Discovery and Quantification of Long-Range RNA Base Pairs in Coronavirus Genomes with SEARCH-MaP and SEISMIC-RNA

Matthew F. Allan, Justin Aruda, Yves Martin, Scott Grote,
Alberic de Lajarte, Jesse Plung, Mateo Valenzuela,
Mark Bathe, Daniel Herschlag, and Silvi Rouskin

February 21, 2024

Introduction

Across all domains of life, RNA molecules perform myriad functions in development **?**, immunity **?**, translation **?**, sensing **??**, epigenetics **?**, cancer **?**, and more. RNA also constitutes the genomes of many threatening viruses **?**, including influenza viruses **?** and coronaviruses **?**. The capabilities of an RNA molecule depend not only on its sequence (primary structure) but also on its base pairs (secondary structure) and three-dimensional shape (tertiary structure) **?**.

Although high-quality tertiary structures provide the most information, resolving them often proves difficult or impossible with mainstay methods used for proteins **?**. Consequently, the world's largest database of tertiary structures – the Protein Data Bank **?** – has accumulated only 1,839 structures of RNAs (compared to 198,506 of proteins) as of February 2024. Worse, most of those RNAs are short: only 119 are longer than 200 nt; of those, only 24 are not ribosomal RNAs or group I/II introns. Due partly to the paucity of non-redundant long RNA structures, methods of predicting tertiary structures for RNAs lag far behind those for proteins **?**.

The situation is only marginally better for RNA secondary structures. If a diverse set of homologous RNA sequences is available, a consensus secondary structure can often be predicted using comparative sequence analysis, which has accurately modeled ribosomal and transfer RNAs, among others **?**. A formalization known as the covariance model **?** underlies the widely-used Rfam database **?** of consensus secondary structures for 4,170 RNA families (as of version 14.10). Although extensive, Rfam contains no protein-coding sequences (with some exceptions such as frameshift stimulating elements) and provides only one secondary structure for each family, even though many RNAs fold into multiple functional structures **??**). Each family also models only a short segment of a full RNA sequence; for coronaviruses, existing families encompass the 5' and 3' untranslated regions, the frameshift stimulating element, and the packaging signal, which collectively constitute only 3% of the genomic RNA.

Predicting secondary structures faces two major obstacles due to the scarcity of high-quality RNA structures, particularly for RNAs longer than 200 nt (including long non-coding **?**, messenger **?**, and viral genomic **?** RNAs). First, prediction methods trained on known RNA structures are limited to small, low-diversity training datasets (generally of short sequences), which causes overfitting and hence inaccurate predictions for dissimilar RNAs (including longer sequences) **??**. Second, without known secondary structures of many diverse RNAs, the accuracy of any prediction method cannot be properly benchmarked **??**. For these reasons, and because thermodynamic-based models also tend to be less accurate for longer RNAs **?** and base pairs spanning longer distances **?**, predicting secondary structures of long RNAs remains unreliable.

The most promising methods for determining the structures of long RNAs use experimental data. Chemical probing experiments involve treating RNA with reagents that modify nucleotides depending on the local secondary structure; for instance, dimethyl sulfate (DMS) methylates adenosine (A) and cytidine (C) residues only if they are not base-paired **?**. Modern methods use reverse transcription to encode modifications of the RNA as mutations in the cDNA, followed by next-generation sequencing – a strategy known as mutational profiling (MaP) **??**. A key advantage of MaP is that the sequencing reads can be clustered to detect multiple secondary structures in an ensemble **??**. Determining the base pairs in those structures still requires structure prediction **?**, although incorporating chemical probing data does improve accuracy **??**.

Several experimental methods have been developed to find base pairs directly, with minimal reliance on structure prediction. M2-seq **?** introduces random mutations before chemical probing to detect correlated mutations between pairs of bases, which indicates the bases interact. However, alternative structures complicate the data analysis **?**, and detectible base pairs can be no longer than the sequencing reads (typically 300 nt). For long-range base pairs, many methods involving crosslinking, proximity ligation, and sequencing have been developed **?**. These methods can find base pairs spanning arbitrarily long distances – as well as between different RNA

molecules – but cannot resolve single base pairs or alternative structures. Detecting, resolving, and quantifying alternative structures with base pairs that span arbitrarily long distances remains an open challenge.

