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

Evolutionary Bioinformatics
Volume 17: 1–18
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769343211014167
SAGE

Sean P Ryder, Brittany R Morgan, Peren Coskun, Katianna Antkowiak and Francesca Massi

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

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 sets of SARS-CoV-2 genomic sequences to identify emerging variants within structured cis-regulatory elements of the SARS-CoV-2 genome. Overall, 20 variants are present at a minor allele frequency of at least 0.5%. Several enhance the stability of Stem Loop 1 in the 5′ untranslated region (UTR), including a group 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-ss molecular switch in the 3′UTR. Finally, 5 variants destabilize structured elements within the 3′UTR hypervariable region, including the S2M (stem loop 2 m) selfish genetic element, 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. Our analysis has implications for the development of therapeutics that target viral cis-regulatory RNA structures or sequences.

KEYWORDS: SARS, COVID-19, RNA structure, coronavirus, phylogeny

RECEIVED: February 1, 2021. **ACCEPTED:** March 29, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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 R01GM139316 to F.M. and S.P.R.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Sean P Ryder, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, LRB-906, Worcester, MA 01605, USA. Email: Sean.Ryder@umassmed.edu

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,[1,2] it is responsible for 3 major outbreaks of severe acute respiratory syndrome since the turn of the century[3-7] (SARS: severe acute respiratory syndrome, MERS: middle east respiratory syndrome, and COVID-19: coronavirus disease 2019). 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.[5,8,9] A novel virus termed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified as the cause of this disease.[5,10] 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.[11,12] As with SARS-CoV, the SARS-CoV-2 infectious virion attaches to a target cell through specific interactions between the viral Spike (S) protein and the host angiotensin converting enzyme 2 (ACE2) receptor on the cell surface.[13] The structure of the S protein receptor binding domain in complex with the ACE2 has been solved by X-ray crystallography, revealing a high degree of similarity to the SARS-CoV-ACE2 complex with a few salient differences.[14,15] Proteolytic cleavage of S by the host peptidase TMPRSS2 drives a conformational change that leads to fusion between the host and viral membranes and subsequent release of the virion into the cell.[16] The viral genomic RNA contains a 5′ cap and a polyA tail and is a substrate for protein synthesis by host ribosomes.[12] The virus produces 2 abundant polyproteins encoded by ORF1a and ORF1b, which comprise about half of the genome. The ratio of ORF1a and ORF1b synthesis is controlled by programmed frame shifting, which requires an RNA pseudoknot to mediate a negative 1 nucleotide shift in the reading frame used by the ribosome.[17,18] The polyproteins encode the viral replication and transcription complex (RTC), proteases, and several accessory proteins required for efficient viral replication.[12] The RTC is a multiprotein complex with RNA-dependent RNA polymerase activity that produces full-length antigenomic RNA 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.[19] Subgenomic RNA transcription requires discontinuous RNA synthesis mediated by a number of conserved sequences and structures in 5′ and 3′ untranslated regions (UTRs) of the viral genomic RNA. As such, each subgenomic transcript contains the same 5′ cap and leader sequence found in the intact genomic RNA.

The stem loop structures found in other coronaviridae are also present in both coding and noncoding regions the SARS-CoV-2 RNA genome.[20] They cluster in the 5′UTR, the N-terminal portion of ORF1a, 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.[21,22] 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.[22-27]

Several effective vaccines are now available that make use of the viral S protein to initiate an immune response.[28-30] However, there remains a need to develop broadly effective antiviral therapies to treat patients already infected with the virus or to prepare for the eventual emergence of vaccine escaping variants of SARS-CoV-2. Only 1 antiviral drug, Remdesivir, is approved by the FDA in the United States for treatment of COVID-19.[31] The majority of ongoing efforts target viral entry and replication, screening for candidate therapeutics that inhibit viral (S protein, the RTC complex, or viral proteases) or key host (ACE2 and TMPRSS2) factors.[32-36] Other efforts have focused on repurposing already approved compounds to improve disease ouctcomes.[37,38] There is also interest in targeting the viral genome directly, using compounds that target conserved RNA structures such as the frame-shifting pseudoknot,[39,40] or using informational drugs such as RNA interference triggers or antisense oligonucleotides that directly recognize the viral genome sequence.[41] As such, it is critically important that we understand how sequence and structure of viral cis-regulatory elements changes as the virus spreads. This understanding will help guide new therapeutic development to the most conserved sequences and structures, limiting the opportunity for the emergence of resistant variants in the future.

DNA sequencing technology has progressed remarkably since the SARS outbreak of 2003.[42,43] It is now routine to determine the sequence of the ~30 kilobase viral genome using high throughput sequencing technology.[10] As a result, scientists and medical professionals from around the world have sequenced the SARS-CoV-2 genome from patient isolates and disseminated their findings through data repositories (eg, the GISAID EpiCoV database) at unprecedented speed.[10,44-46] 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.[47-53] 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.[45,46] 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 genomic sequence sets to identify and characterize emerging variations within the cis-regulatory RNA structures of the virus genome. Our analysis reveals 20 abundant variants, including 2 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 by increasing the length of the paired region. The data also show that structured elements in the 3′UTR hypervariable region, including the enigmatic S2M loop, contain 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 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.[10] 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.[54,55] 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).[56] The MSA was then input into WebLogo 3 to calculate the positional occupancy, entropy, and allele frequency for each query (Supplementary Table 1).[57] 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 (Figure 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 400 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. At this cutoff, the least prevalent variant is observed in at least 5 different genomes (U12G near the 5′ terminus), while most variants are observed in 20 genomes or more (Supplemental Table 1). As such, it is unlikely that these variations are sequencing artifacts.

To extend this analysis, we repeated the study with a second set of SARS-CoV-2 sequences recovered from the GISAID database on May 13, 2020.[45,46] All sequences were downloaded from the database, converted into a blast library, then queried and analyzed as above with the NCBI set. An average of 23 900 ± 630 hits were recovered from each query. As with the NCBI set, occupancy is high (>90%) from position 55 through position 29 829 (Figure 1B, Supplementary Table 1). Due to the large size of the GISAID set—6.6 times the size of the NCBI set—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 (0.5%) were identified in the GISAID set, eleven of which were also identified in the NCBI set (Figure 1C, Table 1). Combining the 2 analyses yields a total of 20 emerging variants in the structured regions of the viral genome. Of these, thirteen are transversions and 7 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 set, 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 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.[58,59] Stem Loop 2 (SL2) contains a highly conserved sequence and structural elements thought to play a role in sgRNA synthesis.[23,60-62] 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.[22,59,63] 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.[64-66] The precise role of these structures in SARS-CoV-2 infection is not known, but recently released structural predictions reveal all 4 stem loops are present in the SARS-CoV-2 genome (Figure 2).[20]

SL1 and flanking single-stranded regions contain 9 of the 20 variants identified in this analysis. In contrast, no abundant 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 (Figure 2B).[67] 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 ($\Delta G = -8.50$ 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 ($\Delta G = -8.40$ 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 ($\Delta G = -10.2$ kcal/mol$_{A12U}$). By contrast, A31U (U = 0.005/7525 GISAID, U = 0.029/1314 NCBI) is strongly destabilizing ($\Delta G = -5.40$ kcal/mol), causing disruption of the lower stem.

Four emerging variations are found just downstream of the SL1 stem (Figure 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, 3 of the 4 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 3 base pairs, stabilizing the duplex by 2.4 kcal/mol ($\Delta G = -10.9$ kcal/mol K) (Figure 2C). 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 ($\Delta G = -11.9$ kcal/mol K). Of these 4 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.

