Analysis of Rapidly Emerging Variants in Structured Regions of the SARS-CoV-2 Genome

Sean P. Ryder*, Brittany R. Morgan, and Francesca Massi

Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA

***Corresponding Author**

Mailing Address 364 Plantation Street, LRB-906

Worcester, MA, 01605

Phone: 508-856-1372

Email: Sean.Ryder@umassmed.edu

Running Title *SARS-CoV-2 variants in RNA structure*

Key Words SARS, COVID-19, RNA Structure, Coronavirus, Phylogeny

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has motivated a widespread effort to understand its epidemiology and pathogenic mechanisms. Modern high-throughput sequencing technology has led to the deposition of vast numbers of SARS-CoV-2 genome sequences in curated repositories, which have been useful in mapping the spread of the virus around the globe. They also provide a unique opportunity to observe virus evolution in real time. Here, we evaluate two cohorts of SARS-CoV-2 genomic sequences to identify rapidly emerging variants within structured cis-regulatory elements of the SARS-CoV-2 genome. Overall, twenty variants are present at a minor allele frequency of at least 0.5%. Several enhance the stability of Stem Loop 1 in the 5´UTR, including a set of co-occurring variants that extend its length. One appears to modulate the stability of the frameshifting pseudoknot between ORF1a and ORF1b, and another perturbs a bi-stable molecular switch in the 3´UTR. Finally, five variants destabilize structured elements within the 3´UTR hypervariable region, including the S2M stem loop, raising questions as to the functional relevance of these structures in viral replication. Two of the most abundant variants appear to be caused by RNA editing, suggesting host-viral defense contributes to SARS-CoV-2 genome heterogeneity. This analysis has implications for the development of therapeutics that target viral cis-regulatory RNA structures or sequences, as rapidly emerging variations in these regions could lead to drug resistance.

Introduction

The betacoronaviridae are non-segmented single-stranded positive sense viruses with an RNA genome of approximately thirty kilobases in length. This family poses a significant threat to human health. In addition to causing approximately 30% of annual upper respiratory infections (Stadler et al. 2003; Su et al. 2016), it is responsible for three major outbreaks of severe acute respiratory syndrome (SARS, MERS, and COVID-19) since the turn of the century (Drosten et al. 2003; Ksiazek et al. 2003; Zhong et al. 2003; Zaki et al. 2012; Wang et al. 2020). COVID-19 is a unique form of pneumonia characterized by high fever, dry cough, and occasionally catastrophic hypoxia. It was first described in the city of Wuhan, Huibei Province, in the fall of 2019 (Chan et al. 2020; Li et al. 2020; Wang et al. 2020). A novel virus termed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified as the cause of this disease (Wang et al. 2020; Wu et al. 2020). The rapid spread of the virus led to a global pandemic that caused significant morbidity and mortality and disruption of daily life for millions of people. The extraordinary impact of this virus fueled strong interest in understanding its pathophysiology and epidemiology with the hope of developing new treatments and approaches to limit its spread.

The SARS-CoV-2 infection cycle is similar to that of other betacoronaviridae (Zheng 2020). Following attachment of the virus to the host cell and membrane fusion, viral genomic RNA is introduced to the host cell where it is translated to produce a polyprotein encoding the viral replicase, proteases, and several accessory proteins. The replicase is an RNA-dependent RNA polymerase that produces full-length antigenomic sequence that serves as a template for the production of additional copies of the viral genome and several nested subgenomic RNAs (sgRNAs) that encode the structural components of the virion.

Conserved stem loop structures are present in both coding and noncoding regions the SARS-CoV-2 RNA genome (Ramya Rangan 2020). They cluster in the 5´UTR, the N-terminal portion of ORF-1a, at the junction of ORF1a and ORF1b, and in the 3´UTR. While their precise role is not known, their function can be inferred from studies of related elements in mouse hepatitis virus (MHV) and other coronaviridae (Yang and Leibowitz 2015; Madhugiri et al. 2016). The structured elements have regulatory roles in various aspects of viral replication, sgRNA synthesis, and translation. Though they are divergent in sequence, the structures appear to be conserved, and in some cases elements from SARS-CoV can functionally substitute for those in MHV with little impact on viral replication. (Goebel et al. 2004; Kang et al. 2006a; Kang et al. 2006b; Zust et al. 2008; Chen and Olsthoorn 2010; Yang and Leibowitz 2015).

DNA sequencing technology has progressed remarkably since the SARS outbreak of 2003 (Geoghegan and Holmes 2018; Zhang et al. 2018). It is now routine to determine the sequence of the ~30 kilobase viral genome using high throughput sequencing technology (Wu et al. 2020). As a result, scientists and medical professionals from around world have sequenced the SARS-CoV-2 genome from patient isolates and disseminated their findings through data repositories (e.g. the GISAID EpiCoV database) at unprecedented speed (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017; Coronaviridae Study Group of the International Committee on Taxonomy of 2020; Wu et al. 2020). This has enabled the construction of molecular phylogenies that have guided our understanding of the virus transmission history, its basal mutation rate, and its potential to evade emerging therapeutics and vaccines (Bedford et al. 2020; Chu et al. 2020; Forster et al. 2020; Kim et al. 2020b; Lv et al. 2020; Pachetti et al. 2020; Pinto et al. 2020). At the time of this writing, almost 25,000 SARS-CoV-2 genome sequences have been deposited in the GISAID EpiCoV database (www.gisaid.org) and are available through a database access agreement (Elbe and

Buckland-Merrett 2017; Shu and McCauley 2017). Over 3500 SARS-CoV-2 genome sequences have been deposited into the National Center for Biotechnology Information (NCBI) Genbank (ncbi.nih.nlv.gov/genbank/sars-cov-2-seqs) and are freely available to the public.

Here, we analyzed both cohorts to identify and characterize rapidly emerging variations within the cis-regulatory RNA structures of the virus genome. Our analysis reveals twenty rapidly emerging variants including two that likely arose through RNA editing. The data identify SL1 of the 5´UTR as a hot spot for viral mutation, where most mutations stabilize the stem loop structure. The data also show that structured elements in the 3´UTR hypervariable region, including the enigmatic S2M loop, contain rapidly emerging variations predicted to be destabilizing. The results provide insight into the relevance of the proposed viral RNA structures, and present a roadmap to avoid potential confounds to RNA therapeutic development.

Results

Identification of rapidly emerging variants in structured regions of the SARS-CoV-2 genome

The genome sequence for viral isolate Wuhan-Hu-1 (Genbank MN908947) was used as a reference genome (Wu et al. 2020). The 5´UTR (1–265), the structured region of ORF1a (266–450), the frameshifting pseudoknot (13,457–13,546), or the 3´UTR (29,543–29,903) were used as queries in a BLASTn search of the NCBI Betacoronavirus database filtered for SARS-CoV-2 (Altschul et al. 1990; Camacho et al. 2009). An average of 3600 ± 160 hits were recovered from each query. The sequences recovered from BLASTn were aligned with MAFFT using the FFT-NS-2 algorithm to produce a multiple sequence alignment (MSA) (Katoh et al. 2002). The MSA was then input into WebLogo 3 to calculate the positional occupancy, entropy, and allele frequency for each

query (Supplementary Table 1) (Crooks et al. 2004). The occupancy defines the number of A, C, G, or U bases observed at each position (denoted as weight in the WebLogo3 output), the entropy defines the positional information content (lower value equals more variation), and the allele frequency defines the fractional occupancy of each nucleotide at each position. The results reveal high occupancy (>90%) from position 57 of the 5´UTR through position 29,836 of the 3´UTR, but the occupancy drops off significantly near the 5´ and 3´ ends of the genome (Fig. 1A), dipping below 20%. This is presumably due to difficulty of capturing the ends of the genome in sequencing library production. Nevertheless, even the extreme termini have coverage of more than 300 genomes. The positional entropy scores identify multiple variations in both low and high occupancy regions suggesting that variant entropy is not overly skewed by the terminal deficiencies in the genomic sequencing data. In total, fourteen variants with a minor allele frequency (MAF) of greater than 0.005 (0.5%) were identified by this approach.

To extend this analysis, we repeated the study with a second cohort of SARS-CoV-2 sequences recovered from the GISAID database on May 13, 2020 (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017). All sequences were downloaded from the database, converted into a blast library, then queried and analyzed as above with the NCBI cohort. An average of 23,900 ± 630 hits were recovered from each query. As with the NCBI cohort, occupancy is high (>90%) from position 55 through position 29,829 (Fig. 1B, Supplementary Table 1). Due to the large size of the GISAID cohort—6.6 times the size of the NCBI cohort—the termini are covered by thousands of genomes despite the relatively low occupancy. In total, seventeen variants with a MAF of at least 0.005 were identified in the GISAID cohort, eleven of which were also identified in the NCBI cohort (Fig. 1C, Supplementary Table 2). Combining the two analyses yields a total of twenty rapidly emerging variants in the structured regions of the viral genome. Of these, thirteen are transversions and seven are transitions. Eighteen are in noncoding regions

(2.9%, 18/602 positions evaluated), and the remaining two are silent mutations within the coding sequence of ORF1a or ORF1b (0.7%, 2/275 positions evaluated). Considering the larger GISAID cohort, there are 80 invariant residues (30.1%) in the 5´UTR, 90 (48.9%) in the ORF1a structured region, 58 (64.4%) in the frameshifting pseudoknot, and 131 (38.9%) in the 3´UTR. Thus, as expected, structures in the coding region seem to show a higher degree of conservation and fewer rapidly emerging alleles than non-coding regions, presumably due to the selective pressure of maintaining the protein coding sequence.

Variations in SL1 through SL4 of the 5´UTR

In MHV, stem loop 1 (SL1) plays a critical role in virus replication and is proposed to form long-range interactions with the 3´UTR (Zuniga et al. 2004; Li et al. 2008). Stem Loop 2 (SL2) contains a highly conserved sequence and structural elements thought to play a role in sgRNA synthesis (Liu et al. 2007; Liu et al. 2009; Chen and Olsthoorn 2010; Lee et al. 2011). The structure of Stem loop 3 (SL3) is less well conserved, but it contains the leftmost transcription regulatory sequence (TRS-L) required for template switching in sgRNA production (Zuniga et al. 2004; Sola et al. 2005; Yang and Leibowitz 2015). Stem loop 4 (SL4) contains an upstream open reading frame (uORF) that could reduce translation initiation at the ORF1a start codon and/or act as a spacer between 5´ structured elements and ORF1a (Raman et al. 2003; Yang et al. 2011; Wu et al. 2014). The precise role of these structures in SARS-CoV-2 infection is not known, but recently released structural predictions reveal all four stem loops are present in the SARS-CoV-2 genome (Fig. 2) (Ramya Rangan 2020).

SL1 and flanking single-stranded regions contain nine of the twenty variants identified in this analysis. In contrast, no rapidly emerging variants (MAF >0.005) are found in SL2 through SL4. To determine how the variations in SL1 influence the

secondary structure, we used RNAfold to calculate the most favored energy structure for each variant (Fig. 2D-E) (Lorenz et al. 2011). U2A (A=0.006/2519 GISAID, A=0.014/419 NCBI) and A4U (U=0.004/1799 GISAID, U=0.016/419 NCBI) have no influence on the stability of SL1 (ΔG=–8.50 kcal/mol for reference and both variants). U11G (G=0.001/4154 GISAID, G= 0.006/772 NCBI) had a small effect on the predicted stability (ΔG=–8.40 kcal/mol$_{U11G}$) due to loss of a stem terminal A-U pair. The A12U variant (U=0.002/4484 GISAID, U=0.011/829 NCBI) stabilized the predicted structure through formation of an additional base pair (ΔG=–10.2 kcal/mol$_{A12U}$). By contrast, A31U (U=0.005/7525 GISAID, U=0.029/1314 NCBI) is strongly destabilizing (ΔG=–5.40 kcal/mol), causing disruption of the lower stem.