Here, we introduce "Structure Ensemble Ablation by Reverse Complement Hybridization with Mutational Profiling" (SEARCH-MaP), an experimental method to discover RNA base pairs spanning arbitrarily long distances. We also develop the software "Structure Ensemble Inference by Sequencing, Mutation Identification, and Clustering of RNA" (SEISMIC-RNA) to analyze MaP data and resolve alternative structures. Using SEARCH-MaP and SEISMIC-RNA, we discover an RNA structure in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that comprises dozens of long-range base pairs and folds in nearly half of genomic RNA molecules. We show that it inhibits the folding of a pseudoknot that stimulates ribosomal frameshifting **??**, hinting a role in regulating viral protein synthesis. We find similar structures in other SARS-related viruses and transmissible gastroenteritis virus (TGEV), suggesting that long-range base pairs involving the frameshift stimulation element are a general feature of coronaviruses. In addition to revealing new structures in coronaviral genomes, our findings show how SEARCH-MaP and SEISMIC-RNA can resolve secondary structure ensembles of long RNA molecules – a necessary step towards a true "AlphaFold for RNA" **?**.

Results

Workflow of SEARCH-MaP and SEISMIC-RNA

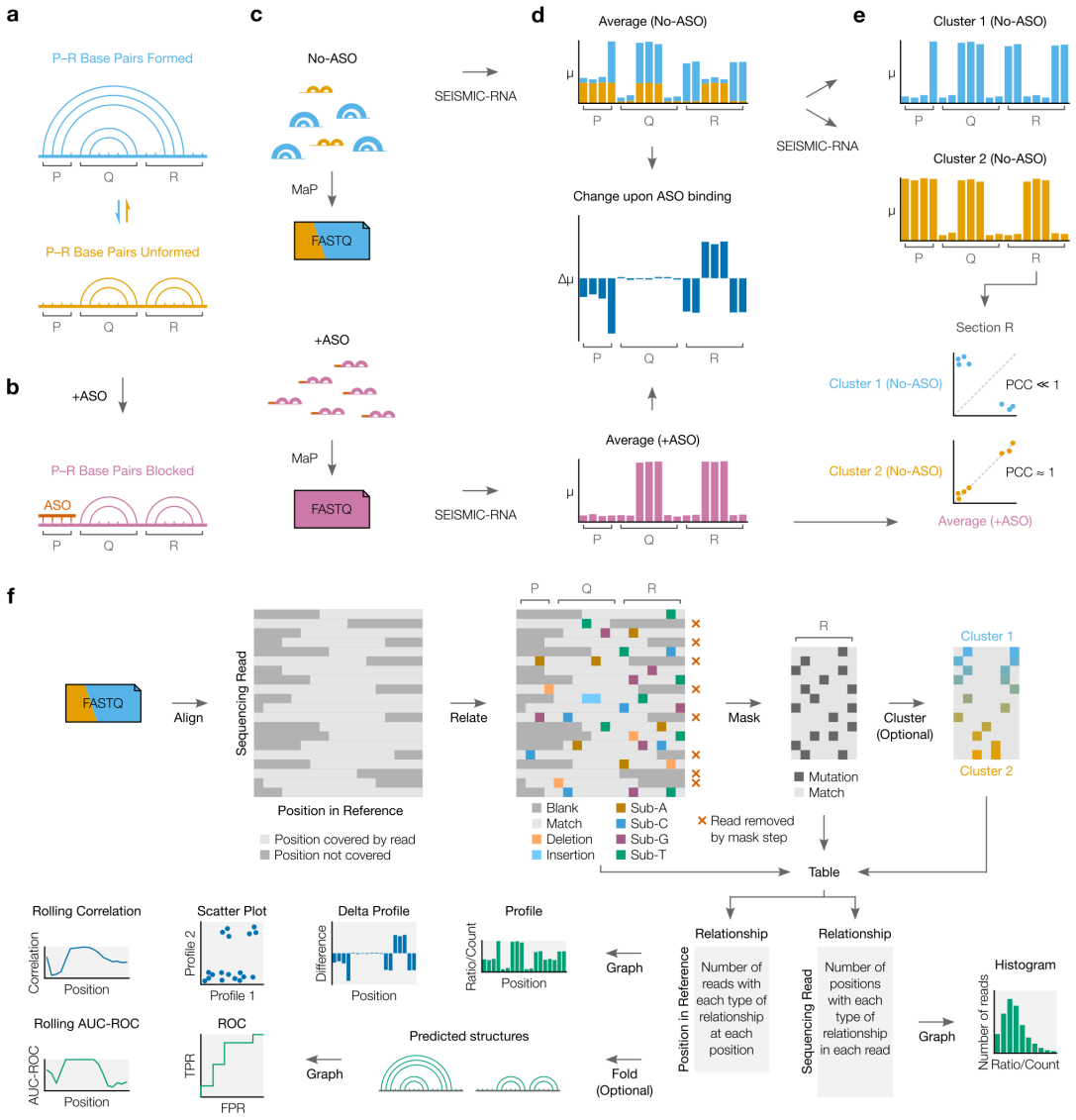


Figure 1: **The workflow of SEARCH-MaP and SEISMIC-RNA.** (Continued on next page.)

Figure 1: (Continued from previous page.) **(a)** This toy RNA is partitioned into three sections (P, Q, and R) and folds into an ensemble of two structures: one in which base pairs between P and R form and one in which they do not. **(b)** Hybridizing an ASO to P blocks it from base-pairing with R. **(c)** A SEARCH-MaP experiment entails separate chemical probing and mutational profiling (MaP) with (+ASO) and without (no-ASO) the ASO, followed by sequencing to generate FASTQ files. The RNA molecules and FASTQ files use the same color scheme as in (a) and are illustrated/colored in proportion to their abundances in the ensemble. **(d)** Mutational profiles with (+ASO) and without (no-ASO) the ASO, computed as ensemble averages with SEISMIC-RNA. The x-axis is the position in the RNA sequence; the y-axis is the fraction of mutated bases (μ) at the position. Each bar in the no-ASO profile is drawn in two colors merely to illustrate how many mutations at each position come from each structure; in a real experiment, this information would not exist before clustering. The change upon ASO binding indicates the difference in the fraction of mutated bases ($\Delta\mu$) between the +ASO and no-ASO conditions. **(e)** Mutational profiles of two