To test whether SL1 is stabilized by the combination variations identified by this approach, we performed thermal UV denaturation experiments with model RNA oligonucleotides corresponding to the reference sequence in Figure 2B and both combination variants in Figure 2C. The increase in absorbance of single stranded nucleic acid (hyperchromicity) compared to folded RNA defines the extent of structure present and can be fitted to a model to extract thermodynamic parameters (see materials and methods equations 1-4).[68,69] The $\Delta H°$ and T_m were determined using a 2 state unimolecular denaturation model,[68] and the $\Delta G°$ and $\Delta S°$ calculated from baseline subtracted data using standard thermodynamic relationships.[69] The absorbance data were normalized to the local maxima and minima (to facilitate comparisons) and plotted as a function of temperature in Figure 2D. The data reveal strong favorable stabilization in the standard enthalpy change for both combination variants ($\Delta\Delta H°$ UUUC/WT = −22 kcal/mol K, $\Delta\Delta H°$

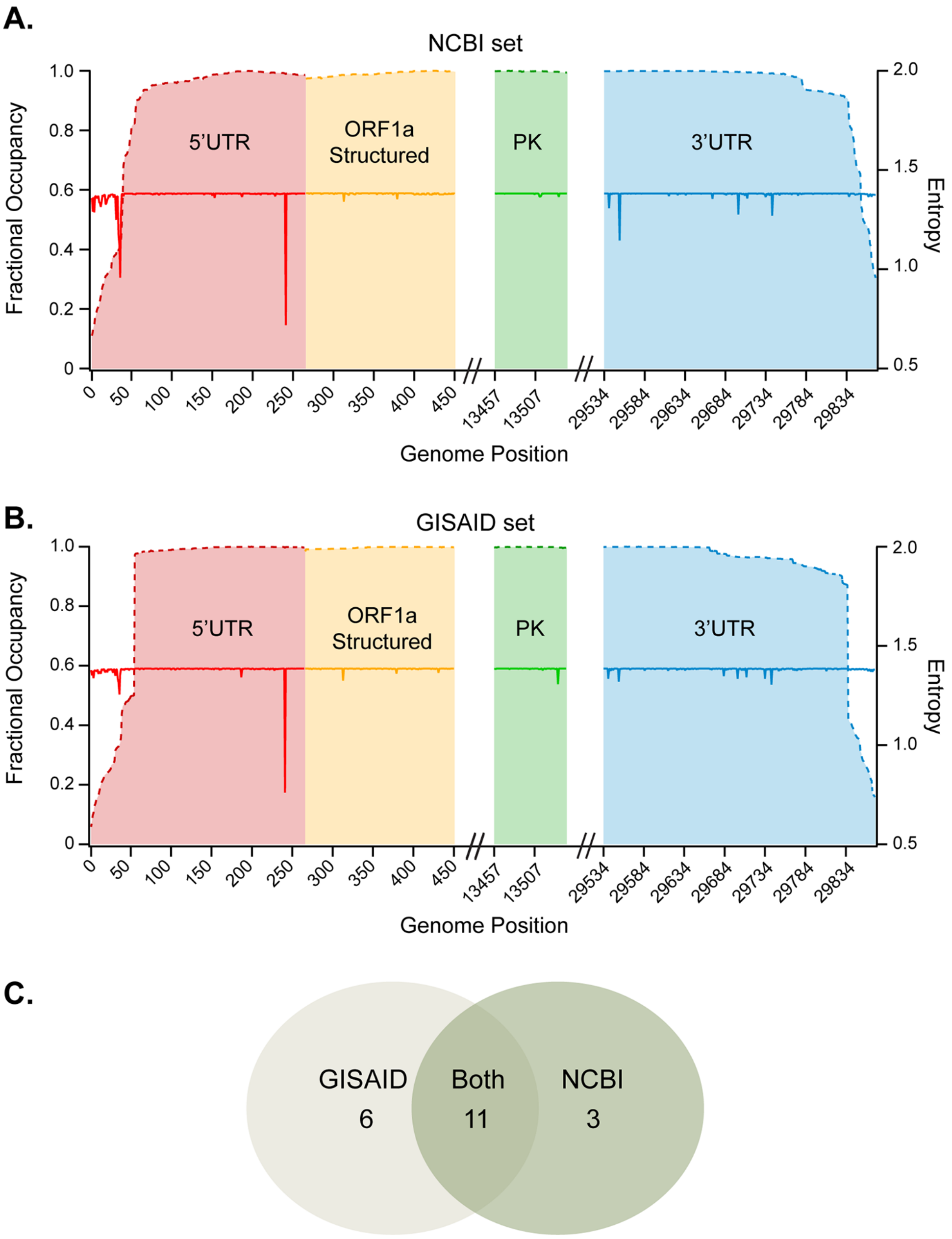


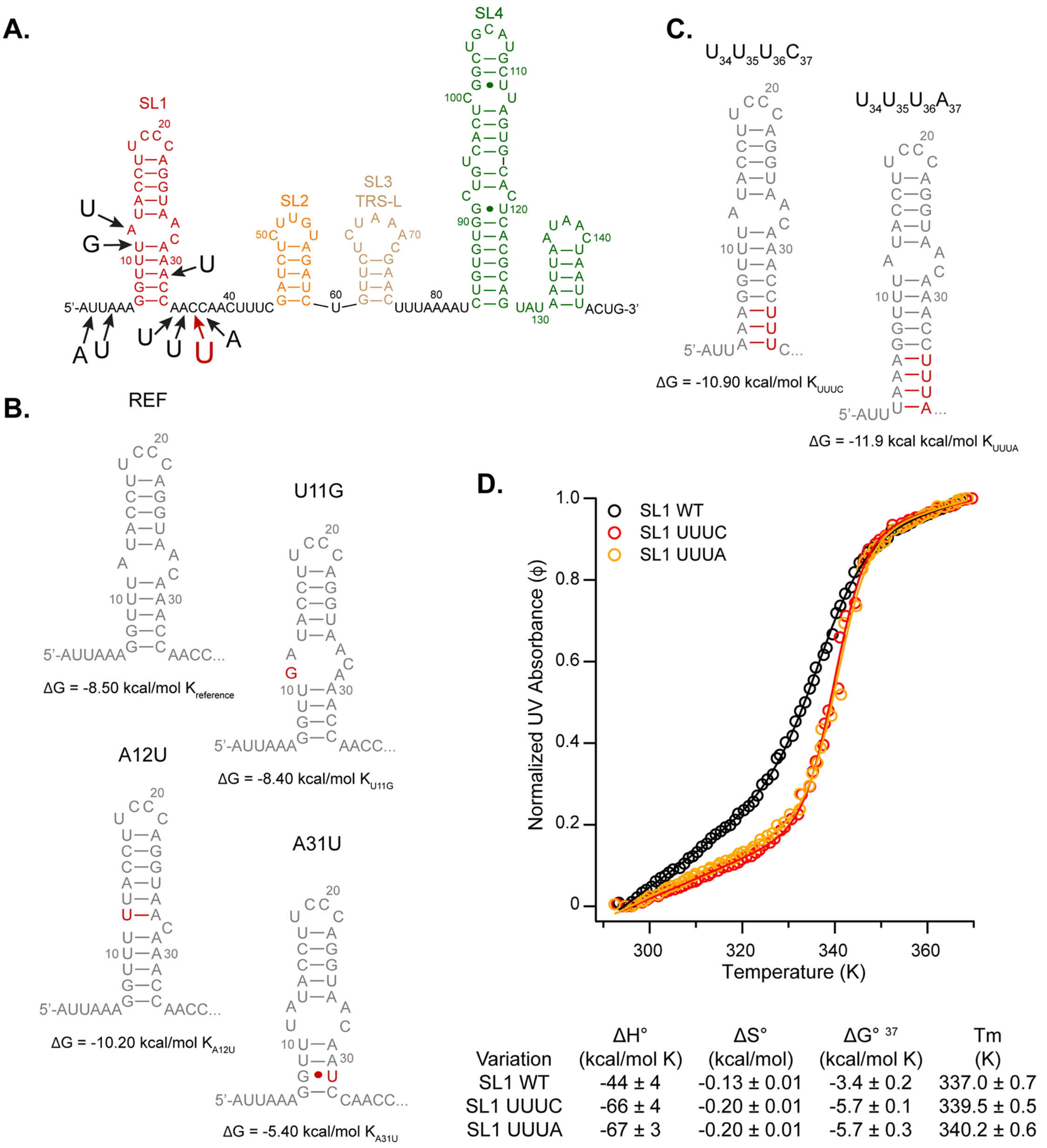
Figure 1. SARS-CoV-2 sequence occupancy and entropy: (A) fractional occupancy (left axis, dashed lines) and positional entropy (right axis, solid lines) of the NCBI set calculated by WebLogo3 as displayed as a function of SARS-CoV-2 genome coordinates. This analysis focused only on well characterized 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 database, and (C) venn diagram of emerging variations in the GISAID set, NCBI set, or both.

