whole and provides a new perspective on variations currently found in SARS-CoV-2.

Key words: coronavirus; repetitive structural motifs; structural phylogenetics; SARS-CoV-2.

1. Introduction

Coronaviruses (CoVs) are positive-strand RNA viruses that belong to the subfamily *Orthocoronavirinae* within the *Coronaviridae*. They were not considered highly pathogenic to humans as they mostly cause mild symptoms, until the outbreak of severe acute respiratory syndrome (SARS)-CoVs in

2002–03 (Ksiazek et al. 2003; Rota et al. 2003; Sorensen et al. 2006; Cui et al. 2019). Over the past decade, more highly pathogenic novel CoVs, such as Middle East respiratory syndrome (MERS)-CoV, Swine acute diarrhoea syndrome-CoV, and the recent SARS-CoV-2, have emerged, seriously impacting global health and economy (Guan et al. 2003; Hemida et al. 2013; Zhou et al. 2018; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses 2020; Huang et al. 2020). Before 2003, there were only ten CoVs with complete genome sequences available (Woo et al. 2009). Since the outbreak of SARS-CoV, extensive studies have driven discovery of unidentified CoVs globally (Perlman and Netland 2009; Woo et al. 2009; Woo et al. 2012a,b; Graham et al. 2013; Hu et al. 2015; Su et al. 2016; Forni et al. 2017). To date, numerous CoVs have been identified and classified into four genera, *Alpha-*, *Beta-*, *Gamma-*, and *Deltacoronavirus*, in the subfamily *Orthocoronaviridae*. Infections of alpha- or beta-CoVs usually cause respiratory illness and gastroenteritis in mammals, while gamma- and delta-CoVs are mostly infecting birds (Woo et al. 2009; Woo et al. 2012a; Wang et al. 2015; Cui et al. 2019; Han et al. 2019; So et al. 2019). Recently, the four genera of CoVs were further classified into twenty-three subgenera basing on phylogenetic relationships and genome structures according to CoVs Study Group of International Committee for Taxonomy of Viruses.

Previously, we have reported that the 5′-proximal sequences of alpha- and beta-CoV genomic RNAs (gRNAs) consist of several structural elements, of which the locations, sizes, variations, homology, etc., were found highly conserved in a lineage-specific manner (Chen and Olsthoorn 2010). These RNA elements, designated as stem-loop (SL) 1 to 6, have been correlated to their *cis*-acting regulatory functions in viral protein synthesis, genome replication, and genome packaging (Escors et al. 2003; Raman et al. 2003; Raman and Brian 2005; Wu et al. 2006; Brown et al. 2007; Liu et al. 2007; Li et al. 2008; Liu et al. 2009; Guan et al. 2012; Keane et al. 2012; Tan et al. 2012; Yang and Leibowitz 2015; Madhugiri et al. 2016). Among SL1 to 6, SL5 is particularly interesting for that its tripartite apical substructures, SL5a-c, displays conserved 5’-UUYCGU-3’ sequences in nearly all group I (alpha-) and II (beta-) CoVs discovered before 2010 (Chen and Olsthoorn 2010). Exceptions are the group IIa beta-CoVs (*Embecovirus* spp.), such as Murine hepatitis virus (MHV) and Bovine CoV (BCoV), which are lacking these 5’-UUYCGU-3’ containing hairpins altogether. Instead, group IIa beta-CoVs exclusively possess a long stem-bulge element in the coding sequences of non-structural protein (NSP) 15 of open reading frame (ORF) 1ab, which encompasses multiple copies of AA-bulges and has been reported to serve as the packaging signal (PS) (Fosmire et al. 1992; Molenkamp and Spaan 1997; Woo et al. 1997; Cologna and Hogue 1998; Chen et al. 2007b; Kuo and Masters 2013). These findings suggest that RNA elements encompassing short structured repeats, termed repetitive structural motifs (RSM), are generally present in CoV gRNAs as they may facilitate conserved vital processes in CoVs’ life cycle, e.g. genome packaging (Kuo and Masters 2013; Masters 2019), by specific interaction with protein and/or RNA factors. Thus, knowledge of the homology and diversity of RSMs and their lineage-specificity in CoVs are of importance in both functional and evolutionary perspectives. Although the diversity of CoVs has been profoundly explored over the past decade, however, our general understanding of their RSMs is limited.

In this study, we have performed a comprehensive structural phylogenetic analysis for gRNAs of all species in subfamily *Orthocoronavirinae* to globally explore RSMs and the encompassing elements. Structural features of these elements were studied for their general homology, diversity, and lineage-specificity. In particular, conservation and variation of the RSMs were also verified in the currently pandemic SARS-CoV-2.

2. Material and methods

2.1 Structural phylogenetic analysis of CoV gRNA sequences

Complete gRNA sequences of CoVs were acquired from the nucleotide database at the National Center for Biotechnology Information (NCBI). Multiple sequence alignments for protein and RNA were processed by Clustal Omega (Sievers et al. 2011) provided by European Bioinformatics Institute (EMBL-EBI) web server. RNA secondary structure was primarily predicted by web server Mfold (Zuker 2003) and R-scape (RNA significant covrination above phylogenetic expectation) (Rivas et al. 2017, 2020; Rivas 2020).

2.2 Pipeline for the identification of RSMs in alpha-, beta-, delta-, and gamma-CoVs

To exploit RSMs in CoVs, we aligned 1,292, 28,775, 163, and 452 complete gRNA sequences of alpha-, beta-, delta-, and gamma-CoVs, respectively, for the discovery of sequence insertions and RSMs in 5′-UTRs and coding regions. Each discovered insertions and their flanking regions were subjected to RNA secondary structure predictions by Mfold web server (Zuker 2003). In parallel, particular inter-domain sequences located in between two highly conserved protein domains were also subjected to RNA structure prediction. Predicted RNA structures were then manually refined according to structural co-variations found in closely related strains/isolates/species of CoVs. All the refined secondary structures were documented for the exploration of RSMs and structural phylogenetic analysis.

2.3 Conservation of RSMs and the protein variations concurrent with variant SL5b

To verify the conservation of RSMs in SARS-CoV, MERS-CoV, and SARS-CoV-2, 340, 598, and 19,120 sequences corresponding to SL5s located in 5′-UTRs were aligned and statistically analyzed, respectively. To identify potential concurrent variations between SARS-CoV-2 SL5 and the coding sequences, 14,118 complete genomic sequences of SARS-CoV-2 were analyzed. These gRNA sequences were clustered according to the variations of RSM located in SL5a-c. Each cluster was analyzed for recurring variations in the coding region that are concurrent with the variant SL5b.

3. Results

3.1 Presence of repetitive structural RNA motifs in gRNAs of CoVs

The existence of RSMs is a common feature in the four genera of CoVs we have studied. However, the conserved sequences, copy numbers, locations in the gRNA, etc., of these RSMs are divergent among different lineages (Table 1). The consensus secondary structures of RSM-encompassing elements are demonstrated for each subgenus in genus *Alpha-*, *Beta-*, *Gamma-*,

Table 1. RSM in CoVs.

Genus	Subgenus	Species (representative)	Hosts	RSMs	Copies	Location in gRNA
Alphacoronavirus	*Colacovirus*	Bat CoV CDPHE15	Bats	gUUCCGUc	2	5’-UTR
	Decacovirus	Bat CoV HKU10	Bats	gUUCCGUc	3	5’-UTR (SL5)
	Duvinacovirus	Human CoV 229E	Humans	gUUCCGUc	3	5’-UTR (SL5)
	Luchacovirus	Lucheng Rn rat CoV	Rats	gUUCCGUc	3-4	5’-UTR (SL5, SL5.1)
	Minacovirus	Mink CoV 1	Minks, ferrets	gUUCCGUc	3	5’-UTR (SL5)
	Minunacovirus	Miniopterus bat CoV HKU8	Bats	gUUCCGUc	3	5’-UTR (SL5)
	Myotacovirus	Myotis ricketti alphaCoV Sax-2011	Bats	gUUCCGUc	3	5’-UTR (SL5)
	Nyctacovirus	BtNv-AlphaCoV/ SC2013	Bats	gUUCCGUc	3	5’-UTR (SL5)
	Pedacovirus	Porcine epidemic diarrhea virus	Bats, pigs	gUUCCGUc	3	5’-UTR (SL5)
	Rhinacovirus	Rhinolophus bat CoV HKU2	Bats	gUUCCGUc	3	5’-UTR (SL5)
	Setracovirus	Human CoV NL63	Bats, humans	gUUCCGUc	3	5’-UTR (SL5)
	Tegacovirus	Transmissible gastroenteritis virus	Pigs, dogs, cats	gUUCCGUc	3	5’-UTR (SL5)
Betacoronavirus	*Embecovirus*	Murine hepatitis virus	Mice, rats, bats, humans, deers, horses, camels, rabbits, *etc.*	agc/guAAu	2-8	ORF1ab(nsp15, nsp3)
	Hibecovirus	Bat Hp-betaCoV Zhejiang2013	Bats	gUUUCGUc	3	5’-UTR (SL5)
	Merbecovirus	Middle East respiratory syndrome-related CoV	Bats, camels, humans	gUUUCGUc	2	5’-UTR (SL5)
	Nobecovirus	Rousettus bat CoV HKU9	Bats	gUUUCGUc	2-3	5’-UTR (SL5, SL5.1)
	Sarbecovirus	Severe acute respiratory syndrome-related CoV	Bats, humans	gUUUCGUc	2	5’-UTR (SL5)
Gammacoronavirus	*Cegacovirus*	Beluga whale CoV SW1	Whales, dophins	uacUUCGgug	2	ORF1ab(nsp3)
	Igacovirus	Avian infectious bronchitis virus	Birds	uGCUAa	2-3	ORF1ab(nsp3)
Deltacoronavirus	*Andecovirus*	Wigeon CoV HKU20	Birds	uGGUa	3	ORF1ab(nsp13/14)
	Buldecovirus	Thrush CoV HKU12-600	Birds	aGUACu, ac/gAGUu	5, 2	ORF1ab(nsp13/ 14)’ 5’-UTR
	Herdecovirus	Night heron CoV HKU19	Birds	aGUACu	4	ORF1ab(nsp13/14)

The RSMs identified in gRNAs of CoVs in each subgenus are listed.

and *Deltacoronavirus*, respectively, with multiple sequence alignments of the corresponding regions that show the locations of these elements in gRNA (Figs. 1–6).

3.2 The RSM-encompassing SL5s located in 5′-proximal regions of alpha-CoV gRNAs

Structural elements consisting of RSMs were generally identified in the 5′-proximal sequences of all alpha-CoV gRNAs, which were assigned as SL5 for being the fifth rigid structural element identified in CoVs from the 5′-end of gRNAs (Chen and Olsthoorn 2010). Figure 1 shows the consensus secondary structures of SL5s in *Decacovirus*, *Luchacovirus*, *Minacovirus*, *Myotacovirus*, and *Nyctacovirus*. In general, these alpha-CoVs' SL5s consist of a long basal stem at the bottom encompassing the AUG start codon of ORF1ab and apical tripartite hairpins (assigned as SL5a, SL5b, and SL5c) that encompass conserved 5′-gUUCCGUc-3′ RSMs. The presence of 5′-gUUCCGUc-3′ RSMs in SL5a-c is highly consistent in all alpha-CoVs, while co-variations are generally found supportive for base-pairing in the