Four rapidly emerging variations are found just downstream of the SL1 stem (Fig. 2A). A34U (U=0.009/7795 GISAID, U= 0.050/1350 NCBI), A35U (U=0.013/7846 GISAID, U=0.071/1380 NCBI), C36U (U=0.028/7967 GISAID, U=0.150/1400 NCBI) and C37A (A=0.003/8091 GISAID, A=0.018/1474 NCBI) variants frequently occur in combination. In the most common combination, three of the four positions ($A_{34}A_{35}C_{36}C_{37}$) are simultaneously replaced ($U_{34}U_{35}U_{36}C_{37}$). This variant has an allele frequency of UUUC=0.004/7795 (GISAID) and UUUC=0.023/1350 (NCBI). The variation extends the lower stem of SL1 by three base pairs, stabilizing the duplex by 2.4 kcal/mol (ΔG=–10.9 kcal/mol) (Fig. 2E). The second most frequent combination ($U_{34}U_{35}U_{36}A_{37}$) is present at an allele frequency of UUUA=0.003/7795 (GISAID) and UUUA=0.018/1350 (NCBI). This variant extends the SL1 lower stem by yet another base pair, increasing its overall stability by 3.4 kcal/mol (ΔG=–11.9 kcal/mol). Of these four positions, only C36U frequently exists as a single variation (U=0.013/7967 GISAID, 0.071/1400 NCBI). RNAfold analysis reveals no change in the stem loop structure or stability for the C36U variant.

Both combination variations are only found in samples sequenced from the United States, with the majority of them coming from the state of Washington. To better assess the relatedness between genomes containing $U_{34}U_{35}U_{36}C_{37}$ and $U_{34}U_{35}U_{36}A_{37}$ variants, we recovered the entire genomes of each example containing either extended SL1 stem variation from the GISAID cohort and aligned them using MAFFT (Katoh et al. 2002). The reference genome (Wuhan-Hu-1) was used as an outgroup. A radial maximal likelihood phylogenetic tree was calculated using the Tamura-Nei model in MEGAX, and the results plotted (Supp. Fig. 1) (Tamura and Nei 1993; Stecher et al. 2020). The phylogenetic relationship shows that both variants are represented in two different branches, but the variants tend to cluster separately within those branches. In one case, thirteen $U_{34}U_{35}U_{36}A_{37}$ genomes cluster within a node that is otherwise occupied $U_{34}U_{35}U_{36}C_{37}$, suggesting that $U_{34}U_{35}U_{36}$ variation arose first, and A_{37} arose as a secondary mutation. The impact of these variations on viral fitness or patient outcomes is not known.

Most of the rapidly emerging variants in SL1 enhance the stem loop structure. This suggests that SL1 stabilization is not overly deleterious to virus replication. In MHV, by contrast, destabilizing mutations of the lower stem are well tolerated in a cell model of virus replication, but mutations that increase the stability the lower stem block replication (Li et al. 2008). It is important to note that there is significant sequence divergence in this region between the two viruses that may explain this apparent dichotomy. Interestingly, the combination $U_{34}U_{35}U_{36}C_{37}$ variation co-occurs with the destabilizing A31U mutation 48.5% of the time, suggesting a potential compensatory role. Consistent with this hypothesis, the extension in SL1 rescues the destabilizing A31U variation by 2.1 kcal/mol (ΔG=–7.5kcal/mol). However, we note that none of the combination $U_{34}U_{35}U_{36}A_{37}$ variant genomes harbor A31U, so it is clear that the SL1 extension can exist in the absence of a compensatory destabilizing mutation. There are no rapidly

emerging variations within the upper stem or the loop of SL1, suggesting this region could be important to infection. Consistent with that hypothesis, mutations that destabilize the upper stem of MHV SL1 block virus replication (Li et al. 2008).

Variations in SL5 through SL10 at the 5´UTR/ORF1a junction

A large branched helical structure termed Stem Loop 5 is predicted to form at the interface between the 5´UTR and the N-terminal region of ORF1a (Fig. 2B) (Ramya Rangan 2020). This region contains three stems (SL5a, SL5b, and SL5c) connected by a helical junction. There is considerable sequence divergence among the coronaviridae in this structure, but the overall fold is largely preserved (Chen and Olsthoorn 2010). In SARS-CoV-2, the SL5a stem occludes the initiation codon for ORF1a, suggesting this structure must open prior to translation initiation. However, the SL5a stem is essential for virus replication in a bovine coronavirus (BCoV) model (Brown et al. 2007). The role of SL5c is more controversial, with one study demonstrating that the stem is dispensable (Yang and Leibowitz 2015), while a previous study showed that it is required (Brown et al. 2007).

We observed two rapidly emerging variants within the SL5 structured region with a minor allele frequency of greater than 0.005. The first, A187C (C=0.007/23832 GISAID, C=0.003/3407 NCBI), occurs within a bulged nucleotide of SL5a and is therefore not expected to alter the structure. The second, C241U (U=0.682/23760 GISAID, U=0.616/3376 NCBI) is in SL5b loop and is the most abundant rapidly emerging variant by far. There are no rapidly emerging mutations with an allele frequency of >0.005 in SL5c in either cohort.

Four additional stem loop structures (SL6-SL10) have been proposed within ORF1a (Fig. 2C) (Ramya Rangan 2020). The presence of SL6 and SL7 is observed in other coronaviridae, but the structures do not appear to have an important function

(Brown et al. 2007; Yang et al. 2015). There is one rapidly emerging variant within this region. C313U occurs within an internal loop region of SL6. The minor allele frequency of this variant is U=0.011/24227 GISAID, U=0.007/3732 NCBI. The variant is a silent mutation, converting a CUC^{Leu} codon to a CUU^{Leu} codon. As it occurs in an internal loop, it is expected to have no impact on the stem loop structure.

Variations in the 5´UTR that could have arisen through RNA editing

The two most abundant variants in the 5´UTR are both C to U transitions. C36U is observed in 2.8% of the sequences from the GISAID cohort, and C241U is observed in 68%. Excluding singletons, the average frequency of C to U transitions at all other positions in the 5´UTR is 0.04%. It is possible that the C36U and C241U variations arise repeatedly during virus replication, or they may have occurred early during the outbreak, or both. The type of transition and the relative abundance of the C36U and C241U variations suggest they might be hot spots for viral genome editing by host defense enzymes. The apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) enzymes are host encoded cytidine deaminases that edit cytidine to uridine in host nucleic acids (Lerner et al. 2018; Silvas and Schiffer 2019). They also target single stranded RNA and DNA virus genomes to affect an antiviral response.

If C36U and C241U substitutions arose at such high frequency because of C to U RNA editing, it might be possible to observe both nucleotides in the same sample of genomic RNA. cDNA produced from a mixed population of viral RNA harvested from an individual would be expected to include a weighted average of C and U in the sequencing reads that could be indicated as a degenerate Y (pyrimidine) in sequencing data, especially if there are near equal reads of each variation. Because WebLogo3 does not consider degenerate sequencing calls in its calculation of allele frequency (Crooks et al. 2004), we used SNP-sites v2.5.1 and VCFtools v0.1.7 to recalculate the

allele frequency inclusive of degenerate bases (Danecek et al. 2011; Page et al. 2016). We calculated the average frequency of all C to Y transitions in the 5´UTR using the larger GISAID cohort, excluding the two candidate editing sites (C36U and C241U) and singletons. The average C to Y transition frequency is 0.014%. By contrast, the frequency of C36Y is 0.063%, 4.5-fold greater than the average, and the frequency of C241Y is 0.18%, 12.9-fold greater than the average. This apparent increase in pyrimidine degeneracy is consistent with the possibility that APOBEC enzymes edit both positions. However, we cannot formally rule out the possibility that some people were co-infected with both variants leading to the degenerate base call, or that C36U and C241U frequently arise via some other mechanism during viral replication. The impact of either variation on viral fitness remains to be determined.

Variations in the frameshifting pseudoknot at the ORF1a/ORF1b junction

An RNA pseudoknot is found at the junction of ORF1a and ORF1b (Fig. 3A) (Ramya Rangan 2020). This structure is involved in -1 programmed ribosome frameshifting, where translating ribosomes shift frame by one nucleotide to the left. Efficient frameshifting requires both a “slippery” sequence and a downstream stable RNA structure (Brierley et al. 1989; Brierley et al. 1992). Like SARS-CoV and MHV, the SARS-CoV-2 pseudoknot has three stems instead of two typically found in pseudoknot structures (Brierley and Dos Ramos 2006; Giedroc and Cornish 2009). A previous study comparing SARS-CoV, MHV, and hybrid variants found that both viral pseudoknots led to approximately the same extent of programmed frameshifting (~20%), but hybrid mutant variants in loop 3 that stabilize the pseudoknot structure increased frame shifting up to 90% (Plant et al. 2010). The same study revealed that silent mutations in the SARS-CoV slippery site reduced programmed frame shifting by three-fold and also blocked viral infection in a cell culture model. Thus, the function of the slippery sequence

and the pseudoknot structure is to ensure that production of ORF1a and ORF1ab polyproteins occurs at appropriate stoichiometric ratios, critical to viral fitness.

We identified one rapidly emerging variant in the frameshifting pseudoknot. C13536U (U=0.015/23306 GISAID, U=0.003/3440 NCBI) is a silent mutation (UAC^{Tyr}:UAU^{Tyr}) located within stem 2 (Fig. 4). C13536 normally forms a Watson-Crick pair with G13493. Mutation to U is expected to cause the formation of a U13536:G13493 wobble pair, which has comparable stability to a Watson-Crick pair but alters the backbone geometry shifting the G residue into the minor groove. To get a better understanding of how this U-G pair might impact the tertiary structure and thus the function of the frameshifting pseudoknot, we used RNAcomposer to build a three-dimensional model of the reference sequence and the C13536U variant (Fig 4B) (Popenda et al. 2012). In the reference model, G13485 forms a base triple with the C13536:G13493 pair (Fig. 4A). The exocyclic amine of C13536 donates a hydrogen bond to the O6 of G13485 in loop 1. In the C13536U model, this base triple cannot form as the hydrogen bond donor is lost. This could conceivably reduce the stability of stem 2, which would be expected to cause less efficient -1 programmed ribosomal frameshifting. More work will be necessary to define exactly how this variation perturbs the structure and if it alters the stoichiometry of viral protein synthesis.

Variations in the 3´UTR in the BSL and PK

Betacoronaviridae 3´UTRs contain a bi-stable molecular switch formed by two mutually exclusive structural conformers, including one that extends the lower stem of the bulged stem loop (BSL), and a second that folds into a pseudoknot (PK, Fig. 3B) (Goebel et al. 2004). Both structured elements are present in MHV, SARS, and MERS, though the sequence diverges significantly between them (Hsue and Masters 1997; Hsue et al. 2000; Goebel et al. 2004). In MHV and BCoV, the BSL and the PK structure

are required for viral replication (Hsue and Masters 1997; Williams et al. 1999; Hsue et al. 2000; Goebel et al. 2004). Mutations that stabilize one form over the other prevent replication. It is proposed that competition between the two structures plays a regulatory role in antigenomic RNA synthesis, but the exact mechanism remains to be determined (Goebel et al. 2004).

There are two rapidly emerging variants in the 5´ portion of the 3´UTR. G29540A is present at a MAF of A=0.008/24313 (GISAID) and A=0.014/3550 (NCBI). This variant lies within a single-stranded region that precedes the BSL structure and as such is not predicted to affect the structure or the molecular switch. In contrast, the G29553A variant (A=0.012/24216 GISAID, A=0.064/3551 NCBI) disrupts a G:C pair in the extended BSL molecular switch conformer that could potentially favor the alternate PK structure. Alternatively, the A substitution may pair with the otherwise bulged U29607 nucleotide, partially compensating for the loss in of the G:C pair. Consistent with the latter possibility, RNAfold predicts that the stability of the reference BSL conformer is -20.20 kcal/mol, while the stability of the G29553A variant conformer is -18.30 kcal/mol and includes a newly formed A:U pair. It remains to be determined how modulation of the internal equilibrium of the molecular switch affects SARS-CoV-2 pathogenesis.