UUUA/WT = −23 kcal/mol K). The change in the standard free energy change ($\Delta\Delta G^\circ$, 37 °C) at body (infection) temperature is -2.3 kcal/mol UUUC/WT and −2.3 kcal/mol UUUA/WT, comparable to the RNAfold predictions described above ($\Delta\Delta G^\circ = -2.4$ kcal/mol UUUC/WT, $\Delta\Delta G^\circ = 3.4$ kcal/mol UUUA/WT).

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 set and aligned them using MAFFT.[56] 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 (Supplemental Figure 1).[70,71] The phylogenetic relationship shows that both variants are represented in 2 different branches, but the variants tend to cluster separately within those branches. In 1 case, 13 $U_{34}U_{35}U_{36}A_{37}$ 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

Table 1. Emerging variants in structured regions of the SARS-CoV-2 genome.

INDEX	COORDINATES	WUHAN-1	MAF (GISAID)	MAF (NCBI)	TYPE	LOCATION	CATEGORY	EDITING
1	2	U	A=0.006/2519	A=0.014/419	Noncoding	5′UTR	Transversion	None
2	4	A	U=0.004/1799	U=0.016/494	Noncoding	5′UTR	Transversion	None
3	11	U	G=0.006/2519	G=0.001/4154	Noncoding	5′UTR	Transversion	None
4	12	A	U=0.002/4484	U=0.011/829	Noncoding	5′UTR	Transversion	None
5	31	A	U=0.005/7525	U=0.029/1314	Noncoding	5′UTR	Transversion	None
6	34	A	U=0.009/7795	U=0.050/1350	Noncoding	5′UTR	Transversion	None
7	35	A	U=0.013/7846	U=0.071/1380	Noncoding	5′UTR	Transversion	None
8	36	C	U=0.028/7967	U=0.150/1400	Noncoding	5′UTR	Transition	APOBEC
9	37	C	A=0.003/8091	A=0.018/1474	Noncoding	5′UTR	Transversion	None
10	187	A	C=0.007/23832	C=0.003/3407	Noncoding	5′UTR	Transversion	None
11	241	C	U=0.682/23760	U=0.616/3376	Noncoding	5′UTR	Transition	APOBEC
12	313	C	U=0.011/24227	U=0.007/3732	Silent	ORF1A	Transition	None
13	13536	C	U=0.015/23306	U=0.003/3440	Silent	ORF1B	Transition	None
14	29540	G	A=0.008/24313	A=0.014/3550	Noncoding	3′UTR	Transition	None
15	29553	G	A=0.012/24216	A=0.064/3551	Noncoding	3′UTR	Transition	None
16	29683	A	U=0.006/23540	U=0.0003/3545	Noncoding	3′UTR	Transversion	None
17	29700	A	G=0.009/23403	G=0.022/3541	Noncoding	3′UTR	Transition	None
18	29711	G	U=0.007/23366	U=0.004/3537	Noncoding	3′UTR	Transversion	None
19	29734	G	C=0.008/23335	C=0.003/3525	Noncoding	3′UTR	Transversion	None
20	29742	G	U=0.009/23285	U=0.019/3526	Noncoding	3′UTR	Transversion	None
21	34-37	CCAA	UUUC=0.005/7372	UUUC=0.023/1350	Noncoding	5′UTR	Multiple	None
22	34-37	CCAA	UUUA=0.004/7372	UUUA=0.018/1350	Noncoding	5′UTR	Multiple	None


A.
SL1
SL2
SL3
TRS-L
SL4
5'-AUUAAA
AACCAACUUUC
UUUAAAAU
ACUG-3'
B.
REF
ΔG = -8.50 kcal/mol K$_{reference}$
U11G
ΔG = -8.40 kcal/mol K$_{U11G}$
A12U
ΔG = -10.20 kcal/mol K$_{A12U}$
A31U
ΔG = -5.40 kcal/mol K$_{A31U}$
C.
$U_{34}U_{35}U_{36}C_{37}$
ΔG = -10.90 kcal/mol K$_{UUUC}$
$U_{34}U_{35}U_{36}A_{37}$
ΔG = -11.9 kcal kcal/mol K$_{UUUA}$
D.
○ SL1 WT
○ SL1 UUUC
○ SL1 UUUA
Normalized UV Absorbance (ϕ)
Temperature (K)
0
0.2
0.4
0.6
0.8
1.0
300
320
340
360

Variation	ΔH° (kcal/mol K)	ΔS° (kcal/mol)	ΔG° 37 (kcal/mol K)	Tm (K)
SL1 WT	-44 ± 4	-0.13 ± 0.01	-3.4 ± 0.2	337.0 ± 0.7
SL1 UUUC	-66 ± 4	-0.20 ± 0.01	-5.7 ± 0.1	339.5 ± 0.5
SL1 UUUA	-67 ± 3	-0.20 ± 0.01	-5.7 ± 0.3	340.2 ± 0.6

Figure 2. Emerging variations in SL1-SL4 of the 5'UTR: (A) the predicted secondary structure of SL1 through SL4 is shown. The position and identity of emerging variants is denoted by an arrow and a letter, (B) 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. (C) same as in (B), but for the two prevalent combination variants that extend the length of SL1, and (D) thermal denaturation curves of SL1 reference sequence and both combination variants. The ΔH°, ΔS°, ΔG°,^{37} and T$_m$ values shown are the average and standard deviation of three experiments.

impact of these variations on viral fitness or patient outcomes is not known.

Most of the 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 of the lower stem block replication.[58] It is important to note that there is significant sequence divergence in this region between the 2 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 = –7.5 kcal/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 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.[58]

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 (Figure 3A).[20] This region contains 3 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.[23] 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.[72] The role of SL5c is more controversial, with 1 study demonstrating that the stem is dispensable,[22] while a previous study showed that it is required.[72]

We observed 2 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 variant by far. There are no emerging mutations with an allele frequency of >0.005 in SL5c in either set.

Four additional stem loop structures (SL6-SL10) have been proposed within ORF1a (Figure 3C).[20] The presence of SL6 and SL7 is observed in other coronaviridae, but the structures do not appear to have an important function.[22,72] There is 1 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 2 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 set, 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.[73,74] 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,[57] we used SNP-sites v2.5.1 and VCFtools v0.1.7 to recalculate the allele frequency inclusive of degenerate bases.[75,76] We calculated the average frequency of all C to Y transitions in the 5′UTR using the larger GISAID set, excluding the 2 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 (Figure 3C).[20] This structure is involved in -1 programmed ribosome frameshifting, where translating ribosomes shift frame by 1 nucleotide to the left. Efficient frameshifting requires both a "slippery" sequence and a downstream stable RNA structure.[17,77] Like SARS-CoV and MHV, the SARS-CoV-2 pseudoknot has 3 stems instead of 2 typically found in pseudoknot structures.[18,78] 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%.[79] 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 1 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 (Figure 3C). 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 (Figure 3D).[80] In the reference model, G13485 forms a base triple with the C13536:G13493 pair (Figure 3E). 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.