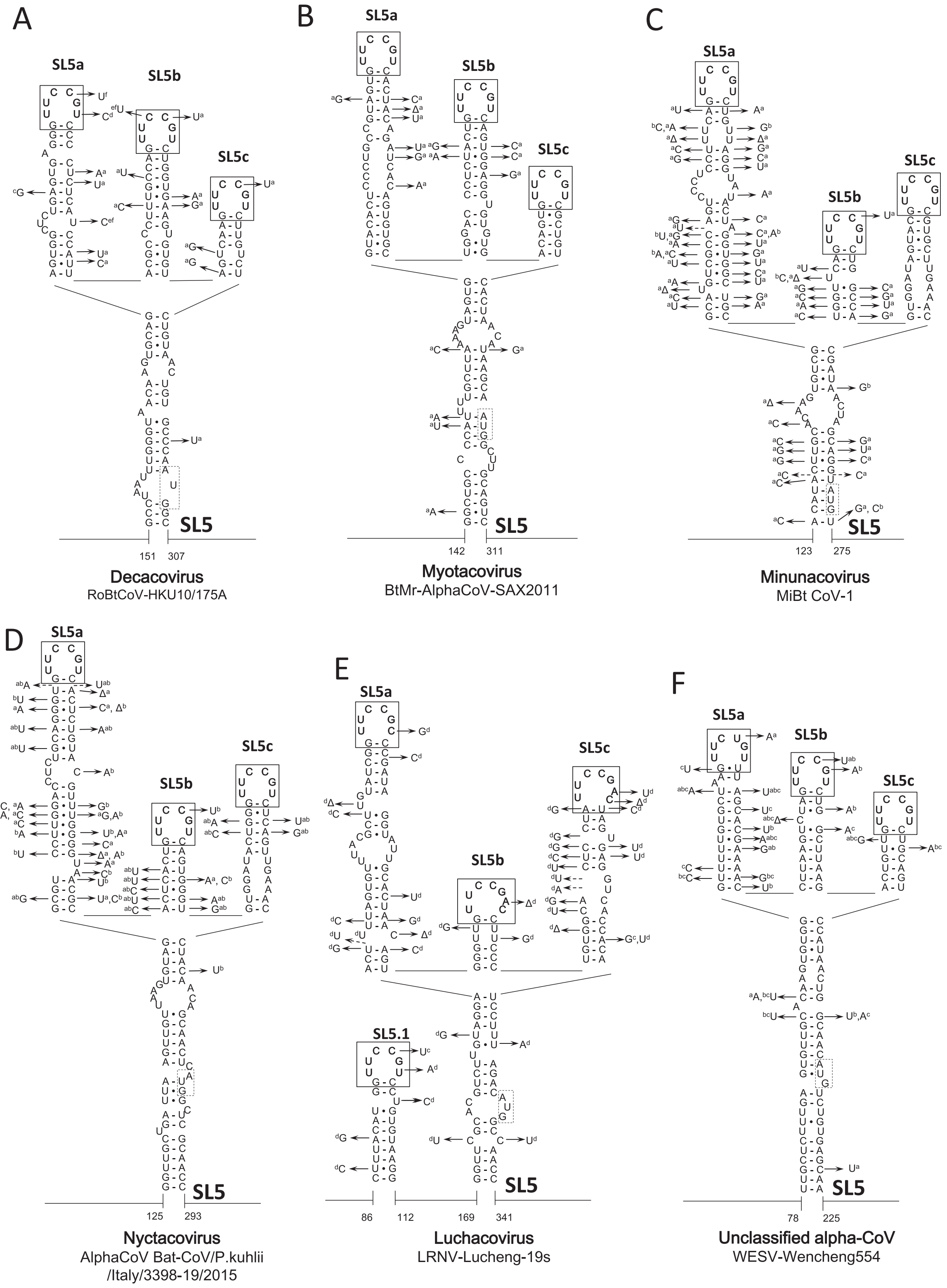

Figure 1. Secondary structures of RSM-encompassing SL5s in the 5′-proximal sequences of alpha-CoV gRNAs. Consensus secondary structures of alpha-CoV SL5s are shown, in which the conserved 5′-gUUCCGUc-3′ RSMs are boxed with solid lines. The start codons of ORF1ab are boxed with dashed lines. Sequence variations between the representative viruses and related species are indicated by arrows to show the variations. Nucleotide deletions are indicated with symbol 'Δ', while insertions are indicated by dashed-line arrows. Numbering represent the positions of the indicated nts counting from the 5′-end of gRNAs. (A) *Decacovirus*: Rousettus bat CoV (RoBtCoV)-HKU10/175A (JQ989271.1); [a]Hipposideros Pomona bat CoV (HiBtCoV)-HKU10/160942 (MN611523.1); [b]RoBtCoV-HKU10/183A (JQ989270.1); [c]HiBtCoV-HKU10/TLC1347A (JQ989273.1); [d]HiBtCoV-HKU10/TLC1343A (JQ989272.1); [e]HiBtCoV-HKU10/LSH5A(JQ989269.1); [f]HiBtCoV-HKU10/TLC1310A (JQ98969.1). (B) *Myotacovirus*: BtMr-AlphaCoV/SAX2011 (KJ473806.1); [a]BtCoV Anlong-57 (KY770851.1). (C) *Minunacovirus*: Miniopterus bat CoV (MiBtCoV)-1 (EU420138.1); [a]Bat CoV (BtCoV)-HKU8/AFCD77 (EU420139.1); [b]Bat CoV (BtCoV)-1B/AFCD307 (EU420137.1). (D) *Nyctacovirus*: AlphaCoV Bat-CoV/P.kuhlii/Italy/3398-19/2015 (NC_046964.1); [a]AlphaCoV Bat-CoV/P.kuhlii/Italy/206645-41/2011(MH938448.1); [b]BtNv-AlphaCoV/SC2013 (KJ473809.1). I *Luchacovirus*: Lucheng Rn rat coronavirus (LRNV) Lucheng-19 (NC_032730.1); [a]Rodent coronavirus isolate RtRl-CoV/FJ2015 (KY370050.1); [b]Alphacoronavirus_UKRn3 (MK163627.1); [c]Rodent CoV (RtCoV)-RtMruf-CoV-1/JL2014 (KY370045.1); [d]Coronavirus AcCoV-JC34 (KX964649.1). (F) Unclassified alphacoronaviruses; Wencheng Sm shrew CoV (WESV)-Xingguo74 (NC_048211.1); [a]WESV-Yudu76 (KY967723.1); [b]WESV-Ruian90 (KY967725.1); [c]WESV-Wencheng554 (KY967733.1).

stems beneath which potentially stabilize the apical RSMs (Fig. 1). Suggested by recent NMR studies, the 5′-UUCCGU-3′ sequence may adopt a conformation similar to 5′-UUCG-3′ tetraloop (Wacker et al. 2020; Schnieders et al. 2021).

An additional copy of RSM was found in *Luchacovirus*, which is located in SL5.1 upstream to SL5 (Fig. 1E). Possession of such an additional copy of RSM in SL5.1 could potentially make some of these CoVs more tolerant to the imperfection of RSM present in SL5b and c. In some unclassified alpha-CoVs listed in NCBI database, homologous structural motifs could also be identified. For instance, Figure 1F shows the secondary structure of SL5 for the unclassified Wencheng Sm Shrew coronavirus (WESV), which also consists of three copies of the 5′-gUUCCGUc-3′ RSMs. Besides the five subgenera shown in Fig. 1, the lineage-specific structural feature of CoV SL5 in other subgenera, including *Duvinacovirus*, *Pedacovirus*, *Rhinacovirus*, *Setracovirus*, and *Tegacovirus*, has been proposed previously, which all encompass the alpha-CoV-specific RSMs (Chen and Olsthoorn 2010). Collectively, these results suggest that SL5 and its tripartite hairpins SL5a–c are universally present in all species of *Alphacoronavirus*, encompassing conserved 5′-gUUCCGUc-3′ RSMs.

3.3 The RSM-encompassing SL5s in beta-CoV gRNAs

In the 5′-proximal gRNA sequences of beta-CoVs, we have also universally identified RNA elements similar to alpha-CoVs' SL5s (Fig. 2). However, the RSMs found in beta-CoVs are predominantly 5′-gUUUCGUc'-3, unlike 5′-gUUCCGUc-3′ discovered in alpha-CoVs. And the features of SL5s are more diverse in beta-CoVs than in alpha-CoVs, in terms the size, the predicted secondary structures, and the copy numbers of RSMs (Fig. 2). *Hibecovirus* is the only subgenus within the genus *Betacoronavirus* that possesses complete three copies of RSMs in SL5 (Fig. 2A). *Sarbecovirus* spp., including the newly emerged SARS-CoV2, possess only two copies of the beta-CoV-specific RSMs in SL5a and b, respectively, while SL5c exhibits 5′-GAAA-3′ tetra-loop (Fig. 2B). This general feature is in agreement with what we have previously found in SARS-CoV and others reported for SARS-CoV-2 (Chen and Olsthoorn 2010; Miao et al. 2020; Rangan et al. 2020). Likewise, in *Merbecovirus* two copies of RSMs are found situating apically in SL5a and b, respectively, while the SL5c displays 5′-AAGGUGC-3′ hepta-loop (Fig. 2C). These findings suggest that two copies of RSMs could be sufficient for virus propagation of *Sarbecovirus* spp. and *Merbecovirus* spp. On the other hand, it is possible that the 5′-GAAA-3′ and 5′-AAGGUGC-3′ sequences in SL5c mediate specific function(s) for CoVs in *Sarbecovirus* and *Merbecovirus*, respectively, since these two sequences are strictly conserved within the same subgenus. In addition, these two SL5c loop sequences are closed with C-G base-pair, unlike the conserved RSMs in SL5a-b that are closed with G-C pairing. Besides the divergent features of SL5c, we have noticed that in *Merbecovirus* the AUG start codon of ORF1ab is located at the stem region of SL5b, instead of at the upper part of the basal stem of SL5 like other beta-CoVs (Fig. 2C). Interestingly, some species in *Merbecovirus*, *e.g.* BtVs-BetaCoVs and HpBtCoVs, possess a downstream in-frame AUG in SL5 basal stem (Fig. 2C). However, these CoVs still preserve the start codons in SL5b like all the other *Merbecovirus* spp., showing no sequential evidence that the downstream AUG could be used for translation initiation.

In *Nobecovirus*, there are only two copies of RSMs present in SL5 (Fig. 2D). Interestingly, an additional RSM containing element, SL5.1, is located upstream to SL5, similar to what we have found in *Luchacovirus* spp in *Alphacoronavirus*. Possession of this extra copy of RSM potentially explains how these CoVs evolved with a missing SL5c and/or a 1-nt deletion in SL5b, which is an extremely rare occasion that cannot be found in other alpha- or beta-CoVs having no additional RSMs in SL5.1. In *Embecovirus*, SL5, which encompasses the AUG start codon of ORF1, is relatively short compared with other beta-CoVs (Fig. 2E), missing the tripartite SL5a-c that encompass RSMs. Alternatively, a larger secondary structure, assigned as SL5* (Supplementary Fig. S1), was predicted, if the long-range RNA-RNA interaction between 5′UTR and NSP1 sequences is taken into consideration (Guan et al. 2012; Yang et al. 2015). Nonetheless, the apical regions of the tripartite SL5-7 in SL5* are divergent and do not contain any forms of RSMs (Supplementary Fig. S1).

We conclude that all alpha- and beta-CoVs (Figs. 1 and 2), except for *Embecovirus* spp. (Fig. 2E), possess 5′-gUUYCGUc-3′ RSMs in the apical regions of SL5a-c. Supplementary Fig. S2 shows the alignments of SL5a-c corresponding sequences in reference alpha- and beta-CoVs and the structural model predicted by Cascade variation/covariation Constrained Folding (CaCoFold) algorithm with significant co-varying pairs reported by R-scape web server. As for CoVs in subgenus *Embecovirus*, a distinct type of RSMs was exclusively found (see below).

3.4 The RSMs located in ORF1ab of gRNAs found in *Embecovirus*

We successfully identified conserved RSM-encompassing elements in all recently discovered CoVs classified in subgenus *Embecovirus* and some unclassified beta-CoVs, including Rabbit CoV (RbCoV)-HKU14, Dromedary camel CoV (DcCoV)-HKU23, Longquan Aa mouse CoV (LAMV), and Longquan Rl rat CoV (LRLV), Rodent CoV (RtCoV), etc. (Lau et al. 2012; So et al. 2019; Wang et al. 2015) (Fig. 3). These elements contain multiple copies of RSMs made up of 5′-agc-3′/5′-guAAu-3′ and 5′-uWWc-3′/5′-gg-3′, displaying four 2-nt bulges (predominantly AA) at the 3′-side and another two (predominantly AU, AA, or UA) at the 5′ side, respectively (Fig. 3A–C). The structural features of these elements are highly similar to what we have reported for the PS of MHV (Fig. 3D) (Chen et al. 2007b), of which the structural integrity is crucial for selective genome packaging (Kuo and Masters 2013; Athmer et al. 2018; Masters 2019). Sequence alignment of the corresponding regions further showed that all of these recently identified elements are located in NSP15 coding regions where the MHV PS is situated (Fig. 3H) (Xu et al. 2006; Chen et al. 2007b). Thus, these RSM containing elements are structural homologues that perfectly resemble the canonical PS of the type beta-CoV, MHV, displaying four and two copies of the 2-nt-bulge on the 3′ and 5′ sides, respectively, and the strictly conserved 5′-CACAA-3′ apical loop sequence (Fig. 3A–D). Interestingly, non-canonical PSs were found in some CoVs recently discovered in rodents, which are truncated and exhibit less copies of RSMs. Masters (2019) has proposed that the PS in Equine CoV (EqCoV) strain NC99 has precise deletions of the central bulge and one of the RSM, yet the conserved 5′-CACAA-3′ apical loop sequence is still present (Fig. 3E). In the two recently isolated Rodent CoVs (RtCoVs), we have found that their apical regions of PSs are partially truncated (Fig. 3F and G), preserving less copies of RSMs. Notably, these two RtCoV PSs are the exceptional examples of PSs having non-5′-CACAA-3′ apical loop sequences. Figure 3H shows the alignment of partial poly-protein 1ab sequences corresponding to the PSs. Clearly, the sequence insertions corresponding to either canonical or non-canonical PSs are exclusively present in NSP15 coding regions of *Embecoviruses* spp. but absent in all the other four subgenera (Fig. 3H).

In addition to the PSs described above, it has been suggested that the *Embecoviruses*-specific RSMs, the 5′-agc-3′/5′-guAAu-3′