RNA editing by APOBEC enzymes could lead to rapidly emerging G to A transitions if the antigenomic strand is edited during viral replication. Antigenomic cytidine deamination recodes C to U, which would be read as an A during replication of the genomic strand. To assess this possibility that the G29540A and G29553A variations arose through RNA editing, we looked for the degenerate base “R” (either purine base) in both sequencing cohorts using SNP-sites and VCFtools as described above (Danecek et al. 2011; Page et al. 2016). There were no degenerate R alleles in the GISAID or NCBI databases at either position, suggesting that neither is produced through frequent APOBEC-mediated editing of the antigenomic strand.

Variants of the hypervariable region, the S2M structure, and the S3 and S4 stems

An extended multiple stem loop structure exists downstream of the 3´UTR pseudoknot (Fig 3C) (Ramya Rangan 2020). This structure contains a hypervariable region (HVR) that folds into a bulged stem loop. The HVR is highly divergent in coronaviridae with the exception of a strictly conserved single-stranded 8-mer sequence referred to as the octanucleotide motif (Goebel et al. 2007; Madhugiri et al. 2014). The function of this region is not well understood, but deletion of the HVR including the conserved 8mer element has no effect on MHV replication in cultured cells (Goebel et al. 2007). An apparent selfish genetic element, termed S2M, exists within the bulged stem loop of the HVR (Ramya Rangan 2020). This element is found in many but not all coronaviridae, and is also found in many other families of positive ssRNA viruses, suggesting it can be horizontally transferred (Tengs et al. 2013; Tengs and Jonassen 2016). The sequence is highly conserved in all viruses where it is found. This element is not present in MHV, and its function (if any) is unknown. Two shorter stems, termed S4 and S3, are also present. Mutations that disrupt S4 have no effect on MHV replication, but S3 appears to be important (Liu et al. 2013).

Five rapidly emerging variants are found in this region of the SARS-CoV-2 genome. Two disrupt Watson-Crick pairs in the HVR bulged stem loop. The A29683U variation is present at a MAF of U=0.006/23540 (GISAID) and U=$3x10^{-4}$/3545 (NCBI), while the A29700G is present at a MAF of G=0.009/23403 (GISAID) and G=0.022/3541 (NCBI). Both variants reduce the stability of a simplified model HVR structure that eliminates the S2M region in RNAfold calculations (ΔG=–24.20 kcal/mol_{ref}, –22.20 kcal/$mol_{A29683U}$, –23.9 kcal/$mol_{A29700G}$, Supplemental Fig. 2), with the A29700U variant forming a compensatory G:U wobble pair. The G29711U variant is present at a MAF of U=0.007/23366 (GISAID), U=0.004/3537 (NCBI). This variant disrupts the GNRA class

tetraloop structure in the loop of the HVR bulged stem structure, and is predicted to modestly destabilize the fold (ΔG = -23.5 kcal/mol$_{G29711U}$). The presence of multiple disruptive variations in this region of the SARS-CoV-2 3´UTR, coupled to previous reports that the HVR is dispensable for MHV replication, suggests that structures are not critical to viral replication. More work will be needed to understand whether the structures in the HVR contribute to SARS-CoV-2 replication or viral fitness.

The presence of the rapidly emerging A29700G transition suggests the possibility that it might arise through adenosine deaminase acting on RNA (ADAR) RNA editing activity. ADARs convert adenosine residues to inosine in double stranded regions of RNA (Keegan et al. 2017). As such, they can play an important role in antiviral response, targeting double stranded RNA viruses and other viruses (including betacoronaviruses) that go through a double stranded RNA intermediate (Tomaselli et al. 2015). During viral replication, inosine residues in the genomic strand would template the incorporation of a C in place of a U during minus strand synthesis, leading to A to G transitions during viral replication. As above, we used SNP-sites and VCFtools to measure the frequency of the degenerate R base at A29700G (Danecek et al. 2011; Page et al. 2016). No degenerate R nucleotides are present in the GISAID cohort, suggesting that frequent RNA editing by ADAR enzymes is not responsible for rapid A29700G emergence.

The final two rapidly emerging variations lie within the enigmatic S2M loop. The structure of the S2M loop from SARS-CoV has been solved by X-ray crystallography (Fig. 5) (Robertson et al. 2005). Its prevalence in positive strand ssRNA viral genomes, its position near the 3´-terminus, and its high degree of sequence conservation all imply a functional role (Tengs and Jonassen 2016). However, not all betacoronaviruses have the S2M loop, and swapping an S2M-containing region from the SARS-CoV 3´UTR with an S2M-deficient MHV region did not alter or improve virus replication in vitro (Goebel et al. 2007). As such, its role in viral replication is unclear.

The first rapidly emerging variation in S2M is G29734C (C=0.008/23285 GISAID, C=0.003/3525 NCBI). In the SARS-CoV S2M crystal structure, this position forms a non-canonical G:A pair (Fig. 5B) (Robertson et al. 2005). The N2 exocyclic amine donates a hydrogen bond to the N1 position of its adenosine partner, and the 2´-hydroxyl group donates a hydrogen bond to the N3 moiety. Substitution of a C in place of G is incompatible with the hydrogen bonds formed in the G:A pair and as such is likely to destabilize the S2M tertiary structure. The second rapidly emerging variation in S2M is G29742U (U=0.009/28235 GISAID, U=0.019/3526 NCBI). This base is involved in a base quadruple, pairing through its Watson-Crick face with a cytidine residue, but also interacting with the C of a parallel G:C pair packed tightly into its minor groove (Fig. 5B) (Robertson et al. 2005). The U variation is incompatible with both the canonical and non-canonical pairings at this position and is likely to be highly destabilizing to the fold.

To evaluate the effect of both of the rapidly emerging variations on the structure of the S2M loop, we performed molecular dynamics (MD) simulations of the S2M loop from SARS-CoV and of both variants. The X-ray structure of the SARS-CoV S2M loop served as initial conformation for all simulations (Robertson et al. 2005). The results confirmed our predictions: both variations destabilize the structure of the S2M domain. In each case we observed that the structure of the S2M loop undergoes a transition quickly after equilibration (within the first 60 ns) to a less compact conformation as shown in Figure 6A-F. To quantify the extent of the transition, we measured the length of the maximum dimension of the S2M loop and observed an extension (Fig. 6I-K). Surprisingly, we observed that the S2M loop of SARS-CoV also samples similarly extended conformations (Fig. 6I), although this happens more slowly and with a lower frequency than in the other two variants, in only two out of the four trajectories that we collected (Supplemental Fig. 3A-C). Analysis of the MD trajectories showed that one important element for the stability of the overall structure of the S2M loop is the

stabilizing interaction between G29734 and A29756. When these two nucleotides are in close proximity to one another, the stem tertiary structure is reorganized by bending. This interaction is destabilized in the G29734C variant because C cannot form hydrogen bonds with A29756. However, we observed that the formation of an A-G base pair is not stable in all three sequence variations (Fig. 6G). In the absence of this base pair, A29756 can move far from G/C29734 (Fig. 6L-N) leading to the sampling of more extended conformations of the loop as shown in figure Fig6 C-F. With the exception of the tails that unfold and refold, the secondary structure of the S2M loop is generally well preserved throughout the simulations, even when sampling more extended conformations as a result of the loss of tertiary interactions (Fig. 6G). A major difference between these variants is that the base pairing is not as stable in the stem-loop structure in the presence of the G29742U variation (Fig. 6G-H). This variation in sequence affects the stability of the base quartet described above (Fig. 5A), and ultimately impacts the stability of the adjacent base pairs (Fig. 6H). The base quartet is at the junction between the two helices and is important to set their relative orientations. The G29742U variation causes a local melting of the secondary structure and a disruption of this junction, ultimately destabilizing the compact conformation of the S2M loop (Fig. 6C).

Discussion

The global SARS-CoV-2 pandemic has led to an explosion in whole genome sequencing of naturally occurring viral isolates. These data have been useful in the identification of rapidly emerging variations that impact viral protein structure and function (Kim et al. 2020b; Pachetti et al. 2020). They have also been used to monitor the spread of the virus through molecular phylogeny (Chu et al. 2020; Coronaviridae Study Group of the International Committee on Taxonomy of 2020; Forster et al. 2020).

Here, we have used available data to investigate how rapidly emerging variants could impact structured cis-regulatory elements in the virus genome. These elements govern viral replication, subgenomic RNA synthesis, and translation control in other betacoronaviruses (Yang and Leibowitz 2015; Madhugiri et al. 2016). Rapidly emerging variants could enhance or dampen viral pathogenesis and overall fitness, which could affect the extent and duration of the outbreak. As such, it is critically important to understand how such variations arise, and what regions of the genome are most prone to mutation.

Due to the burden of the SARS-CoV-2 outbreak, there is renewed interest in the development of novel strategies to treat betacoronavirus infections. Functional RNA structures in the viral genome could provide new targets for small molecule therapeutic development. Many antibiotics work through interactions with ribosomal RNA structure, and RNA targeting small molecule drugs are currently approved or in development for a variety of infectious and genetic diseases (Guan and Disney 2012). The SARS-CoV-2 genome has many structured elements that could be targeted, including SL1-SL4 in the 5´UTR, the frameshifting pseudoknot at the ORF1a and ORF1b boundary, and the molecular switch in the 3´UTR. The results presented here suggest that the hypervariable region, including the S2M structure, might be less well suited to targeted drug development. Structures with rapidly emerging variations are problematic for drug development as well, as the relatively high viral mutation rate, coupled to its potential to be edited by APOBEC and ADAR enzymes, could lead to the rapid evolution of resistant variants.

Similarly, hybridization-guided therapeutics, such as antisense oligonucleotides, small interfering RNAs, and CRISPR-derived drugs could potentially be targeted to the SARS-CoV-2 genome. Unstructured regions in noncoding regions of the viral genome make particularly compelling targets, as access will not be blocked by RNA structure or

transit of the ribosome. However, because these strategies rely on base complementarity to achieve target specificity, rapid virus evolution could prove their Achilles' heel. The data presented here identify regions less prone to variation, making them better candidates for RNA-guided therapeutics.

The observation that SL1 is prone to rapidly emerging variations is interesting, as this region is not only present on the positive strand of the viral genome, but is also found on all subgenomic RNAs (Kim et al. 2020a). Moreover, the complement to SL1 in antigenomic RNA is likely recognized by viral RNA-dependent RNA polymerase (RdRP) to produce genomic copies of the viral RNA. As such, it could make a good target for therapeutic development. However, the presence of multiple variations, often in combination, makes strategies that rely upon base pairing unlikely to be effective for all virus subtypes. The diversity of variations that enhance the stability of SL1, including variations that lengthen the stem, suggests that SL1 stability is important to SARS-CoV-2 replication. But if stability matters more than sequence identity, we can expect the evolution of rapid resistance to therapeutics designed to modulate SL1 stability.

The bi-stable molecular switch in the 3´UTR is potentially the most compelling structure for targeted drug discovery. It is conceptually straightforward to design antisense oligonucleotides that lock the switch into one conformer or the other. Both conformers are necessary for MHV replication, and only one rapidly emerging variant of minimal consequence was identified in this region. It is likely that this switch plays a role in SARS-CoV-2 replication, as has been observed in other betacoronaviruses. More work will be necessary to assess its potential as a drug target.

RNA editing appears to play a role in two rapidly emerging variations near stem loop structures in the 5´UTR. The prevalence of RNA editing of the viral genome is not known, and it remains unclear whether editing affects viral fitness or pathogenesis. It will

be interesting to assess the extent of RNA editing during active infection, a task that would probably be best achieved through direct RNA sequencing (Kim et al. 2020a).