Variations in the 3′UTR in the BSL and PK

Betacoronaviridae 3′UTRs contain a bi-stable molecular switch formed by 2 mutually exclusive structural conformers, including 1 that extends the lower stem of the bulged stem loop (BSL), and a second that folds into a pseudoknot (PK, Figure 4A).[24] Both structured elements are present in MHV, SARS, and MERS, though the sequence diverges significantly between them.[24,81,82] In MHV and BCoV, the BSL and the PK structure are required for viral replication.[24,81-83] Mutations that stabilize 1 form over the other prevent replication. It is proposed that competition between the 2 structures plays a regulatory role in antigenomic RNA synthesis, but the exact mechanism remains to be determined.[24]

There are 2 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 K, while the stability of the G29553A variant conformer is −18.30 kcal/mol K 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 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 sets using SNP-sites and VCFtools as described above.[75,76] 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 (Figure 5A).[20] 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.[84,85] The function of this region is not well understood, but deletion of the HVR including the conserved 8-mer element has no effect on MHV replication in cultured cells.[84] An apparent selfish genetic element, termed S2M, exists within the bulged stem loop of the HVR.[20] 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.[86,87] 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.[88]

Five 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 = 3×10^{-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 ($\Delta G = -24.20\,\text{kcal/mol}_{\text{ref}}$, $-22.20\,\text{kcal/mol}_{\text{A29683U}}$, $-23.9\,\text{kcal/mol}_{\text{A29700G}}$, Supplemental Figure 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 ($\Delta G = -23.5\,\text{kcal/mol}_{\text{G29711U}}$). 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 the HVR is 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 emerging A29700G transition suggests the possibility that it might arise through adenosine deaminase

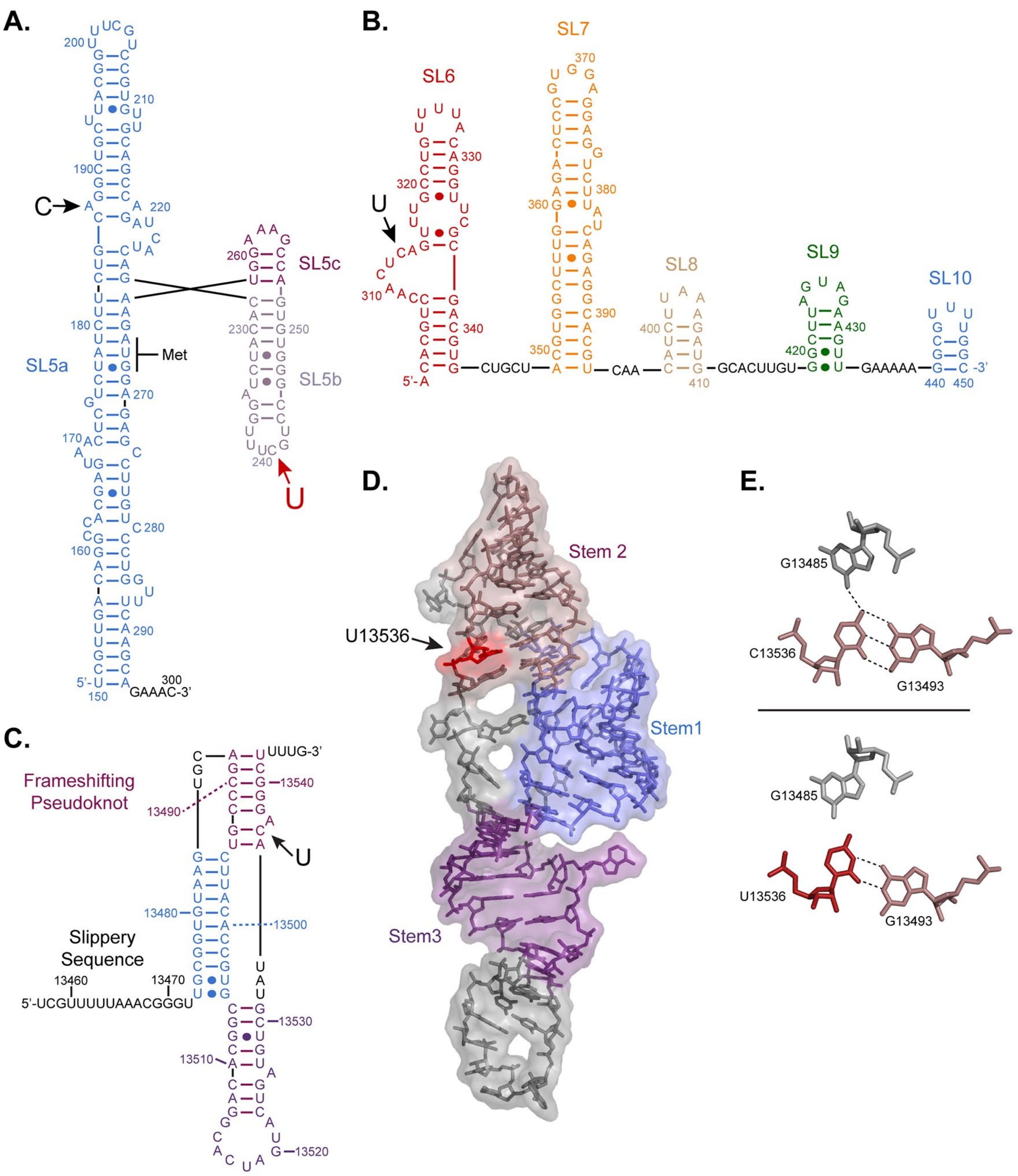


Figure 3. Emerging variations in ORF1A stems and the frameshifting pseudoknot: (A) the predicted secondary structure of SL5 is shown. Emerging variants are denoted by an arrow, and the identity of the variation is given next to the arrow. The position of the ORF1a start codon is labeled, (B) the predicted secondary structure of SL6-SL10 is shown. Variations are labeled as in panel A, (C) the secondary structure of the frameshifting pseudoknot is shown. The position and identity of emerging variants are denoted by an arrow and a letter, (D) the molecular model of the frameshifting pseudoknot calculated by RNAcomposer is shown. Stems 1 to 3 are labeled in colors corresponding to those shown on the secondary structure in panel C, and (E) 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.

acting on RNA (ADAR) RNA editing activity. ADARs convert adenosine residues to inosine in double stranded regions of RNA.[89] 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.[90] 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.[75,76] No degenerate R nucleotides are present in the GISAID set, suggesting that frequent RNA editing by ADAR enzymes is not responsible for rapid A29700G emergence.