clusters (top) obtained by clustering the no-ASO ensemble in (d) using SEISMIC-RNA, and scatter plots comparing the mutational profiles (bottom) between the +ASO ensemble average (x-axis) and each cluster (y-axis); each point represents one base in section R. The expected Pearson correlation coefficient (PCC) is shown beside each scatter plot. **(f)** The workflow of SEISMIC-RNA. First, sequencing reads (in FASTQ files) are aligned to reference sequence(s). For every read, the relationship to each base in the reference sequence (i.e. match, substitution, deletion, insertion) is determined. In the next step, relationships are called as mutated, matched, or uninformative; and positions and reads failing to meet certain criteria are masked out. Optionally, masked reads can be clustered to reveal alternative structures. The types of relationships at each position and in each read are then counted and tabulated. SEISMIC-RNA can use these tables to predict RNA secondary structures or draw a variety of graphs including mutational profiles, scatter plots, and receiver operating characteristic (ROC) curves.

We illustrate SEARCH-MaP with an RNA comprising three sections (P, Q, and R) that folds into an ensemble of two structures: one in which base pairs between P and R form and one in which they do not (Figure **??**a). Searching for base pairs involving section P begins by blocking P with an antisense oligonucleotide (ASO), which ablates the base pairs between P and R (Figure **??**b). The RNA is chemically probed separately with (+ASO) and without (no-ASO) the ASO, followed by mutational profiling (MaP) and sequencing, e.g. using DMS-MaPseq **?** (Figure **??**c).

SEISMIC-RNA can detect base pairs by comparing the +ASO and no-ASO mutational profiles. Theoretically, each structure has its own mutational profile **?**, but the mutational profile of a single structure is not directly observable because all structures are physically mixed during the experiment (Figure **??**c, top). Instead, the directly observable mutational profile is of the "ensemble average" – the average

of the structures' (unobserved) mutational profiles, weighted by the their (unobserved) proportions (Figure **??**d, top). Because the mutational profile of section R changes when it base-pairs with P, the ensemble averages of R differ between the +ASO and no-ASO conditions (Figure **??**d, middle). However, the ASO has little effect on section Q because this section does not base-pair with P (Figure **??**d, middle). Therefore, one can deduce that P interacts with R – but not with Q – because hybridizing an ASO to P alters the mutational profile of R but not of Q.