The analyses presented in this study will only improve as more sequencing data are added to available repositories (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017). It is possible that identification of more rapidly emerging variants will clarify some of the remaining ambiguities. The results presented here highlight the power of high-throughput sequencing of viral genomes to define viral cis-regulatory elements, and stand as a testament to the researchers collecting, sequencing, and sharing viral genomic data to help quell the impact of this tragic and overwhelming pandemic.

Materials and Methods

Calculation of allele frequency and occupancy

The Wuhan-Hu-1 isolate of SARS-CoV-2 (GenBank Accession Number MN908947) was used as a reference genome. The sequences corresponding to the 5´UTR (1-265), the ORF1a structured region (266-450), the frameshifting pseudoknot (13457-13546), and the 3´UTR (29534-29870) were used as queries in a BLASTN search (Altschul et al. 1990). For the NCBI cohort, BLASTN searches were performed against the NCBI betacoronavirus database of 11,495 (as of May 14th, 2020) betacoronavirus sequences. Searches were performed using the web portal with default parameters except “max target sequences” was set to 20,000. BLAST hits were filtered by organism for “severe acute respiratory syndrome coronavirus 2”, and the remaining hits were downloaded as a hit table and aligned sequences. A multiple sequence alignment was prepared using a locally installed copy of MAFFT version 7.464 using the default FFT-NS-2 algorithm (Katoh et al. 2002). The output file was then analyzed with a locally installed copy of WebLogo3 version 3.6.0 (Crooks et al. 2004). The resultant logo data table contains the calculated sequence entropy, the occupancy (weight), and the

count number for each base at each position. The allele frequency was then calculated by dividing the count number by the sum of all counts for all four bases. The minor allele frequency is defined as the frequency of the second most abundant allele and is typically represented by the format variant=frequency/counts.

For the GISAID cohort, 24,468 curated SARS-CoV-2 genomic sequences were downloaded from the GISAID Initiative EpiCoV database (on May 13th, 2020) under the terms of their data access agreement (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017). The genomic sequences were compiled into a blast library using a locally installed copy of BLAST+ version 2.8.1, and queried using the command line tool blastn as describe above with the exception that the max_target_seqs flag was set to 30,000 (Camacho et al. 2009). Aligned sequences were recovered from the resulting hit table using a custom shell command, then analyzed using MAFFT and WebLogo3 as described for the NCBI cohort above.

Calculation of minimum free energy structures:

The sequence corresponding to SL1 and flanking nucleotides (1-37), the BSL and flanking nucleotides (29,547-29,643), or variations thereof were input into the web server for RNAfold using the default parameters (Lorenz et al. 2011). The calculated ΔG for the minimum energy structure, the ensemble free energy, the frequency of the minimum free energy structure in the ensemble, the ensemble diversity, and the secondary structure in dot-bracket notation were recorded in Supplementary Table 3. The bulged stem loop in the HVR (29,627-29,834) and variants thereof were analyzed by the same approach, except nucleotides 29,721 through 29,800 were removed to simplify the overall structure. RNAfold was not able to accurately calculate the secondary structure of the region surrounding the s2m structure.

Phylogenetic analysis of SL1 variants

Examples of the specific combination variants $U_{34}U_{35}U_{36}C_{37}$ and $U_{34}U_{35}U_{36}A_{37}$ were recovered from the GISAID cohort 5´UTR BLASTn hits by searching for the variation plus two invariant nucleotides on either side using custom shell commands. Each variant combination was searched using this approach to count the number of occurrences and to recover the sequence. Following alignment, the hits were inspected to ensure the correct pattern match, and in one instance, manually edited to remove an example where the search pattern identified a match at the incorrect position. The sequence IDs were then used to recover the intact genomic sequence from the GISAID cohort library. MAFFT was then used to generate multiple sequence alignments of the entire genome using the procedure outlined above. Output files were loaded into MEGAX version 10.1.8 (for Mac), and the maximum likelihood tree was calculated using the Tamura-Nei model (Tamura and Nei 1993; Stecher et al. 2020).

Degenerate base frequency analysis

Because WebLogo3 does not consider degenerate base calls, the MAFFT-generated MSA files outlined above were converted into VCF format using a locally installed copy of SNP-sites version 2.5.1. The allele frequencies were then re-analyzed using VCFtools version 0.1.17 (Danecek et al. 2011; Page et al. 2016). The abundance of Y or R degenerate base calls for specific positions was calculated from the overall frequency each base, excluding counts for symbols that denote the absence of a base at the given position.

Molecular modeling of the frameshifting pseudoknot and variants

Three-dimensional molecular models of the frame shifting pseudoknot (13,472-13,543) and variants thereof were calculated using the RNAcomposer web server. The modeling

algorithm was guided using dot-bracket notation to match the recently published secondary structure of SARS-CoV-2 (Popenda et al. 2012; Ramya Rangan 2020). The output PDB files were visualized and analyzed in Pymol version 1.7.6.0.

Molecular Dynamics simulations

Molecular dynamics simulations were performed with NAMD 2.13 (Phillips et al. 2005) using the CHARMM 36 force field (MacKerell et al. 1998). The X-ray structure of SARS-CoV S2M RNA (PDB code 1XJR) was used as the initial structure for the SARS-CoV and the G29734C and G29742U variants after performing the respective mutations using the Mutator plug-in of VMD (Humphrey et al. 1996). In VMD, each structure was solvated with a water box with explicit TIP3P (Jorgensen et al. 1983; MacKerell et al. 1998) solvent and an ionic concentration of 0.15 M NaCl. The box was much larger than the initial structure (cubic box of 106 Å) to allow for extension of the RNA. Four independent trajectories of each system (WT, G11C, and G19U) were generated with the following procedure. The solvated structures were first minimized using the conjugate gradient method for 500 steps to relax any high energy contacts or unphysical geometries. An additional 2,000 steps of conjugate gradient minimization were performed with the heavy atoms of the RNA and the two Mg^{2+} ions restrained with a harmonic constraint force of 10 kcal/mol^{-1} Å^{2}. Next, particle velocities were randomly assigned from the Maxwell distribution and equilibration was performed in the isothermal-isobaric ensemble gradually decreasing the restraints to zero (using 9 stages of 50 ps each). The pressure and temperature were maintained at 1 atm and 298 K using Langevin dynamics and the Nosé-Hoover Langevin piston method. The SHAKE constraint algorithm (Ryckaert et al. 1977) was used to allow a 2 fs time step. The particle mesh Ewald method (Essmann et al. 1995) was used to calculate electrostatic interactions with periodic boundary conditions. Production trajectories were then

collected using the isothermal-isobaric ensemble for 180 ns. The Mg^{2+} ions remained stably coordinated throughout the trajectories.

Trajectory analysis was performed with VMD (Humphrey et al. 1996) and structures were visualized with VMD using the STRIDE algorithm for secondary structure identification (Frishman and Argos 1995). A hydrogen bond is defined by a donor-acceptor distance of less than 3.5 Å and a donor-hydrogen-acceptor angle of $130° < \theta < 180°$. The ring distance between a pair of nucleotides was calculated as the center of mass distance between the atoms N1, C2, N3, C4, C5, and C6. The maximum dimension of the RNA structure was calculated by aligning the trajectories with the starting structure (where the longest axis of the RNA was aligned with the y-axis) and finding the maximum dimension of the box needed to fully contain the RNA.

Acknowledgements:

We thank and gratefully acknowledge the originating laboratories responsible for obtaining specimens and the submitting laboratories that generated genetic sequencing data shared through the GISAID Initiative. A full list of these laboratories is presented in the supplementary table 4. We thank Jeremy Luban, Brian Kelch, and members of the Ryder laboratory for commenting on the manuscript and for helpful discussions. This work was supported in part by the COVID19 Pandemic Research Fund at UMass Medical School. Research in our labs is funded by NIH grant R21HG011001 and R01GM117237 to S.P.R. and NIH grant R01 GM139316 to F.M. and S.P.R.

Author Contributions:

SPR performed the sequence analyses and molecular modeling. BRM collected the MD trajectories. BRM and FM performed the MD trajectory analyses. SPR wrote the paper under the advisement of all authors.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L, Nalla A, Pepper G, Reinhardt A, Xie H et al. 2020. Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv*.

Brierley I, Digard P, Inglis SC. 1989. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell* **57**: 537-547.

Brierley I, Dos Ramos FJ. 2006. Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res* **119**: 29-42.

Brierley I, Jenner AJ, Inglis SC. 1992. Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *J Mol Biol* **227**: 463-479.

Brown CG, Nixon KS, Senanayake SD, Brian DA. 2007. An RNA stem-loop within the bovine coronavirus nsp1 coding region is a cis-acting element in defective interfering RNA replication. *J Virol* **81**: 7716-7724.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chan JF, Yuan S, Kok KH, To KK, Chu H, Yang J, Xing F, Liu J, Yip CC, Poon RW et al. 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**: 514-523.

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

Chu HY, Englund JA, Starita LM, Famulare M, Brandstetter E, Nickerson DA, Rieder MJ, Adler A, Lacombe K, Kim AE et al. 2020. Early Detection of Covid-19 through a Citywide Pandemic Surveillance Platform. *N Engl J Med*.

Coronaviridae Study Group of the International Committee on Taxonomy of V. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5**: 536-544.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188-1190.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA et al. 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* **348**: 1967-1976.

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**: 33-46.

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. 1995. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**: 8577-8593.

Forster P, Forster L, Renfrew C, Forster M. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* **117**: 9241-9243.

Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23**: 566-579.

Geoghegan JL, Holmes EC. 2018. Evolutionary Virology at 40. *Genetics* **210**: 1151-1162.

Giedroc DP, Cornish PV. 2009. Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res* **139**: 193-208.

Goebel SJ, Miller TB, Bennett CJ, Bernard KA, Masters PS. 2007. A hypervariable region within the 3' cis-acting element of the murine coronavirus genome is nonessential for RNA synthesis but affects pathogenesis. *J Virol* **81**: 1274-1287.

Goebel SJ, Taylor J, Masters PS. 2004. The 3' cis-acting genomic replication element of the severe acute respiratory syndrome coronavirus can function in the murine coronavirus genome. *J Virol* **78**: 7846-7851.

Guan L, Disney MD. 2012. Recent advances in developing small molecules targeting RNA. *ACS Chem Biol* **7**: 73-86.

Hsue B, Hartshorne T, Masters PS. 2000. Characterization of an essential RNA secondary structure in the 3' untranslated region of the murine coronavirus genome. *J Virol* **74**: 6911-6921.

Hsue B, Masters PS. 1997. A bulged stem-loop structure in the 3' untranslated region of the genome of the coronavirus mouse hepatitis virus is essential for replication. *J Virol* **71**: 7567-7578.

Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph* **14**: 33-38, 27-38.

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**: 926-935.

Kang H, Feng M, Schroeder ME, Giedroc DP, Leibowitz JL. 2006a. Putative cis-acting stem-loops in the 5' untranslated region of the severe acute respiratory syndrome coronavirus can substitute for their mouse hepatitis virus counterparts. *J Virol* **80**: 10600-10614.

-. 2006b. Stem-loop 1 in the 5' UTR of the SARS coronavirus can substitute for its counterpart in mouse hepatitis virus. *Adv Exp Med Biol* **581**: 105-108.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

Keegan L, Khan A, Vukic D, O'Connell M. 2017. ADAR RNA editing below the backbone. *RNA* **23**: 1317-1328.

Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. 2020a. The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**: 914-921 e910.

Kim SJ, Nguyen VG, Park YH, Park BK, Chung HC. 2020b. A Novel Synonymous Mutation of SARS-CoV-2: Is This Possible to Affect Their Antigenicity and Immunogenicity? *Vaccines (Basel)* **8**.

Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W et al. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* **348**: 1953-1966.

Lee CW, Li L, Giedroc DP. 2011. The solution structure of coronaviral stem-loop 2 (SL2) reveals a canonical CUYG tetraloop fold. *FEBS Lett* **585**: 1049-1053.