The final 2 emerging variations lie within the enigmatic S2M loop. The structure of the S2M loop from SARS-CoV

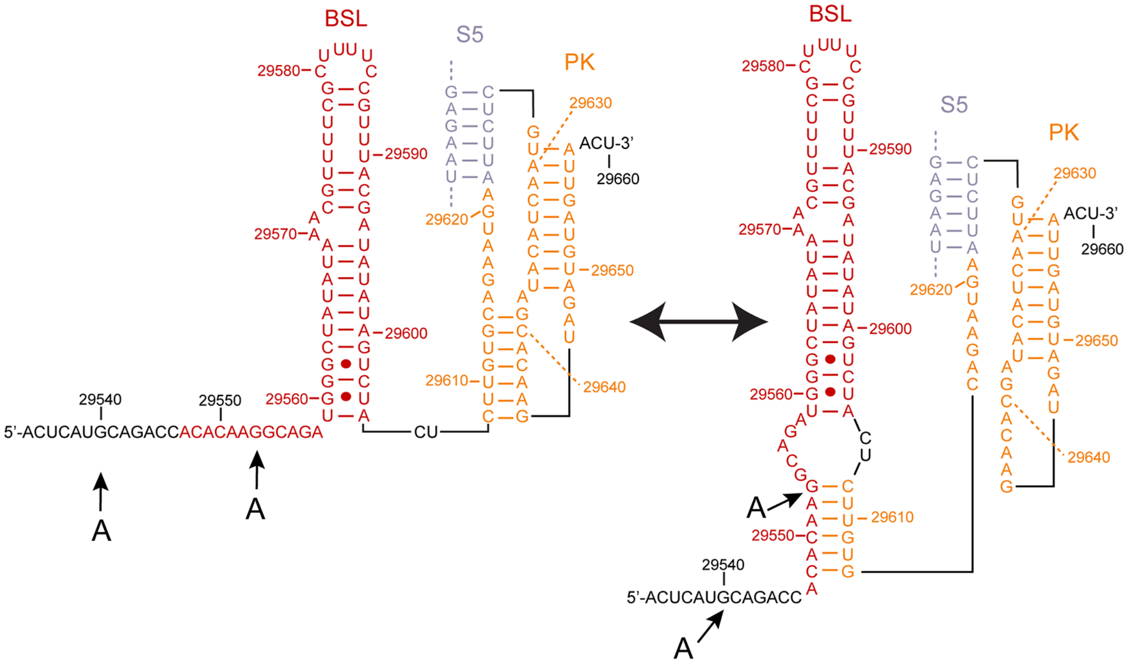


Figure 4. Emerging variations in the bi-stable molecular switch in the 3′UTR: The secondary structure of the 3′UTR bi-stable molecular switch in both predicted conformers is shown. The position and identity of emerging variants is denoted by an arrow and a letter.

has been solved by X-ray crystallography (Figure 5B).[91] 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.[86] 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.[84] As such, its role in viral replication is unclear.

The first 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 (Figure 5C).[91] 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 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 (Figure 5D).[91] 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 further evaluate the effect of both of the 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.[91] 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 to F. To quantify the extent of the transition, we measured the length of the maximum dimension of the S2M loop and observed an extension (Figure 6I–K). Surprisingly, we observed that the S2M loop of SARS-CoV also samples similarly extended conformations (Figure 6I), although this happens more slowly and with a lower frequency than in the other 2 variants, in only 2 out of the 4 trajectories that we collected (Supplemental Figure 3A–C). Analysis of the MD trajectories showed that 1 important element for the stability of the overall structure of the S2M loop is the stabilizing interaction between G29734 and A29756. When these 2 nucleotides are in close proximity to 1 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 (Figure 6G). In the absence of this base pair, A29756 can move far from G/C29734 (Figure 6L–N) leading to the sampling of more extended conformations of the loop as shown in Figure 6C to 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 (Figure 6G).

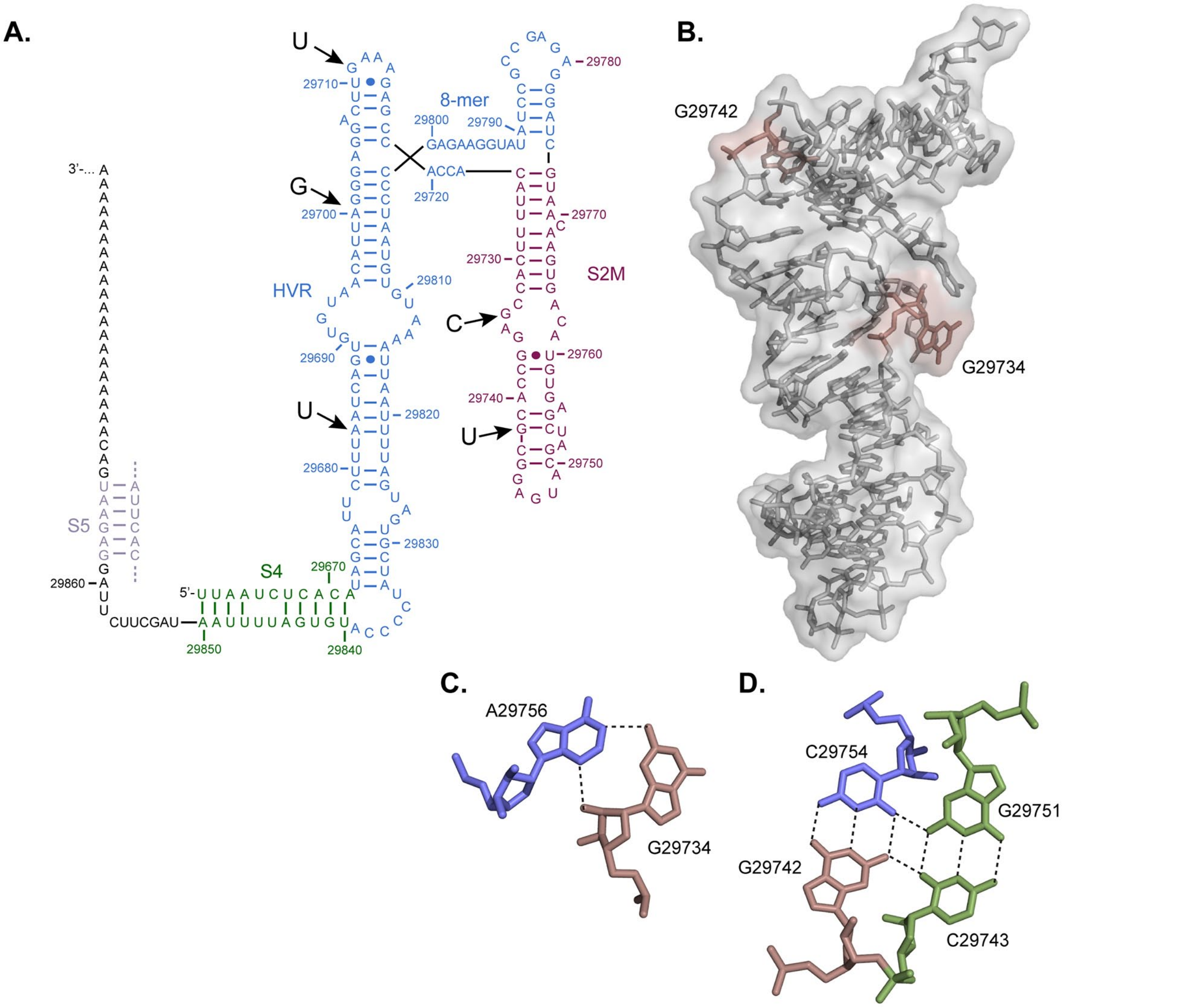


Figure 5. Emerging variations in the 3′UTR: (A) the secondary structure of the 3′ half of the SARS-CoV-2 genome is shown. This region includes the HVR, the octanucleotide motif (8-mer), the S2M structure, and the S4 and S5 stems. In all panels, the position of emerging variants is labeled as in Figure 2, (B) the crystal structure of the S2M region from SARS-CoV is shown. The position of 2 emerging variations in SARS-CoV-2 are shown in red adjacent to corresponding nucleotides in the SARS-CoV structure, (C) 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, and (D) G29742 is involved in a base quadruple with 3 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.

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 (Figure 6G and H). This variation in sequence affects the stability of the base quartet described above (Figure 5D), and ultimately impacts the stability of the adjacent base pairs (Figure 6H). The base quartet is at the junction between the 2 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 (Figure 6C).

