Structure of the full SARS-CoV-2 RNA genome in infected cells

Tammy C. T. Lan[1], Matthew F. Allan[1, 2, 3], Lauren E. Malsick[4], Stuti Khandwala[1, 5, 6], Sherry S. Y. Nyeo[1, 5, 6], Mark Bathe[3], Anthony Griffiths[4], Silvi Rouskin[1†]

[1] Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA
[2] Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[3] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4] National Emerging Infectious Diseases Laboratories, Boston University School of Medicine, Boston University, Boston, Massachusetts, USA
[5] Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[6] Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

† To whom correspondence should be addressed: E-mail: srouskin@wi.mit.edu

SUMMARY

SARS-CoV-2 is a betacoronavirus with a single-stranded, positive-sense, 30-kilobase RNA genome responsible for the ongoing COVID-19 pandemic. Currently, there are no antiviral drugs or vaccines with proven efficacy, and development of these treatments are hampered by our limited understanding of the molecular and structural biology of the virus. Like many other RNA viruses, RNA structures in coronaviruses regulate gene expression and are crucial for viral replication. Although genome and transcriptome data were recently reported, there is to date little experimental data on predicted RNA structures in SARS-CoV-2 and most putative regulatory sequences are uncharacterized. Here we report the secondary structure of the entire SARS-CoV-2 genome in infected cells at single nucleotide resolution using dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq). Our results reveal previously undescribed structures within critical regulatory elements such as the genomic transcription-regulating sequences (TRSs). Contrary to previous studies, our in-cell data show that the structure of the frameshift element, which is a major drug target, is drastically different from prevailing *in vitro* models. The genomic structure detailed here lays the groundwork for coronavirus RNA biology and will guide the design of SARS-CoV-2 RNA-based therapeutics.

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of the current coronavirus disease 2019 (COVID-19), recently declared a global pandemic by the World Health Organization (WHO). SARS-CoV-2 is an enveloped virus belonging to the genus *betacoronavirus*, which also includes SARS-CoV, the virus responsible for the 2003 SARS outbreak, and Middle East respiratory syndrome coronavirus (MERS-CoV), the virus responsible for the 2012 MERS outbreak. Despite the devastating effects these viruses have had on public health and the economy, currently no effective antivirals treatment or vaccines exist. There is therefore an urgent need to understand their unique RNA biology and develop new therapeutics against this class of viruses.

Coronaviruses (CoVs) have single-stranded and positive-sense genomes that are the largest of all known RNA viruses (27 – 32 kb) (Masters, 2006). Previous studies on coronavirus structures have focused on several conserved regions that are important for viral replication. For several of these regions, such as the 5' UTR, the 3' UTR, and the frameshift element (FSE), structures have been predicted computationally with supportive experimental data from RNase probing and Nuclear magnetic resonance (NMR) spectroscopy (Plant *et al.*, 2005; Yang and Leibowitz, 2015). Functional studies have revealed the importance of their secondary structures for viral transcription and replication (Brierley, Digard and Inglis, 1989; Liu *et al.*, 2007; Li *et al.*, 2008; Yang and Leibowitz, 2015).

The FSE is located near the boundary of ORF1a and ORF1b and causes the ribosome to "slip" and shift register by -1 nt in order to bypass a canonical stop codon at the end of ORF1a and translate the viral RNA-dependent RNA polymerase (RdRP) and other proteins in ORF1b (Brierley *et al.*, 1987; Plant *et al.*, 2005). Studies on multiple viruses have shown that an optimal frameshifting rate is critical, and small differences in percentage of frameshifting lead to dramatic differences in genomic RNA production and infection dose (Plant *et al.*, 2010). Therefore, the FSE has emerged as a major drug target for binding of small molecules that can influence the rate of ribosome slippage and be used as a treatment against SARS-CoV-2.

The structure of the FSE of SARS-CoV (whose sequence differs from the SARS-CoV-2 FSE by just one nucleotide), was solved by NMR to be a 3-stem pseudoknot (Plant *et al.*, 2005; Rangan, Zheludev and Das, 2020). The prevailing mechanism is that the 3-stem pseudoknot causes the ribosome to pause at the slippery sequence and backtrack by 1 nt to release mechanical tension (Plant and Dinman, 2008). However, the frameshifting rate of SARS-CoV-2 during infection is unknown, and none of the RNA structure models for the FSE has been validated in cells.

Furthermore, the structures of many physiologically important elements of coronavirus genomes remain uncharacterized, including most of the transcription-regulating sequences (TRSs). The TRSs are short sequences that are critical for the transcription of subgenomic RNAs (sgRNAs), from which all proteins not encoded by ORF1ab are translated (Sola *et al.*, 2015). The SARS-CoV-2 genome contains ten TRSs: one "leader

TRS" (TRS-L) located in the 5' UTR, and nine "body TRSs" (TRS-B), one preceding each open reading frame other than ORF1ab (Kim *et al.*, 2020).

The nine TRS-Bs give rise to nine sgRNAs via discontinuous transcription (Sola *et al.*, 2015). In this process, transcription of a minus-strand sgRNA begins at the 3' end of the genomic RNA. When transcription reaches one of the TRS-Bs, the nascent transcript dissociates from the TRS-B and re-anneals to the TRS-L. Transcription resumes at the TRS-L and finishes at the 5' end of the genomic RNA. Thus, each sgRNA contains a short 5' "leader" sequence (69 nt from the 5' end of the genome to TRS-L) followed by the TRS-B and the portion of the genome downstream of the TRS-B.

Nine of the ten TRSs in the SARS-CoV-2 genome contain the same core sequence (CS) of 6 nt (ACGAAC) flanked by short sequences that vary among the TRSs (Kim *et al.*, 2020). The one exception is the TRS of ORF7b, whose CS is AAGAAC. Interestingly, while the CSs of SARS-CoV-2 and SARS-CoV are identical, the sequence differs substantially from the CSs of other closely related coronaviruses (Yount *et al.*, 2003).

Since the sequence of the SARS-CoV-2 genome became available, several groups have computationally predicted structures at multiple genomic regions (Andrews *et al.*, 2020; Rangan, Zheludev and Das, 2020). This work provided important first insights into the secondary structure of SARS-CoV-2 using algorithms such as RNAz (Gruber *et al.*, 2010) and Contrafold (Do, Woods and Batzoglou, 2006). However, these algorithms are limited by "breaking up" the full-length RNA into small windows (~120 nt) and folding each window independently of the rest of the genomic sequence. This strategy makes the final structure model sensitive to the choice of the exact start and end of each independent window. Further, *in silico* generated models cannot account for the physiological context in which an RNA folds within cells.

Over the past ten years, the drawbacks of *in silico* models have been largely overcome by the development of genome-wide strategies to chemically probe RNA structure in cells. The most popular chemical probes are dimethyl sulfate (DMS) (Rouskin *et al.*, 2014), SHAPE (Siegfried *et al.*, 2014), and icSHAPE (Spitale *et al.*, 2015), all of which provide critical *in vivo* measurements that are used by computational algorithms to generate accurate models of RNA structures *in vivo*.

In this study, we perform DMS mutational profiling with sequencing (DMS-MaPseq) (Zubradt *et al.*, 2016) on infected Vero cells to generate the first experimentally determined genome-wide secondary structure of SARS-CoV-2. Our results reveal major differences with *in silico* predictions and highlight the physiological structures of known functional elements. Our work provides experimental data on the structural biology of RNA viruses and will inform efforts in the development of RNA-based diagnostics and therapeutics for SARS-CoV-2.

RESULTS

The genome-wide structure of SARS-CoV-2 in cells

To determine the intracellular genome-wide structure of SARS-CoV-2, we added dimethyl sulfate (DMS) to infected Vero cells and performed mutational profiling with sequencing (DMS-MapSeq) (Zubradt *et al.*, 2016) (FIGURE 1A). DMS rapidly and specifically modifies unpaired adenines and cytosines *in vivo* at their Watson-Crick faces. Our results were highly reproducible between independent biological replicates (r = 0.89; Figure 1B). Combined, a total of ~1.58 million paired reads mapped to the coronavirus genome (Figure 1C), representing ~40% of total cellular RNA (post ribosomal RNA depletion). This large fraction of coronavirus reads from total intracellular RNA is consistent with previous literature using SARS-CoV-2 infected Vero cells (Kim *et al.*, 2020). DMS treated samples had high signal to noise ratio, with adenines and cytosines having a mutation rate ~9-fold higher than the background (guanines and uracils). In contrast, in untreated samples the mutation rate on all four bases (0.10%) was slightly lower than previously reported average sequencing error rates of 0.24% (Pfeiffer *et al.*, 2018) (Figure 1D).

We used the DMS-MaPseq data as constraints in RNAstructure (Mathews, 2004) to fold the entire SARS-CoV-2 genomic RNA (Supplementary Figure 1). We first examined the results at the 265 nt 5' UTR (Figure 1E), one of the best studied regions in the coronavirus genome (Yang and Leibowitz, 2015; Madhugiri *et al.*, 2018). The 5' UTRs of multiple coronaviruses have been characterized extensively in terms of their structures and roles in viral replication. In agreement with previously published structures of this highly conserved region (Yang and Leibowitz, 2015; Madhugiri *et al.*, 2018), we found five stem loops (SL1 – 5) within the SARS-CoV-2 5' UTR and three stem loops (SL6 – 8) near the beginning of ORF1a.

Stem loop 1 (SL1) is vital for viral replication. A previous study on murine hepatitis virus (MHV) found that mutations that destabilize the upper portion (near the loop) of SL1 or stabilize the lower portion (near the base) prevent viral replication (Li *et al.*, 2008). In addition, viable viruses with a single-base deletion in the lower portion also had mutations in the 3' UTR. Together, these observations suggested that the lower portion of SL1 must be able to unfold and likely interacts with the 3' UTR to assist in sgRNA transcription (Li *et al.*, 2008). In our structure, the lower portion of SL1 has low DMS reactivity, consistent with a pairing between these bases either within the stem or with the 3' UTR. Further work is needed to test a potential alternative structure for SL1 that would involve long range interactions between the 5' and the 3' UTRs of SARS-CoV-2.