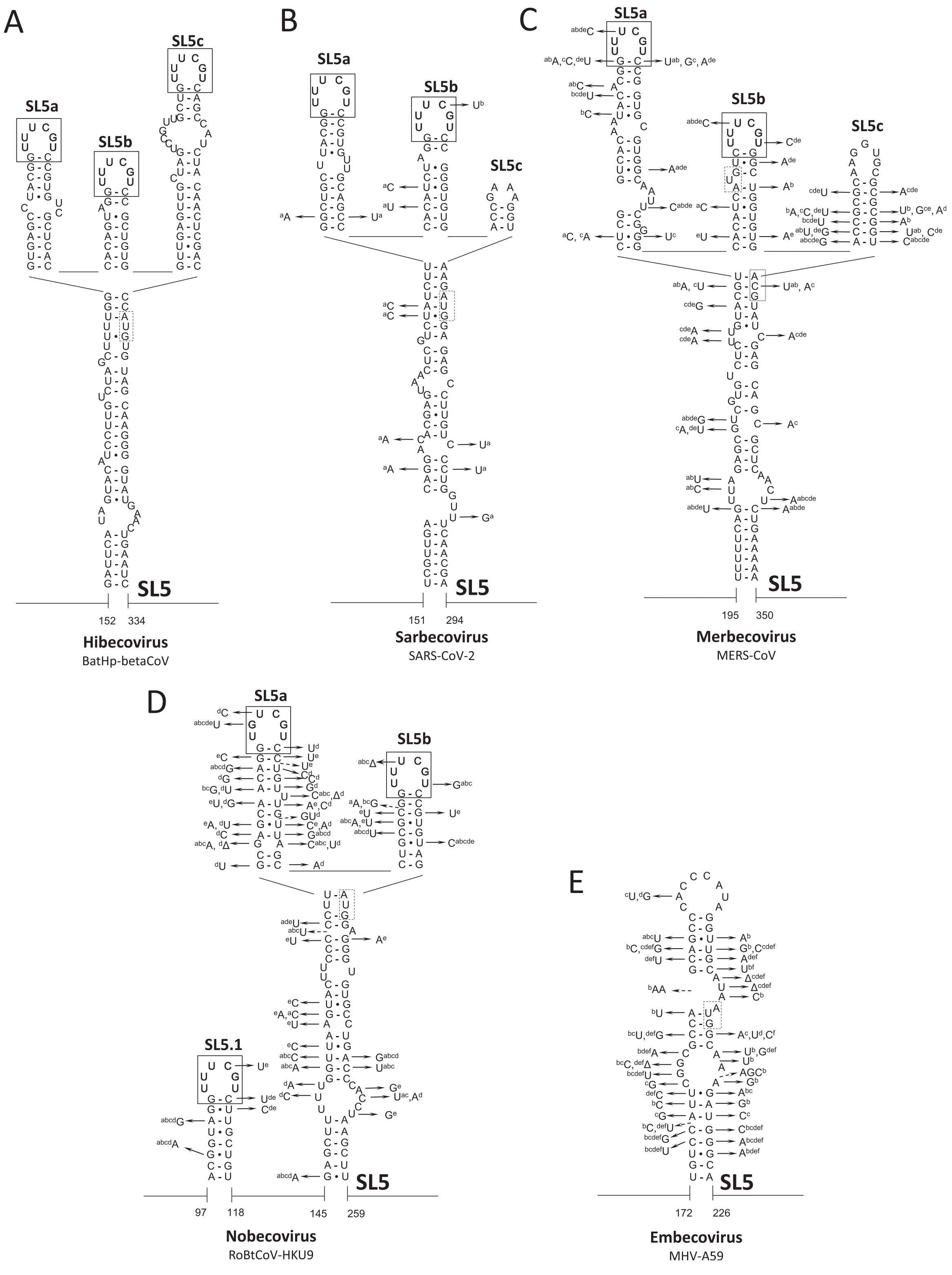

Figure 2. Secondary structures of RSM containing SL5s in the 5′-proximal sequences of beta-CoV gRNAs. Consensus secondary structures of beta-CoV SL5s are shown, in which the conserved 5′-gUUUCGUc-3′ RSMs are boxed with solid lines. Start codons of ORF1ab are boxed with dashed lines, while the downstream in-frame AUG codons found in few *Merbecovirus* spp. Are boxed with dot-lines. Sequence variations between the representative viruses and related species are indicated by arrows to show the variations. Nucleotide deletions are indicated with symbol 'Δ', while insertions are indicated by dashed-line arrows. Numbering represent the positions of the indicated nts counting from the 5′-end of gRNAs. (A) *Hibecovirus*: Bat Hp-betaCoV/Zhejiang2013 (NC_025217.1). (B) *Sarbecovirus*: SARS CoV-2/Wuhan-Hu-1 (NC_045512.2); ^{a}SARS-CoV/Tor2 (NC_004718.3); ^{b}SARS CoV-2/human/USA/Oklahoma-ADDL-1/2020 (MT998442.1). (C) *Merbecovirus*: MERS-CoV/EMC/2012 (NC_019843.3); ^{a}BtVs-BetaCoV/SC2013 (KJ473821.1); ^{b}Hypsugo bat CoV (HpBtCoV)-HKU25/NL140462 (KX442565.1); ^{c}BtPa-BetaCoV/GD2013 (KJ473820.1); ^{d}Erinaceus hedgehog CoV (EaHedCoV)-HKU31/Rs13 (MK907287.1); ^{e}BetaCoV-Erinaceus/2012-174/GER/2012 (NC_039207.1). (D)*Nobecovirus*: Rousettus bat CoV (RoBtCoV)-HKU9 (NC_009021.1); ^{a}BtCoV-HKU9-1 (EF065515.1); ^{b}BtCoV-HKU9-10-1 (HM211100.1); ^{c}BtCoV-HKU9-5-1 (HM211098.1); ^{d}RoBtCoV-GCCDC1/365 (NC-030886.1); ^{e}BtRt-BetaCoV/GX2018 (MK211379.1). (E) *Embecovirus*: Mouse hepatitis virus (MHV)-A59 (NC_048217.1); ^{a}RatCoV-Parker (NC_012936.1); ^{b}BetaCoV-HKU24/R05005I (NC_017083.1); ^{c}Human CoV (HuCoV)-HKU1 (NC_006577.2); ^{d}HuCoV-OC43/ATCC-VR759 (NC_006213.1); ^{e}Bovine CoV (BCoV)-ENT (NC_003045.1); ^{f}Rabbit CoV (RbCoV)-HKU14 (NC_017083.1). (F)MHV-A59 (NC_048217.1); ^{a}HuCoV-HKU1 (NC_006577.2); ^{b}BCoV-2014-13 (KX982264.1); ^{c}Dromedary camel CoV (DcCoV)-HKU23; ^{d}HuCoV-OC43-ATCC/VR759 (NC_006213.1); ^{e}EqCoV-NC99 (EF446615.1); ^{f}RbCoV-HKU14 (NC_017083.1); ^{g}MHV-3 (FJ647224.1); ^{h}BetaCoV-HKU24-R05005I (NC_026011.1).

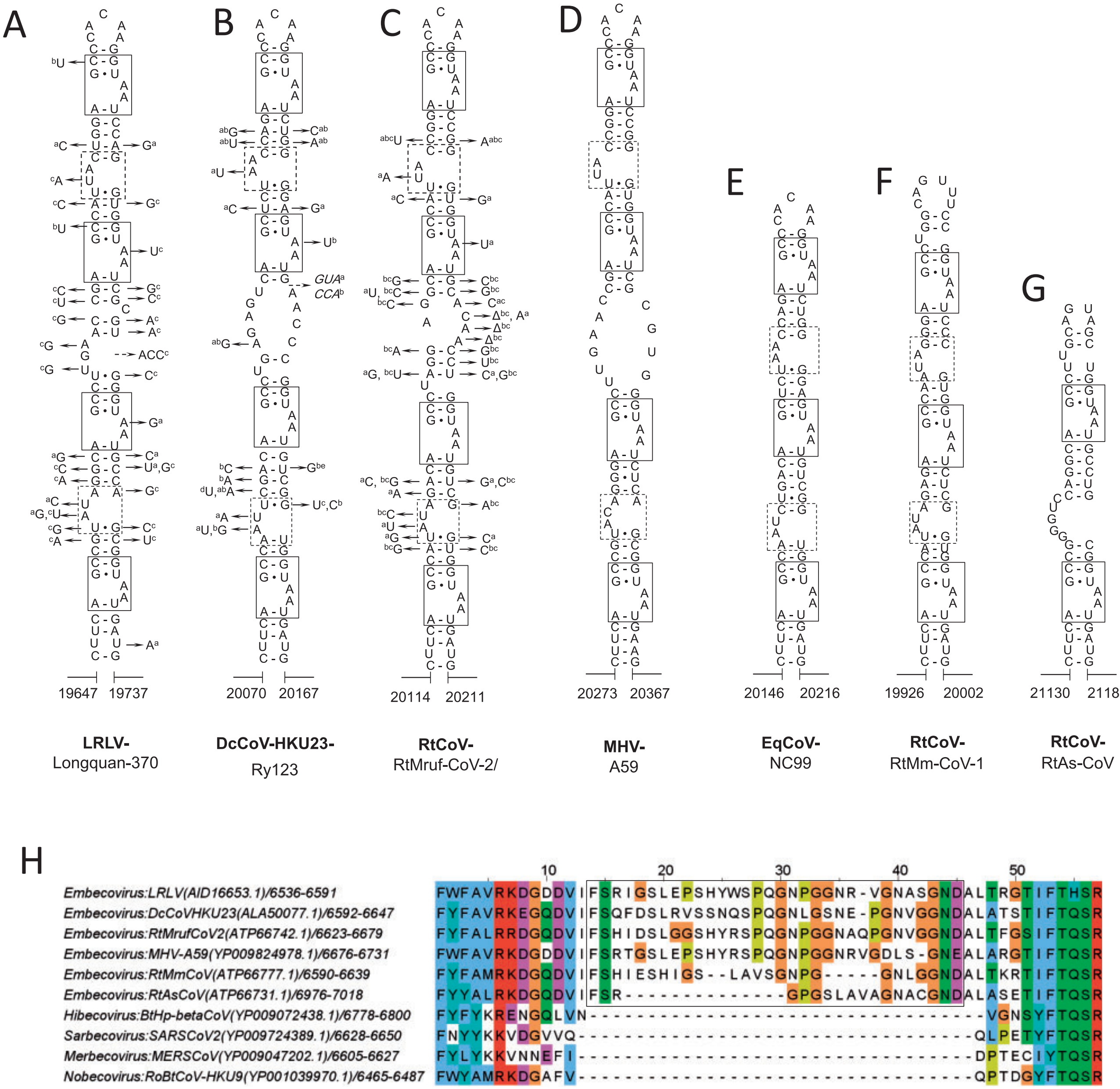
A
B
C
D
E
F
G
19647 19737
20070 20167
20114 20211
20273 20367
20146 20216
19926 20002
21130 21182
LRLV-
Longquan-370
DcCoV-HKU23-
Ry123
RtCoV-
RtMruf-CoV-2/
MHV-
A59
EqCoV-
NC99
RtCoV-
RtMm-CoV-1
RtCoV-
RtAs-CoV
H
10 20 30 40 50
Embecovirus:LRLV(AID16653.1)/6536-6591 FWFAVRKDGDDVIFSRIGSLEPSHYWSPQGNPGGNR-VGNASGNDALTRGTIFTHSR
Embecovirus:DcCoVHKU23(ALA50077.1)/6592-6647 FYFAVRKEGQDVIFSQFDSLRVSSNQSPQGNLGSNE-PGNVGGNDALATSTIFTQSR
Embecovirus:RtMrufCoV2(ATP66742.1)/6623-6679 FYFALRRDGQDVIFSHIDSLGGSHYRSPQGNPGGNAQPGNVGGNDALTFGSIFTQSR
Embecovirus:MHV-A59(YP009824978.1)/6676-6731 FWFAVRKDGDDVIFSRTGSLEPSHYRSPQGNPGGNRVGDLS-GNEALARGTIFTQSR
Embecovirus:RtMmCoV(ATP66777.1)/6590-6639 FYFAMRKDGQDVIFSHIESHIGS--LAVSGNPG-----GNLGGNDALTKRTIFTQSR
Embecovirus:RtAsCoV(ATP66731.1)/6976-7018 FYYALRKDGQDVIFSR--------------GPGSLAVAGNACGNDALASETIFTQSR
Hibecovirus:BtHp-betaCoV(YP009072438.1)/6778-6800 FYFYKRENGQLVN----------------------------------VGNSYFTQSR
Sarbecovirus:SARSCoV2(YP009724389.1)/6628-6650 FNYYKKVDGVVQ----------------------------------QLPETYFTQSR
Merbecovirus:MERSCoV(YP009047202.1)/6605-6627 FYLYKKVNNEFI----------------------------------DPTECIYTQSR
Nobecovirus:RoBtCoV-HKU9(YP001039970.1)/6465-6487 FWYAMRKDGAFV----------------------------------QPTDGYFTQSR

Figure 3. Secondary structures of RSM-encompassing PS located in NSP15 coding region in gRNAs of *Embecovirus* spp. Consensus secondary structures of the RNA elements identified in NSP15 coding sequences of several newly discovered beta-CoVs are shown. The conserved 5′-agc-3′/5′-guAAu-3′ and 5′-uWWc-3′/5′-gg-3′ RSMs are boxed with solid and dashed lines, respectively, each displaying a 2-nt bulge at the 3′ or the 5′-side of the element. Sequence variations between the representative viruses and related species are indicated by arrows to show the variations. Nucleotide deletions are indicated with symbol 'Δ', while insertions are indicated by dashed-line arrows. Numbering represent the positions of indicated nts counting from the 5′-end of gRNAs. (A) Longquan Rl rat coronavirus (LRLV)-Longquan-370 (KF294371.1); [a]LRLV/Longquan-708 (KF294372.1); [b]LRLV/Longquan-189 (KF294370); [c]Longquan Aa mouse coronavirus (LAMV)-Longquan-343 (KF294357.1). (B) Dromedary camel CoV (DcCoV)-HKU23-Ry123; [a]RbCoV-HKU14-1 (JN874559.1); [b]BetaCoV-HKU24/R05010I (KM349744.1); [c]CoV-HKU23/NV1097(MN514966.1); [d]DcCoV-HKU23/NV1385 (MN514967.1). (C) RtCoV-RtMrufCoV-2/JL201(NC_046954); [a]RtCoV-RtApCoV/Tibet2014 (KY370047); [b]RtCoV-RtNnCoV/SAX2015 (KY370049); [c]RtCoV-RtRnCoV/YN2013 (KY370043), [d]RtCoV-RtMmCoV/GD2015 (KY370048), [e]RtCoV-RtBiCoV/FJ2015 (KY370051). (D) Packaging Signal of MHV-A59. (E) Equine CoV/NC99 (EF446615.1). (F) RtCoV-RtMmCoV-1/IM2014 (KY370052.1). (G) RtCoV-RtAsCoV/IM2014 (KY370044.1). (H) Multiple sequence alignment for partial sequences of poly-protein 1ab is shown to demonstrate the sequence insertions corresponding to the RSM-encompassing PS in NSP15 coding region. The corresponding regions of PSs in NSP15 are boxed.

units, are present elsewhere in one and three isolates of EqCoVs and RbCoVs, exhibiting one and three additional copies of the AA-bulges, respectively (Masters 2019). In fact, by our thorough survey more RSMs can be found in all these RbCoVs and EqCoVs. Figure 4 shows the RNA secondary structures of particular regions of *Embecoviruses* spp. gRNAs, which is located in between nucleic-acid binding (NAB) and beta-CoV-specific (βSM) domains of NSP3 coding region. In RaCoV HKU14-1 and related species, we found four copies of RSMs at the 5′-side and two at the 3′-side (Fig. 4A). Such structural feature is highly similar to the canonical MHV PS, that is, a two-fold quasi-symmetry stem-bulge element (Fig. 3A–D). However, the upper and the lower halves of the MHV PSs are separated by ∼6 mismach internal loops, but in these newly identified PS-like elements, there are ∼30 Watson–Crick (WC) and non-WC base pairs in between the upper and the lower halves in RbCoVs (Fig. 4A) and EqCoVs (Fig.