Lerner T, Papavasiliou FN, Pecori R. 2018. RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-to-U RNA Editing Machinery and Its Implication in Human Disease. *Genes (Basel)* **10**.

Li L, Kang H, Liu P, Makkinje N, Williamson ST, Leibowitz JL, Giedroc DP. 2008. Structural lability in stem-loop 1 drives a 5' UTR-3' UTR interaction in coronavirus replication. *J Mol Biol* **377**: 790-803.

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY et al. 2020. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **382**: 1199-1207.

Liu P, Li L, Keane SC, Yang D, Leibowitz JL, Giedroc DP. 2009. Mouse hepatitis virus stem-loop 2 adopts a uYNMG(U)a-like tetraloop structure that is highly functionally tolerant of base substitutions. *J Virol* **83**: 12084-12093.

Liu P, Li L, Millership JJ, Kang H, Leibowitz JL, Giedroc DP. 2007. A U-turn motif-containing stem-loop in the coronavirus 5' untranslated region plays a functional role in replication. *RNA* **13**: 763-780.

Liu P, Yang D, Carter K, Masud F, Leibowitz JL. 2013. Functional analysis of the stem loop S3 and S4 structures in the coronavirus 3'UTR. *Virology* **443**: 40-47.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Lv H, Wu NC, Tak-Yin Tsang O, Yuan M, Perera R, Leung WS, So RTY, Chun Chan JM, Yip GK, Hong Chik TS et al. 2020. Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections. *Cell Rep*: 107725.

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**: 3586-3616.

Madhugiri R, Fricke M, Marz M, Ziebuhr J. 2014. RNA structure analysis of alphacoronavirus terminal genome regions. *Virus Res* **194**: 76-89.

-. 2016. Coronavirus cis-Acting RNA Elements. *Adv Virus Res* **96**: 127-163.

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC et al. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med* **18**: 179.

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**: e000056.

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. 2005. Scalable molecular dynamics with NAMD. *J Comput Chem* **26**: 1781-1802.

Pinto D, Park YJ, Beltramello M, Walls AC, Tortorici MA, Bianchi S, Jaconi S, Culap K, Zatta F, De Marco A et al. 2020. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*.

Plant EP, Rakauskaite R, Taylor DR, Dinman JD. 2010. Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. *J Virol* **84**: 4330-4340.

Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW. 2012. Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40**: e112.

Raman S, Bouma P, Williams GD, Brian DA. 2003. Stem-loop III in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *J Virol* **77**: 6720-6730.

Ramya Rangan INZ, and Rhiju Das. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA*.

Robertson MP, Igel H, Baertsch R, Haussler D, Ares M, Jr., Scott WG. 2005. The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol* **3**: e5.

Ryckaert J-P, Ciccotti G, Berendsen HJC. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**: 327-341.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**.

Silvas TV, Schiffer CA. 2019. APOBEC3s: DNA-editing human cytidine deaminases. *Protein Sci* **28**: 1552-1566.

Sola I, Moreno JL, Zuniga S, Alonso S, Enjuanes L. 2005. Role of nucleotides immediately flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA synthesis. *J Virol* **79**: 2506-2516.

Stadler K, Masignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, Rappuoli R. 2003. SARS--beginning to understand a new virus. *Nat Rev Microbiol* **1**: 209-218.

Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* **37**: 1237-1239.

Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF. 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol* **24**: 490-502.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512-526.

Tengs T, Jonassen CM. 2016. Distribution and Evolutionary History of the Mobile Genetic Element s2m in Coronaviruses. *Diseases* **4**.

Tengs T, Kristoffersen AB, Bachvaroff TR, Jonassen CM. 2013. A mobile genetic element with unknown function found in distantly related viruses. *Virol J* **10**: 132.

Tomaselli S, Galeano F, Locatelli F, Gallo A. 2015. ADARs and the Balance Game between Virus Infection and Innate Immune Cell Response. *Curr Issues Mol Biol* **17**: 37-51.

Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y et al. 2020. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA*.

Williams GD, Chang RY, Brian DA. 1999. A phylogenetically conserved hairpin-type 3' untranslated region pseudoknot functions in coronavirus RNA replication. *J Virol* **73**: 8349-8355.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* **579**: 265-269.

Wu HY, Guan BJ, Su YP, Fan YH, Brian DA. 2014. Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5'-untranslated-region mutants. *J Virol* **88**: 846-858.

Yang D, Leibowitz JL. 2015. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res* **206**: 120-133.

Yang D, Liu P, Giedroc DP, Leibowitz J. 2011. Mouse hepatitis virus stem-loop 4 functions as a spacer element required to drive subgenomic RNA synthesis. *J Virol* **85**: 9199-9209.

Yang D, Liu P, Wudeck EV, Giedroc DP, Leibowitz JL. 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.

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* **367**: 1814-1820.

Zhang YZ, Shi M, Holmes EC. 2018. Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **172**: 1168-1172.

Zheng J. 2020. SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat. *Int J Biol Sci* **16**: 1678-1685.

Zhong NS, Zheng BJ, Li YM, Poon, Xie ZH, Chan KH, Li PH, Tan SY, Chang Q, Xie JP et al. 2003. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet* **362**: 1353-1358.

Zuniga S, Sola I, Alonso S, Enjuanes L. 2004. Sequence motifs involved in the regulation of discontinuous coronavirus subgenomic RNA synthesis. *J Virol* **78**: 980-994.

Zust R, Miller TB, Goebel SJ, Thiel V, Masters PS. 2008. Genetic interactions between an essential 3' cis-acting RNA pseudoknot, replicase gene products, and the extreme 3' end of the mouse coronavirus genome. *J Virol* **82**: 1214-1228.

Figure Legends

Figure 1. *SARS-CoV-2 sequence occupancy and entropy* **A**. Fractional occupancy (left axis, dashed lines) and positional entropy (right axis, solid lines) of the NCBI cohort calculated by WebLogo3 as displayed as a function of SARS-CoV-2 genome coordinates. This analysis focused only on well characterized betacoronavirus structured elements. The relative positional relationships of each region are marked. Hash marks donate areas of the entire genome that were not considered in this study. **B.** The same representation as in (A), but calculated using the GISAID EpiCoV cohort. **C.** Venn diagram of rapidly emerging variations in the GISAID cohort, NCBI cohort, or both.

Figure 2. *Rapidly emerging variants in SL1-SL10 of the 5´UTR and ORF1a:* **A.** The predicted secondary structure of SL1 through SL4 is shown. The position and identity of rapidly emerging variants is denoted by an arrow and a letter. The minor allele frequency for each variant is given in the form variant=frequency/total (cohort). **B.** The predicted secondary structure of SL5 is shown. Rapidly emerging variants are denoted by an arrow, and the identity of the variation is given next to the arrow. The position of the ORF1a start codon is labeled. The minor allele frequency for each variation is given. **C.** The predicted secondary structure of SL6-SL10 is shown. Variations are labeled as in panel A. The minor allele frequency for the single variation is shown above, and includes the identity of the resultant silent codon change. **D.** The structures of single SL1 variants are shown. The specific variant is shown in red. The variant ID is given above the structure. The RNAfold calculated minimum free energy structure is presented in the diagram, and its thermodynamic stability is given below. **E.** Same as in D, but for the two prevalent combination variants that extend the length of SL1.

Figure 3. *Rapidly emerging of the frameshifting pseudoknot and the 3´UTR:* **A.** The secondary structure of the frameshifting pseudoknot is shown. **B.** The secondary structure of the pseudoknot conformer or the extended BSL conformer is shown of the bi-stable molecular switch in the 3´UTR is shown. **C.** The secondary structure of the 3´ half of the SARS-CoV-2 genome is shown. This region includes the HVR, the octanucleotide motif (8-mer), the S2M structure, and the S4 and S5 stems. In all paneles, the position of rapidly emerging variants is labeled as in figure 2.

Figure 4. *Structure of the frameshifting pseudoknot* **A.** Comparison of the base triple observed in the reference model (top) and in the U13536 variation model (bottom). Hydrogen bonds are denoted by dashed lines. The U13536 variant is colored in red. **B.** The molecular model of the frame shifting pseudoknot calculated by RNAcomposer is shown. Stems 1-3 are labeled in colors corresponding to those shown on the secondary structure in figure 3.

Figure 5. *Rapidly emerging variations in S2M disrupt key tertiary interactions:* **A.** The crystal structure of the S2M region from SARS-CoV is shown. The position of two rapidly emerging variations in SARS-CoV-2 are shown in red adjacent to corresponding nucleotides in the SARS-CoV structure. **B.** G29734 is involved in a non-Watson-Crick pair with A29756. The hydrogen bonding pattern is denoted with dashed lines. Both nucleotides are conserved between SARS-CoV and SARS-CoV-2. The variant position is marked with red **C.** G29742 is involved in a base quadruple with three residues conserved between SARS-CoV and SARS-CoV-2. The Watson-Crick partner of G29743 is in blue. A G:C pair that packs into the minor groove is shown in green. The position of the variant base is denoted by red. Hydrogen bonds between the bases are shown as dashed lines.

Figure 6. *Molecular Dynamics simulations of S2M variations* Overlay of the structures from a single representative 180 ns trajectory of the SARS-CoV S2M loop (**A-B**), G29734C (**C-D**) and G29742U (**E-F**) variants. Front and back orientations show the following residues as sticks: G/U29742 and C29754 in A, C and D; G/C29734 and A29756 in B, D and F. Structures are colored as a function of time (blue = 0 ns, red =180 ns). Hydrogen bond frequency between the base pairs of the S2M loop is shown in **G** for SARS-CoV (blue), G29734C (red) and G29742U (green) variants. Hydrogen bond frequency for the interacting nucleotides in the quartet (highlighted in bold font) and base pairs around G/U29742 is shown for SARS-CoV (blue), G29734C (red) and G29742U (green) variants in **H**. The hydrogen bond frequency is calculated over the four 180 ns trajectories in G and H. Histograms of the largest dimension of the S2M loop measured for the four 180 ns trajectories of SARS-CoV (**I**), G29734C (**J**) and G29742U (**K**) variants. Histograms of the base distance measured between G/C29734 (G/C11) and A29756 (A33) for the four 180 ns trajectories of SARS-CoV (**L**), G29734C (**M**) and G29742U (**N**) variants. In all panels the nucleotides are numbered as in the X-ray structure of SARS CoV S2M RNA (PDB code 1XJR), the corresponding number in the reference genome can be obtained by adding 29,723.

Supplemental Figure Legends

Supp. Figure 1 Maximal Likelihood phylogenetic tree calculated using the entire genome sequence of isolates carrying the SL1 variations in Fig. 1E. UUUA variants are colored blue, and UUUC variants are colored green.

Supp. Figure 2 Most HVR variations destabilize the bulged stem loop structure. The secondary structure of a simplified HVR stem is shown. The region containing the S2M motif and 8-mer region have been deleted to simplify folding calculations (see methods). The predicted secondary structure and thermodynamic stability of this model stem loop and variations are shown. The positions of variations are marked in red.

Supp. Figure 3 Overlay of the structures from the MD trajectories of the SARS-CoV S2M loop (**A**), G29734C (**B**) and G29742U (**C**) variants. The overlay includes structures from all four 180 ns trajectories collected for each model. Twenty-six structures were selected at equally spaced intervals from the combined trajectories and are colored as a function of time and trajectory from blue to red.