The most evolutionarily conserved secondary structure in the 5' UTR is stem loop 2 (SL2) and the structure of the MHV SL2 has been solved by NMR (Liu *et al.*, 2009). The sequences of SARS-CoV-2 and MHV SL2 are identical, and our in-cell model of SL2 is consistent with the NMR structure, with every base in the stem showing low DMS reactivity. Although C50 within the loop of SL2 has low reactivity, this is explained by the NMR structure, in which C50 and G53 (as numbered here) are paired. Interestingly, the secondary structure but not the sequence of the stem was shown to be essential for

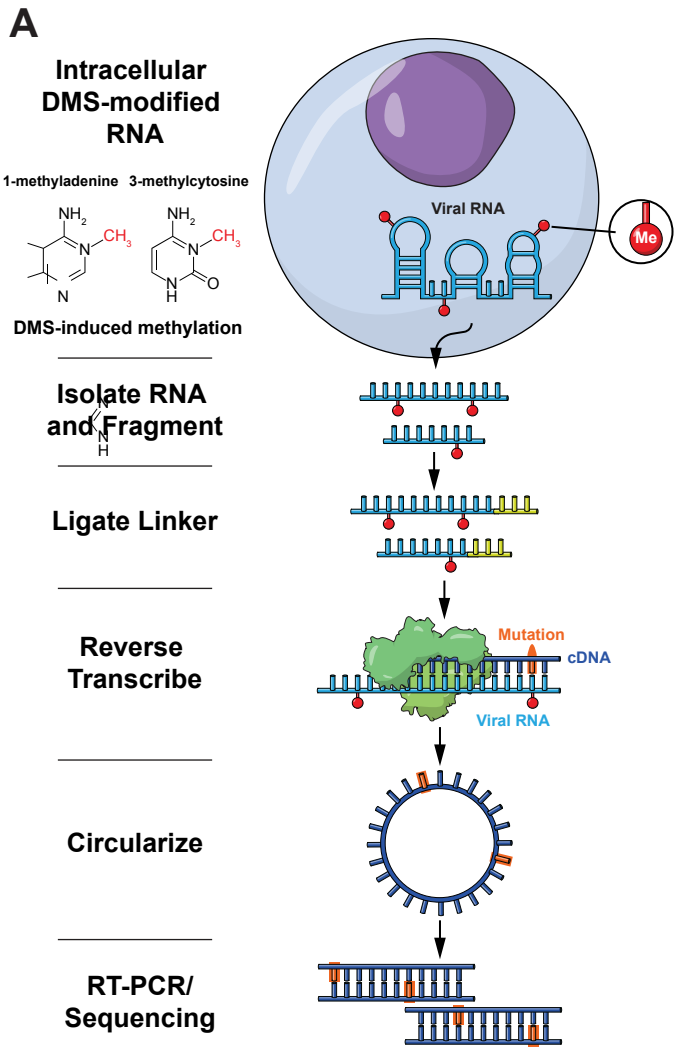
replication: in a previous study, disruption of SL2 by three mutations on either side of the stem prevented MHV-A59 replication, but compensatory mutations restored nearly wild-type growth (Liu *et al.*, 2007). Mutants of MHV-A59 with non-functional SL2 were able to transcribe minus-strand genomic RNA in cells but not minus-strand subgenomic RNAs, suggesting that SL2 plays a role in discontinuous transcription. Disruption of SL2 also reduced the rate of *in vitro* translation of a downstream luciferase reporter (Liu *et al.*, 2007). To the best of our knowledge, the precise mechanism through which SL2 is required for viral replication remains unknown.

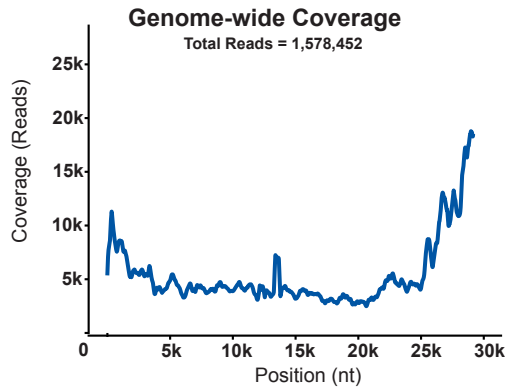
Stem loop 3 (SL3) contains the leader TRS (TRS-L) involved in discontinuous transcription (Yang and Leibowitz, 2015), the structure and function of which will be discussed in Figure 3.

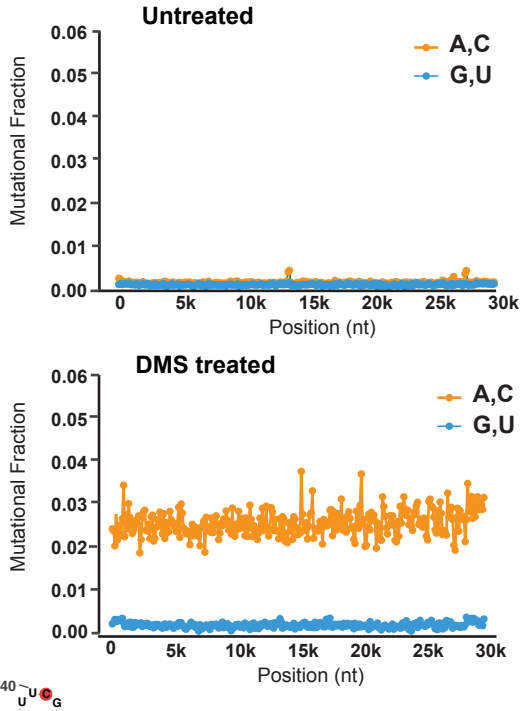
Stem loop 4 (SL4) is a bulged stem loop downstream of TRS-L. A previous study in MHV found that SL4 was required for sgRNA synthesis and that deleting the entire SL4 was lethal (Yang *et al.*, 2011). However, mutation or deletion of either the lower (SL4a) or upper (SL4b) part of SL4 merely impairs replication, with disruption of SL4a causing greater impairment (Yang *et al.*, 2011). In addition, replacing the entire SL4 with a shorter, unrelated sequence that was predicted to fold into a stable stem loop impaired MHV growth but was not lethal (Yang *et al.*, 2011). Based on these results, Yang et al. proposed that discontinuous transcription requires proper spacing between the stem loops upstream and downstream of the TRS, and SL4 maintains this space (Yang *et al.*, 2011). SL4 also contains the start codon of a short open reading frame (nine codons in SARS-CoV-2) known as the “upstream” ORF (uORF) (Madhugiri *et al.*, 2014). In a study on MHV, disruption of the uORF start codon increased translation of ORF1a and modestly decreased viral replication (Madhugiri *et al.*, 2014).

Stem loop 5 (SL5) contains four branches (SL5, SL5A, SL5B, SL5C), and its structure in MHV-A59 was predicted using chemical probing with SHAPE (Yang and Leibowitz, 2015). Disruption of SL5C has been shown to prevent replication of viral defective interfering particles (Yang and Leibowitz, 2015) but not of whole MHV (Yang *et al.*, 2015). SL5 contains the start codon of ORF1a, although in different coronaviruses the start codon can be located in SL5A, SL5B, or the stem 3’ of SL5C (Yang and Leibowitz, 2015). Our data support a three-branched SL5 for SARS-CoV-2, similar to previous models of SARS-CoV (Yang and Leibowitz, 2015) and SARS-CoV-2 (Rangan, Zheludev and Das, 2020) in which the start codon is located in the main SL5 stem, just downstream of SL5C.

Stem loops 6, 7, and 8 (SL6 – 8) lie downstream of the 5’ UTR, within the coding sequence of nsp1. The structures of these stem loops vary among betacoronaviruses, with SARS-CoV predicted to have all three stem loops, MHV only SL6 and SL7, and bovine coronavirus (BCoV) a structure in which SL6 and SL7 are branches of another stem (Yang *et al.*, 2015). As predicted for SARS-CoV, our model also has all three stem loops, though this contrasts with an *in silico* model of SARS-CoV-2 that features three short stem loops in place of SL8 (Rangan, Zheludev and Das, 2020). Mutations of SL6 that disrupt its secondary structure but preserve the amino acid sequence of nsp1 were not lethal in MHV, suggesting that SL6 is not essential (Yang *et al.*, 2015)


A
Intracellular DMS-modified RNA
1-methyladenine
3-methylcytosine
DMS-induced methylation
Viral RNA
Me
Isolate RNA and Fragment
Ligate Linker
Reverse Transcribe
Mutation
cDNA
Viral RNA
Circularize
RT-PCR/ Sequencing


C
Genome-wide Coverage
Total Reads = 1,578,452
Coverage (Reads)
25k
20k
15k
10k
5k
0
5k
10k
15k
20k
25k
30k
Position (nt)


D
Untreated
A,C
G,U
Mutational Fraction
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
5k
10k
15k
20k
25k
30k
Position (nt)
DMS treated
A,C
G,U
Mutational Fraction
Position (nt)

B
r = 0.89
in cells replicate 2 DMS signal
in cells replicate 1 DMS signal
0.00
0.05
0.10
0.15
0.20
0.25
0.30

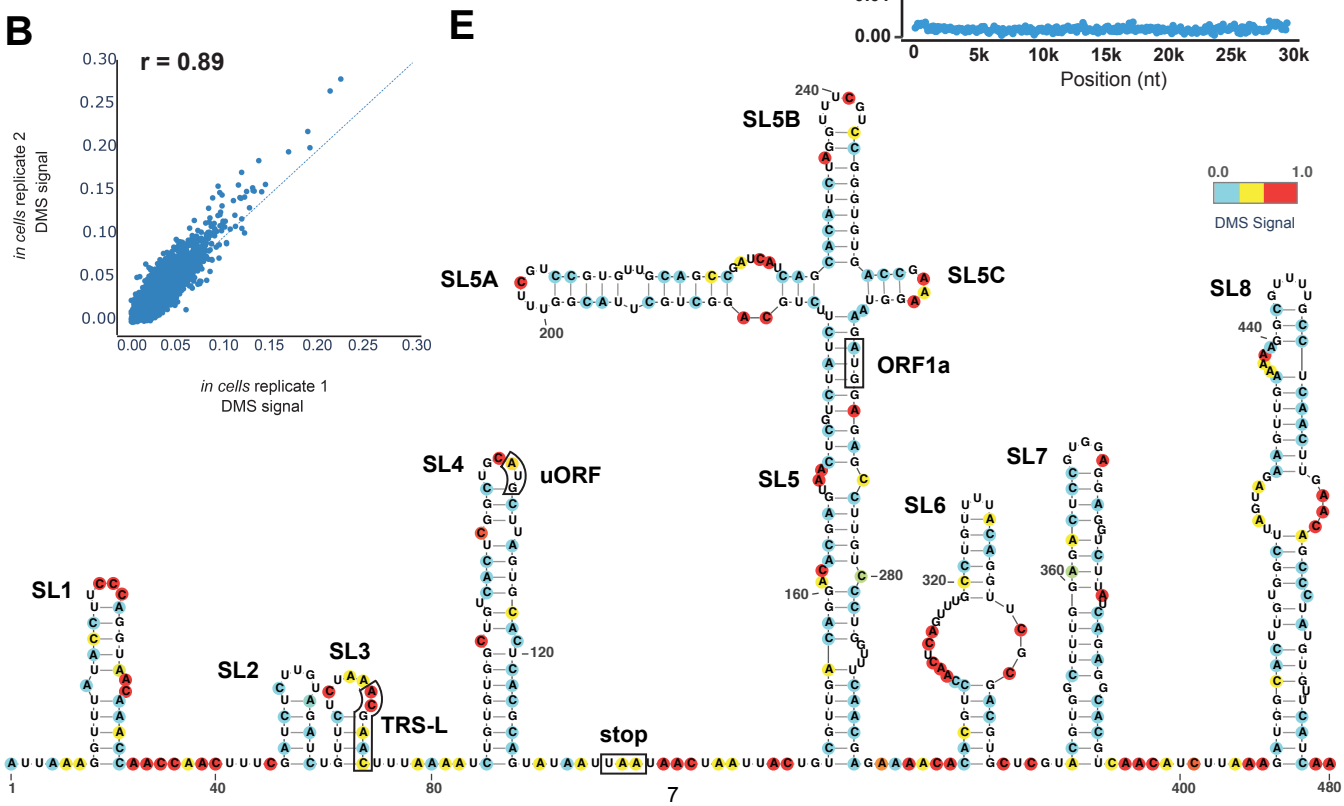

E
SL5B
SL5A
SL5C
ORF1a
SL8
SL4
uORF
SL5
SL6
SL7
SL1
SL2
SL3
TRS-L
stop
0.0
1.0
DMS Signal
1
40
80
400
480

Figure 1: Genome-wide probing of intracellular SARS-CoV-2 RNA structure with DMS-MaPseq. **(A)** Schematic of the experimental protocol for probing viral RNA structures with DMS-MaPseq. **(B)** Correlation of DMS reactivities for each base between two biological replicates. **(C)** Genome-wide coverage as a function of position. Coverage at each position represents the average coverage over a 400 nt window. **(D)** Signal and noise as a function of genome position for untreated and DMS-treated RNA. Signal (mutation rate for A and C) and noise (mutation rate for G and U) at each position was plotted as the average of 100 nt window. Mutational Fraction of 0.01 at a given position represents 1% of reads having a mismatch or deletion at that position. **(E)** In-cell model of the 5' UTR and beginning of ORF1a structure. Bases are colored by their DMS signal; bases that are not DMS reactive are colored white.