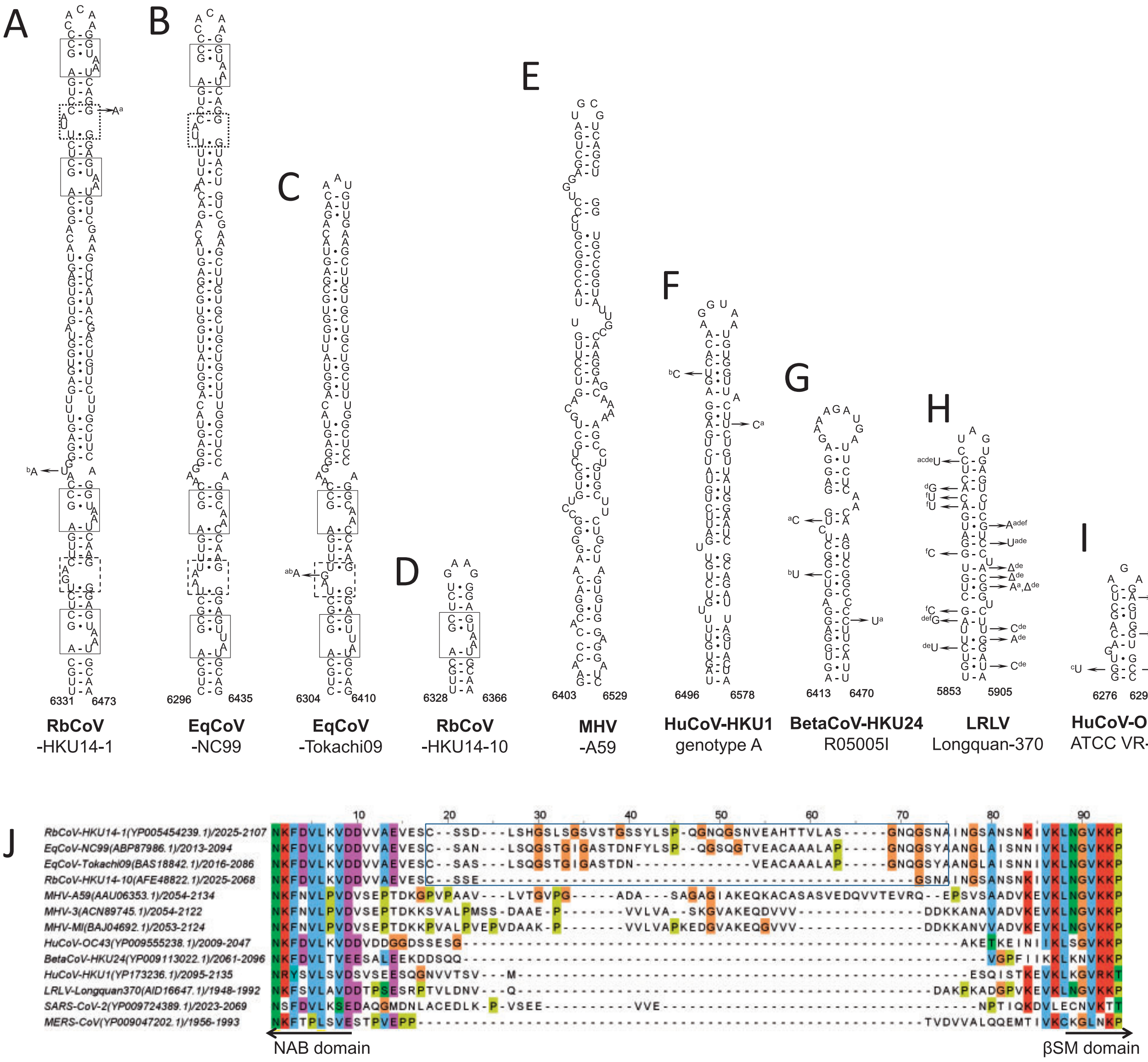
A
B
C
D
E
F
G
H
I
J
6331 6473
6296 6435
6304 6410
6328 6366
6403 6529
6496 6578
6413 6470
5853 5905
6276 6299
RbCoV
-HKU14-1
EqCoV
-NC99
EqCoV
-Tokachi09
RbCoV
-HKU14-10
MHV
-A59
HuCoV-HKU1
genotype A
BetaCoV-HKU24
R05005I
LRLV
Longquan-370
HuCoV-OC43
ATCC VR-759
RbCoV-HKU14-1(YP005454239.1)/2025-2107
EqCoV-NC99(ABP87986.1)/2013-2094
EqCoV-Tokachi09(BAS18842.1)/2016-2086
RbCoV-HKU14-10(AFE48822.1)/2025-2068
MHV-A59(AAU06353.1)/2054-2134
MHV-3(ACN89745.1)/2054-2122
MHV-MI(BAJ04692.1)/2053-2124
HuCoV-OC43(YP009555238.1)/2009-2047
BetaCoV-HKU24(YP009113022.1)/2061-2096
HuCoV-HKU1(YP173236.1)/2095-2135
LRLV-Longquan370(AID16647.1)/1948-1992
SARS-CoV-2(YP009724389.1)/2023-2069
MERS-CoV(YP009047202.1)/1956-1993
10
20
30
40
50
60
70
80
90
NAB domain
βSM domain

Figure 4. Secondary structures of PS-like elements located in between NAB and βSM domains of NSP3 coding regions in gRNAs of *Embecovirus* spp. Secondary structures of the PS-like elements identified in NSP3 coding sequences are shown. The conserved 5′-agc-3′/5′-guAAu-3′ and 5′-uWWc-3′/5′-gg-3′ RSMs are boxed with solid and dashed lines, respectively, each displaying a 2-nt bulge at the 3′ or the 5′-side of the element. Sequence variations between the representative viruses and related species are indicated by arrows to show the variations. Nucleotide deletions are indicated with symbol ‘Δ’, while insertions are indicated by dashed-line arrows. Numbering represent the positions of indicated nts counting from the 5′-end of gRNAs. (A) RbCoV-HKU14-1 (JN874559.1); [a]RbCoV-HKU14-3 (JN874560.1); [b]RbCoV-HKU14-8 (JN874561.1). (B) EqCoV-NC99 (EF446615.1). (C) EqCoV-Tokachi09 (LC061272.1); [a]EqCoV-Obihiro12-1 (LC061273.1); [b]EqCoV-Obihiro12-2 (LC061274.1). (D) RbCoV-HKU14-10 (JN874562.1). (E) MHV-A59 (NC_048217); MHV-3 (FJ647224.1). (F) HuCoV-HKU1/genotype A(NC_006577.2); [a]HuCoV-CHKU1/N091605B (KY674943.1); [b]HuCoV-HKU1/genotype B/N08-87 (KY674921.1). (G) BetaCoV-HKU24/R05010I (KM349742.1); [a]LAMV-Longquan-343 (KF294357.1); [b]RtCoV-RtApCoV/Tibet2014 (KY370047.1). (H) LRLV-370 (KF294371.1); [a]LRLV-708 (KF294372.1); [b]LRLV-189 (KF294370.1); [c]RtBiCoV/FJ2015 (KY370051.1); [d]RtRnCoV/YN2013 (KY370043.1); [e]RtMmCoV/GD2015 (KY370048.1). (I) HuCoV-OC43 (NC_006213.1); [a]HuCoV-OC43/Seattle/USA/SC2476/2015 (KY967360.1); [b]HuCoV-OC43/HZ-459 (MG197723.1); [c]HuCoV-OC43/human/USA/9211-43/1992 (KF530097.1). (J) Multiple sequence alignment of partial poly-protein 1ab at the inter-domain region between NAB and βSM domains of NSP3. The corresponding regions of RSM-encompassing PS-like elements in NSP3 coding sequence are boxed.

4B). We further found that the apical region of the elements seemed to be deleted in EqCoV-Tokachi09, -Obihiro12-1, and -obihiro12-2, resulting in truncated elements encompassing fewer RSMs at the basal stems (Fig. 4C). In RbCoV-HKU14-10, the majority of the element was deleted, preserving only one RSM in the basal stem (Fig. 4D). In other *Embecovirus* spp., sequence insertions at this particular region are divergent in length, yet no potential RSMs were identified (Fig. 4E–I). Besides RbCoVs and EqCoVs, sequence insertions at this particular region are also present in MHV and related species (Fig. 4J), which all fold into long stem-bulge elements with multiple non-WC base pairs, mismatches, and unpaired As (Supplementary Fig. S3). However, none of these elements contain the canonical 2-nt-bulge-displaying RSMs. This finding indicates that the inserted sequences there do not necessarily form any conserved structural motifs, implying that such insertions can be an early event during evolution of *Embecovirus* spp.

Notably, in some recently discovered beta-CoVs isolated from rodents (Wang et al. 2015), including RtApCoVs, BetaCoV-HKU24, and LAMVs, another PS-like element was identified in a distinct location of Nsp3 coding region (Supplementary Fig. S4). This element very much resembles the structural feature of the canonical PS of MHV, particularly the upper part, exhibiting two and one 2-nt bulges at the 3′- and 5′-side, respectively (Supplementary Fig. S4A). Sequence alignment showed that this element is present in the inter-domain region of acidic hypervariable region (HVR) and papain-like protease 1 ($PL1^{pro}$) domains (Supplementary Fig. S4B).

To sum up, possession of RNA elements encompassing the 2-nt-bulge-displaying RSMs is an exclusive hallmark of beta-CoVs in *Embecovirus* (Fig. 3). Such elements are consistently located in Nsp15 coding region corresponding to MHV PS. In some *Embecovirus* spp., additional RSM-encompassing elements are found located in the inter-domain regions of NSP3 NAB/βSM and/or HVR/$PL1^{pro}$ domains (Fig. 4 and Supplementary Fig. S4), exhibiting structural homology to PS.

3.5 The two distinct RSMs found in gamma-CoVs

The genus *Gammacoronavirus* consists of two subgenera, namely *Cegacovirus* and *Igacovirus*. Species in the former have been found in whales and dolphins, while the latter are mostly associated with birds. The RSMs found in *Cegacovirus* spp, for example, Beluga whale CoV (BWCoV)-SW1 and Bottlenose dolphin CoV (BDCoV)-HKU22, are the 5′-uacUUCGgug-3′ apical structural motifs with 5′-CAGG-3′/5′-AA-3′ internal loops (Fig. 5A). In subgenus *Igacovirus*, the 5′-uGCUAa-3′ RSMs were predominantly identified in most species (Fig. 5B), except in the recently discovered Canada goose (CG)-CoV, which has been reported to be divergent from other avian CoVs identified previously (Papineau et al. 2019). Multiple sequence alignment shows that the RSMs

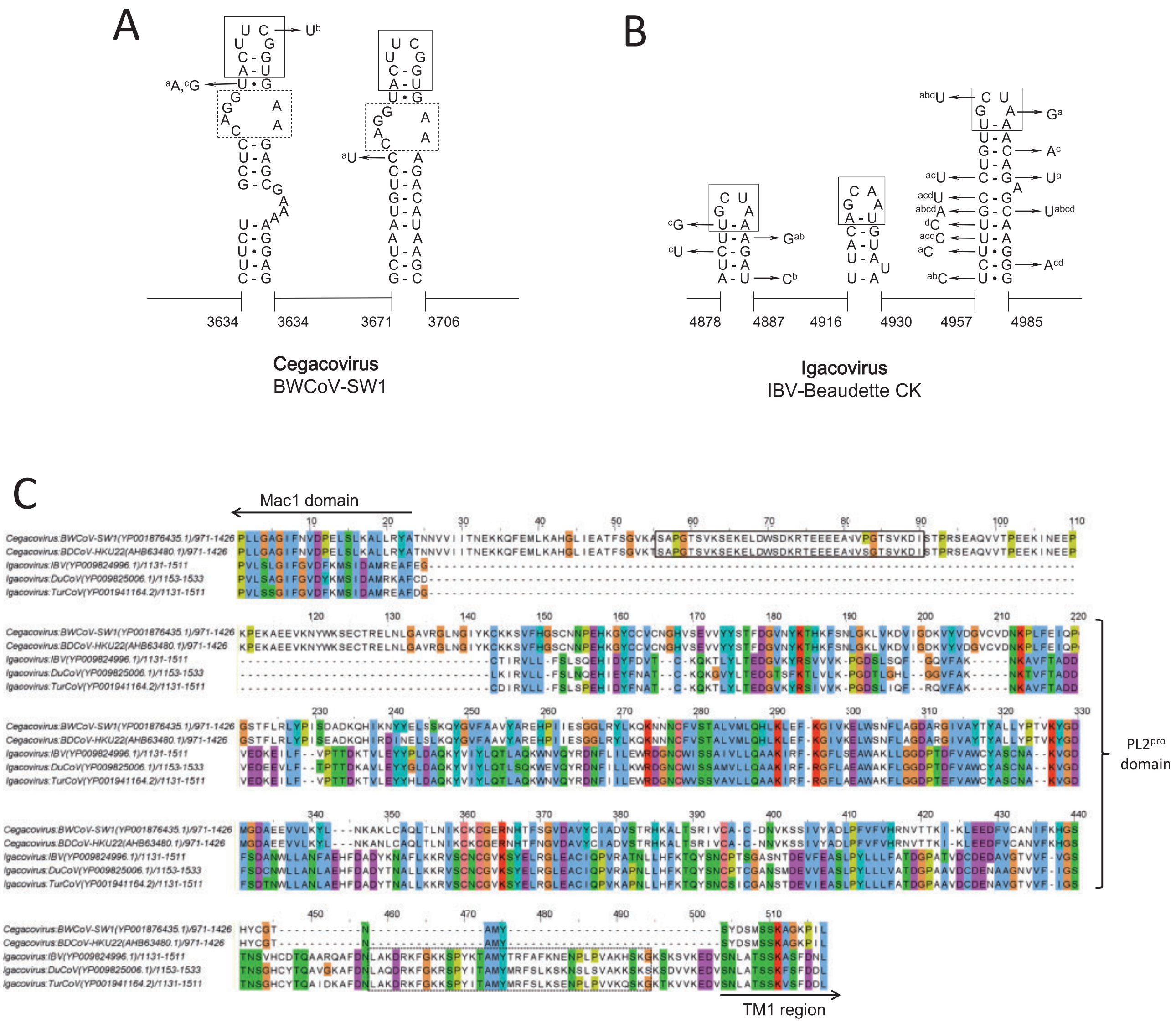
A
3634
3634
3671
3706
Cegacovirus
BWCoV-SW1
B
4878
4887
4916
4930
4957
4985
Igacovirus
IBV-Beaudette CK
C
Mac1 domain
Cegacovirus:BWCoV-SW1(YP001876435.1)/971-1426
Cegacovirus:BDCoV-HKU22(AHB63480.1)/971-1426
Igacovirus:IBV(YP009824996.1)/1131-1511
Igacovirus:DuCoV(YP009825006.1)/1153-1533
Igacovirus:TurCoV(YP001941164.2)/1131-1511
PL2pro domain
TM1 region

Figure 5. Conserved RSMs in gamma-CoVs. Secondary structures of the conserved RSMs found in gamma-CoVs are shown. (A) *Cegacovirus*: Beluga Whale CoV (BWCoV)-SW1 (NS_010646.1); [a]BDCoV-HKU22/CF090331 (KF793826.1); [b]BDCoV/37112-4 (MN690611.1). The apical 5′-acUUCGgu-3′ RSMs are boxed with solid lines, while the internal bulges 5′-CAGG-3′/5′-AA-3′ are boxed with dashed-lines. (B) *Igacovirus*: Avian infectious bronchitis virus (IBV)-Beaudette/CK (AJ311317.1); [a]IBV-Peafowl/GD/Q6/2003 (AY641576.1); [b]TuCoV-MG10 (EU095850.1); [c]Duck CoV (DuCoV)-DK/CH/HN/ZZ2004 (JF705860.1); [d]IBV-CK/Poland/G103/2016 (MK581207.1). The 5′-uGCUAa-3′ RSMs are boxed. (C) Multiple sequence aliment of partial poly-protein 1ab corresponding to $PL2^{pro}$ domain and the flanking regions of Nsp3 is shown. The corresponding regions of RNA structures comprised of conserved RSMs present in *Cegacovirus* and *Igacovirus* are boxed with solid and doted lines, respectively.

specific to *Cegacovirus* and *Igacovirus* are located at two distinct positions in Nsp3 coding region, respectively (Fig. 5C). The RSMs found in *Cegacovirus* spp., such as BWCoV-SW1, are located in the inter-domain region between Marcro (Mac1) and Viral protease 2 (PL2^{pro}) domains, while RSMs of *Igacovirus* spp., including Avian infectious bronchitis viruses (IBVs), Turkey CoVs (TuCoVs), Duck CoVs (DuCoVs), etc., are found in between PL2^{pro} domain and Trans-membrane (TM)-1 region (Fig. 5C). These findings demonstrate how structural RNA elements evolve in CoV gRNAs without disrupting functional protein domains by situating at inter-domain regions. Interestingly, upstream to the RSMs described above an inserted sequence with five tandem repeats was discovered in between HVR and Mac1 domains in some *Igacovirus* spp., including Avian CoVs (ACoVs) and Duck CoVs (DuCoVs), in which five hairpins displaying 5′-CAAA-3′ tetra-loops can be formed (Supplementary Fig. S5A). These additional RSMs may serve specific functions, if any, in these CoVs. Another possibility is that these repeats were conserved at the protein level, since the corresponding sequence encodes identical PQK-containing peptides (Supplementary Fig. S5C). However, we have noticed that no wobble mutations were found in these repeats, while the highly conserved upstream PQK tri-peptide are encoded by different codons forming no particular RNA structures (Supplementary Fig. S5B). Thus, it is likely that this inserted sequence underwent a selection driven by conservation of RNA structure. Nonetheless, more phylogenetic and experimental evidence is needed to prove this assumption. It is also worthwhile to note that in Human CoV (HCoV)-HKU1, an alpha-CoV, various lengths (60–480 nts) of NDDEDVVTGD tandem repeats were found upstream to its PL1^{pro} domain of Nsp3 (Woo et al. 2006), though no significant RNA structures could be identified by our approach. Nevertheless, acquisitions of additional sequence potentially increase the sequence space of CoVs for developing fitness, under the constraint of either protein sequences or RNA structures (Holmes 2003; Moreno et al. 2014; Wang et al. 2018). To sum up, we have identified particular RSM-encompassing elements in gamma-CoVs, which are located in between the coding sequences of Mac1, PL2^{pro}, and TM1 domains of NSP3.