To test these predictions experimentally, we performed UV thermal denaturation experiments with model RNA oligonucleotides corresponding to the reference S2M stem sequence and both emerging variations. The data were analyzed and plotted as per the SL1 stem experiments described above (Figure 7A). Consistent with our MD simulations, the S2M WT structure appears to be marginally stable, with a standard enthalpy change (ΔH°) of −35 ± 7 kcal/molK and an apparent T_m of 328 ± 1 K (Figure 7B), considerably weaker than the SL1 structure. Both the S2M G29734C and G29742U variants fit poorly to the 2 state denaturation model, with virtually no cooperativity in the transition from low to high temperatures (Figure 7A, compare WT to variant plot profiles). This suggests that both variants are mostly unfolded even at low temperature. Together, our data are consistent with the

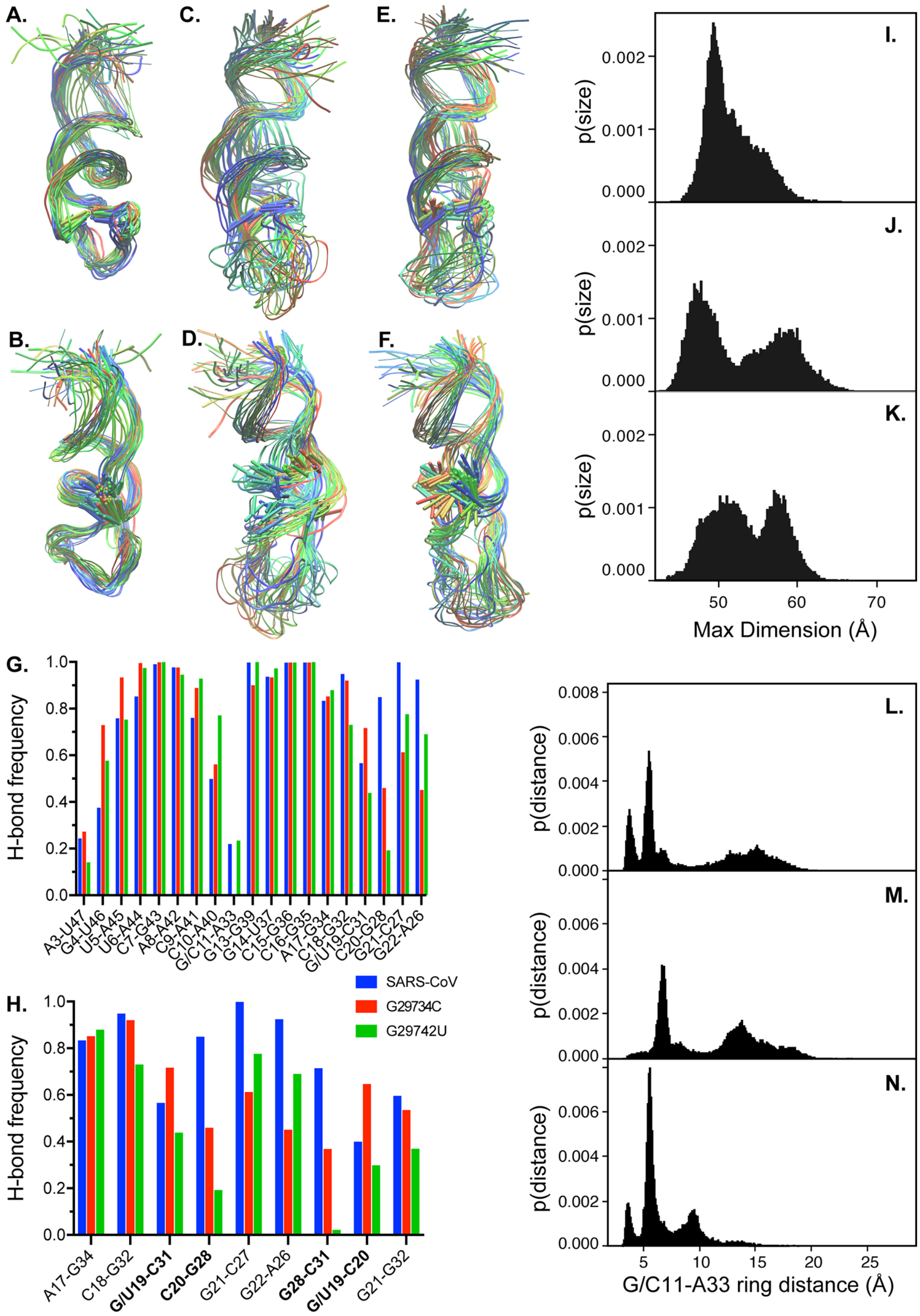
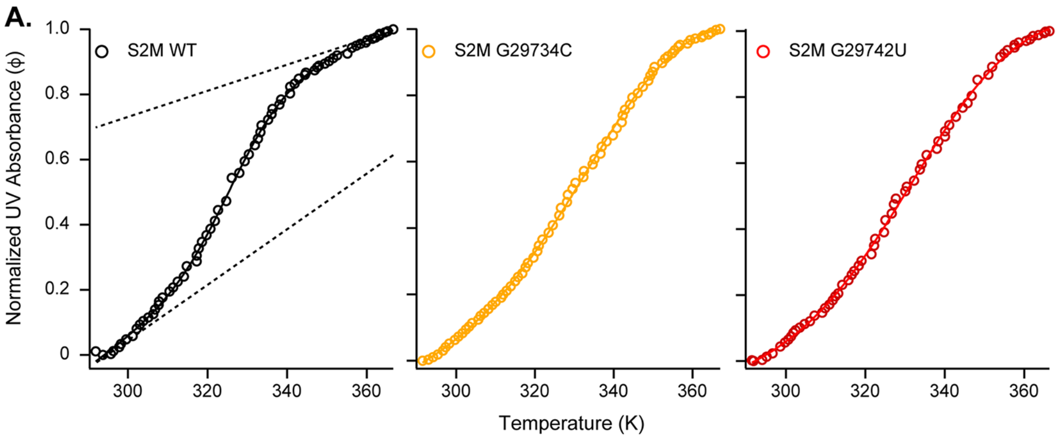


Figure 6. Molecular dynamics simulations of S2M variations: overlay of the structures from a single representative 180 ns trajectory of the SARS-CoV S2M loop (A and B), G29734C (C and D) and G29742U (E and F) variants. Front and back orientations show the following residues as sticks: G/U29742 and C29754 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 4 180 ns trajectories in (G) and (H). Histograms of the largest dimension of the S2M loop measured for the 4 180 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 4 180 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.

prediction from MD simulations that the S2M structure is marginally stable at 37°C and that the G29742U but not G29743C further reduces that stability. These results are consistent with the hypothesis that the S2M structure is not critical to viral replication, but more work will be needed to directly test that hypothesis in an infection model.



B.

Variation	ΔH° (kcal/mol K)	ΔS° (kcal/mol)	ΔG° ^{37} (kcal/mol K)	Tm (K)
S2M WT	-35 ± 7	-0.11 ± 0.02	-1.9 ± 0.5	328 ± 1
G29734C	n.d.	n.d.	n.d.	n.d.
G29742U	n.d.	n.d.	n.d.	n.d.

Figure 7. The S2M stem loop and both emerging variations are marginally stable at infection temperature: (A) normalized thermal UV denaturation curves with the S2M region and variations thereof are plotted as a function of temperature. The data are represented and analyzed as in Figure 2, with lines describing the upper and lower baselines of S2M WT RNA shown to facilitate visualization of the unfolding transition, and (B) fitted parameters presented in the table below are the average and standard deviation of 3 experiments. Values presented as n.d. could not be determined from fits to the 2 state denaturation model described in the methods, presumably because the S2M UUUC and UUUA variants are not folded throughout the temperature range used in the experiment.