Structured and accessible regions in the SARS-CoV-2 genome

To broadly locate regions of potential interest in the SARS-CoV-2 genome, we identified all structured and accessible regions, which we define as stretches of at least 10 consecutive paired bases or 14 consecutive unpaired bases, respectively (Figure 2A). Structured regions could be functional elements of the genome that are potential targets of therapeutics. We identify 215 structured regions covering a total of 2,359 bases (8.5% of the genome). These regions include structures with known functions, such as SL5 in the 5' UTR and a bulged stem loop in the 3' UTR (Yang and Leibowitz, 2015), as well as many structures whose functions are unknown. The longest stretch of consecutive paired bases is 25 nt (positions 11,306 - 11,330) and falls within a previously uncharacterized set of two bulged stem loops in ORF1a. These results provide a starting point for characterizing additional functional structures in the viral life cycle.

Accessible regions in viral genomes are potential targets of antisense oligonucleotide therapeutics (Eckardt, Romby and Sczakiel, 1997; Ding, 2001; Rangan, Zheludev and Das, 2020) through the mechanism of RNA interference (Wilson and Doudna, 2013). We identify 261 accessible regions covering 6,305 bases (21% of the entire genome). As there are 11 such regions within ORF-N, which is present in every sgRNA, every sgRNA may be targetable with antisense oligos. Additionally, every open reading frame except ORF-E contains at least one of these predicted accessible regions, raising the possibility of targeting individual sgRNAs.

At a more stringent threshold of 25 consecutive nucleotides predicted to be unpaired, we find 78 accessible regions, one of which is located within ORF-N (34 nt, positions 28,721 - 28,754) and hence in every genomic and subgenomic RNA. The two longest stretches of consecutive predicted unpaired bases are both 122 nt long and occupy positions 9,602 - 9,723 (within ORF1a) and 21,573 - 21,694 (within ORF-S). These regions may offer multiple binding sites for antisense oligonucleotides.

Transcription-Regulating Sequences (TRSs) lie within stem loops

As the transcription-regulating sequences (TRSs) are necessary for the synthesis of sgRNAs, we analyzed our structural models of the leader TRS (TRS-L) and the nine body TRSs (TRS-B). The leader TRS (TRS-L) is the central component of the 5' UTR involved in discontinuous transcription (Sola *et al.*, 2015). *In silico* models for several alpha and betacoronaviruses variously place TRS-L in stem loop 3 (SL3) or in an unpaired stretch of nucleotides (Liu *et al.*, 2007; Yang and Leibowitz, 2015). The TRS-L of SARS-CoV and

of SARS-CoV-2 was predicted to lie in the 3' side of the stem of SL3, which is consistent with our in-cell model (Liu et al., 2007; Yang and Leibowitz, 2015; Rangan, Zheludev and Das, 2020). In our data, the stem of SL3 contains two bases with medium reactivity (Figure 2B), which suggests that SL3 transitions between folded and unfolded states, as is hypothesized for the alphacoronavirus transmissible gastroenteritis virus (TGEV) (Madhugiri et al., 2018).

Of the nine body TRSs, we find that seven (all but the TRSs of ORF7a and ORF7b) lie within a stem loop. Of these, all but one TRS (N) place the core sequence on the 5' side of the stem. Four body TRSs (M, ORF6, ORF8, and N) are predicted to lie in stem loops with two or three bulges, with the core sequence spanning one of the internal bulges. The other three structured body TRSs (S, ORF3a, and E) lie in stem loops without bulges, with the final paired base in the 5' side of the stem contained in the core sequence. Strikingly, the entire core sequence is paired in two body TRSs (S and M), and partially exposed in a loop or bulge in the other five (Figure 2B).

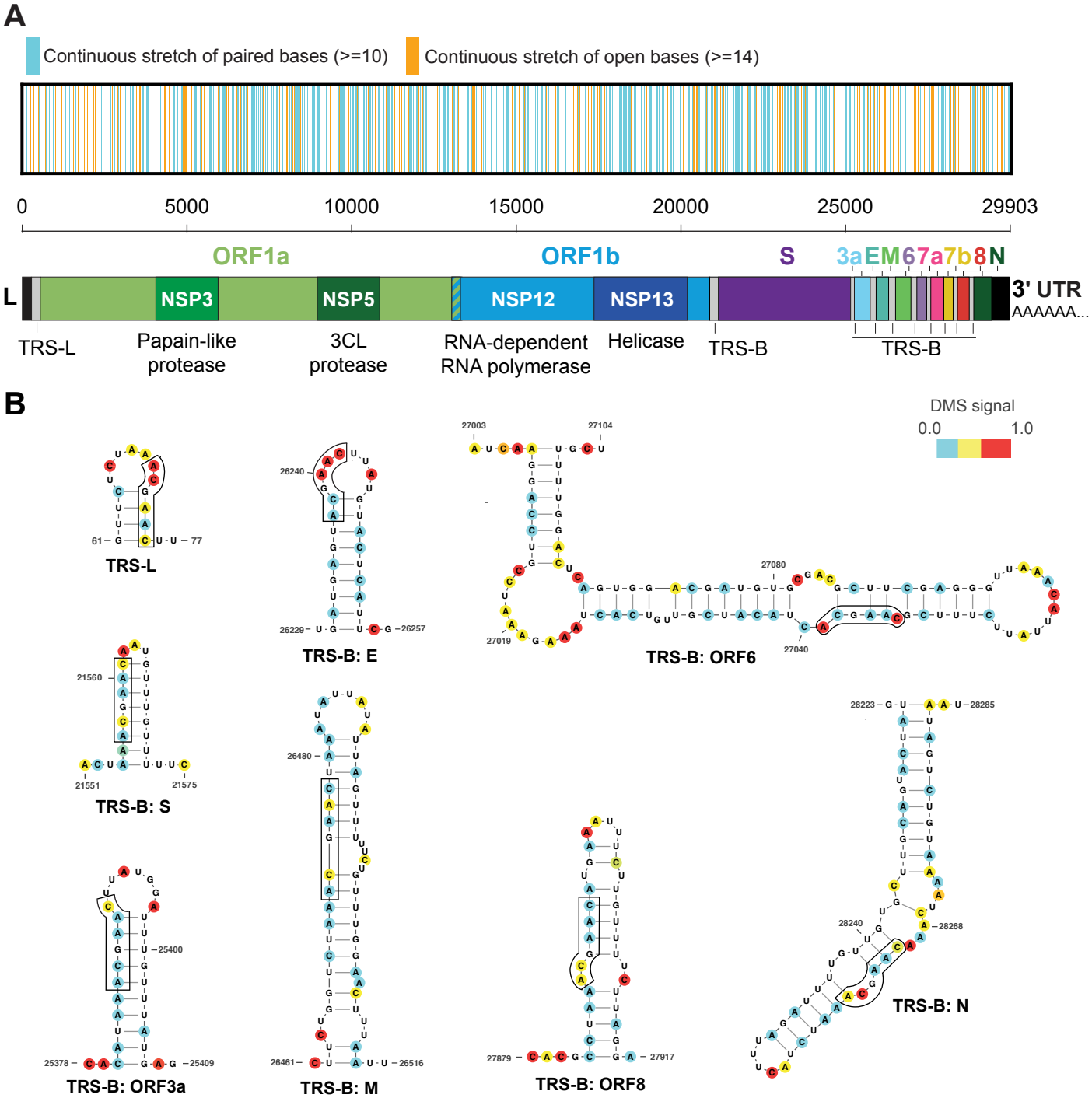


Figure 2: Structured and unstructured regions in the SARS-CoV-2 genome. (A) Locations of highly structured and unstructured regions in the SARS-CoV-2 genome. Highly structured regions are defined are stretches of at least 10 consecutive paired bases; unstructured regions shown are stretches of at least 14 consecutive unpaired bases. The thickness of each bar is proportional to the number of consecutive paired (blue) or unpaired (orange) bases. The data is plotted over a schematic of the genome, highlighting the organization of open reading frames (ORFs) and the transcription regulatory sequences (TRS). **(B)** In-cell model of each of the eight TRSs predicted to lie within a stem loop. The core sequence (CS) of each TRS is outlined in black. Models are arranged in genomic order from top-to-bottom, left-to-right.

In-cell data-derived model differs from previous *in silico* predictions

We evaluated the robustness of our in-cell data derived genome-wide model by varying two critical RNA folding parameters used by RNAstructure: 1) the maximum allowed distance for base pairing and 2) the threshold for DMS signal normalization.

A previous *in silico* approach for folding RNA found that limiting base pairs to be 100 to 150 nt apart was optimal to avoid overpredicting structured regions (Lange *et al.*, 2012). However, some RNA viruses contain known essential structures wherein bases over 300 nt apart are paired (e.g. the Rev response element in HIV-1 spans approximately 350 nt (Watts *et al.*, 2009)). We therefore varied the maximum distance (md) allowed for base pairing from 120 nt to 350 nt. We computed the agreement between the resulting structures using a modified version of the Fowlkes-Mallows index (Fowlkes and Mallows, 1983) that compares base pairing partners as well as unpaired bases (Methods). Overall, there was high agreement while varying the md from 120 nt to 350 nt, suggesting that long-distance (i.e. >120 nt) interactions across the SARS-CoV-2 RNA have a small effect on the identity of local structures. The genome structure folded with an md of 120 nt was 97.5% identical to the structure with an md of 350 nt, and in the latter structure only 3.8% of base pairs spanned >120 nt. Next, we proceeded with the md limit of 350 nt and tested two different DMS signal thresholds that normalize reactivity to either the median of the top 5 % or top 10% of the most reactive bases. We found that the structure models produced with the two normalization approaches were highly similar, with 93.6% identity (Figure 3A). Thus, within the ranges that we tested, our genome-wide data-derived model was robust to variation in the parameters of RNAstructure (Mathews, 2004).

We proceeded with the whole genome structure modelled with a md of 350 nt and a DMS signal normalization of 5% for further analysis (Supplementary Figure). Previous studies that computationally predicted genome-wide SARS-Cov-2 RNA structures used 1) RNAz, a thermodynamic-based model that additionally takes sequence alignment and considers base pairing conservation (Gruber *et al.*, 2010; Rangan, Zheludev and Das, 2020), and 2) Contrafold, which predicts RNA secondary structures without physics-based models and instead uses learned parameters based on known structures (Do, Woods and Batzoglou, 2006). These recent studies predicted 228 structures with RNAz with lengths ranging from 90 to 120 nt, and 79 structures with Contrafold with lengths ranging from 55 to 111 nt (Rangan, Zheludev and Das, 2020). For each of these structures, we computed the agreement between the different models (Figure 3B). We report the agreement using the mFMI while either excluding external bases pairs or including these pairs (Methods). As expected, agreement with the structures from purely computational prediction is higher when excluding external base pairs (average 76.3% for RNAz, 69.3% for Contrafold) than

when including them (average 71.2% for RNAz, 54.0% for Contrafold). Since our goal is to compare the overall similarity of two structures, we chose the inclusion of external base pairs as the more accurate metric for comparing the structures. Our predictions overall agreed more with those from RNAz (mean 71.2%, median 75.2%) than Contrafold (mean 54.0%, median 54.4%). We report the agreement between our structure and the RNAz structures across the entire genome (Figure 3C). Most structures are 60 to 80% identical, with several short regions that disagree substantially.