3.6 Conserved RSMs identified in ORF1ab and/or 5′UTRs of delta-CoV gRNAs

Deltacoronavirus is a genus recently defined in the subfamily *Orthocoronavirinae*, which consists of subgenera *Buldecovirus*, *Andecovirus*, and *Herdecovirus*. In *Buldecovirus*, four recently identified CoVs, Thrush CoV (ThCoV)-HKU12, Munia CoV (MunCoV)-HKU13, Bulbul CoV (BuCoV)-HKU11, and Magpie-robin CoV (MR)-HKU18 (Woo et al. 2012a), have sequence insertions in their 5′UTRs, encompassing the 5′-ac-3′/5′-gAGUu-3′ RSMs (Supplementary Fig. S6). Besides in the 5′-UTR, we have also found particular sequence insertions in ORF1ab that harbour distinct RSMs (Fig. 6). In Wigeon CoV (WigCoV)-HKU20, the sole species in subgenus *Andecovirus*, we have found three copies of 5′-uGGUa-3′ RSMs (Fig. 6A). In Night-heron CoV (NHCoV)-HKU19, a species in *Herdecovirus*, four RSMs displaying 5′-GUAC-3′sequences were found, in which three were located in the apical loops of structured hairpins (Fig. 6B). In the recently discovered *Buldecovirus* spp., Thrush CoV (ThCoV)-HKU12, Common-moorhen CoV (CMCoV)-HKU21, White-eye CoV (WECoV)-HKU16, Falcon CoV (FaCoV)-HKU27, and Pigeon CoV (PiCoV)-HKU29 (Woo et al. 2012a; Lau et al. 2018), five RSMs encompassing 5′-GUAC-3′ sequences were found, in which two can have alternative conformations (Fig. 6C). Multiple sequence alignment revealed that the corresponding region of these RSMs was located in between the putative helicase (Hel) NSP13 and exoribonuclease (ExoN) NSP14 (Fig. 6D) (Woo et al. 2012a). At this particular inter-domain region, a homologous RNA element has been reported to act as potential PS in Bulbul CoV (BuCoV)-HKU11 (Masters 2019). Interestingly, four other *Buldecovirs* spp., including Munia CoV (MunCoV)-HKU13, Magpie-robin CoV (MR) CoV-HKU18, Porcine CoV (PorCoV)-HKU15, and Sparrow CoV (SpCoV)-HKU17, contain no such sequence insertion at this Nsp13/14 junction (Fig. 6D). Instead, the former two are possessors of the 5′-proximal RSMs (Supplementary Fig. S6), while in the latter two we were unable to identify particular RSMs. To sum up, these newly identified RNA elements in delta-CoVs suggest that the existence of RSMs in gRNAs is not limited to alpha-, beta-, and gamma-CoVs but a common feature in all CoVs.

3.7 Conservation and variation of the RSMs in SARS-CoV-2 SL5

Since the pandemic of SARS-CoV-2 in early 2020, thousands of gRNA samples have been sequenced, allowing us to deeply study the conservation and variation of the RSMs located in SL5 of SARS-CoV-2. Sequences of the apical motifs of SL5a, b, and c were extracted from 19,120 complete genomes of SARS-CoV-2 for alignments, which were acquired from NCBI database in September 2020. For comparison, 340 SARS-CoV and 598 MERS-CoV gRNA sequences were also analyzed. Figure 7 shows the partial secondary structure of SL5a-c, which is highly conserved within each lineage. In SARS-CoVs, 327 out of 340 isolates share identical sequences which are shown in Fig. 7A. The apical parts of SL5a–c are highly conserved among all the SARS-CoVs, though four U-to-C variations were found at the third position of the loop sequence of SL5a and b while three A-to-C variations were found in SL5c (Fig. 7A). The U-to-C variations in the apical loops make these SARS-CoV SL5a-b contain alpha-CoV-specific 5′-UUCCGU-3′ loop sequences instead of sequences specific to beta-CVs. In MERS-CoVs, the apical regions of SL5a-c are also highly conserved, showing 588 identical sequences out of 598 isolates (Fig. 7B). Four U-to-C variations were found at the third position of the loop sequences of SL5a and b, while the 5′-AAGAUGC-3′ loop sequence of SL5c is absolutely conserved in all the MERS-CoV gRNA sequences. Figure 7C shows partial SL5a-c secondary structure of the reference SARS-CoV-2 (isolate Wuhan-Hu-1, NC_045512.2), exhibiting the canonical beta-CoV specific 5′-UUUCGU-3′ hexa-loop in both SL5a and b, and the *sarbecovirus*-specific 5′-GAAA-3′ tetra-loop in SL5c, respectively. Few positions on the loops of SL5a-c were found having mutations. In particular, the fourth position of SL5b loop exhibits C-to-U variation (C241U) in 15,891 out of 19,120 isolates (Fig. 7C). We further analyzed how these variant and canonical loop sequences were distributed in SL5a-c of SARS-CoV, MERS-CoV, and SARS-CoV2, respectively. Supplementary Table S1 listed that in SARS-CoV and MERS-CoV both SL5a and b predominantly contains the canonical 5′-UUUCGU-3′ loop sequences, while in SARS-CoV-2 the majority possesses 5′-UUUCGU-3′ and 5′-UUU$\underline{\text{U}}$GU-3′ in SL5a and SL5b, respectively. This suggests that the variant RSM in SL5b has been adapted during the transmission worldwide (Kannan et al. 2020; Wang et al. 2020b; Yang et al. 2020).

3.8 Variations in the viral protein coding regions concurrent with variant SL5b in SARS-CoV-2

To study if any protein variations were concurrent with the variant SL5b, we refined the dataset of SARS-CoV-2 sequences to a

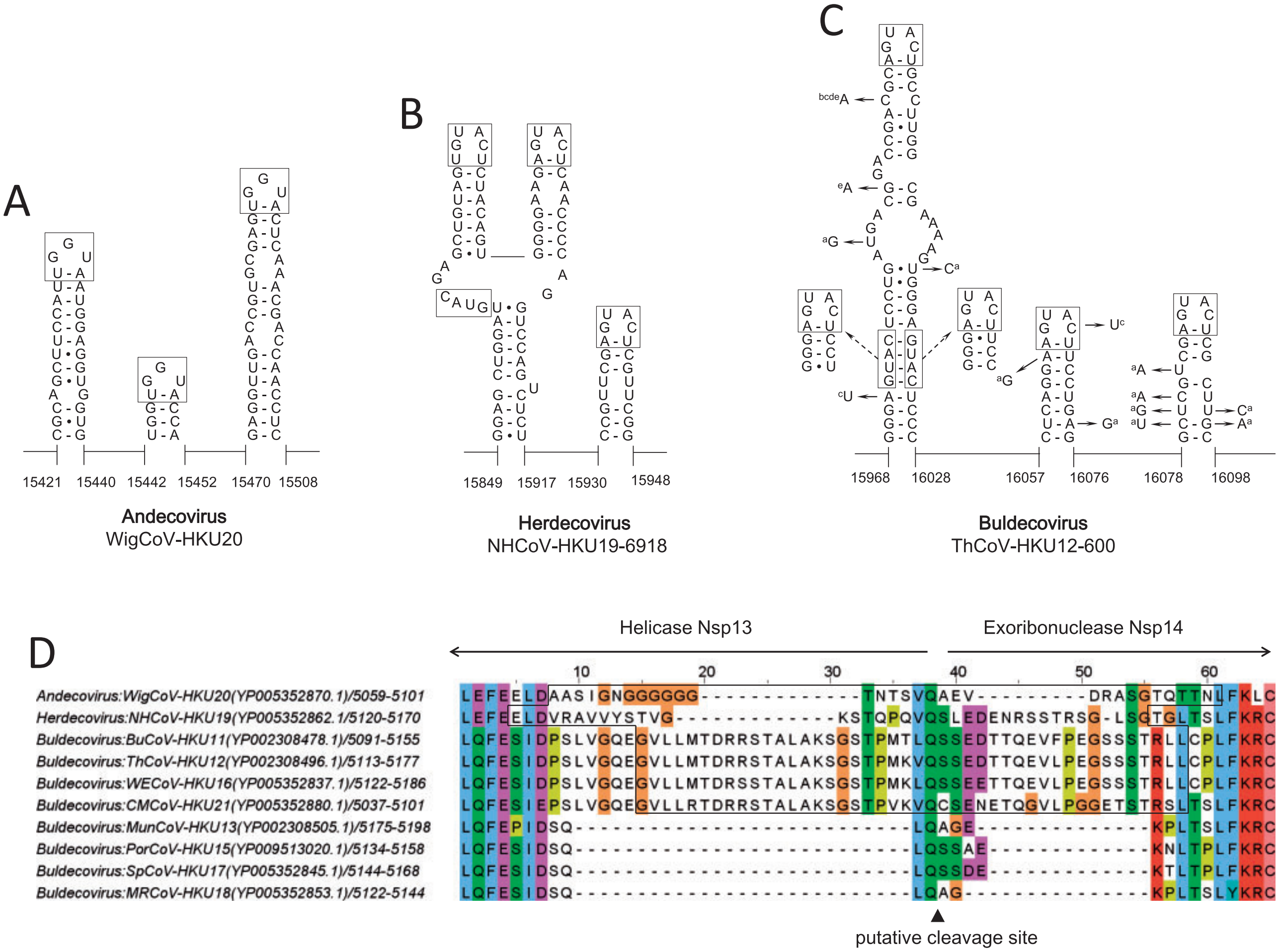
A
15421 15440 15442 15452 15470 15508
Andecovirus
WigCoV-HKU20
B
15849 15917 15930 15948
Herdecovirus
NHCoV-HKU19-6918
C
15968 16028 16057 16076 16078 16098
Buldecovirus
ThCoV-HKU12-600
D
Helicase Nsp13
Exoribonuclease Nsp14
10 20 30 40 50 60
Andecovirus:WigCoV-HKU20(YP005352870.1)/5059-5101 LEFEELDAASIGNGGGGGG-------------TNTSVQAEV---------DRASGTQTTNLFKLC
Herdecovirus:NHCoV-HKU19(YP005352862.1/5120-5170 LEFEELDVRAVVYSTVG--------------KSTQPQVQSLEDENRSSTRSG-LSGTGLTSLFKRC
Buldecovirus:BuCoV-HKU11(YP002308478.1)/5091-5155 LQFESIDPSLVGQEGVLLMTDRRSTALAKSGSTPMTLQSSEDTTQEVFPEGSSSTRLLCPLFKRC
Buldecovirus:ThCoV-HKU12(YP002308496.1)/5113-5177 LQFESIDPSLVGQEGVLLMTDRRSTALAKSGSTPMKVQSSEDTTQEVLPEGSSSTRLLCPLFKRC
Buldecovirus:WECoV-HKU16(YP005352837.1)/5122-5186 LQFESIDPSLVGQEGVLLMTDRSSTALAKSGSTPMKLQSSEETTQEVLPEGSSSTRLLCPLFKRC
Buldecovirus:CMCoV-HKU21(YP005352880.1)/5037-5101 LQFESIEPSLVGQEGVLLRTDRRSTALAKSGSTPVKVQCSENETQGVLPGGETSTRSLTSLFKRC
Buldecovirus:MunCoV-HKU13(YP002308505.1)/5175-5198 LQFEPIDSQ-----------------------------LQAGE--------------KPLTSLFKRC
Buldecovirus:PorCoV-HKU15(YP009513020.1)/5134-5158 LQFESIDSQ-----------------------------LQSSAE--------------KNLTPLFKRC
Buldecovirus:SpCoV-HKU17(YP005352845.1)/5144-5168 LQFESIDSQ-----------------------------LQSSDE--------------KTLTPLFKRC
Buldecovirus:MRCoV-HKU18(YP005352853.1)/5122-5144 LQFESIDSQ-----------------------------LQAG----------------KPLTSLYKRC
putative cleavage site

Figure 6. Conserved RSMs in delta-CoVs. Secondary structures of the RNA elements encompassing conserved RSMs (boxed with solid lines) found in the inter-domain region between Nsp13 and Nsp14 are shown. (A) *Andecovirus*: Wigeon CoV (WigCoV)-HKU20 (JQ065048.1). (B) *Herdecovirus*: Night-heron CoV (NHCoV)-HKU19 (JQ065047.1). (C) *Buldecovirus*: Thruch CoV (ThCoV)-HKU12-600 (FJ376621.1); [a]Common-moorhen CoV (CMCoV)-HKU21-8295 (JQ065049.1); [b]White-eye CoV (WECoV)-HKU16-6847 (JQ065044.1); [c]Bulbul CoV (BuCoV)-HKU11-796 (FJ376620.1); [d]Falcon CoV (FaCoV)-HKU27 (LC364342.1); [e]Pigeon CoV (PiCoV)-HKU29 (LC364344.1). Alternative structures are indicated with dash-line arrows. (D) Multiple sequence alignment of partial poly-protein 1ab at the inter-domain region between NSP13 and NSP14. Sequences corresponding to the RSM-encompassing elements were boxed with solid lines. The putative Nsp13/14 cleavage site is indicated by '▲'.

set of 14,118 complete gRNA sequences. According to the presence of either the canonical or variant loop sequences in SL5a-b, SARS-CoV-2 can be grouped into eight clusters (Table 2 and Supplementary Table S2–S4), in which cluster A (both SL5a and SL5b contains canonical 5′-UUUCGU-3′sequences) and B (SL5b contains 5′-UUUUGU-3′ variant sequence) overwhelmingly account for a large proportion. Table 2 lists the variations located in the coding regions that differ from cluster A and B. In cluster A, A1163, U7540, G16647, G22992, and G23401 are 100% conserved, while in cluster B these positions are much more variable (Table 2). The opposite situation is found at positions A17858, G26144, and C28863, which are perfectly conserved in cluster B but variable in cluster A. These findings suggest that mutations in cluster A and B have followed different evolutionary paths to a certain extent. Among the variations listed in Table 2, synonymous mutations at distinct positions, *e.g.* C3037U, C8782U, G16647U, C18060U, G23401A, and U28144C, are found differentially occurred in the two clusters. It has been reported that evolutionary constraints imposed by RNA secondary structure can result in accumulating synonymous mutations (Holmes 2003), and several recent reports have proposed that various RNA structures are present in the coding region of SARS-CoV-2 (Huston et al. 2020; Rangan et al. 2020; Simmonds 2020; Wacker et al. 2020; Ziv et al. 2020). However, how these synonymous mutations potentially affect those RNA structures in a SL5-related manner needs to be elucidated in the future.