Going one step further, one can resolve the mutational profile where P and R base-pair, even without knowing the exact base pairs. This step uses SEISMIC-RNA to cluster the no-ASO ensemble into two mutational profiles over section R – each corresponding to one structure – and comparing them to the +ASO ensemble average (Figure **??**e). Because the ASO blocks the P–R base pairs, the +ASO mutational profile will correlate better with that of the structure where P and R do not base-pair; in this case, cluster 2 correlates better. Therefore, the mutational profile of cluster 1 corresponds to the structure where P and R base-pair.

SEARCH-MaP finds base pairs in ribosomal RNA

We first validated SEARCH-MaP using 23S ribosomal RNA (rRNA) from *E. coli*. We obtained ground truth structure models from the Comparative RNA Web **?** and selected two known stems. For each stem, we designed two ASOs, one targeting each side.

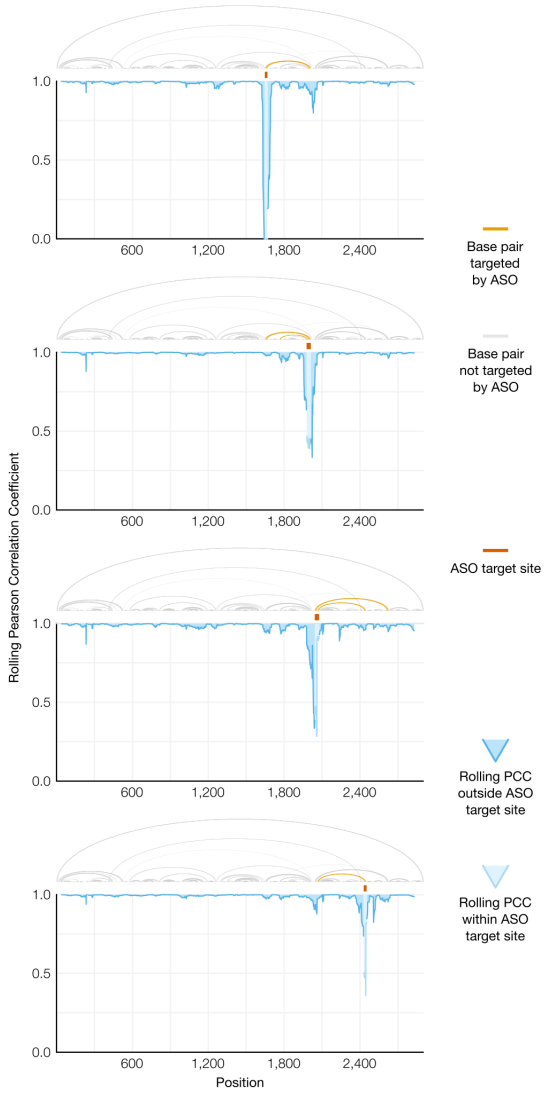


Figure 2: **Validation of SEARCH-MaP on 23S ribosomal RNA from *E. coli*.** Each graph shows the rolling (window = 45 nt) Pearson correlation coefficient (PCC) between the rRNA to which one ASO was added and a no-ASO control. The known secondary structure of the 23S rRNA **?** is drawn above the graph; base pairs with one side within the ASO target site are highlighted.

We folded the 23S rRNA with each ASO, performed DMS-MaPseq over the entire transcripts, and compared ensemble average mutational profiles with and without ASOs using SEISMIC-RNA (Figure **??**). Every ASO caused a prominent dip in the rolling Pearson correlation coefficient (PCC) at its target site and the immediate vicinity, confirming that each ASO bound properly to the RNA. The ASO targeting positions 1,647-1,668 also caused a smaller dip in PCC around positions 1,987-2,045 – coinciding with the 3' side of a stem whose 5' side was targeted by the ASO – showing that this stem could be detected with SEARCH-MaP. Conversely, targeting

the 3' side of this stem with an ASO binding positions 1,978-2,010 caused a small – though still above-baseline – dip in PCC around position 1,670 (near the stem's 5' side) and another around position 1,800 (the 5' side of another stem targeted by the ASO). Shifting the ASO slightly downstream to target positions 2,042-2,076 maintained the dips in PCC around 1,670 and 1,800 while introducing new dips around positions 2,245, 2,435, and 2,630, which correspond to the 3' sides of three stems within or close to the ASO target site. Binding an ASO to 2,429-2,452 – the 3' side of one such stem – caused the PCC to dip around position 2,055 at the 5' end of this stem. These results show that SEARCH-MaP could detect multiple stems ranging from 200 to 600 nt in 23S rRNA from *E. coli*.