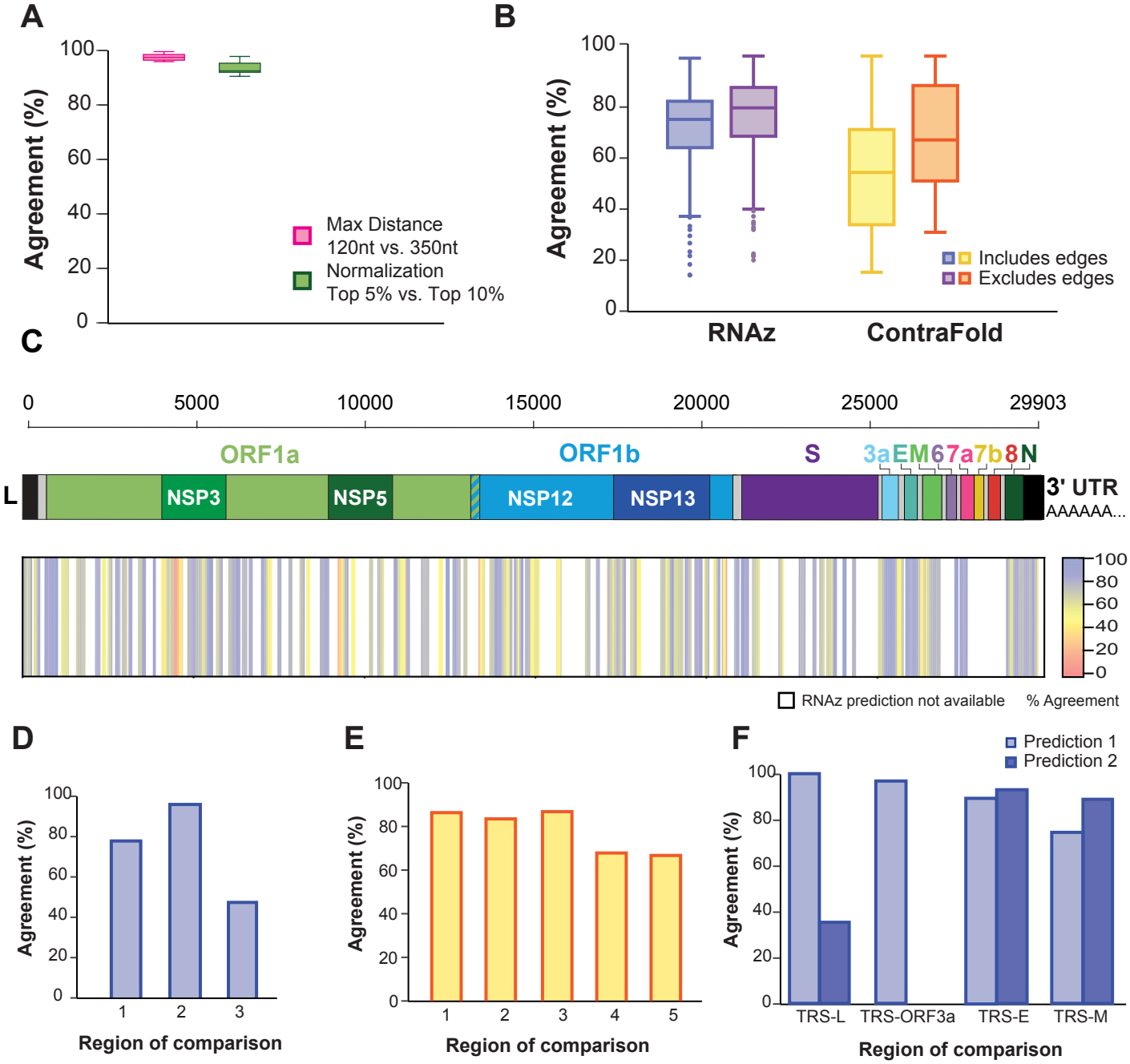


Figure 3: Comparison of our in-cell genome-wide structure model with previous models. (A) Consistency of our in-cell structure models. Agreement is given between our structure models predicted using a maximum distance limit of 120 nt and 350 nt between paired bases at 5% signal normalization and between our predictions using 5% and 10% DMS normalization at 350 nt maximum allowed base pair distance. **(B)** Agreement of our structure model with all predicted structures from RNAz and Contrafold. Agreement is given for both excluding and including external base pairs. **(C)** Agreement of our structure with a previous model from RNAz across the genome. At positions for which multiple RNAz model exists, the average agreement with all models is given. **(D)** Agreement of our model with RNAz predicted structures with the three highest P-values in regions with previously unannotated structures. **(E)** Agreement of our model with Contrafold predicted structures with the five highest maximum expected accuracies in evolutionarily conserved regions. **(F)** Agreement of our TRS structure models to RNAz predicted structures. For TRSs for which multiple RNAz models exist, agreement with each prediction is shown.

In addition, we computed the similarity of our model compared to the structures with the three highest P-values predicted with RNAz that do not overlap known structures in the Rfam database (Kalvari *et al.*, 2018; Rangan, Zheludev and Das, 2020) (Figure 3D). We noted that in all three cases, the structure at the center of the window was nearly identical to ours, and most of the disagreements arose at the edges, presumably due to the effects of the windows from RNAz being limited to 120 nt. Of the five structures predicted with Contrafold that had the largest maximum expected accuracies, our agreement ranged from 66.0% to 86.1%, well above the genome-wide mean (54.0%), suggesting that these structures are indeed more accurate than the average Contrafold structure (Figure 3E).

Finally, we compared the structures at the TRS elements to those predicted by RNAz (Rangan, Zheludev and Das, 2020) (Figure 3F). To remove the effects of external base pairs, we focused only on the complete structural element (e.g. a stem loop) in which the TRS was located. RNAz predicted structure for four TRSs. Our model for TRS-L was identical to the first predicted window from RNAz but differed significantly (35.3% agreement) from the next prediction of the same TRS-L element within a different folding window, indicating that the choice of folding window can have a large effect on the RNAz structure model. For the other three TRS elements for which RNAz predicted at least one structure for, our agreement ranges from 74.4% to 96.8%, above the genome-wide average of 71.2%, lending support to both models.

Uncovering an unexpected structure at Frameshift Element (FSE)

The frameshift element (FSE) causes the ribosome to slip and shift register by -1 nt in order to bypass a canonical stop codon and translate the viral RNA-dependent RNA polymerase (RdRP) (Plant and Dinman, 2008). Previous studies on coronaviruses and other viruses have shown that an optimal frameshifting rate is critical and small differences in percentage of frameshifting lead to dramatic differences in genomic RNA production and infection dose (Plant *et al.*, 2010). Therefore, the FSE has emerged as a major drug target for small molecule binding that could influence the rate of frameshifting and be used as a treatment against SARS-CoV-2. To date, there is little experimental data on the structure of SARS-CoV-2 FSE and the prevailing model is a 3-stem pseudoknot forming downstream of the slippery site, which is thought to pause the ribosome and allow frameshifting to occur (Plant and Dinman, 2008).

To closely examine the FSE structure in cells, we used DMS-MaPseq target specific protocol (Zubradt *et al.*, 2016). We designed primers targeting 283 nt surrounding the FSE and amplified this region from cells infected with SARS-CoV-2 that were treated with DMS. Our analysis revealed a strikingly different structure than the prevailing model (Plant *et al.*, 2005; Rangan, Zheludev and Das, 2020) (Figure 4A). Our in-cell model does not include the expected pseudoknot formation downstream of the slippery sequence. Instead, half of the canonical stem 1 (Figure 4A, purple) finds an alternative pairing partner (pink) driven by 10 complementary bases upstream of the slippery site (Figure 4A, pink). We call this pairing Alternative Stem 1 (AS1).

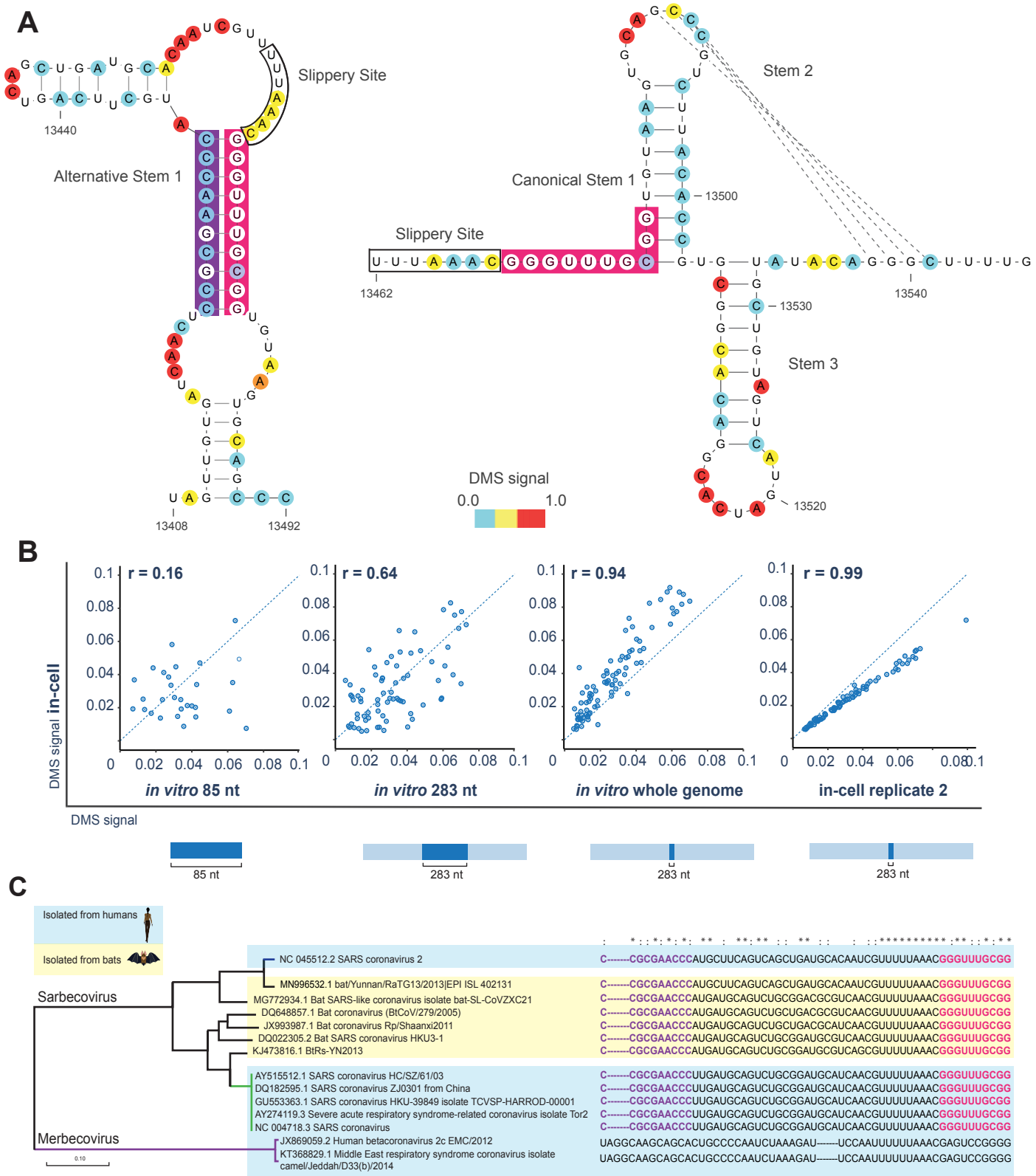


Figure 4: Frameshifting element (FSE) adopts an unexpected structure in cells. **(A)** Structural model of FSE derived from DMS-MaPseq from (left) infected VERO cells and (right) in vitro-transcribed RNA. Nucleotides are color-coded by normalized DMS signal. The 5' of the canonical stem 1 is highlighted in pink, the complement pairing (starting 46nt downstream) is shown in purple and the slippery site boxed in white. Structural model of in vitro-transcribed 85 nt FSE shown is the major cluster structure from DREEM clustering. **(B)** Scatter plots comparing FSE structures in different contexts. Comparison of DMS signals of in-cell replicate 1 with (leftmost) in vitro refolded 85 nt., (middle-left) in vitro refolded 283 nt., (middle-right) in vitro refolded whole genome, and (rightmost) in-cell replicate 2. The blue dotted line is the identity line; R is Pearson's coefficient. **(C)** Sequence conservation of FSE alternative pairing. The 5' sequence of canonical stem 1 is highlighted in pink and the complement sequence is highlighted in purple. Symbols above the sequences indicate perfect conservation among all viruses in the alignment (*) or perfect conservation among only the sarbecoviruses (:).