Several non-synonymous variations are found differential between cluster A and B, for example, C14408U, A23403G, and G28881A-G28882A-G28883C resulting in Nsp12 P323L, S protein D614G, and N protein R203K-G204R, were found differ, respectively (Table 2). Over 99 per cent of SARS-CoVs in cluster A have Pro at the 323 position of Nsp12, while in cluster B over 99 per cent prefer Leu at position 323 (Table 2). Interestingly, in Nsp12 such a variation is located at one of its NSP8 interacting sites, while NSP8, besides NSP7, is a crucial co-factor to form the RNA-dependent RNA polymerase (RdRp) complex with NSP12 (Kirchdoerfer and Ward 2019; Gao et al. 2020; Lu et al. 2020a; Wang et al. 2020a). A potential correlation between variant SL5b and polymerase complex warrants further research. Likewise, the D614G variation in S proteins is highly constraint in cluster B, which has been recently suggested to increase infectivity of SARS-CoV-2 (Kannan et al. 2020; Korber et al. 2020). However, no

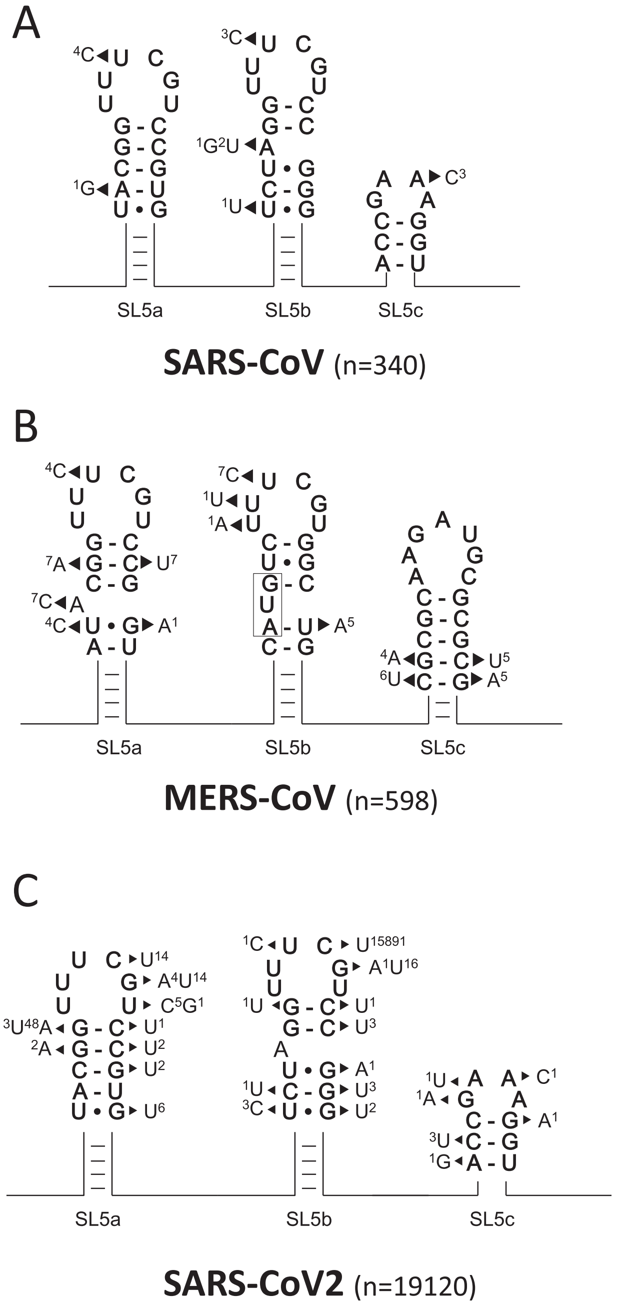

Figure 7. Variation hot-spots for the apical regions of SL5a-c in SARS-CoV, MERS-CoV, and SARS-CoV-2. Secondary structures of the apical parts of SL5a-c are shown for (A) SARS-CoV, (B) MERS-CoV, and (C) SARS-CoV-2. Counts for the nucleotide variations at each position on the apical region of SL5a-c are indicated. Total analyzed gRNA sequences are 340, 598, and 19,120, respectively.

evidence of a direct association between S proteins and 5′-UTR of CoV gRNAs has been reported so far. Thus, why variant SL5b is highly concurrent with S-protein D614G mutations and whether such concurrency contributes to the infectivity should be studied further in the future. The concurrent changes found in N protein are particularly interesting, though its concurrency is less significant compared with those in S and Nsp12, since that N protein is highly conserved and well-known for its RNA binding affinity (McBride et al. 2014; Hurst et al. 2009; Masters 2019). Curiously, deletion of the RGTSPA sequence in N protein is exclusively found in Cluster B (Table 2), and the R203K-G204R variations are highly constrained in Cluster B (Table 2) and E (Supplementary Table S3). These findings suggest that the C to U variation in SL5b is correlated to these N protein variations. The R203K-G204R variations are located in the Ser/Arg (S/R)-rich region in N protein, which is part of the disordered central linker region (CLR) in between the N-terminal RNA binding domain and the C-terminal dimerization domain (Chen et al. 2007a; Surjit and Lal 2009; Peng et al. 2020). Not until recently, the functional importance of the SR-rich region and non-structured CLR for CoV genome packaging has been proposed (Tylor et al. 2009; Carlson et al. 2020a; Cascarina and Ross 2020; Lu et al. 2020b; Nikolakaki and Giannakouros 2020; Perdikari et al. 2020; Savastano et al. 2020; Ye et al. 2020), yet the correlation between SL5 and the S/R-rich region of N proteins remains to be determined experimentally. To sum up, identification of variations in the viral protein coding sequences that differentially concur with variant RSMs provide fresh perspective on the functional relevance and the potential interplay between SL5 and viral proteins in SARS-CoV-2.

4. Discussion

4.1 Coronavirus RSMs are generally located in the 5′-UTRs and/or the inter-domain regions of ORF1ab

RNA viruses have a potentially large sequence space for the evolution of new variants due to the high mutation rate in replication process of gRNA (Holmes 2003). Conservation of functional proteins/domains is one of the major constraints for RNA viruses sequence space during evolution, while another constraint, yet inadequately studied, is RNA structure (Holmes 2003; Wang et al. 2018). In the exploration of CoV RSMs, we have seen how these two constraints apply to the evolution of CoVs. The RSM corresponding sequences that we have previously identified in SARS-CoV and MHV were found to be ancillary sequences that do not alter overall folding of SL5 basal stem and endonuclease NSP15, respectively (Xu et al. 2006; Chen and Olsthoorn 2010; Deng and Baker 2018). In alpha-, beta-, and some delta-CoVs, RSM-encompassing RNA elements were identified with apparent sequence insertions that the exact RSM-corresponding sequence is flanked by highly conserved regions (Figs. 1–3 and 6). Other RSM corresponding region that do not exhibit clear borders in primary sequence were otherwise identified through general predictions of RNA secondary structures (Fig. 5 and Supplementary Fig. S5). Remarkably, all of these RSMs are found present in 5′-UTRs and/or in the inter-domain regions of ORF1ab. These findings support the assumption that sequences strictly obey the conservation of RNA structure should preferentially be located in the inter-domain regions to minimize their alterations to the integrity of functional domains in a (poly-)protein. On the other hand, inter-domain RNA structures have been suggested to modulate ribosome elongation and promote protein folding (Watts et al. 2009; Faure et al. 2016, 2017). Such modulation effect could be another factor making highly structured RSMs to be preferentially located in domain junctions, regardless of their specific *cis*-acting functions.

4.2 Are PS-like elements located in NSP3 coding region the origin of canonical PS in *Embecovirus* spp.?

NSP3 is the largest and relatively variable protein encoded by CoV genomes (Neuman et al. 2008; Serrano et al. 2009; Neuman et al. 2014; Neuman, 2016; Lei et al. 2018), suggesting that the

Table 2. Differential variations found in cluster A and B SARS-CoV-2.

Total gRNA sequences N = 14,118			Cluster A SL5a (UUUCGU) SL5b (UUUCGU) $n = 2,431$				Cluster B SL5a (UUUCGU) SL5b (UUUUGU) $n = 11,646$		
Coding proteins	Positions		Counts (a.a.) porportions				Counts (a.a.) porportions		
		C	U	A	G	C	U	A	G
nsp2	1059	2,425 (T) 99.8%	6 (I) 0.2%			8,390 (T) 72.0%	3,256 (I) 28.0%		
	1163			2,431 (I) 100%			3,326 (F) 28.6%	8,320 (I) 71.4%	
nsp3	3037	2,408 (F) 99.1%	23 (F) 0.9%			43 (F) 0.4%	11,603 (F) 99.6%		
	3177	2,266 (P) 93.2%	165 (L) 6.8%			11,636 (P) 99.9%	7 (L) 0.06%	3 (H) 0.03	
	7540		2,431 (T) 100%			3,069 (T) 26.4%	8,577 (T) 73.6%		
nsp4	8072			2,430 (N) 99.9%	1 (D) 0.04%			11,635 (N) 99.9%	11 (D) 0.1%
	8782	894 (S) 36.8%	1,537 (S) 63.2%			11,640 (S) 99.9%	6 (S) 0.1%		
nsp6	11083	1 (F) 0.04%	391 (F) 16.1%		2,039 (L) 83.8%		199 (F) 1.7%		11,447 (L) 98.3%
nsp12	14408	2,416 (P) 99.4%	15 (L) 0.6%			103 (P) 0.9%	11,542 (L) 99.1%	1 (H) 0.009%	
	14805	2,167 (Y) 89.1%	264 (Y) 10.9%			11,627 (Y) 99.8%	19 (Y) 0.2%		
nsp13	16647				2,431 (T) 100%	1 (T) 0.009%	3,070 (T) 26.4%		8,575 (T) 73.6%
	17858			1,270 (Y) 52.2%	1,161 (C) 47.8%			11,646 (Y) 100%	
nsp14	18060	1,258 (L) 51.7%	1,173 (L) 48.3%			11,643 (L) 99.99%	3 (L) 0.03%		
	18555	2,431 (D) 100%				8,556 (D) 73.5%	3,090 (D) 26.5%		
	18877	2,421 (L) 99.6%	10 (L) 0.4%			10,789 (L) 92.6%	857 (L) 7.4%		
S protein	22992				2,431 (S) 100%			3,085 (N) 26.5%	8,561 (S) 73.5%
	23401				2,431 (Q) 100%		3 (H) 0.03%	3,068 (Q) 26.3%	8,575 (Q) 73.6%
	23403			2,409 (D) 99.1%	22 (G) 0.9%			11 (D) 0.1%	11,635 (G) 99.9%
ORF3a	25563		10 (H) 0.4%		2,421 (Q) 99.6%	3 (H) 0.03%	4,741 (H) 40.7%		6,902 (Q) 59.3%
	26144		207 (V) 8.5%		2,224 (G) 91.5%				11,646 (G) 100%
ORF8	28144	1,539 (S) 63.3%	891 (L) 36.7%			6 (S) 0.1%	11,638 (L) 99.9%		
N protein	28863	2,357 (S) 97.0%	74 (L) 3.0%			11,646 (S) 100%			
	28878		1 (I) 0.04%	60 (N) 2.5%	2,370 (S) 97.5%	3 (S) 0.03%	2 (I) 0.02%	2 (N)	11,639 (S)
	28896	2,372 (A) 97.6%			59 (G) 2.4%	11,590 (A) 99.99%			1 (G) 0.009%’
3’UTR	29700			2,268 (n.a.) 93.3%	163 (n.a.) 6.7%			11,631 (n.a.) 99.9%	15 (n.a.) 0.1%
N protein	28881–28883	GGG 2,428 (RG) 99.9%	AAC 3 (KR) 0.1%	Others 0 (n.a.) 0.0%	Deletion 0 (RGTSPA) 0.0%	GGG 6,151 (RG) 52.8%	AAC 5,420 (KR) 46.5%	Others 26 (n.a.) 0.2%	Deletion 51 (RGTSPA) 0.4%

Sequence variations in the viral protein coding regions in cluster A and B SARS-CoV-2 are listed.

sequence space is less constraint for the evolution of RNA 3D motifs in this region (Wang et al. 2018). In CoVs, there were about fifteen domains reported in NSP3 of *Embecovirus* spp., speculating that more inter-domain regions could harbor functional RNA elements (Woo et al. 2006; Woo et al. 2009; Lei et al. 2018). On the other hand, (repetitive) sequence insertions are usually found in the NSP3 coding region, particularly adjacent to the HVR domain. Besides the canonical PS found in NSP15 coding region, we have identified homologous PS-like elements in two inter-domain regions of NSP3 (Fig.4 and Supplementary Fig. S4). It has been suggested that the highly conserved canonical PS in *Embecovirus* spp. is a late acquisition during the evolution of the CoV family, since sequences corresponding to PS are apparently insertions and completely absent in other beta-CoVs (Joseph et al. 2007; Zhang et al. 2018; Masters 2019). However, how the canonical PS was acquired remains unsolved. Interestingly, the structural features of the PS-like element in NSP3 coding region very much meet the evolutionary criterion for RNA 3D structures, including the free energy-driven formation of WC-pairing helices/stems and the molecular interaction-driven formation of non-WC-pairing recurrent motifs (Fig. 4 and Supplementary Figs. S4 and 5) (Wang et al. 2018). Presumably, the PS could have evolved at the inter-domains of NSP3 coding region, followed by genome recombination to the canonical location in NSP15. Once the PS had been preserved in NSP15 coding region, the PS-like elements in NSP3 could potentially degenerate due to functional redundancy. This may explain why the structural diversity of the PS-like elements in NSP3 is high but the canonical PS in NSP15 coding region is much more conserved.