SEARCH-MaP detects and quantifies long-range base pairing in SARS-CoV-2

Aside from ribosomes, many of the best-characterized functional long-range RNA base pairs occur in the genomes of RNA viruses **?**. Coronaviruses regulate translation of their first open reading frame (ORF1) using programmed ribosomal frameshifting **?**. In the middle of ORF1, a switch called a frameshift stimulation element (FSE) makes a fraction of ribosomes slip backwards into the -1 reading frame. Ribosomes that maintain reading frame terminate at a stop codon shortly after the FSE, while those that frameshift bypass that stop codon and reach the end of ORF1. Why coronaviruses need a frameshifting mechanism remains an open question **?**, yet all have FSEs **?**.

Every coronaviral FSE contains a "slippery site" (UUUAAAC) and a structure characterized as a pseudoknot in multiple species **???**. Indeed, the isolated core of the FSE in SARS-CoV-2 was shown to fold into a pseudoknot with three stems **???**. However, we discovered that when FSE is in its natural place in the SARS-CoV-2 genome, pseudoknot stem 1 is disassembled while an alternative stem 1 folds **?**. A 283 nt segment of the RNA genome – containing both the FSE and alternative stem 1 – failed to fully mimic the DMS reactivities of the full virus (PCC = 0.75). A

2,924 nt segment came closer (PCC = 0.93), suggesting that – only in the context of this longer sequence – the FSE adopts yet another structure, presumably long-range base-pairing **?**.

We used SEARCH-MaP and SEISMIC-RNA to find the long-range base pairs formed by the FSE. We hypothesized they would match a structure another group had discovered and named the "FSE-arch" **?**. If so, the structure of the FSE would be perturbed by – and only by – ASOs targeting either side of the putative FSE-arch. To investigate, we added (separately) thirteen groups of DNA ASOs to the 2,924 nt segment (Figure **??**a). Each group contained four or five ASOs targeting a contiguous 213-244 nt section of the RNA; target sites of adjacent groups abutted without overlapping. After adding each group of ASOs, we performed DMS-MaPseq **?** with two pairs of RT-PCR primers: flanking the ASO target site (to confirm binding) and flanking the 5' FSE-arch (to detect structural changes). We obtained data for every ASO group except 13 (Supplementary Figure **??**). All ASO groups bound properly, evidenced by suppression of DMS reactivities over their target sites (Supplementary Figure **??**).

To quantify structural changes over the 5' FSE-arch, we calculated the rolling Pearson correlation coefficient (PCC) of the DMS reactivities between each sample and a no-ASO control (Figure **??**b). The rolling PCC of a no-ASO replicate remained between 0.93 and 1.00 (mean = 0.97), confirming the DMS reactivities were reproducible. ASO group 9 – targeting both 3' inner stems of the FSE-arch – caused the rolling PCC to dip below 0.5 over both 5' inner stems, exactly as expected if the inner stems of the FSE-arch existed. The only other ASO groups with substantial effects were 3, 4, and 5, which overlapped or abutted the FSE and presumably perturbed short-range base pairs; the outer stem of the FSE-arch (targeted by ASO group 10) did not apparently form. These results suggest both inner stems of the FSE-arch exist and are the predominant long-range base pairs involving the immediate vicinity of the FSE.