The prevailing model of the SARS-CoV-2 FSE is based on previous studies of the SARS-CoV FSE, as they only differ in sequence by a single nucleotide located in a putative loop (Rangan, Zheludev and Das, 2020). Nuclease mapping and Nuclear Magnetic Resonance (NMR) analysis of the SARS-CoV FSE solved the structure of an *in vitro* refolded, truncated 85 nt region starting at the slippery site (Plant *et al.*, 2005). This structure did not include the sequence upstream of the slippery site and formed a 3-stem pseudoknot.

Interestingly, *in silico* predictions of the RNA structure of the SARS-CoV-2 genome using RNAz (Rangan, Zheludev and Das, 2020) and ScanFold (Andrews *et al.*, 2020) do not find the 3-stem pseudoknot but instead support our in-cell model of Alternative Stem 1. In SARS-CoV-2, ScanFold not only predicted the AS1 but also found that it was more stable relative to random sequences than any other structure in the entire frameshift element (Andrews *et al.*, 2020). Indeed, three conceptually varied methods (DMS-MaPseq, RNAz, and ScanFold) aimed at identifying functional structures, run independently by different research groups all converge on the Alternative Stem 1 as a central structure at the FSE.

In order to directly compare our in-cell findings with the reports of the 3-stem pseudoknot, we *in vitro*-transcribed, refolded, and DMS-probed the same 85 nt sequence as analyzed by NMR (Plant *et al.*, 2005). Our in vitro-data driven model agrees well with the NMR model (87.1% identical) and finds all three canonical stems, including the pseudoknot.

FSE structure is dependent on the sequence context

The major differences we observed in the structure of the FSE in cells vs. *in vitro* could either be due to 1) length of the *in vitro* refolded viral RNA or 2) factors in the cellular environment that are absent *in vitro*. To distinguish between these two possibilities, we re-folded the FSE in the context of longer native sequences.

We found that as we increased the length of the *in vitro* re-folded construct by including more of its native sequence, from 85 nt to 283 nt to 30 kb, the DMS reactivity patterns became progressively more similar to the pattern we observed in cells (Figure 4B). Indeed, in the context of the full ~30 kb genomic RNA, the structure of the FSE is nearly identical to the structure in physiological conditions during SARS-CoV-2 infection in cells ($r = 0.94$). These results indicate that the length of the entire RNA molecule is important for correctly folding the FSE. Strikingly, at a length of 283 nt and above, the main structure forming is Alternative Stem 1 rather than the 3-stem pseudoknot. Our data indicate that given the full range of pairing possibilities in the genome, AS1 is more favorable and the predominant structure in cells.

Alternative Stem 1 pairing sequence is conserved across sarbecoviruses

To determine if other coronaviruses may have a similar alternative structure of the frameshift element, we searched for the sequence that pairs with canonical stem 1 in a set of curated coronaviruses (Ceraolo and Giorgi, 2020). This set contains 53 isolates of SARS-CoV-2, 12 other sarbecoviruses (including the SARS-CoV reference genome), and

2 merbecoviruses. The 10 nt complement (CCGCGAACCC) to a sequence overlapping canonical stem 1 of the FSE (GGGUUUGCGG) was perfectly conserved in all 12 of the sarbecoviruses, six of which were isolated from bats (Figure 4C). However, the 10 nt complement was not present in either merbecovirus. Aligning the sequences of all 20 betacoronaviruses with complete genomes in RefSeq revealed that the 10 nt complement was conserved in all of and only the three sarbecoviruses in RefSeq: SARS-CoV, SARS-CoV-2, and BtCoV BM48-31 (data not shown). These results suggest that AS1 is unique to the sarbecoviruses.

The Frameshift Element (FSE) forms two distinct structures in cells

In the past, RNA structures based on experimental data have been limited to a population average structure which assumes that every individual molecule are in folded in the same conformation. However, previous studies have identified the existence of biologically relevant alternative structures in viral genomic RNA that play vital roles in the viruses' life cycles. An example of this is the HIV-1 Rev responsive element (RRE) and 5′ untranslated region (UTR) which regulate viral RNA export and packaging. As aforementioned, the rate of frameshifting in coronaviruses is critical for the virus and set at a specific percentage. While the mechanism of how this frameshifting rate is maintained remains unidentified, studies have proposed that alternative RNA structure may play a role in promoting frameshifting.

To determine whether the SARS-CoV-2 FSE forms alternative structures, we applied the "detection of RNA folding ensembles using expectation-maximization" (DREEM) algorithm on our in-cell DMS-MaPseq data (Tomezsko *et al.*, 2020). In brief, the DREEM algorithm groups DMS reactivities into distinct clusters by considering the likelihood of the co-occurrence of reactive bases. (i.e. if two bases are highly mutated in the population average but never mutated together on a single read, we can assume that at least two conformations are present). DREEM directly clusters experimental data without relying on generating a thermodynamic based model, thus enabling the discovery of new RNA structures based purely on chemical probing data from cells.

We analyzed modified intracellular RNA of a 283 nt region surrounding the FSE specifically amplified from two biological replicates. We found two distinct patterns of DMS reactivities (Figure 5A), suggesting that the RNA folds into at least two distinct conformations in this region. In both biological replicates, clusters 1 and 2 (corresponding to Structure 1 and 2) separate at a reproducible ratio (~55% vs. 45%) where Structure 1 is significantly different from Structure 2 ($r = 0.46$) but highly similar to the corresponding cluster in biological replicates ($r = 0.99$) (Figure 5B). Both structures have the Alternative Stem 1 pairing spanning the slippery sequence. However, Structure 2 forms a large 75 nt stem immediately downstream of Alternative Stem 1 whereas Structure 1 does not (Figure 5C). Further studies measuring frameshifting efficiency are needed to determine whether this 75 nt stem specifically promotes frameshifting by stalling the ribosome right at the slippery sequence.

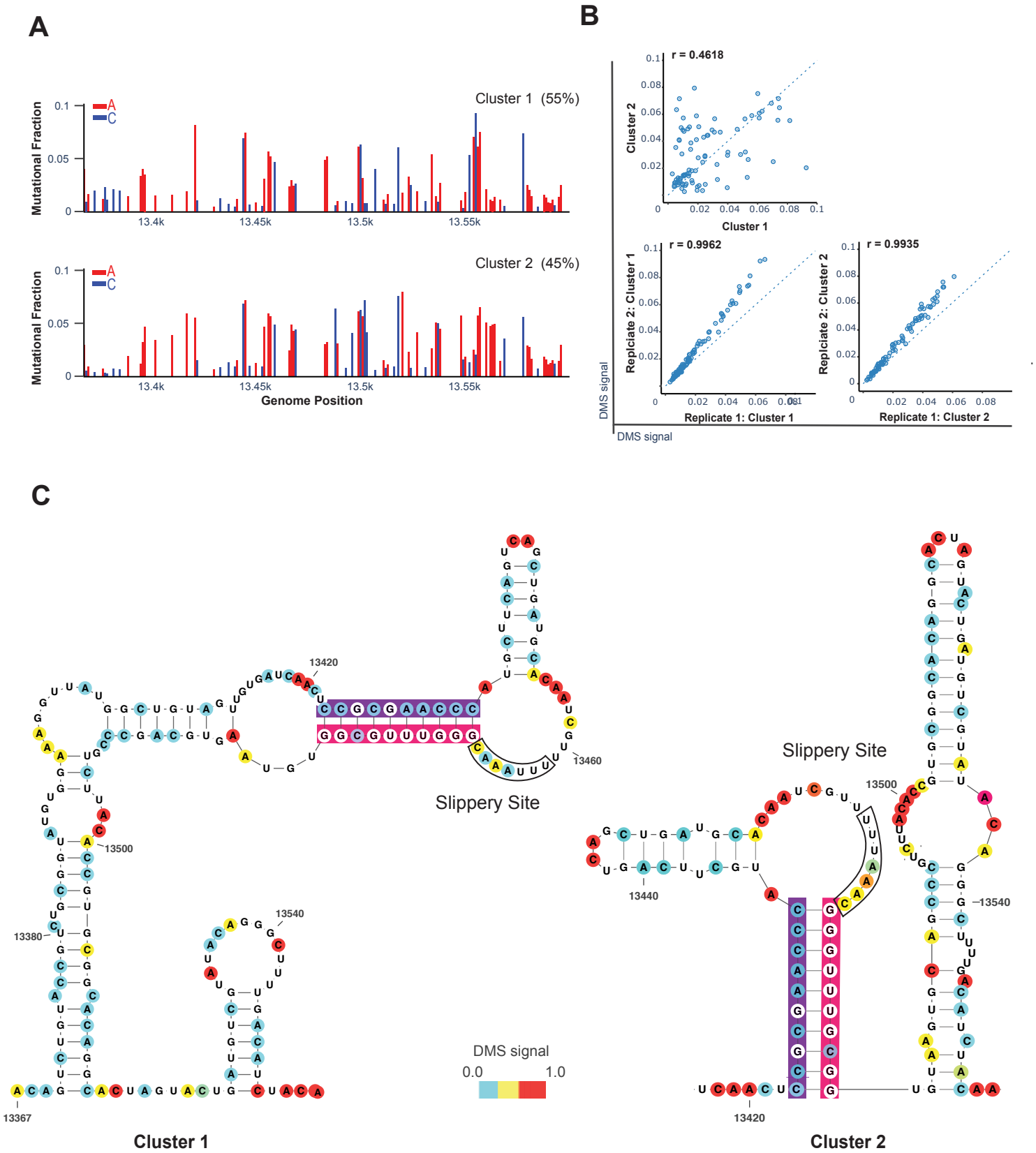


Figure 5: Alternative conformations of the frameshift element (FSE) derived from in-cell DMS-MaPseq data. **(A)** DMS signal for 283 nt surrounding the FSE structure from infected VERO cells, identified by DREEM clustering. Percentages for each cluster are determined by DREEM from representative sample of n = 2. **(B)** Comparison between clusters; (left) scatter plots of DMS signal between cluster 1 and 2 of two biological replicates; (middle, right) scatter plot of the variation in DMS signal for the same cluster between two biological replicates. The blue dotted line is the identity line; R is Pearson's coefficient. **(C)** Structural model of FSE from infected VERO cells derived from DMS-MaPseq followed by DREEM clustering. Nucleotides are color-coded by normalized DMS signal. Alternative stem 1 is highlighted, the 5' pairing complement is shown in purple, the 3' sequence (or 5' of canonical stem 1) is shown in pink and the slippery site boxed in white.

DISCUSSION

Here we present, to our knowledge, the first secondary structure of the entire RNA genome of SARS-CoV-2 in infected cells, based on chemical probing with DMS-MaPseq. Importantly, we find that many genomic regions fold differently than in silico-based predictions. We attribute these differences in large parts to the sequence context used for folding (how much native sequence is taken upstream and downstream for a given region of interest). Our data underscore the contribution of the full-length RNA molecule towards structure formation at local regions.