4.3 The functions of RSM-encompassing elements in CoV gRNAs

The RSM-encompassing elements identified in this study exhibit the hallmark of encapsidation signals which mediate selective packaging of ssRNAs with multiple capsid-protein binding sites in many RNA viruses (Knaus and Nassal 1993; Kramvis and Kew 1998; Flodell et al. 2002; Kim et al. 2011; Dykeman et al. 2013; Stockley et al. 2013; Rolfsson et al. 2016; Shi and Suzuki 2018; Comas-Garcia 2019; Rein 2019; Ding et al. 2020). In CoVs, the most studied RSM-containing element is the PS of MHV, of which the structural integrity has been shown crucial to be recognized by N, M, or N–M protein complex and direct selective packaging of gRNA (Narayanan et al. 2003; Chen et al. 2007b; Chang et al. 2009; Narayanan and Makino 2001; Kuo and Masters 2013; Kuo et al. 2014; Kuo et al. 2016; Masters 2019). It is possible that other RSM-encompassing elements play a similar role, although the structures of these elements may be divergent in different (sub)genera. For instance, the first 598 nucleotides (nt) from the 5' end of transmissible gastroenteritis virus gRNAs were reported to be essential for selective packaging (Escors et al. 2003; Morales et al. 2013). Recently, two independent studies also showed that the first 400 and 1000-nt 5′-proximal sequences of SARS-CoV-2 gRNA selectively promote liquid–liquid phase separation (LLPS) of N proteins (Carlson et al. 2020b; Iserman et al. 2020). And since the very 5′ leader regions upstream to transcriptional regulatory sequence (TRS), which are present also in all sub-genomic (sg)RNAs, unlikely contributes to the selectivity of genome packaging, these reports suggested that the PSs locate in the region between the first 75 to 400-nt gRNAs in these CoVs, where the structured SL4 and 5 are situated (Morales et al. 2013; Chen and Olsthoorn 2010; Iserman et al. 2020; Miao et al. 2020; Rangan et al. 2020). Based on *in vitro* ribonucleoprotein networks analyzed by mutational profiling (RNP-MaP) (Weidmann et al. 2020), Iseman et al. have suggested that N proteins primarily interact with terminal non-structured regions of gRNA, such as the 5′-proximal unstructured regions flanked by SL4/5 and the 3′-proximal N protein-coding sequences, respectively, promoting LLPS (Iserman et al. 2020). If this also applies *in vivo*, the formation of SL5 may help the unstructured regions to be oriented for direct N protein interactions, while the RSMs located in tripartite SL5a-c could also play a role in selective genome packaging of alpha- and beta-CoVs by direct or indirect interactions with N, M, or N/M complexes (Masters 2019). The only exceptional beta-CoVs are *embecovirus* spp., in which the tripartite SL5a-c are absent (Fig. 2E) but exclusively possess the PS in NSP15 coding region (Fig. 3). On the other hand, all the CoV species investigated in this study have been found to possess predominantly only one type of RSM, supporting that all these RSMs may be functionally equivalent to a certain extent. And notably, all the RSMs were exclusively found in either 5′-UTR or ORF1ab but not the regions encompassing sub-gRNAs. This nature of RSMs is in agreement with the selectiveness of full-length gRNA packaging. Anyhow, to verify that if the 5′-proximal SL5, and other RSM containing elements, plays a general role in gRNA packaging, viruses with reverse-genetically engineered mutations in these RSMs should be further studied.

Besides genome packaging, RSM containing elements may also mediate other functions, directly or indirectly. For instance, some characteristic features of the SL5 present in alpha- and beta-CoVs have potentially correlated the RSMs to translation regulation, including that the SL5 is generally located downstream to the uORF and the AUG start codon of ORF1ab is predominately locates at the 3′-side of the basal stem (Fig. 1 and 2). These conserved features preferentially support the assignment for SL5 of *Embecovirus* spp. to be what is shown in Fig. 2E but not the SL5* which encompasses relatively large portion of Nsp1 coding sequence (Supplementary Fig. S1). The rigid secondary structure of SL5 makes the tripartite SL5a-c to precisely locate in between the uORF and the translation initiation site in a specific 3D orientation. Direct or indirect interactions between these RSMs and their protein binders, if any, may moderate the translation as well as replication processes. For instance, it has been proposed that N proteins facilitate interactions between 5′- and 3′-ends of CoV genomes, leading circularization of the gRNA for the recruitment of viral replicase and cellular proteins (Lo et al. 2019). On the other hand, PS can possibly associated with NSP3 by N protein-NSP3 interaction (Hurst et al. 2009), playing a role in the regulation of replication and translation. Recently, the *cis*-acting SL1 in 5′-UTR of SARS-CoV-2 gRNA has been found responsible for the evasion of NSP1-mediated translation inhibition (Schubert et al. 2020; Shi et al. 2020; Thoms et al. 2020; Tidu et al. 2020). These discoveries highlight the importance of 5′-UTR in translational regulation, although the 5′-proximal RSMs may not necessarily directly participate in all of these regulations.

4.4 Structural phylogenetics of RSMs in CoVs

In this study, we have shown that a variety of RSMs is present in all CoVs in a lineage-specific manner. In another word, the structural features of RSM-containing elements predict the lineage of CoVs. These elements are generally located in the 5′-UTRs and/or in the inter-domain regions of ORF1ab. Presumably, RSM-containing elements could be functionally equivalent to each other mediating selective gRNA packaging and/or other conserved vital processes in CoVs' life cycle. We emphasize that the impact of RNA elements to CoV evolution and transmission should be more adequately studied in the future. Particularly, to understand functional relevance of the 5′-proximal SL5 is of great importance, since the highly pathogenic human CoVs, including SARS-CoV, MERS-CoV, and SARS-CoV-2, harbor highly conserved SL5s. In conclusion, our comprehensive exploration revealed the conservation and the diversity of RSMs, providing better insights into the structural RNA motifs in all CoVs as a whole. Basing on the general knowledge of RSMs, a variety of approach could be jointly facilitated in the face of present and future threat of pandemic CoVs.

Acknowledgements

We thank Dr. E. Rivas from Harvard University for the general discussion of structural co-variations and the utility of R-scape.

Funding

S.-C.C. and C.-H.Y. are supported from the Young Scholar Fellowship Program by the Ministry of Science and Technology (MOST) in Taiwan, under MOST Grant for the Columbus Program No. MOST 109-2636-B-006 -005, National Cheng Kung University, Taiwan and was supported in part by Higher Education Sprout Project, Ministry of Education to the Headquarters of University.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Athmer, J. et al. (2018) 'Selective Packaging in Murine Coronavirus Promotes Virulence by Limiting Type I Interferon Responses', *mBio*, 9.

Brown, C. G. et al. (2007) 'An RNA Stem-Loop within the Bovine Coronavirus nsp1 Coding Region is a Cis-Acting Element in Defective Interfering RNA Replication', *Journal of Virology*, 81: 7716–24.

Carlson, C. R. et al. (2020a) 'Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for Its Dual Functions', *Molecular Cell*, 80: 1092–103.e4.

—— et al. (2020b) 'Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for Its Dual Functions', *Molecular Cell*, 80: 1092–1103.e1094.

Cascarina, S. M., and Ross, E. D. (2020) 'A Proposed Role for the SARS-CoV-2 Nucleocapsid Protein in the Formation and Regulation of Biomolecular Condensates', *The Faseb Journal*, 34: 9832–42.

Chang, C. K. et al. (2009) 'Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging', *Journal of Virology*, 83: 2255–64.

Chen, C. Y. et al. (2007a) 'Structure of the SARS Coronavirus Nucleocapsid Protein RNA-Binding Dimerization Domain Suggests a Mechanism for Helical Packaging of Viral RNA', *Journal of Molecular Biology*, 368: 1075–86.

Chen, S. C., and Olsthoorn, R. C. (2010) 'Group-Specific Structural Features of the 5'-Proximal Sequences of Coronavirus Genomic RNAs', *Virology*, 401: 29–41.

—— et al. (2007b) 'New Structure Model for the Packaging Signal in the Genome of Group IIa Coronaviruses', *Journal of Virology*, 81: 6771–4.

Cologna, R., and Hogue, B. G. (1998) 'Coronavirus Nucleocapsid Protein. RNA Interactions', *Advances in Experimental Medicine and Biology*, 440: 355–9.

Comas-Garcia, M. (2019) 'Packaging of Genomic RNA in Positive-Sense Single-Stranded RNA Viruses: A Complex Story', *Viruses*, 11: 253.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020) 'The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2', *Nature Microbiology*, 5: 536–44.

Cui, J., Li, F., and Shi, Z. L. (2019) 'Origin and Evolution of Pathogenic Coronaviruses', *Nature Reviews. Microbiology*, 17: 181–92.

Deng, X., and Baker, S. C. (2018) 'An "Old" Protein with a New Story: Coronavirus Endoribonuclease is Important for Evading Host Antiviral Defenses', *Virology*, 517: 157–63.

Ding, P. et al. (2020) 'Identification of the Initial Nucleocapsid Recognition Element in the HIV-1 RNA Packaging Signal', *Proceedings of the National Academy of Sciences of the United States of America*, 117: 17737–46.

Dykeman, E. C., Stockley, P. G., and Twarock, R. (2013) 'Packaging Signals in Two Single-Stranded RNA Viruses Imply a Conserved Assembly Mechanism and Geometry of the Packaged Genome', *Journal of Molecular Biology*, 425: 3235–49.

Escors, D. et al. (2003) 'Transmissible Gastroenteritis Coronavirus Packaging Signal is Located at the 5' End of the Virus Genome', *Journal of Virology*, 77: 7890–902.

Faure, G. et al. (2016) 'Role of mRNA Structure in the Control of Protein Folding', *Nucleic Acids Research*, 44: 10898–911.

—— et al. (2017) 'Adaptation of mRNA Structure to Control Protein Folding', *RNA Biology*, 14: 1649–54.

Flodell, S. et al. (2002) 'The Apical Stem-Loop of the Hepatitis B Virus Encapsidation Signal Folds into a Stable Tri-Loop with Two Underlying Pyrimidine Bulges', *Nucleic Acids Research*, 30: 4803–11.

Forni, D. et al. (2017) 'Molecular Evolution of Human Coronavirus Genomes', *Trends in Microbiology*, 25: 35–48.

Fosmire, J. A., Hwang, K., and Makino, S. (1992) 'Identification and Characterization of a Coronavirus Packaging Signal', *Journal of Virology*, 66: 3522–30.

Gao, Y. et al. (2020) 'Structure of the RNA-Dependent RNA Polymerase from COVID-19 Virus', *Science*, 368: 779–82.

Graham, R. L., Donaldson, E. F., and Baric, R. S. (2013) 'A Decade after SARS: Strategies for Controlling Emerging Coronaviruses', *Nature Reviews. Microbiology*, 11: 836–48.

Guan, B. J. et al. (2012) 'Genetic Evidence of a Long-Range RNA-RNA Interaction between the Genomic 5' Untranslated Region and the Nonstructural Protein 1 Coding Region in Murine and Bovine Coronaviruses', *Journal of Virology*, 86: 4631–43.

Guan, Y. et al. (2003) 'Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China', *Science*, 302: 276–8.

Han, Y. et al. (2019) 'Identification of Diverse Bat Alphacoronaviruses and Betacoronaviruses in China Provides New Insights into the Evolution and Origin of Coronavirus-Related Diseases', *Frontiers in Microbiology*, 10:1900.

Hemida, M. G. et al. (2013) 'Middle East Respiratory Syndrome (MERS) Coronavirus Seroprevalence in Domestic Livestock in Saudi Arabia, 2010 to 2013′, *Eurosurveillance*, 18: 20659.

Holmes, E. C. (2003) 'Error Thresholds and the Constraints to RNA Virus Evolution', *Trends in Microbiology*, 11: 543–6.

Hu, B. et al. (2015) 'Bat Origin of Human Coronaviruses', *Virology Journal*, 12: 221.

Huang, C. et al. (2020) 'Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan', *The Lancet*, 395: 497–506.

Hurst, K. R., Koetzner, C. A., and Masters, P. S. (2009) 'Identification of in Vivo-Interacting Domains of the Murine Coronavirus Nucleocapsid Protein', *Journal of Virology*, 83: 7221–34.

Huston, N. C. et al. (2020) 'Comprehensive in-Vivo Secondary Structure of the SARS-CoV-2 Genome Reveals Novel Regulatory Motifs and Mechanisms', *bioRxiv*.

Iserman, C. et al. (2020) ' Specific Viral RNA Drives the SARS CoV-2 Nucleocapsid to Phase Separate', *bioRxiv*.

Joseph, J. S. et al. (2007) 'Crystal Structure of a Monomeric Form of Severe Acute Respiratory Syndrome Coronavirus Endonuclease nsp15 Suggests a Role for Hexamerization as an Allosteric Switch', *Journal of Virology*, 81: 6700–8.

Kannan, S. R. et al. (2020) 'Infectivity of SARS-CoV-2: There is Something More than D614G? ', *Journal of Neuroimmune Pharmacology*, 15: 574–7.

Keane, S. C. et al. (2012) 'Functional Transcriptional Regulatory Sequence (TRS) RNA Binding and Helix Destabilizing Determinants of Murine Hepatitis Virus (MHV) Nucleocapsid (N) protein', *The Journal of Biological Chemistry*, 287: 7063–73.

Kim, D. Y. et al. (2011) 'Conservation of a Packaging Signal and the Viral Genome RNA Packaging Mechanism in Alphavirus Evolution', *Journal of Virology*, 85: 8022–36.

Kirchdoerfer, R. N., and Ward, A. B. (2019) 'Structure of the SARS-CoV nsp12 Polymerase Bound to nsp7 and nsp8 co-Factors', *Nature Communications*, 10: 2342.

Knaus, T., and Nassal, M. (1993) 'The Encapsidation Signal on the Hepatitis B Virus RNA Pregenome Forms a Stem-Loop Structure That is Critical for Its Function', *Nucleic Acids Research*, 21: 3967–75.