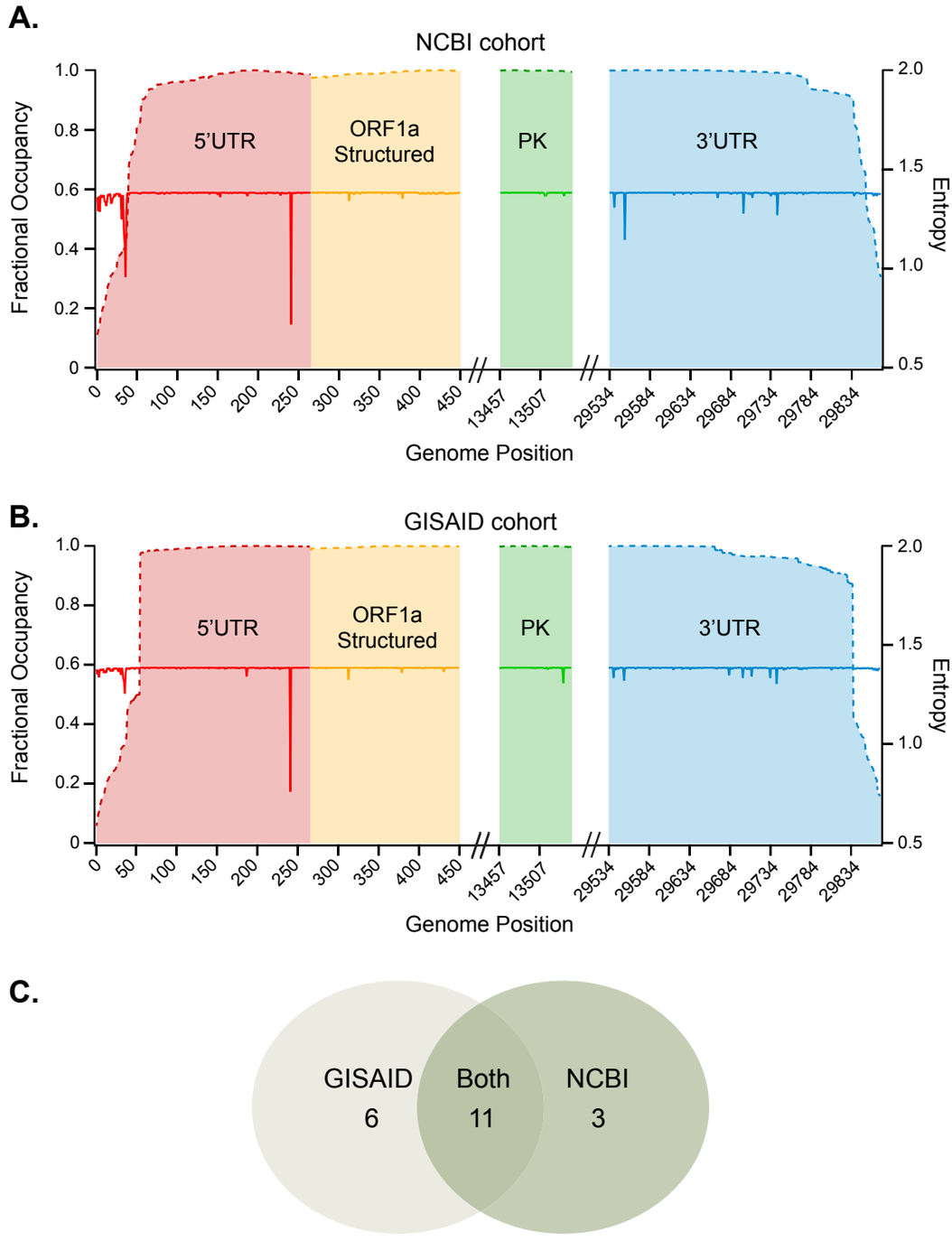

A.
NCBI cohort
Fractional Occupancy
Entropy
5'UTR
ORF1a
Structured
PK
3'UTR
Genome Position
B.
GISAID cohort
Fractional Occupancy
Entropy
5'UTR
ORF1a
Structured
PK
3'UTR
Genome Position
C.
GISAID
6
Both
11
NCBI
3

Figure 1

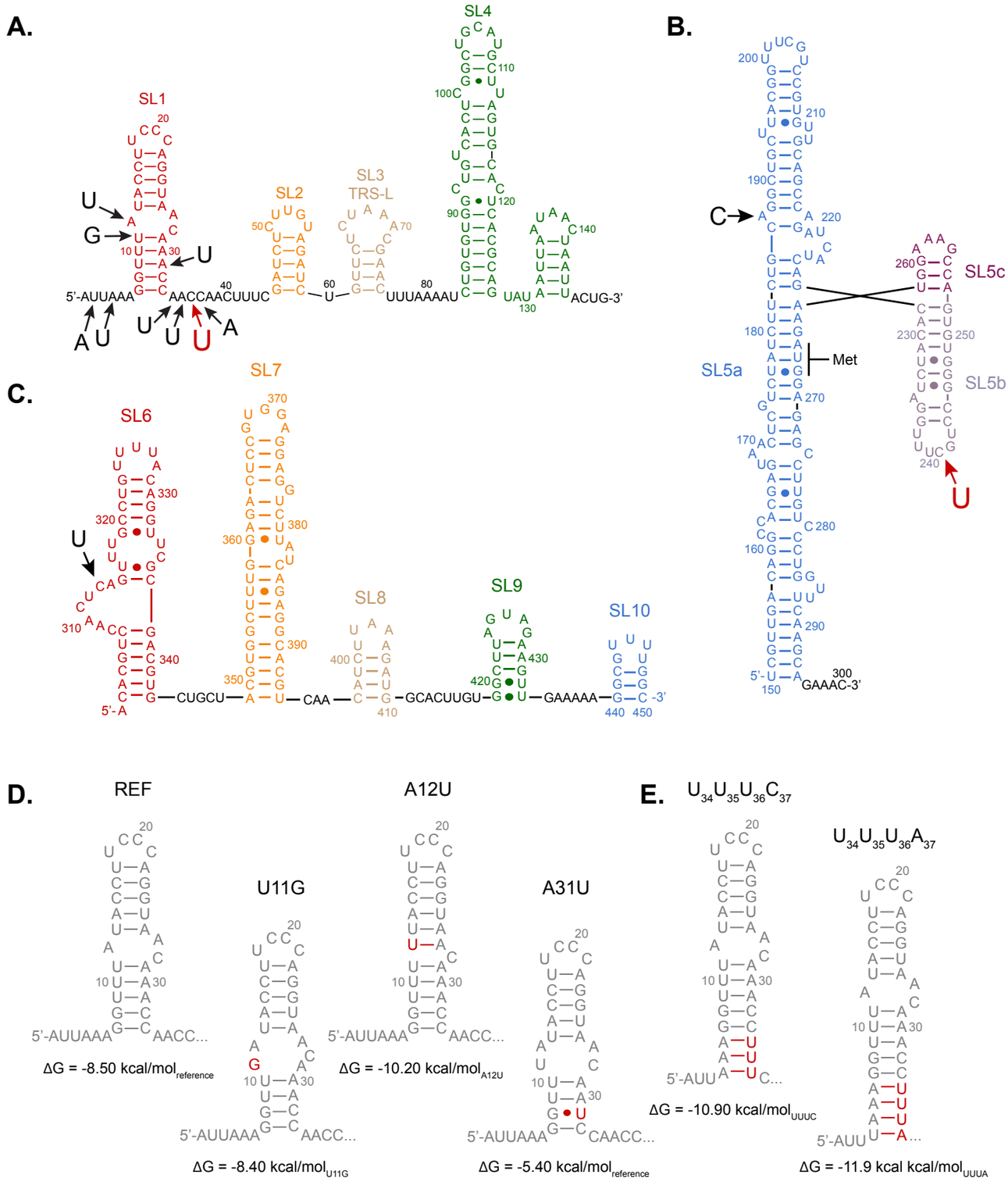

A.
SL1
SL2
SL3
TRS-L
SL4
5'-AUUAAA
AACCAACUUUC
UUUAAAAU
UAU
ACUG-3'
B.
C
SL5a
SL5b
SL5c
Met
5'-U
GAAAC-3'
C.
SL6
SL7
SL8
SL9
SL10
5'-A
CUGCU
CAA
GCACUUGU
GAAAAA
-3'
D.
REF
A12U
U11G
A31U
5'-AUUAAA
AACC...
ΔG = -8.50 kcal/mol reference
ΔG = -10.20 kcal/mol A12U
ΔG = -8.40 kcal/mol U11G
ΔG = -5.40 kcal/mol reference
E.
U34U35U36C37
U34U35U36A37
5'-AUU
C...
ΔG = -10.90 kcal/mol UUUC
ΔG = -11.9 kcal kcal/mol UUUA

Figure 2

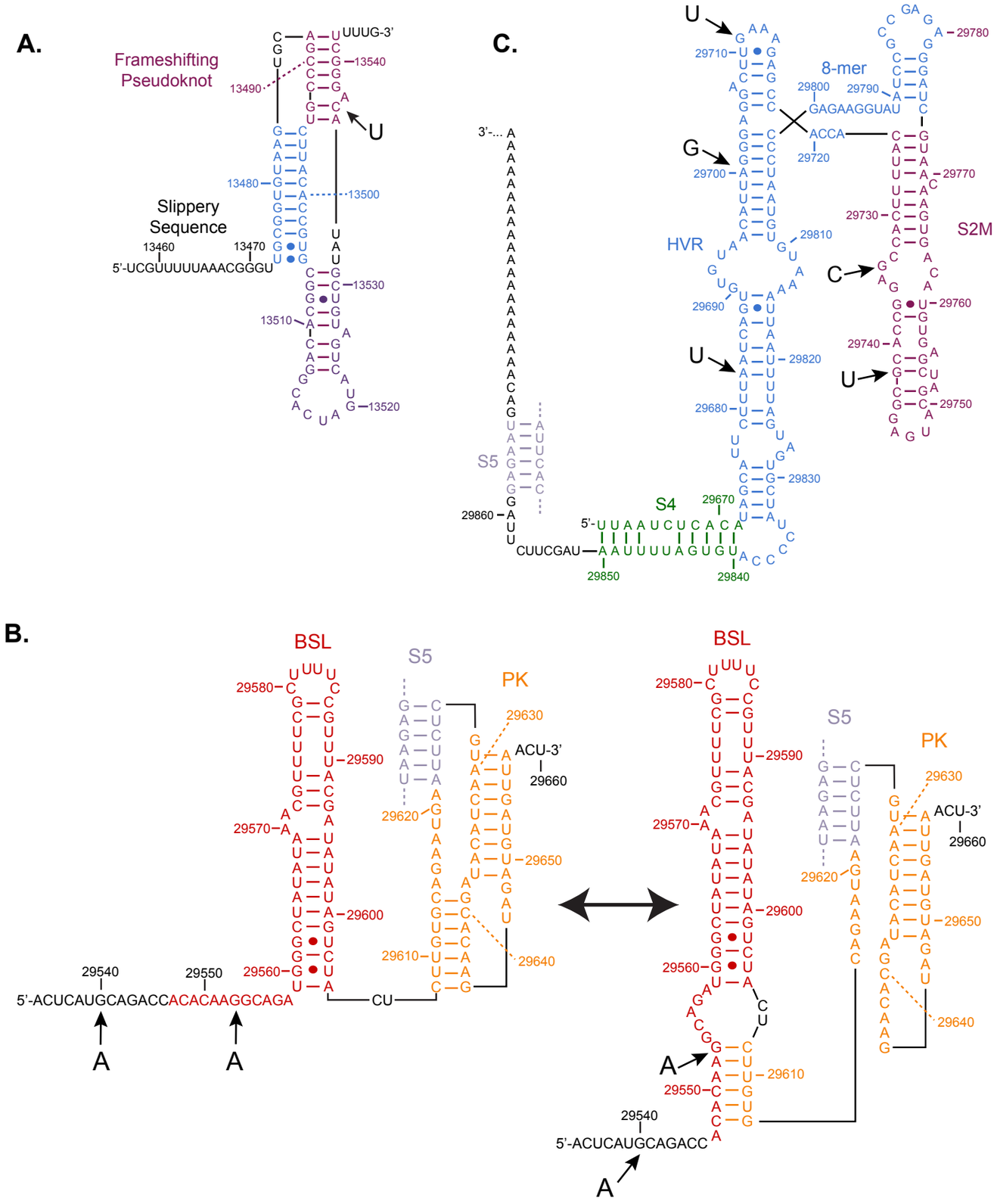

A.
Frameshifting Pseudoknot
Slippery Sequence
5'-UCGUUUUUAAACGGGU
UUUG-3'
13460
13470
13480
13490
13500
13510
13520
13530
13540
U
C.
3'-...
S5
29860
CUUCGAU
5'-UUAAUCUCACA
29850
29840
S4
29670
HVR
29690
29700
29710
29720
29730
29740
29750
29760
29770
29780
29790
29800
29810
29820
29830
8-mer
GAGAAGGUAU
ACCA
S2M
U
G
C
U
B.
BSL
S5
PK
29580
29590
29600
29610
29620
29630
29640
29650
29660
ACU-3'
29570
29560
29550
29540
5'-ACUCAUGCAGACCACACAAGGCAGA
CU
A
A
5'-ACUCAUGCAGACC