Of particular note, our in-cell data reveal an alternative conformation for the frameshift element (FSE) in which the 5' side of stem 1 of the canonical pseudoknot is paired to a complementary 10 nt sequence that lies shortly upstream. We find that this Alternative Stem 1 (AS1) predominates to the extent that we are not able to detect the canonical pseudoknot. In previous work (Tomezsko et al., 2020), we estimated our limit of detection for minor conformations at 6%, suggesting two potential models for frameshifting in SARS-CoV-2. In the first model, the pseudoknot structure causes frameshifting and forms at a level below our limit of detection. AS1 provides a means to downregulate the rate of frameshifting by sequestering the 5' side of the canonical stem 1. In the second model, frameshifting is stimulated by a large 75 nt stem loop immediately downstream of AS1 that blocks the ribosome and causes it to slip. An alternative structure without the 75 nt stem loop does not cause ribosome pausing at the slippery site. These alternative genomic conformations enable SARS-CoV-2 to regulate translation of ORF1ab (Figure 6). If the second model is true, then in sarbecoviruses the ability to form a pseudoknot may be vestigial, as the FSE sequence is highly conserved among coronaviruses (Plant et al., 2005).

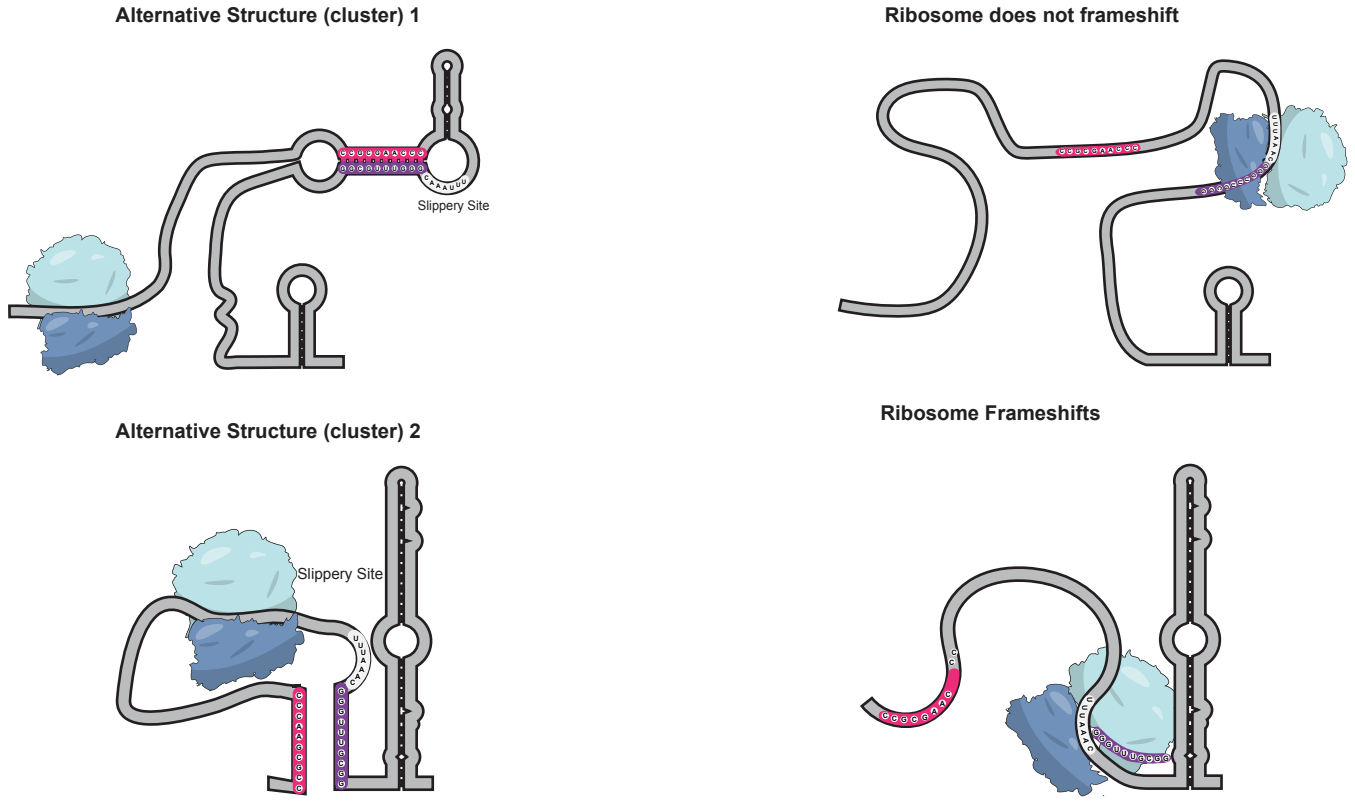


Figure 6: Model of alternative structures regulating frameshifting.
When genomic RNA folds into Alternative Structure 1 (top), the slippery site resides within a loop in the middle of a long stem-loop. As the ribosome starts to unwind the RNA, it may pause at the base of the stem, but this pause is far from the slippery site. By the time the ribosome reaches the slippery site, the structure in front of it has been unwound. As the ribosome continues it will reach the upstream stop codon and terminate translation. In contrast, Alternative Structure 2 (bottom) forms a 75 nt stem loop right in front of the slippery site. This stem loop can cause the ribosome to pause, frameshift -1 nt and bypass the upstream stop codon to continue translation.

Previous functional studies on SARS-CoV report that the pseudoknot increases frameshifting but do not contradict our model of Alternative Stem 1 (Baranov *et al.*, 2005; Plant *et al.*, 2005; Su *et al.*, 2005). Functional assays use short reporter constructs that are likely to fold into a structure with a pseudoknot, similar to what we observed for the 85 nt construct refolded *in vitro*. Indeed, a large body of literature shows that different types of pseudoknots or stable stems placed shortly after the slippery site increase frameshifting rates in reporter assays (Brierley, Digard and Inglis, 1989; Baranov *et al.*, 2005). However, much work is needed to understand which structures play a major role during active infection. To our knowledge, no studies have examined the effects of mutations and measured the translation efficiency of ORF1ab in cells infected with a sarbecovirus.

Our data reveal additional structures in the genome that may be involved in regulating gene expression. For example, previous work in MHV and SARS-CoV suggests that the N protein binds to and unwinds TRS structures to regulate sub genomic gene expression (Grossoehme *et al.*, 2009). Importantly, there is evidence that the stability of the TRS structure can affect its affinity for the N protein. All together, these results indicate that small molecules or antisense oligoes that alter the stability of the TRS structures will disrupt the expression of subgenomic RNAs and can serve as therapeutic strategies.

So far, we have not identified any structures of minus-strand RNAs. Of our 1.58 million paired reads mapping to SARS-CoV-2, only 4,955 (0.31%) came from minus-strand RNA, insufficient coverage to obtain reliable DMS signals. However, as minus-strand RNAs must be transcribed to positive-strand genomic and subgenomic RNAs, they may contain structures that are important for this process and potentially druggable. Higher-throughput genome-wide sequencing or PCR amplification of a targeted region of the minus strand would enable us to discover structures in the minus strands.

In this study we obtained structures of RNA in infected Vero cells. We were limited to studying population average RNA structures, with no temporal resolution or measure of heterogeneity. Our models of stable structures across the genome represent prominent conformers forming in cells. However, it is possible that some of the open regions we identify, which have evenly distributed DMS signal, are in fact composites of alternative structures (Tomezsko *et al.*, 2020). Deeper sequencing data is needed to determine the degree of structure heterogeneity access the coronavirus genome. Further studies will determine how RNA structures change from the time a virion enters a cell to the time when new virions are released.

We show that *in vitro* RNA-refolding of the full-length 30 kb genome can recapitulate the structures formed at the FSE in cells. These results provide evidence that some of the structures we find in cells are largely driven by intrinsic RNA thermodynamics. We expect such structures will remain the same between different cell types, although more work is needed to establish the extent of structure changes as a function of the abundance of various cellular and viral protein factors.

Our in-cell data-derived model of SARS-CoV-2 presents major RNA structures across the entire genome and provides the foundation for further studies. In particular, we propose a new model for the FSE in which the predominant structure forms an Alternative Stem 1. Future work will involve determining by what mechanism and to what extent the alternative structures of the SARS-CoV-2 FSE regulate translation of ORF1ab, as well as whether the FSE can fold into a pseudoknot in cells. Better understanding of the structures and mechanisms of elements of the SARS-CoV-2 genome will enable the design of targeted therapeutics.

REFERENCES

Andrews, R. J. *et al.* (2020) *An in silico map of the SARS-CoV-2 RNA Structurome.*, *bioRxiv : the preprint server for biology*. doi: 10.1101/2020.04.17.045161.

Baranov, P. V. *et al.* (2005) 'Programmed ribosomal frameshifting in decoding the SARS-CoV genome', *Virology*. doi: 10.1016/j.virol.2004.11.038.

Brierley, I. *et al.* (1987) 'An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV.', *The EMBO journal*. doi: 10.1002/j.1460-2075.1987.tb02713.x.

Brierley, I., Digard, P. and Inglis, S. C. (1989) 'Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot', *Cell*. doi: 10.1016/0092-8674(89)90124-4.

Ceraolo, C. and Giorgi, F. M. (2020) 'Genomic variance of the 2019-nCoV coronavirus', *Journal of Medical Virology*. doi: 10.1002/jmv.25700.

Darty, K., Denise, A. and Ponty, Y. (2009) 'VARNA: Interactive drawing and editing of the RNA secondary structure', *Bioinformatics*. doi: 10.1093/bioinformatics/btp250.

Ding, Y. (2001) 'Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond', *Nucleic Acids Research*. doi: 10.1093/nar/29.5.1034.

Dixon, W. J. (1960) 'Simplified Estimation from Censored Normal Samples', *The Annals of Mathematical Statistics*. doi: 10.1214/aoms/1177705900.

Do, C. B., Woods, D. A. and Batzoglou, S. (2006) 'CONTRAfold: RNA secondary structure prediction without physics-based models', in *Bioinformatics*. doi: 10.1093/bioinformatics/btl246.

Eckardt, S., Romby, P. and Sczakiel, G. (1997) 'Implications of RNA structure on the annealing of a potent antisense RNA directed against the human immunodeficiency virus type 1', *Biochemistry*. doi: 10.1021/bi9707234.

Edgar, R. C. (2004) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*. doi: 10.1093/nar/gkh340.

Fowlkes, E. B. and Mallows, C. L. (1983) 'A method for comparing two hierarchical clusterings', *Journal of the American Statistical Association*. doi: 10.1080/01621459.1983.10478008.

Grossoehme, N. E. *et al.* (2009) 'Coronavirus N Protein N-Terminal Domain (NTD) Specifically Binds the Transcriptional Regulatory Sequence (TRS) and Melts TRS-cTRS RNA Duplexes', *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2009.09.040.

Gruber, A. R. *et al.* (2010) 'RNAZ 2.0: Improved noncoding RNA detection', in *Pacific Symposium on Biocomputing 2010, PSB 2010*.

Hajdin, C. E. *et al.* (2013) 'Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1219988110.

Harcourt, J. *et al.* (2020) 'Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States', *Emerging Infectious Diseases*. doi: 10.3201/EID2606.200516.

Kalvari, I. *et al.* (2018) 'Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families', *Nucleic Acids Research*. doi: 10.1093/nar/gkx1038.

Kim, D. *et al.* (2020) 'The Architecture of SARS-CoV-2 Transcriptome', *Cell*. doi: 10.1016/j.cell.2020.04.011.