Korber, B. et al. (2020) 'Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus', *Cell*, 182: 812–827.e819.

Kramvis, A., and Kew, M. C. (1998) 'Structure and Function of the Encapsidation Signal of Hepadnaviridae', *Journal of Viral Hepatitis*, 5: 357–67.

Ksiazek, T. G. et al. (2003) 'A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome', *New England Journal of Medicine*, 348: 1953–66.

Kuo, L. et al. (2014) 'Recognition of the Murine Coronavirus Genomic RNA Packaging Signal Depends on the Second RNA-Binding Domain of the Nucleocapsid Protein', *Journal of Virology*, 88: 4451–65.

——, Koetzner, C. A., and Masters, P. S. (2016) 'A Key Role for the Carboxy-Terminal Tail of the Murine Coronavirus Nucleocapsid Protein in Coordination of Genome Packaging', *Virology*, 494: 100–7.

——, and Masters, P. S. (2013) 'Functional Analysis of the Murine Coronavirus Genomic RNA Packaging Signal', *Journal of Virology*, 87: 5182–92.

Lau, S. K. et al. (2012) 'Isolation and Characterization of a Novel Betacoronavirus Subgroup a Coronavirus, Rabbit Coronavirus HKU14, from Domestic Rabbits', *Journal of Virology*, 86: 5481–96.

Lau, S. K. P. et al. (2018) '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', *Journal of Virology*, 92:e00265–18.

Lei, J., Kusov, Y., and Hilgenfeld, R. (2018) 'Nsp3 of Coronaviruses: Structures and Functions of a Large Multi-Domain Protein', *Antiviral Research*, 149: 58–74.

Li, L. et al. (2008) 'Structural Lability in Stem-Loop 1 Drives a 5' UTR-3' UTR Interaction in Coronavirus Replication', *Journal of Molecular Biology*, 377: 790–803.

Liu, P. et al. (2009) 'Mouse Hepatitis Virus Stem-Loop 2 Adopts a uYNMG(U)a-like Tetraloop Structure That is Highly Functionally Tolerant of Base Substitutions', *Journal of Virology*, 83: 12084–93.

—— et al. (2007) 'A U-Turn Motif-Containing Stem-Loop in the Coronavirus 5' Untranslated Region Plays a Functional Role in Replication', *RNA*, 13: 763–80.

Lo, C. Y. et al. (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', *The FEBS Journal*, 286: 3222–39.

Lu, S. et al. (2020a) ' CDD/SPARCLE: The Conserved Domain Database in 2020', *Nucleic Acids Research*, 48: D265–D268.

—— et al. (2020b) 'The SARS-CoV-2 Nucleocapsid Phosphoprotein Forms Mutually Exclusive Condensates with RNA and the Membrane-Associated M Protein', *bioRxiv*.

Madhugiri, R. et al. (2016) 'Coronavirus cis-Acting RNA Elements', *Advances in Virus Research*, 96: 127–63.

Masters, P. S. (2019) 'Coronavirus Genomic RNA Packaging', *Virology*, 537: 198–207.

McBride, R., van Zyl, M., and Fielding, B. C. (2014) 'The Coronavirus Nucleocapsid is a Multifunctional Protein', *Viruses*, 6: 2991–3018.

Miao, Z. et al. (2020) 'Secondary Structure of the SARS-CoV-2 5'-UTR', *RNA Biology*, 1–10.

Molenkamp, R., and Spaan, W. J. (1997) 'Identification of a Specific Interaction between the Coronavirus Mouse Hepatitis Virus A59 Nucleocapsid Protein and Packaging Signal', *Virology*, 239: 78–86.

Morales, L. et al. (2013) 'Transmissible Gastroenteritis Coronavirus Genome Packaging Signal is Located at the 5' End of the Genome and Promotes Viral RNA Incorporation into Virions in a Replication-Independent Process', *Journal of Virology*, 87: 11579–90.

Moreno, E. et al. (2014) 'Exploration of Sequence Space as the Basis of Viral RNA Genome Segmentation', *Proceedings of the National Academy of Sciences of the United States of America*, 111: 6678–83.

Narayanan, K. et al. (2003) 'Nucleocapsid-Independent Specific Viral RNA Packaging via Viral Envelope Protein and Viral RNA Signal', *Journal of Virology*, 77: 2922–7.

——, and Makino, S. (2001) 'Cooperation of an RNA Packaging Signal and a Viral Envelope Protein in Coronavirus RNA Packaging', *Journal of Virology*, 75: 9059–67.

Neuman, B. W. (2016) 'Bioinformatics and Functional Analyses of Coronavirus Nonstructural Proteins Involved in the Formation of Replicative Organelles', *Antiviral Research*, 135: 97–107.

Neuman, B. W. et al. (2014) 'Atlas of Coronavirus Replicase Structure', *Virus Research*, 194: 49–66.

—— et al. (2008) 'Proteomics Analysis Unravels the Functional Repertoire of Coronavirus Nonstructural Protein 3′, *Journal of Virology*, 82: 5279–94.

Nikolakaki, E., and Giannakouros, T. (2020) 'SR/RS Motifs as Critical Determinants of Coronavirus Life Cycle', *Frontiers in Molecular Biosciences*, 7: 219.

Papineau, A. et al. (2019) 'Genome Organization of Canada Goose Coronavirus, a Novel Species Identified in a Mass Die-off of Canada Geese', *Science Report*, 9: 5954.

Peng, Y. et al. (2020) 'Structures of the SARS-CoV-2 Nucleocapsid and Their Perspectives for Drug Design', *The EMBO Journal*, 39: e105938.

Perdikari, T. M. et al. (2020) 'SARS-CoV-2 Nucleocapsid Protein Undergoes Liquid-Liquid Phase Separation Stimulated by RNA and Partitions into Phases of Human Ribonucleoproteins', *bioRxiv*.

Perlman, S., and Netland, J. (2009) 'Coronaviruses post-SARS: Update on Replication and Pathogenesis', *Nature Reviews. Microbiology*, 7: 439–50.

Raman, S. et al. (2003) 'Stem-Loop III in the 5' Untranslated Region is a Cis-Acting Element in Bovine Coronavirus Defective Interfering RNA Replication', *Journal of Virology*, 77: 6720–30.

——, and Brian, D. A. (2005) 'Stem-Loop IV in the 5' Untranslated Region is a Cis-Acting Element in Bovine Coronavirus Defective Interfering RNA Replication', *Journal of Virology*, 79: 12434–46.

Rangan, R. et al. (2020) 'RNA Genome Conservation and Secondary Structure in SARS-CoV-2 and SARS-Related Viruses: A First Look', *RNA*, 26: 937–59.

Rein, A. (2019) 'RNA Packaging in HIV', *Trends in Microbiology*, 27: 715–23.

Rivas, E. (2020) 'RNA Structure Prediction Using Positive and Negative Evolutionary Information', *PLoS Computational Biology*, 16: e1008387.

——, Clements, J., and Eddy, S. R. (2017) 'A Statistical Test for Conserved RNA Structure Shows Lack of Evidence for Structure in lncRNAs', *Nature Methods*, 14: 45–8.

——, ——, and —— (2020) 'Estimating the Power of Sequence Covariation for Detecting Conserved RNA Structure', *Bioinformatics*, 36: 3072–6.

Rolfsson, O. et al. (2016) 'Direct Evidence for Packaging Signal-Mediated Assembly of Bacteriophage MS2', *Journal of Molecular Biology*, 428: 431–48.

Rota, P. A. et al. (2003) 'Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome', *Science*, 300: 1394–9.

Savastano, A. et al. (2020) 'Nucleocapsid Protein of SARS-CoV-2 Phase Separates into RNA-Rich Polymerase-Containing Condensates', *Nature Communication*, 11: 6041.

Schnieders, R. et al. (2021) '1)H, (13)C and (15)N Chemical Shift Assignment of the Stem-Loop 5a from the 5'-UTR of SARS-CoV-2', *Biomolecular NMR Assignments*, 15: 203–11.

Schubert, K. et al. (2020) 'SARS-CoV-2 Nsp1 Binds the Ribosomal mRNA Channel to Inhibit Translation', *Nature Structural & Molecular Biology*, 27: 959–66.

Serrano, P. et al. (2009) 'Nuclear Magnetic Resonance Structure of the Nucleic Acid-Binding Domain of Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 3′, *Journal of Virology*, 83: 12998–3008.

Shi, G., and Suzuki, T. (2018) 'Molecular Basis of Encapsidation of Hepatitis C Virus Genome', *Frontiers in Microbiology*, 9: 396.

Shi, M. et al. (2020) ' SARS-CoV-2 Nsp1 Suppresses Host but Not Viral Translation through a Bipartite Mechanism', *bioRxiv*.

Sievers, F. et al. (2011) 'Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega', *Molecular Systems Biology*, 7: 539.

Simmonds, P. (2020) 'Pervasive RNA Secondary Structure in the Genomes of SARS-CoV-2 and Other Coronaviruses', *mBio*, 11:

So, R. T. Y. et al. (2019) 'Diversity of Dromedary Camel Coronavirus HKU23 in African Camels Revealed Multiple Recombination Events among Closely Related Betacoronaviruses of the Subgenus Embecovirus', *Journal of Virology*, 93.

Sorensen, M. D. et al. (2006) 'Severe Acute Respiratory Syndrome (SARS): Development of Diagnostics and Antivirals', *Annals of the New York Academy of Sciences*, 1067: 500–5.

Stockley, P. G. et al. (2013) 'Packaging Signals in Single-Stranded RNA Viruses: Nature's Alternative to a Purely Electrostatic Assembly Mechanism', *Journal of Biological Physics*, 39: 277–87.

Su, S. et al. (2016) 'Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses', *Trends Microbiol*, 24: 490–502.

Surjit, M., and Lal, S. K. (2009) 'The Nucleocapsid Protein of the SARS Coronavirus: Structure, Function and Therapeutic Potential', *Molecular Biology of the SARS-Coronavirus*, 129–51.

Tan, Y. W., Hong, W., and Liu, D. X. (2012) 'Binding of the 5'-Untranslated Region of Coronavirus RNA to Zinc Finger CCHC-Type and RNA-Binding Motif 1 Enhances Viral Replication and Transcription', *Nucleic Acids Res*, 40: 5065–77.

Thoms, M. et al. (2020) 'Structural Basis for Translational Shutdown and Immune Evasion by the Nsp1 Protein of SARS-CoV-2', *Science*, 369: 1249–55.

Tidu, A. et al. (2020) 'The Viral Protein NSP1 Acts as a Ribosome Gatekeeper for Shutting down Host Translation and Fostering SARS-CoV-2 Translation', *RNA*. doi: 10.1261/rna.078121.120.

Tylor, S. et al. (2009) 'The SR-Rich Motif in SARS-CoV Nucleocapsid Protein is Important for Virus Replication', *Canadian Journal of Microbiology*, 55: 254–60.

Wacker, A. et al. (2020) ' Secondary Structure Determination of Conserved SARS-CoV-2 RNA Elements by NMR Spectroscopy', *Nucleic Acids Research,* 48:12415–12435.

Wang, Q. et al. (2020a) 'Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase', *Cell*, 182: 417–428.e413.

Wang, R. et al. (2020b) 'Characterizing SARS-CoV-2 Mutations in the United States', *Res Sq, rs.3.rs-49671*.

Wang, W. et al. (2015) 'Discovery, Diversity and Evolution of Novel Coronaviruses Sampled from Rodents in China', *Virology*, 474: 19–27.

Wang, Y. et al. (2018) 'RNA 3-Dimensional Structural Motifs as a Critical Constraint of Viroid RNA Evolution', *PLoS Pathog*, 14: e1006801.

Watts, J. M. et al. (2009) 'Architecture and Secondary Structure of an Entire HIV-1 RNA Genome', *Nature*, 460: 711–6.

Weidmann, C. A. et al. (2021) 'Analysis of RNA-Protein Networks with RNP-MaP Defines Functional Hubs on RNA', *Nature Biotechnology*, 39: 347–56.

Woo, K. et al. (1997) 'Murine Coronavirus Packaging Signal Confers Packaging to Nonviral RNA', *Journal of Virology*, 71: 824–7.

Woo, P. C. et al. (2009) 'Coronavirus Diversity, Phylogeny and Interspecies Jumping', *Experimental Biology and Medicine (Maywood)*, 234: 1117–27.

—— et al. (2012a) 'Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavirus', *Journal of Virology*, 86: 3995–4008.

—— et al. (2012b) 'Genetic Relatedness of the Novel Human Group C Betacoronavirus to Tylonycteris Bat Coronavirus HKU4 and Pipistrellus Bat Coronavirus HKU5′, *Emerging Microbes & Infections*, 1: e35–5.

—— et al. (2006) 'Comparative Analysis of 22 Coronavirus HKU1 Genomes Reveals a Novel Genotype and Evidence of Natural

Recombination in Coronavirus HKU1', *Journal of Virology*, 80: 7136–45.

Wu, H. Y., Ozdarendeli, A., and Brian, D. A. (2006) 'Bovine Coronavirus 5'-Proximal Genomic Acceptor Hotspot for Discontinuous Transcription is 65 Nucleotides Wide', *Journal of Virology*, 80: 2183–93.

Xu, X. et al. (2006) 'New Antiviral Target Revealed by the Hexameric Structure of Mouse Hepatitis Virus Nonstructural Protein nsp15', *Journal of Virology*, 80: 7909–17.

Yang, D., and Leibowitz, J. L. (2015) 'The Structure and Functions of Coronavirus Genomic 3' and 5' Ends', *Virus Res*, 206: 120–33.

—— et al. (2015) 'SHAPE Analysis of the RNA Secondary Structure of the Mouse Hepatitis Virus 5' Untranslated Region and N-Terminal nsp1 Coding Sequences', *Virology*, 475: 15–27.

Yang, H. C. et al. (2020) 'Analysis of Genomic Distributions of SARS-CoV-2 Reveals a Dominant Strain Type with Strong Allelic Associations', *Proceedings of the National Academy of Sciences of the United States of America*, 117: 30679–86.

Ye, Q. et al. (2020) 'Architecture and Self-Assembly of the SARS-CoV-2 Nucleocapsid Protein', *bioRxiv*.

Zhang, L. et al. (2018) 'Structural and Biochemical Characterization of Endoribonuclease Nsp15 Encoded by Middle East Respiratory Syndrome Coronavirus', *Journal of Virology*, 92:

Zhou, P. et al. (2018) 'Fatal Swine Acute Diarrhoea Syndrome Caused by an HKU2-Related Coronavirus of Bat Origin', *Nature*, 556: 255–8.

Ziv, O. et al. (2020) 'The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2', *Molecular Cell*, 80: 1067–1077.e1065.

Zuker, M. (2003) 'Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction', *Nucleic Acids Research*, 31: 3406–15.