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 emerging variations that impact viral protein structure and function.[50,52] They have also been used to monitor the spread of the virus through molecular phylogeny.[44,48,49] Here, we have used available data to investigate how 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.[21,22] 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.[92] 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 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.[41] 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 variations is interesting, as this region is not only present on the positive strand of the viral genome but also found on all subgenomic RNAs.[93] Moreover, the complement to SL1 in antigenomic RNA is likely recognized by viral RNA-dependent RNA polymerase 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 1 conformer or the other. Both conformers are necessary for MHV replication, and only 1 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 2 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.[93]

Our study has limitations that warrant further consideration. It remains to be determined whether the variations presented in this study impact the replication cycle of the SARS-CoV-2 virus. It is conceivable that variations identified in the 5′ or 3′UTRs could impact multiple aspects of the viral life cycle, including viral genome and anti-genome synthesis, subgenomic RNA synthesis, translation of viral mRNA, and viral RNA packaging. Subsequent to the completion of our analysis, a cell-based model system to study viral replication was developed and used to assess the impact of various coding mutations on viral replication and pathogenesis.[94] This system could be used to directly assess the impact of each cis-regulatory variation described here, and further delineate which step (if any) in the viral life cycle is impacted. A second limitation is that the RNA structure predictions, molecular dynamics simulations, and folding free energy measurements contained herein were performed in the absence of cellular factors known to influence RNA stability. It is possible that cellular RNA-binding proteins, macromolecular crowding, RNA modifications, liquid phase separation, or other factors inherent to the cytoplasm could influence the magnitude of the effects observed in our study.[95-97] These questions, too, could be further addressed in cell-based viral replication and structural probing assays such as SHAPE (Selective 2′-Hydroxyl Acylation analyzed by Primer Extension), which was recently used to describe the viral genome structure in cells.[98]

The analysis presented in this study is expected to improve as more sequencing data are added to available repositories.[45,46] It is possible that identification of additional 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

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) recovered in FASTA format were used as queries in a BLASTN search (Altschul et al. 1990). For the NCBI set, 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. Hits were downloaded as a hit table and aligned sequences. A multiple sequence alignment was prepared using a locally installed copy of MAFFT (Multiple Alignment using Fast Fourier Transform) version 7.464 using the default FFT-NS-2 algorithm.[56] The output file was then analyzed with a locally installed copy of WebLogo3 version 3.6.0.[57] 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 4 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 set, 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.[45,46] 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.[55] 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 set 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.[67] 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 2. 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 set 5′UTR BLASTn hits by searching for the variation plus 2 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 1 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 set 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.[70,71]

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.[75,76] 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.[20,80] 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[99] using the CHARMM 36 force field.[100] 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.[101] In VMD, each structure was solvated with a water box with explicit TIP3P[100,102] 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 2000 steps of conjugate gradient minimization were performed with the heavy atoms of the RNA and the 2 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[103] was used to allow a 2 fs time step. The particle mesh Ewald method[104] 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[101] and structures were visualized with VMD using the STRIDE algorithm for secondary structure identification.[105] 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.

Thermal denaturation experiments

RNA oligonucleotides corresponding to $SL1_{ref}$, $SL1_{UUUC}$, $SL1_{UUUA}$, $S2M_{ref}$, $S2M_{G29734C}$, and $S2M_{G29742U}$ were obtained from Integrated DNA Technologies (IDT, Coralville, IA). Sequences are available in Supplementary Table 3. The RNA oligonucleotides were diluted to 1 µM final concentration in 20 mM sodium cacodylate buffer pH 7.0, 100 mM NaCl, and 0.2 mM EDTA, heated to 65°C for 2 minutes, and then allowed to cool to room temperature. The UV absorbance was then measured at 1.0° increments across a temperature range from 20°C to 95°C with a 1°/min ramp speed in a Cary 100 UV spectrophotometer. The absorbance data were fit to a system of equations describing a 2 state unimolecular denaturation model (equation 1 and 2) as previously described to determine the standard enthalpy change (ΔH°) and the melting temperature (T_m).[68]

$$K = e^{\left\{\frac{-\Delta H}{R}\times\left(\frac{1}{Tm}-\frac{1}{T}\right)\right\}} \tag{1}$$

$$Abs = \left(m_u T + b_u\right) + \frac{\left(m_f T + b_f\right) - \left(m_u T + b_u\right)\times K}{1-K} \tag{2}$$

where m_u and m_f are the slopes of the unfolded and folded states, b_u and b_f are the Y intercepts of the unfolded and folded states, and R is the gas constant. The standard entropy change (ΔS°) at 37°C and the standard free energy change (ΔG°) were calculated by first normalizing the absorbance data by subtraction of the upper and lower baselines to determine the fraction unfolded (f) as function of temperature (equation 3), which was then used to define the K_{eq} of unfolding as previously described (equation 4).[69]

$$f = \frac{Abs - (m_u T + b_u)}{(m_f T + b_f) - (m_u T + b_u)} \tag{3}$$

$$K_{eq} = \frac{f}{1-f} \tag{4}$$

The ΔG° and ΔS° at 37°C were then calculated from the ΔH° and K_{eq} using basic thermodynamic relationships.

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. We thank the Watts, Zamore, and Thompson laboratories for sharing equipment.

Author Contributions

SPR performed the sequence analyses and molecular modeling. BRM collected the MD trajectories. BRM and FM analyzed the MD trajectories. PC performed the thermal denaturation experiments and KA prepared necessary reagents. SPR wrote the paper under the advisement of all authors.

ORCID iD

Sean P Ryder https://orcid.org/0000-0003-4960-0739

Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. Stadler K, Masignani V, Eickmann M, et al. SARS—beginning to understand a new virus. *Nat Rev Microbiol*. 2003;1:209-218.
2. Su S, Wong G, Shi W, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol*. 2016;24:490-502.
3. Drosten C, Gunther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med*. 2003;348:1967-1976.
4. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*. 2003;348:1953-1966.
5. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. 2020;17:1061-1069.
6. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med*. 2012;367:1814-1820.
7. Zhong NS, Zheng BJ, Li YM, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet*. 2003;362:1353-1358.
8. Chan JF, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. 2020;395:514-523.
9. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med*. 2020;382:1199-1207.
10. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579:265-269.
11. Zheng J. SARS-CoV-2: an emerging coronavirus that causes a global threat. *Int J Biol Sci*. 2020;16:1678-1685.
12. Hartenian E, Nandakumar D, Lari A, Ly M, Tucker JM, Glaunsinger BA. The molecular virology of coronaviruses. *J Biol Chem*. 2020;295:12910-12934.
13. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. 2020;183:1735.
14. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 2020;581:215-220.
15. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science*. 2005;309:1864-1868.
16. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*. 2020;181:271-280 e8.
17. Brierley I, Digard P, Inglis SC. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell*. 1989;57:537-547.
18. Giedroc DP, Cornish PV. Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res*. 2009;139:193-208.
19. Sola I, Almazan F, Zuniga S, Enjuanes L. Continuous and discontinuous RNA synthesis in coronaviruses. *Annu Rev Virol*. 2015;2:265-288.
20. Rangan R, Zheludev IN, Hagey RJ, et al. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA*. 2020;26:937-959.
21. Madhugiri R, Fricke M, Marz M, Ziebuhr J. Coronavirus cis-acting RNA elements. *Adv Virus Res*. 2016;96:127-163.
22. Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res*. 2015;206:120-133.
23. Chen SC, Olsthoorn RC. Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology*. 2010;401:29-41.
24. Goebel SJ, Taylor J, Masters PS. The 3' cis-acting genomic replication element of the severe acute respiratory syndrome coronavirus can function in the murine coronavirus genome. *J Virol*. 2004;78:7846-7851.
25. Kang H, Feng M, Schroeder ME, Giedroc DP, Leibowitz JL. Stem-loop 1 in the 5' UTR of the SARS coronavirus can substitute for its counterpart in mouse hepatitis virus. *Adv Exp Med Biol*. 2006;581:105-108.
26. Kang H, Feng M, Schroeder ME, Giedroc DP, Leibowitz JL. 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*. 2006;80:10600-10614.
27. Zust R, Miller TB, Goebel SJ, Thiel V, Masters PS. 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*. 2008;82:1214-1228.
28. Stephenson KE, Le Gars M, Sadoff J, et al. Immunogenicity of the Ad26.COV2. S vaccine for COVID-19. *JAMA*. 2021;325:1535-1544.
29. Baden LR, El Sahly HM, Essink B, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med*. 2021;384:403-416.
30. Polack FP, Thomas SJ, Kitchin N, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med*. 2020;383:2603-2615.
31. Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the treatment of Covid-19-final report. *N Engl J Med*. 2020;383:1813-1826.
32. Artese A, Svicher V, Costa G, et al. Current status of antivirals and druggable targets of SARS CoV-2 and other human pathogenic coronaviruses. *Drug Resist Updat*. 2020;53:100721.
33. Murgolo N, Therien AG, Howell B, et al. SARS-CoV-2 tropism, entry, replication, and propagation: considerations for drug discovery and development. *PLoS Pathog*. 2021;17:e1009225.
34. Shyr ZA, Gorshkov K, Chen CZ, Zheng W. Drug discovery strategies for SARS-CoV-2. *J Pharmacol Exp Ther*. 2020;375:127-138.
35. Khan A, Khan M, Saleem S, et al. Phylogenetic analysis and structural perspectives of RNA-dependent RNA-polymerase inhibition from SARs-CoV-2 with natural products. *Interdiscip Sci*. 2020;12:335-348.
36. Khan MT, Ali A, Wang Q, et al. Marine natural compounds as potents inhibitors against the main protease of SARS-CoV-2—a molecular dynamic study. *J Biomol Struct Dyn*. Published online June 1, 2020. doi: 10.1080/07391102.2020.1769733
37. Dittmar M, Lee JS, Whig K, et al. Drug repurposing screens reveal FDA approved drugs active against SARS-Cov-2. *biorxiv*. 2020.