Lange, S. J. *et al.* (2012) 'Global or local? Predicting secondary structure and accessibility in mRNAs', *Nucleic Acids Research*. doi: 10.1093/nar/gks181.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*. doi: 10.1038/nmeth.1923.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*. doi: 10.1093/bioinformatics/btp352.

Li, L. *et al.* (2008) 'Structural Lability in Stem-Loop 1 Drives a 5′ UTR-3′ UTR Interaction in Coronavirus Replication', *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2008.01.068.

Liu, P. *et al.* (2007) 'A U-turn motif-containing stem-loop in the coronavirus 5′ untranslated region plays a functional role in replication', *RNA*. doi: 10.1261/rna.261807.

Liu, P. *et al.* (2009) 'Mouse Hepatitis Virus Stem-Loop 2 Adopts a uYNMG(U)a-Like Tetraloop Structure That Is Highly Functionally Tolerant of Base Substitutions', *Journal of Virology*. doi: 10.1128/jvi.00915-09.

Madhugiri, R. *et al.* (2014) 'RNA structure analysis of alphacoronavirus terminal genome regions', *Virus Research*. doi: 10.1016/j.virusres.2014.10.001.

Madhugiri, R. *et al.* (2018) 'Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions', *Virology*. doi: 10.1016/j.virol.2017.11.025.

Masters, P. S. (2006) 'The Molecular Biology of Coronaviruses', *Advances in Virus Research*. doi: 10.1016/S0065-3527(06)66005-3.

Mathews, D. H. (2004) 'Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization', *RNA*. doi: 10.1261/rna.7650904.

O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*. doi:

10.1093/nar/gkv1189.

Pfeiffer, F. *et al.* (2018) 'Systematic evaluation of error rates and causes in short samples in next-generation sequencing', *Scientific Reports*. doi: 10.1038/s41598-018-29325-6.

Plant, E. P. *et al.* (2005) 'A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal', *PLoS Biology*. doi: 10.1371/journal.pbio.0030172.

Plant, E. P. *et al.* (2010) 'Achieving a Golden Mean: Mechanisms by Which Coronaviruses Ensure Synthesis of the Correct Stoichiometric Ratios of Viral Proteins', *Journal of Virology*. doi: 10.1128/jvi.02480-09.

Plant, E. P. and Dinman, J. D. (2008) 'The role of programmed-1 ribosomal frameshifting in coronavirus propagation', *Frontiers in Bioscience*. doi: 10.2741/3046.

Rangan, R., Zheludev, I. N. and Das, R. (2020) 'RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look', *RNA*. doi: 10.1261/rna.076141.120.

Rouskin, S. *et al.* (2014) 'Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo', *Nature*. doi: 10.1038/nature12894.

Sayers, E. W. *et al.* (2009) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*. doi: 10.1093/nar/gkn741.

Siegfried, N. A. *et al.* (2014) 'RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP)', *Nature methods*. doi: 10.1038/nmeth.3029.

Sola, I. *et al.* (2015) 'Continuous and Discontinuous RNA Synthesis in Coronaviruses', *Annual Review of Virology*. doi: 10.1146/annurev-virology-100114-055218.

Spitale, R. C. *et al.* (2015) 'Structural imprints in vivo decode RNA regulatory mechanisms', *Nature*. doi: 10.1038/nature14263.

Su, M. C. *et al.* (2005) 'An atypical RNA pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of SARS coronavirus', *Nucleic Acids Research*. doi: 10.1093/nar/gki731.

Tomezsko, P. J. *et al.* (2020) 'Determination of RNA structural diversity and its role in HIV-1 RNA splicing', *Nature*. doi: 10.1038/s41586-020-2253-5.

Watts, J. M. *et al.* (2009) 'Architecture and secondary structure of an entire HIV-1 RNA genome', *Nature*. doi: 10.1038/nature08237.

Wilson, R. C. and Doudna, J. A. (2013) 'Molecular Mechanisms of RNA Interference', *Annual Review of Biophysics*. doi: 10.1146/annurev-biophys-083012-130404.

Yang, D. *et al.* (2011) 'Mouse Hepatitis Virus Stem-Loop 4 Functions as a Spacer Element Required To Drive Subgenomic RNA Synthesis', *Journal of Virology*. doi: 10.1128/jvi.05092-11.

Yang, D. *et al.* (2015) 'Shape analysis of the rna secondary structure of the mouse hepatitis virus 5' untranslated region and n-terminal nsp1 coding sequences', *Virology*. doi: 10.1016/j.virol.2014.11.001.

Yang, D. and Leibowitz, J. L. (2015) 'The structure and functions of coronavirus genomic

3' and 5' ends', *Virus Research*. doi: 10.1016/j.virusres.2015.02.025.

Yount, B. *et al.* (2003) 'Reverse genetics with a full-length infectious cDNA of severe acute respiratory syndrome coronavirus', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1735582100.

Zubradt, M. *et al.* (2016) 'DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo', *Nature Methods*. doi: 10.1038/nmeth.4057.

MATERIALS & METHODS

Vero cells infection and DMS modification

SARS-CoV-2 total viral RNA was extracted from Vero cells (ATCC CCL-81) cultured in DMEM (Gibco) supplemented with 10% FBS (Gibco) plated into 100 mm dishes and infected at a MOI of 0.01 with 2019-nCoV/USA-WA1/2020 (Passage 6). Infected cells were incubated at 37 °C, 5% CO_2 and harvested 2 days post infection either with or without DMS treatment. Infected cell pellets were centrifuged at 5000xg for 5 min at 4 °C and resuspended in Trizol (Ambion).

DMS modification of SARS-CoV-2 RNA in infected cells

200 µl DMS (or 2% v/v) was added dropwise to the plated Vero cells 48 h post SARS-CoV-2 infection and incubated for 4 min at 37 °C. DMS was neutralized by adding 15 ml PBS (ThermoFisher Scientific) with 30% β-mercaptoethanol. The cells were centrifuged at 1,000g for 5 min at 4 °C. The cells were washed twice by resuspending the pellet with 15 ml PBS with 30% β-mercaptoethanol and centrifugation to pellet then just once with 15 ml PBS. After washes, the pellet was resuspended in 1 ml Trizol (ThermoFisher Scientific) and RNA was extracted following the manufacturer's specifications. Total RNA was purified using RNA Clean and Concentrator -25 kit (Zymo).

DMS modification of *in vitro*-transcribed RNA

gBlocks were obtained from IDT for the SARS-CoV-2 85nt and 283nt FSE which corresponds to nucleotides 13460-13546 and nucleotides 13,342-13,624 based on 2019-nCoV/USA-WA1/2020. The regions of interest were amplified by PCR with a forward primer that contained the T7 promoter sequence (TAATACGACTCACTATAGGGTT). The PCR product was used for T7 Megascript in vitro transcription (ThermoFisher Scientific) according to manufacturer's instructions with a 16 h incubation time at 37 °C. Subsequently, 1 µl Turbo DNase I (ThermoFisher Scientific) was added to the reaction and incubated at 37 °C for 15 min. The RNA was purified using RNA Clean and Concentrator -5 kit (Zymo). 10 µg of RNA in 10 µl H_2O was denatured at 95 °C for 1 min then placed on ice. On the basis of the DMS concentration used in the next step, 300 mM sodium cacodylate buffer (Electron Microscopy Sciences) with 6 mM $MgCl_2$+ (refolding buffer) was added so that the final volume was 100 µl. (e.g. for 2.5% final DMS concentration: add 87.5 µl refolding buffer and 2.5 µl DMS) Then, 2.5 µl was added and incubated at 37 °C for 5 min while shaking at 500 r.p.m. on a thermomixer. The DMS was

neutralized by adding 60 µl β-mercaptoethanol (Millipore-Sigma). The RNA was purified using RNA Clean and Concentrator -5 kit.

DMS modification of full-length SARS-CoV-2 RNA *in vitro*

Full-length SARS-CoV-2 RNA was extracted from the supernatant of infected Vero cells (as described above), resuspended in 1 ml Trizol (ThermoFisher Scientific) and RNA was extracted following the manufacturer's specifications. The RNA was purified using RNA Clean and Concentrator -5 kit (Zymo) and DMS modified as described above.

Human rRNA subtraction of total cellular RNA

15 µg of total RNA per reaction was used as the input for rRNA subtraction. First, 1 µl rRNA subtraction mix (15 µg/µl) and 2 µl 5× hybridization buffer (end concentration: 200 mM NaCl, 100 mM Tris-HCl, pH 7.4) were added to each reaction, and final volume was then adjusted with water to 10 µl. The samples were denatured at 95 °C for 2 min and then temperature was reduced by 0.1 °C/s until the reaction was at 45 °C. Next, 10 µl RNase H buffer and 2 µl hybridase thermostable RNase H (Lucigen) preheated to 45 ° were added. The samples were incubated at 45 °C for 30 min. The RNA was cleaned with RNA Clean and Concentrator -5, following the manufacturer's instructions and eluted in 45 µl water. Then, 5 µl Turbo DNase buffer and 3 µl Turbo DNase (ThermoFisher Scientific) were added to each reaction and incubated for 30 min at 37 °C. The RNA was purified with RNA Clean and Concentrator -5 (Zymo) following instructions.

RT–PCR and sequencing of DMS-modified RNA

For reverse transcription, 1.5 µg of rRNA subtracted total RNA or 10 µg of in vitro-transcribed RNA was added to 4 µl 5× first strand buffer (ThermoFisher Scientific), 1 µl 10µM reverse primer, 1 µl dNTP, 1 µl 0.1M DTT, 1 µl RNaseOUT and 1 µl TGIRT-III (Ingex). The reverse-transcription reaction was incubated at 60 °C for 1.5 h. 1 µl 4M NaOH was then added and incubated at 95 °C for 3 min to degrade the RNA. The cDNA was purified with Oligo Clean and Concentrator -5 (Zymo) following instructions. PCR amplification was done using Advantage HF 2 DNA polymerase (Takara) for 30 cycles according to the manufacturer's specifications. The PCR product was purified by DNA Clean and Concentrator -5 (Zymo) following manufacturer's instructions. RNA-seq library for 150 bp insert size was constructed following the manufacturer's instruction (NEBNext Ultra™ II DNA Library Prep Kit). The library was loaded on ISEQ-100 Sequencing flow cell with ISEQ-100 High-throughput Sequencing Kit and the library was run on ISEQ-100 (paired-end run,151 x 151 cycles).

Library generation with DMS-modified SARS-CoV-2 RNA

After rRNA subtraction (described above), extracted DMS-modified RNA from SARS-CoV-2 infected Vero cells was fragmented using the RNA Fragmentation kit (ThermoFisher Scientific). 1.5 µg of rRNA subtracted total RNA was fragmented at 70 °C for 2.5 min. The fragmented RNA was mixed with an equal volume 2× Novex TBE-urea

sample buffer (ThermoFisher Scientific) and run on a 10% TBE-urea gel (ThermoFisher Scientific) at 200V for 1 h 15 min for size selection of RNA that is ~150nt. To dephosphorylate and repair the ends of randomly fragmented RNA, 2 µl 10x CutSmart buffer (New England Biolabs), 10 µl shrimp alkaline phosphatase (New England Biolabs), 2 µl RNaseOUT (ThermoFisher Scientific) and water were added to a final volume of 20 µl and 37 °C for 1 h. Next, 4 µl 50% PEG-800 (New England Biolabs), 4 µl 10× T4 RNA ligase buffer (New England Biolabs), 4 µl T4 RNA ligase, truncated KQ (England Biolabs) and 2 µl linker were added to the reaction and incubated for 18 h at 22 °C. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 10 µl water. Excess linker was degraded by adding 2 µl 10× RecJ buffer (Lucigen), 1 µl RecJ exonuclease (Lucigen), 1 µl 5′ deadenylase (New England Biolabs) and 1 µl RNaseOUT, then incubating for 1 h at 30 °C. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions and eluted in 11 µl water.