Figure 3

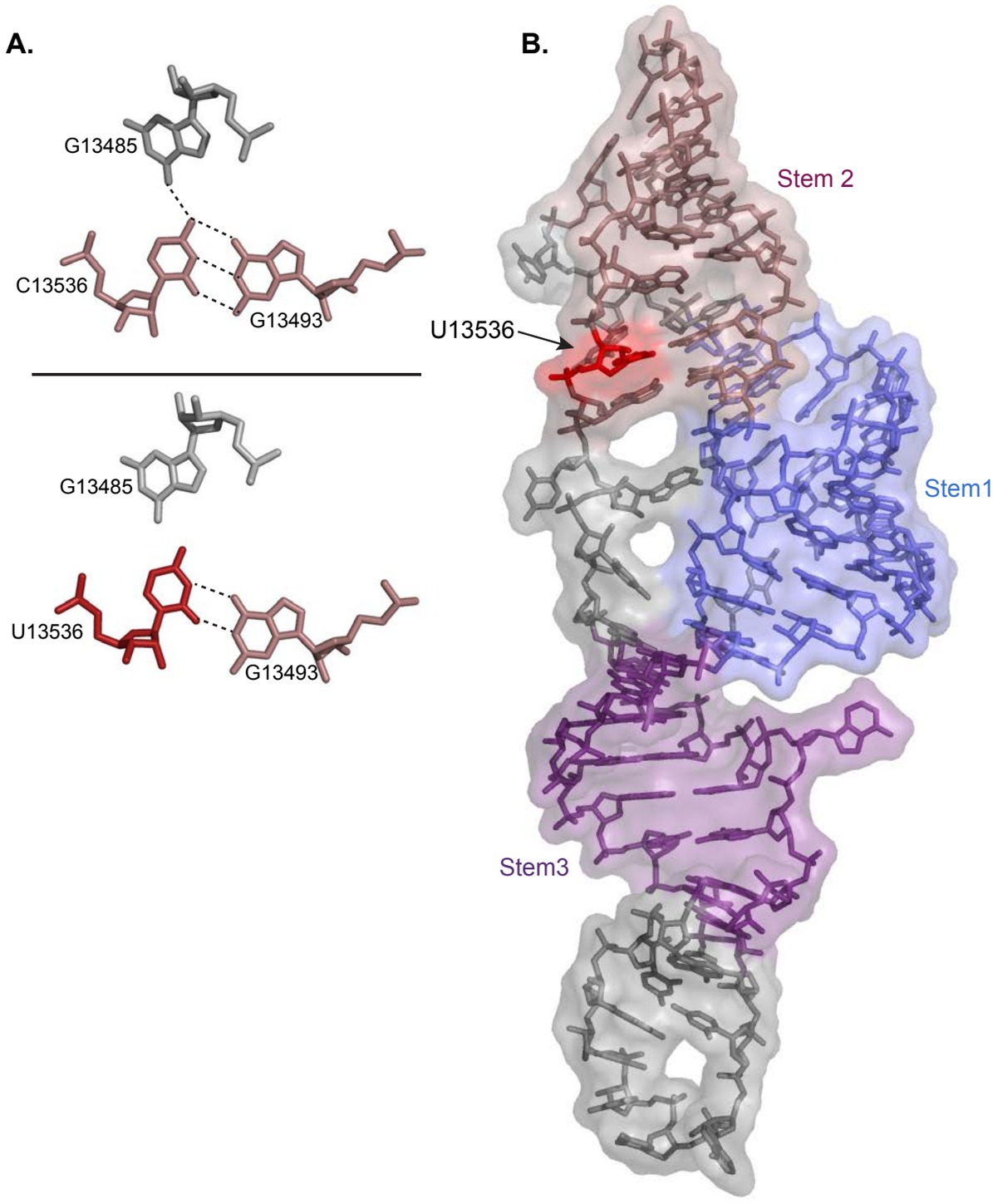

A.
G13485
C13536
G13493
G13485
U13536
G13493
B.
Stem 2
U13536
Stem1
Stem3

Figure 4

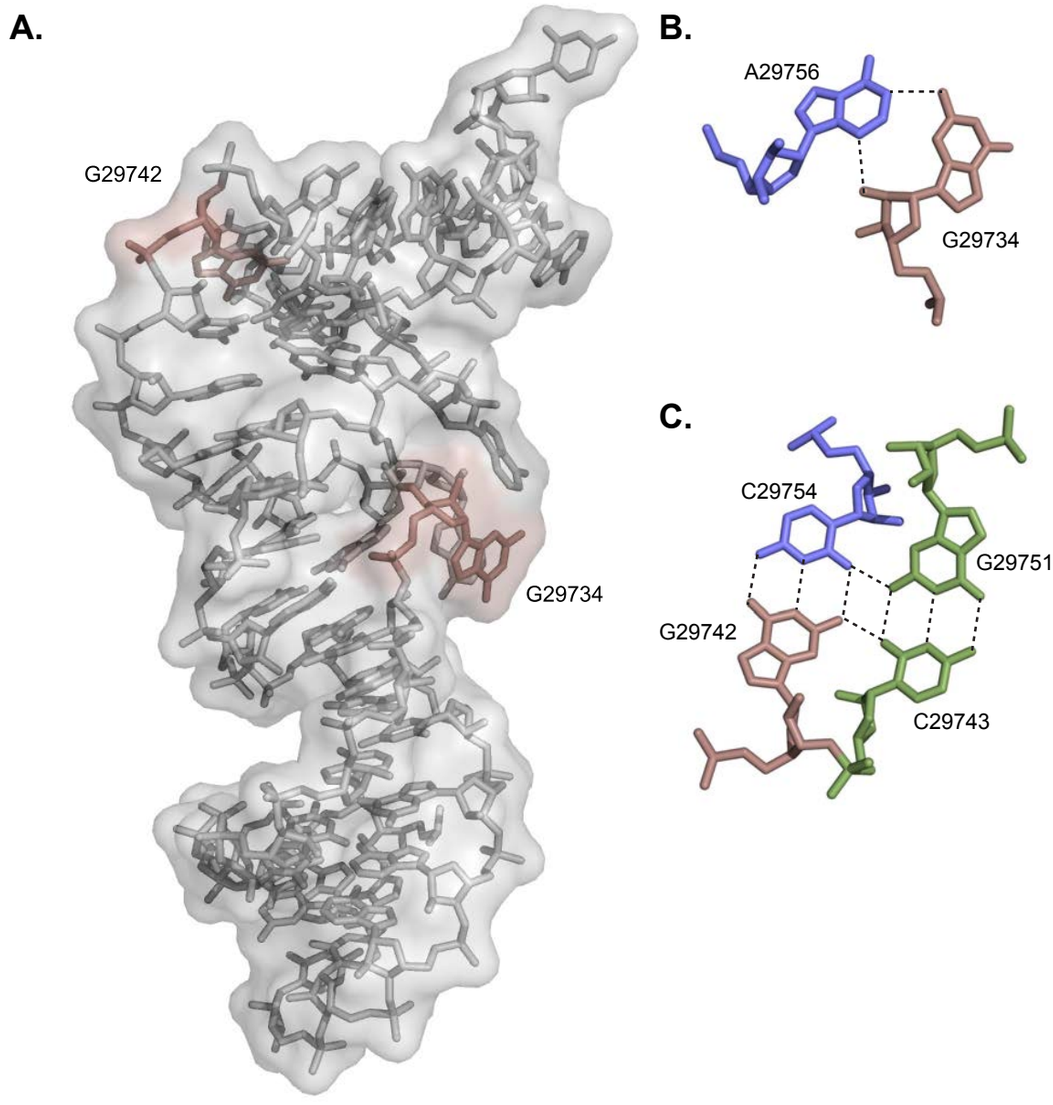

A.
G29742
G29734
B.
A29756
G29734
C.
C29754
G29751
G29742
C29743

Figure 5

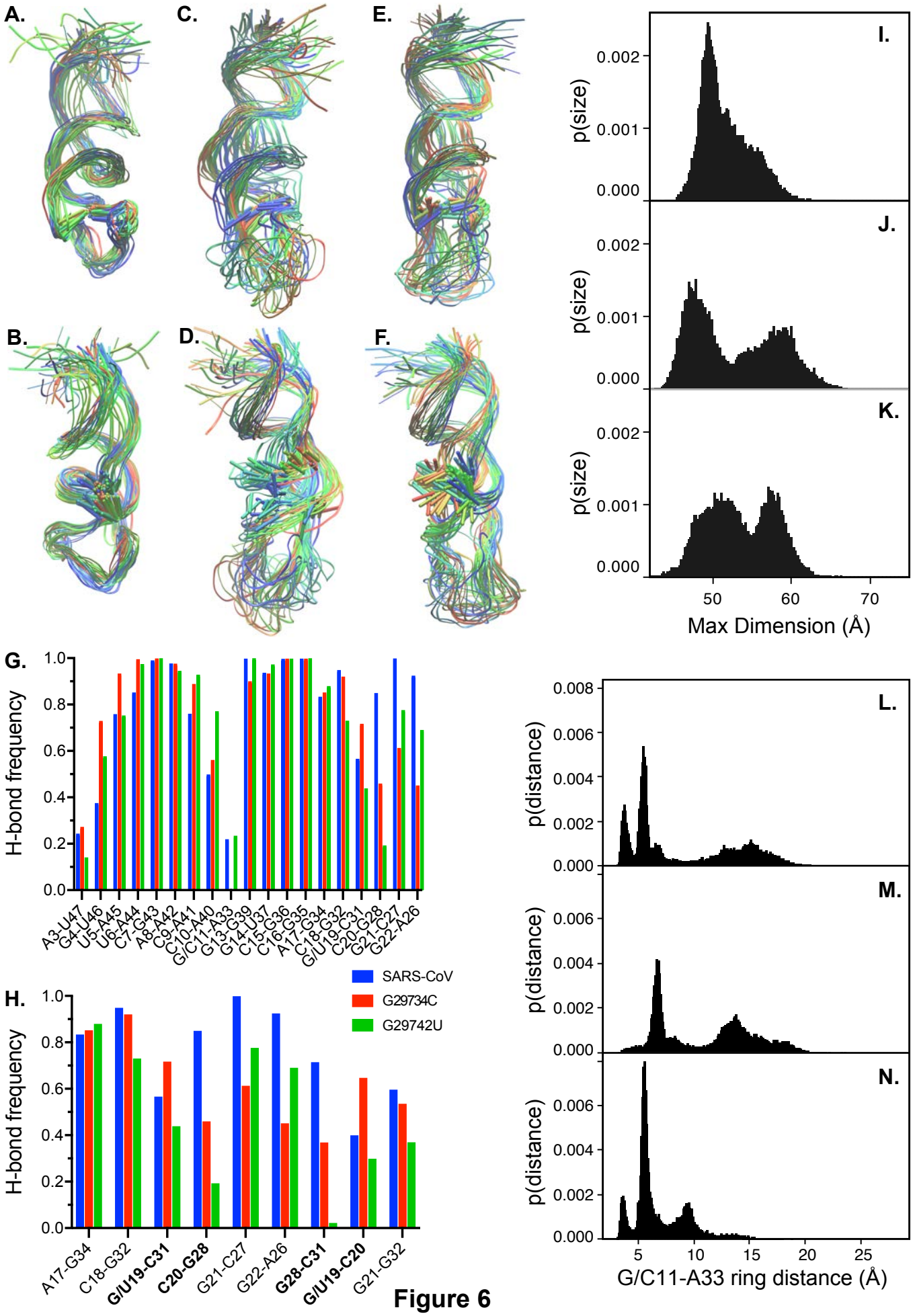
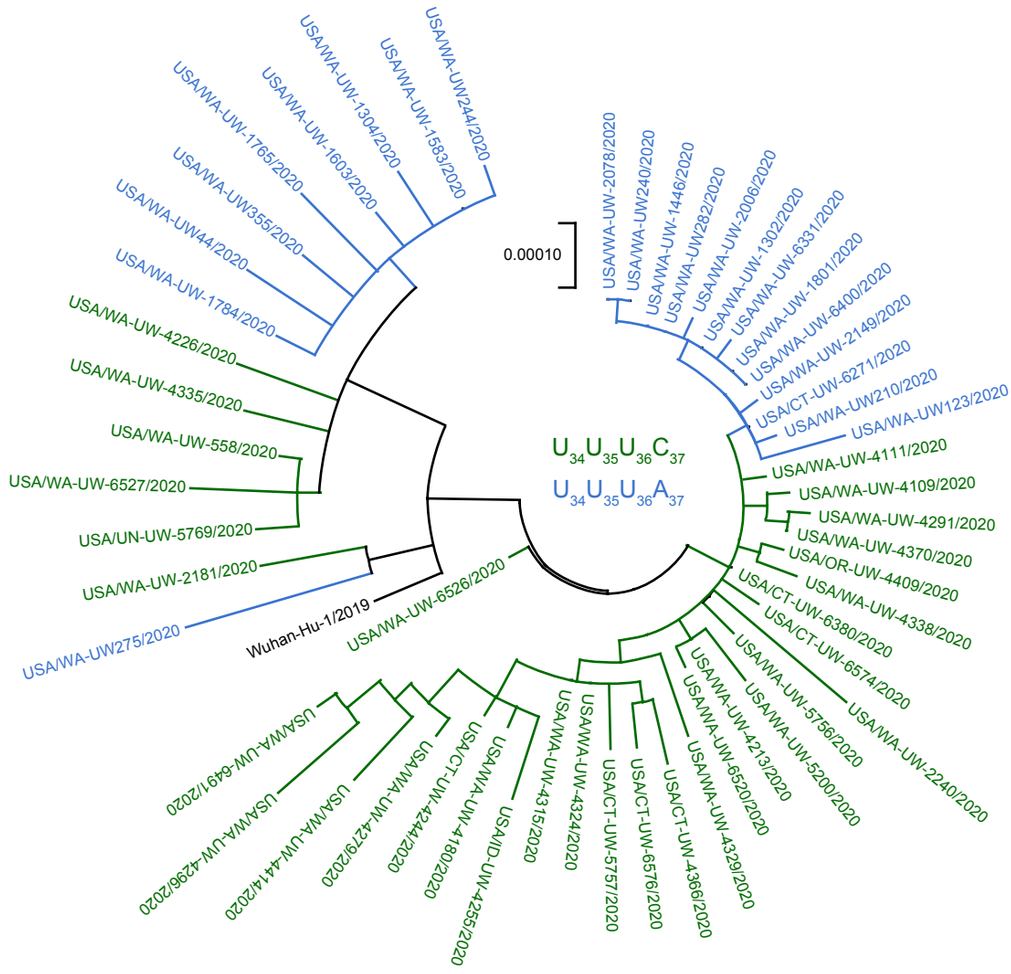
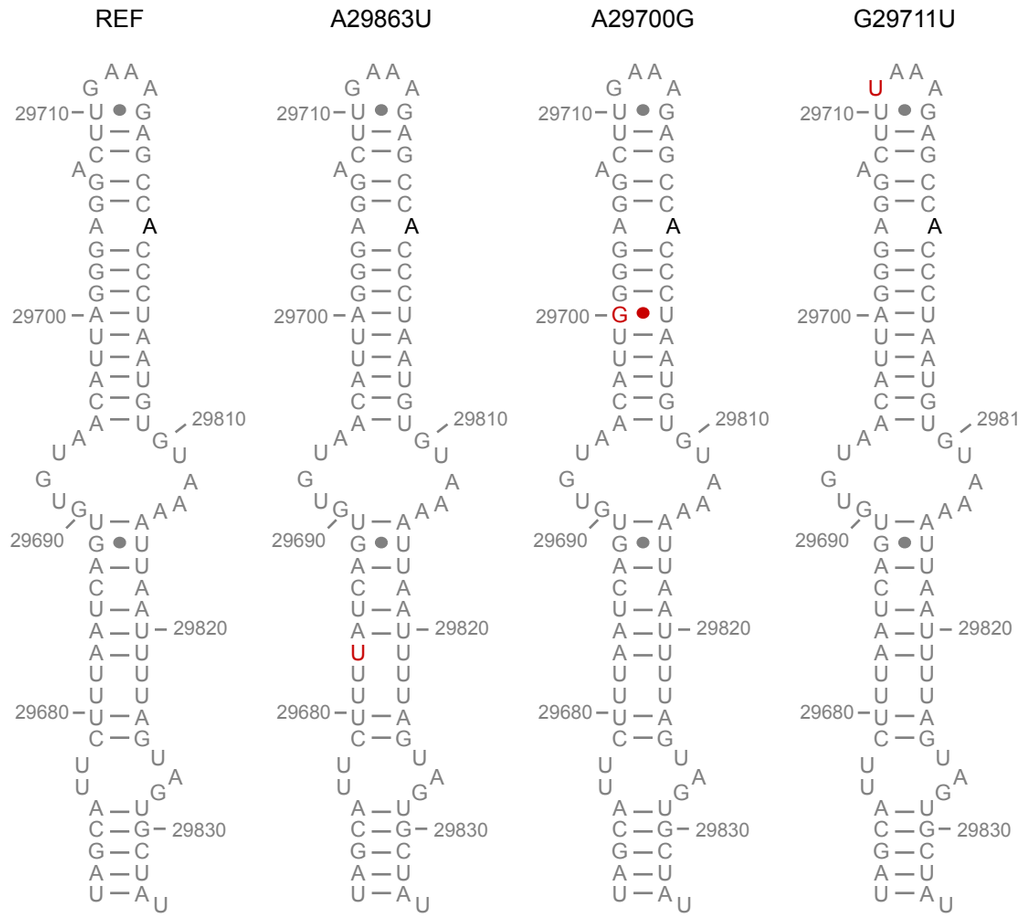

A.
B.
C.
D.
E.
F.
I.
J.
K.
p(size)
0.000
0.001
0.002
50
60
70
Max Dimension (Å)
G.
H-bond frequency
0.0
0.2
0.4
0.6
0.8
1.0
A3-U47
G4-U46
U5-A45
U6-A44
C7-G43
A8-A42