38. Riva L, Yuan S, Yin X, et al. A large-scale drug repositioning survey for SARS-CoV-2 antivirals. *bioRxiv*. 2020.

39. Ahn DG, Lee W, Choi JK, et al. Interference of ribosomal frameshifting by antisense peptide nucleic acids suppresses SARS coronavirus replication. *Antiviral Res*. 2011;91:1-10.

40. Park SJ, Kim YG, Park HJ. Identification of RNA pseudoknot-binding ligand that inhibits the -1 ribosomal frameshifting of SARS-coronavirus by structure-based virtual screening. *J Am Chem Soc*. 2011;133:10094-100100.

41. Le TK, Paris C, Khan KS, Robson F, Ng WL, Rocchi P. Nucleic acid-based technologies targeting coronaviruses. *Trends Biochem Sci*. 2020;46:351-365.

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

43. Zhang YZ, Shi M, Holmes EC. Using metagenomics to characterize an expanding virosphere. *Cell*. 2018;172:1168-1172.

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

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

46. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data-from vision to reality. *Euro Surveill*. 2017;22:30494.

47. Bedford T, Greninger AL, Roychoudhury P, et al. Cryptic transmission of SARS-CoV-2 in Washington state. *Science*. 2020;370:571-575.

48. Chu HY, Englund JA, Starita LM, et al. Early detection of Covid-19 through a citywide pandemic surveillance platform. *N Engl J Med*. 2020;383:185-187.

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

50. Kim SJ, Nguyen VG, Park YH, Park BK, Chung HC. A novel synonymous mutation of SARS-CoV-2: is this possible to affect their antigenicity and immunogenicity? *Vaccines (Basel)*. 2020;8:220.

51. Lv H, Wu NC, Tak-Yin Tsang O, et al. Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections. *Cell Rep*. 2020;31:107725.

52. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med*. 2020;18:179.

53. Pinto D, Park YJ, Beltramello M, et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*. 2020;583:290-295.

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

55. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.

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

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

58. Li L, Kang H, Liu P, et al. Structural lability in stem-loop 1 drives a 5' UTR-3' UTR interaction in coronavirus replication. *J Mol Biol*. 2008;377:790-803.

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

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

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

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

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

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

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

66. Yang D, Liu P, Giedroc DP, Leibowitz J. Mouse hepatitis virus stem-loop 4 functions as a spacer element required to drive subgenomic RNA synthesis. *J Virol*. 2011;85:9199-91209.

67. Lorenz R, Bernhart SH, Zu Siederdissen CH, et al. ViennaRNA package 2.0. *Algorithms Mol Biol*. 2011;6:26.

68. Proctor DJ, Ma H, Kierzek E, Kierzek R, Gruebele M, Bevilacqua PC. Folding thermodynamics and kinetics of YNMG RNA hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. *Biochemistry*. 2004;43:14004-14014.

69. Puglisi JD, Tinoco I Jr. Absorbance melting curves of RNA. *Methods Enzymol*. 1989;180:304-325.

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

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

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

73. Lerner T, Papavasiliou FN, Pecori R. RNA editors, cofactors, and mRNA targets: an overview of the C-to-U RNA editing machinery and its implication in human disease. *Genes (Basel)*. 2018;10:13.

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

75. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156-2158.

76. Page AJ, Taylor B, Delaney AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*. 2016;2:e000056.

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

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

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

80. Popenda M, Szachniuk M, Antczak M, et al. Automated 3D structure composition for large RNAs. *Nucleic Acids Res*. 2012;40:e112.

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

82. Hsue B, Masters PS. 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*. Oct 1997;71:7567-7578.

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

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

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

86. Tengs T, Jonassen CM. Distribution and evolutionary history of the mobile genetic element s2m in coronaviruses. *Diseases*. 2016;4:27.

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

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

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

90. Tomaselli S, Galeano F, Locatelli F, Gallo A. ADARs and the balance game between virus infection and innate immune cell response. *Curr Issues Mol Biol*. 2015;17:37-51.

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

92. Guan BJ, Wu HY, Brian DA. An optimal cis-replication stem-loop IV in the 5' untranslated region of the mouse coronavirus genome extends 16 nucleotides into open reading frame 1. *J Virol*. 2011;85:5593-5605.

93. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell*. 2020;181:914-921 e10.

94. Ogando NS, Dalebout TJ, Zevenhoven-Dobbe JC, et al. SARS-coronavirus-2 replication in Vero E6 cells: replication kinetics, rapid adaptation and cytopathology. *J Gen Virol*. 2020;101:925-940.

95. Leamy KA, Assmann SM, Mathews DH, Bevilacqua PC. Bridging the gap between in vitro and in vivo RNA folding. *Q Rev Biophys*. 2016;49:e10.

96. Mitchell D, 3rd, Assmann SM, Bevilacqua PC. Probing RNA structure in vivo. *Curr Opin Struct Biol*. 2019;59:151-158.

97. Schroeder SJ. Perspectives on viral RNA genomes and the RNA folding problem. *Viruses*. 2020;12:1126.

98. Huston NC, Wan H, Strine MS, de Cesaris Araujo Tavares R, Wilen CB, Pyle AM. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol Cell*. 2021;81:584-598 e5.

99. Phillips JC, Braun R, Wang W, et al. Scalable molecular dynamics with NAMD. *J Comput Chem*. 2005;26:1781-1802.

100. MacKerell AD, Bashford D, Bellott M, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*. 1998;102:3568-3616.

101. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996;14:33-38, 27-28.

102. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79:926-935.

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

104. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *J Chem Phys*. 1995;103:8577-8593.

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