For reverse transcription, 1.5 µg of rRNA subtracted total RNA or 10 µg of in vitro-transcribed RNA was added to 4 µl 5× first strand buffer (ThermoFisher Scientific), 1 µl 10µM reverse primer, 1 µl dNTP, 1 µl 0.1M DTT, 1 µl RNaseOUT and 1 µl TGIRT-III (Ingex). The reverse-transcription reaction was incubated at 60 °C for 1.5 h. 1 µl 4M NaOH was then added and incubated at 95 °C for 3 min to degrade the RNA. The reverse-transcription product was mixed with an equal volume 2× Novex TBE-urea sample buffer (ThermoFisher Scientific) and run on a 10% TBE-urea gel (ThermoFisher Scientific) at 200V for 1 h 15 min for size selection of cDNA that is ~250nt. The size-selected and purified cDNA was circularized using CircLigase ssDNA ligase kit (Lucigen) following manufacture's protocol. 2 µl of the circularized product was then used for PCR amplification using Phusion High-Fidelity DNA Polymerase (NEB) for a maximum of 16 cycles. The PCR product was run on an 8% TBE gel at 180V for 1 h and size-selected for products ~300 nt. The product was then sequenced with iSeq100 (Illumina) to produce either 150×150-nt paired-end reads.

Mapping and quantification of mutations

Fastq files were trimmed using TrimGalore (github.com/FelixKrueger/TrimGalore) to remove Illumina adapters. Trimmed paired reads were mapped to the genome of SARS-CoV-2 isolate SARS-CoV-2/human/USA/USA-WA1/2020 (GenBank: MN985325.1) (Harcourt *et al.*, 2020) using Bowtie2 (Langmead and Salzberg, 2012) with the following parameters: --local --no-unal --no-discordant --no-mixed -L 12 -X 1000. Reads aligning equally well to more than one location were discarded. SAM files from Bowtie2 were converted into BAM files using Picard Tools SamFormatConverter (broadinstitute.github.io/picard).

For each pair of aligned reads, a bit vector the length of the reference sequence was generated using DREEM (Tomezsko *et al.*, 2020). Bit vectors contained a 0 at every position in the reference sequence where the reference sequence matched the read, a 1 at every base at which there was a mismatch or deletion in the read, and no information for every base that was either not in the read or had a Phred score <20. We refer to

positions in a bit vector with a 0 or 1 as “informative bits” and all other positions as “uninformative bits.”

For each position in the reference sequence, the number of bit vectors covering the position and the number of reads with mismatches and deletions at the position were counted using DREEM. The ratio of mismatches plus deletions to total coverage at each position was calculated to obtain the population average mutation rate for each position.

Filtering bit vectors

In cases indicated below, bit vectors were discarded if they had two mutations closer than 4 bases apart, had a mutation next to an uninformative bit, or had more than an allowed total number of mutations (greater than 10% of the length of the bit vector and greater than three standard deviations above the median number of mutations among all bit vectors). The average mutation rate for each position was computed from the filtered bit vectors in the same way as described above.

Normalizing the mutation rates

The mutation rates for all of the bases in the RNA molecule were sorted in numerical order. The greatest 5% or 10% of mutation rates (specified where relevant in the main text) were chosen for normalization. The median among these signals was calculated. All mutation rates were divided by this median to compute the normalized mutation rates. Normalized rates greater than 1.0 were winsorized by setting them to 1.0 (Dixon, 1960).

Computing genome coverage and mutation rates

Genome-wide coverage (Figure 1C) was computed by counting the number of unfiltered bit vectors from the in-cell library that contained an informative bit (0 or 1) at each position. Signal and noise plots (Figure 1D) were generated from the unfiltered population average mutation rate. A total of 103 (0.34%) positions across the genome were discarded for having a noise mutation rate greater than 1% in the untreated sample (likely due to endogenous modifications or “hotspot” reverse transcription errors). The signal and noise were computed every 100 nt, starting at nucleotide 51. For each of these nucleotides, the average mutation rate was computed over the 100 nt window starting 50 bases upstream and ending 49 bases downstream. The “signal” was defined as the average mutation rate of A and C, while the “noise” was defined as the average mutation rate of G and U.

The correlation of mutation rates between biological replicates genome-wide (Figure 1B) was computed using the unfiltered bit vectors. The correlation of mutation rates between different conditions of the FSE (Figure 4B) was computed using the filtered bit vectors. The correlation of mutation rates between clusters and biological replicates for the FSE (Figure 5B) was computed using the filtered bit vectors after clustering into two clusters. For all correlation plots, the Pearson correlation coefficient is given. A total of 6 (0.02%) outliers with >30% mutation rate were removed to prevent inflating the Pearson correlation coefficients.

Folding the entire SARS-CoV-2 genome

The unfiltered population average mutation rate was obtained from the in-cell library reads. The 29,882 nt genome of SARS-CoV-2 was divided into ten segments, each roughly 3 kb the boundaries of which are predicted to be open and accessible by RNAz (Rangan, Zheludev and Das, 2020). For each segment, the population average mutation rate was normalized. The segment was then folded using the Fold algorithm from RNAstructure (Mathews, 2004) with parameters -m 3 to generate the top three structures, -md to specify a maximum base pair distance, and -dms to use the normalized mutation rates as constraints in folding. All mutation rates on G and U bases were set to -999 (unavailable constraints). Connectivity Table files output from Fold were converted to dot bracket format using ct2dot from RNAstructure (Mathews, 2004). The ten dot bracket structures were concatenated into a single genome-wide structure.

The modified Fowlkes-Mallows index

Given two RNA structures of the same length (*L*) in dot-bracket notation, all base pairs in each structure were identified. Each base pair was represented as a tuple of (position of 5' base, position of 3' base). The number of base pairs common to both structures (P_{12}) as well as the number of base pairs unique to the first structure (P_1) and to the second structure (P_2) were computed. Given these quantities, the Fowlkes-Mallows index (a measure of similarity between two binary classifiers) is defined as FMI = $P_{12}/\sqrt{(P_{12}+P_1)(P_{12}+P_2)}$ (Fowlkes and Mallows, 1983). In the case that $(P_{12}+P_1)(P_{12}+P_2)=0$, we let FMI = 0.

As the Fowlkes-Mallows index does not consider positions at which the structures agree on bases that are unpaired, the index needed to be modified; otherwise regions with few base pairs would tend to score too low. Thus, the number of positions at which both sequences contained an unpaired base (*U*) was computed. Two variations of the modified Fowlkes-Mallows index (mFMI) were tested that differed in their treatment of externally paired bases, defined as bases paired to another base outside of the region of the structure being compared. The version of mFMI excluding external base pairs counted all externally paired bases as unpaired when computing *U*. The number of positions containing a paired base (*P*) was computed as $P = L - U$. In this case, mFMI was defined as mFMI = $U/L + P/L \times \mathrm{FMI}$, which weights the Fowlkes-Mallows index by the fraction of paired bases and adds the fraction of unpaired bases (U/L), as the structures agree at all unpaired positions.

To include external base pairs, any position containing an externally paired base was not counted in *U*. The number of positions at which both structures contained an externally paired base with the same orientation (i.e. both facing in the 5' or 3' direction) was computed as the number *E*. The number of positions at which at least one structure contained a base that was paired, but not externally, was computed as *P*. Then, the mFMI was defined as mFMI = $U/L + E/L + P/L \times \mathrm{FMI}$, which weights the Fowlkes-Mallows index by the fraction of positions containing a paired base and considers positions in

which both bases are unpaired as in agreement, but only counts externally paired bases as agreeing if both structures contain an externally paired base at the same position and the base pairs have the same orientation.

Comparisons to previous *in silico* predictions

Excel files from the supplemental material of (Rangan, Zheludev and Das, 2020) were parsed to obtain the coordinates and predicted structures. For each predicted structure, agreement with the region of our structure with the same coordinates was computed using the mFMI, either including or excluding external base pairs (as specified in the text). Box plots of the agreement for each window (Figure 3B) show the minimum, first quartile, median, third quartile, and maximum; data lying more than 1.5 times the interquartile range from the nearest quartile are considered outliers and are plotted as individual points. The numbers of points in each box plot are given in the Results section for Figure 3B.

Folding the frameshift element

Reads from RT-PCR of a 283 nt segment of in-cell RNA spanning the FSE (nucleotides 13,342 - 13,624) were used to generate bit vectors. The bit vectors were filtered as described above, and the filtered average mutation rates were normalized. The RNA was folded using the ShapeKnots algorithm from RNAstructure (Hajdin *et al.*, 2013) with parameters -m 3 to generate three structures and -dms to use the normalized mutation rates as constraints in folding. All signals on G and U bases were set to -999 (unavailable constraints). Connectivity Table files output from ShapeKnots were converted to dot bracket format using ct2dot from RNAstructure (Mathews, 2004).

Coronavirus sequence alignments

Accession numbers of curated sarbecovirus and merbecovrus genomes were obtained from (Ceraolo and Giorgi, 2020) and downloaded from NCBI. The sequences were aligned using the MUSCLE (Edgar, 2004) web service with default parameters. The region of the multiple sequence alignment spanning the two sides of Alternative Stem 1 was located and the sequence conservation computed using custom Python scripts.

For the alignment of all betacoronaviruses with genomes in NCBI RefSeq (O'Leary *et al.*, 2016), all reference genomes of betacoronaviruses were downloaded from RefSeq using the query "betacoronavirus[organism] AND complete genome" with the RefSeq source database as a filter. The sequences were aligned using the MUSCLE (Edgar, 2004) web service with default parameters. The subgenus of betacoronavirus to which each virus belonged was obtained from the NCBI taxonomy database (Sayers *et al.*, 2009).

Detecting alternative structures of the FSE

The filtered bit vectors (the same used to fold the frameshift element) were clustered using the expectation maximization algorithm of DREEM to allow detection of a maximum of two alternative structures (Tomezsko *et al.*, 2020).

Quantification of minus-strand reads

Mapped reads from the in-cell library were classified as minus-strand using a custom Python script if they had the following SAM flags (Li *et al.*, 2009): PAIRED and PROPER_PAIR and ({READ1 and MREVERSE and not REVERSE} or {READ2 and REVERSE and not MREVERSE}) and not (UNMAP or MUNMAP or SECONDARY or QCFAIL or DUP or SUPPLEMENTARY).

Visualizing RNA structures

RNA structures were drawn using VARNA (Darty, Denise and Ponty, 2009). The bases were colored using the normalized DMS signals.

ACKNOWLEDGMENTS

The SARS-CoV-2 starting material was provided by the World Reference Center for Emerging Viruses and Arboviruses (WRCEVA), with Natalie Thornburg (nax3@cdc.gov) as the CDC Principal Investigator." We thank T.B. Faust for manuscript input and E. Smith for illustrator images. This work was supported by The Office of Naval Research Award # N00014-20-1-2084 and the Burroughs Wellcome fund.

AUTHOR CONTRIBUTIONS

T.C.T.L. and S.R. conceived and designed the project. T.C.T.L. carried out all experiments with collaborative contributions from L.E.M. and A.G. M.F.A., T.C.T.L., S.K., S.S.Y.N. and S.R. performed the data analysis. T.C.T.L., M.F.A., and S.R. interpreted the results and wrote the paper with input from S.K., S.S.Y.N., L.E.M, M.B., and A.G.

DECLARATION OF INTERESTS

The authors declare no competing interests