C9-A41
C10-A40
G/C11-A33
G13-G39
G14-U37
C15-G36
C16-G35
A17-G34
C18-G32
G/U19-C31
C20-G28
G21-C27
G22-A26
SARS-CoV
G29734C
G29742U
H.
A17-G34
C18-G32
G/U19-C31
C20-G28
G21-C27
G22-A26
G28-C31
G/U19-C20
G21-G32
L.
M.
N.
p(distance)
0.000
0.002
0.004
0.006
0.008
5
10
15
20
25
G/C11-A33 ring distance (Å)

Figure 6

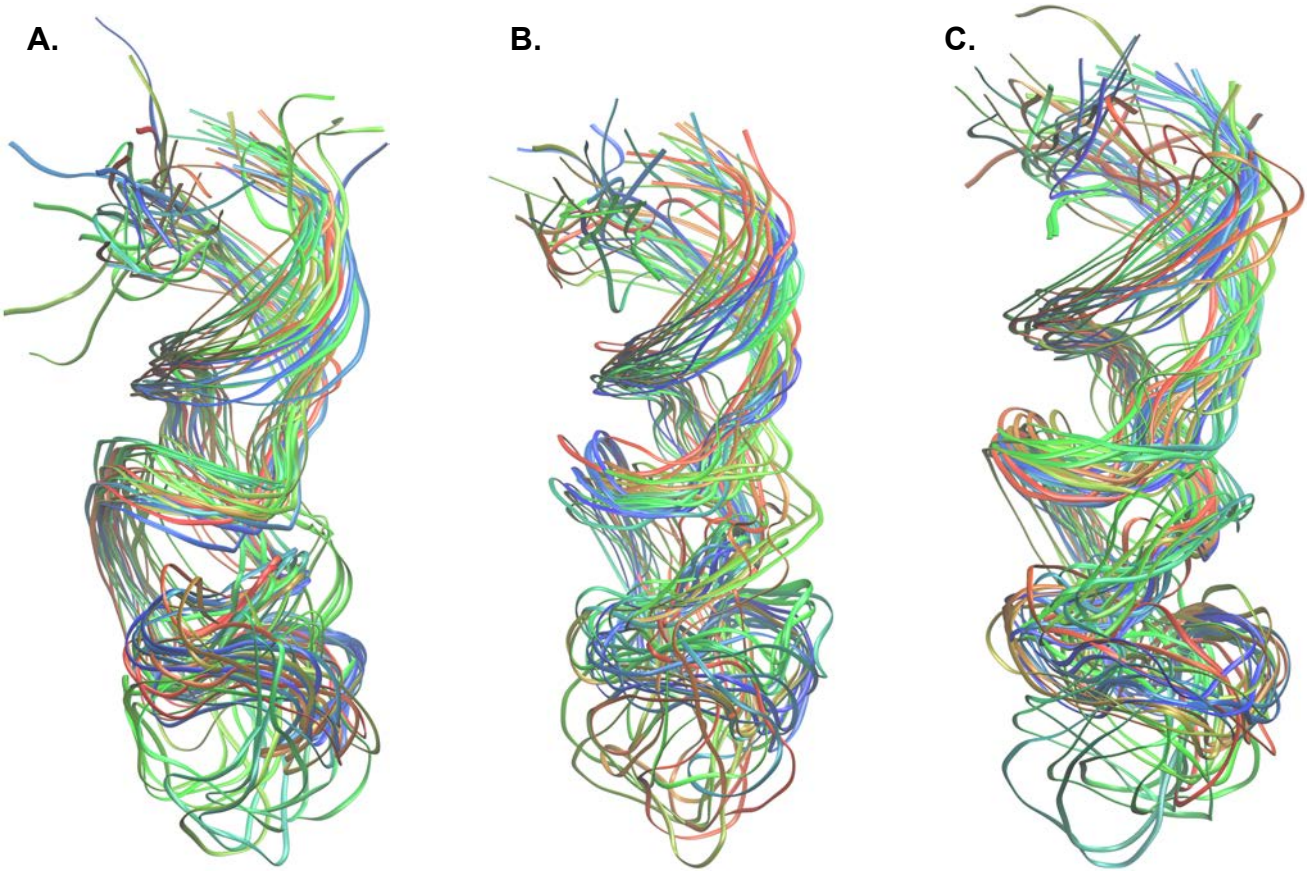


Supp. Figure 1



ΔG = –24.20 kcal/mol$_{reference}$ ΔG = –22.20 kcal/mol$_{A29683U}$ ΔG = –23.90 kcal/mol$_{A29700G}$ ΔG = –23.50 kcal/mol$_{G29711U}$

Supp. Figure 2



Supp. Figure 3