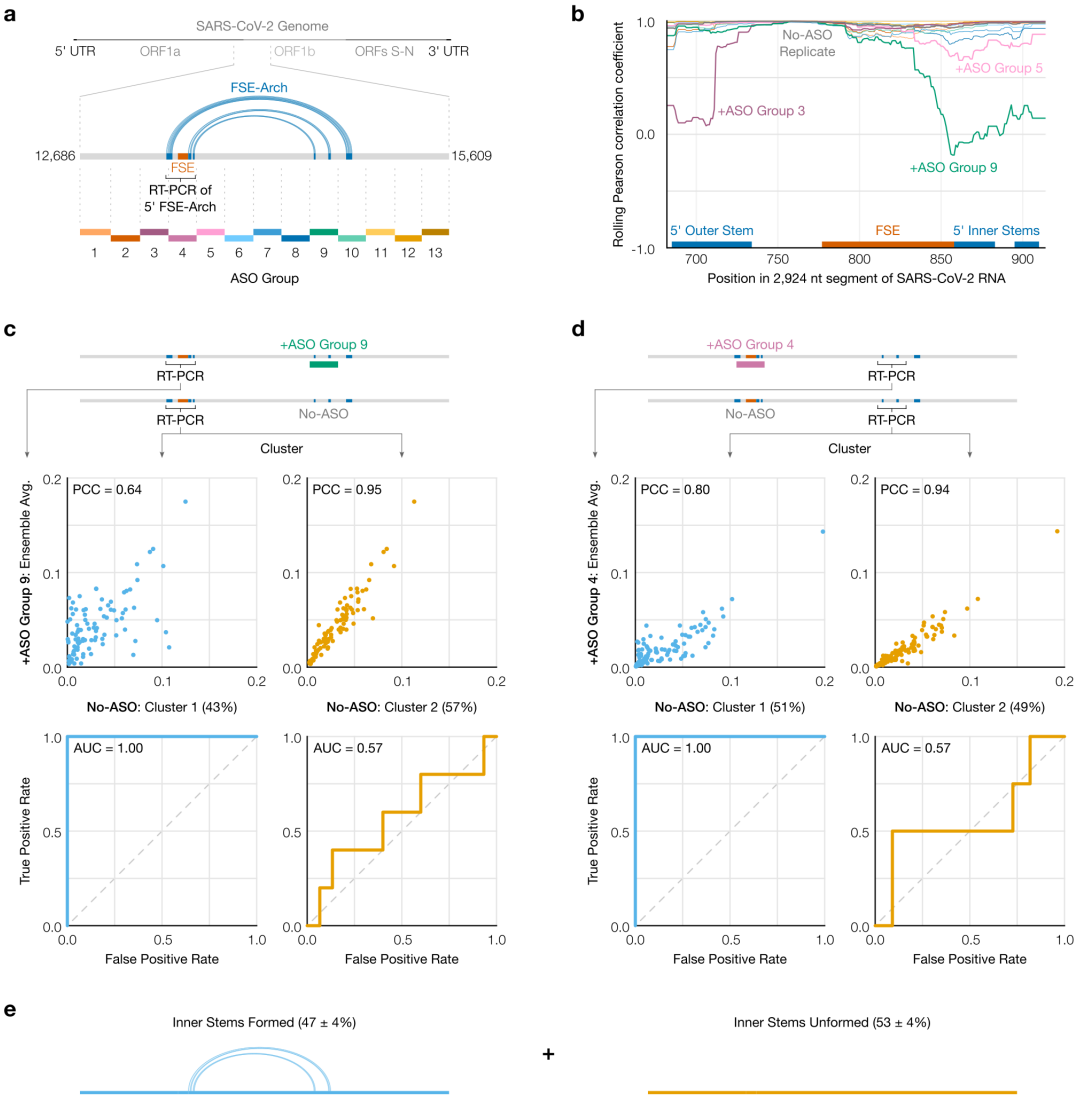


Figure 3: **Search for a long-range base pairs involving the SARS-CoV-2 FSE. (a)** The 2,924 nt segment of the SARS-CoV-2 genome containing the frameshift stimulation element (FSE) and putative FSE-arch **?**. The target site for each group of antisense oligonucleotides (ASOs) is indicated by dotted lines; lengths are to scale. **(b)** Rolling (window = 45 nt) Pearson correlation coefficient (PCC) of DMS reactivities over the 5' FSE-arch between each +ASO sample and a no-ASO control. Each curve represents one ASO group, colored as in (a); groups 4 and 13 are not shown. Locations of the FSE and the outer and inner stems of the 5' FSE-arch are also indicated. **(c)** (Top) Scatter plots of DMS reactivities over the 5' FSE-arch comparing each cluster of the no-ASO sample to the sample with ASO group 9; each point is one position in the 5' FSE-arch. (Bottom) Receiver operating characteristic (ROC) curves comparing each cluster of the no-ASO sample to the two inner stems of the FSE-arch, with area under the curve (AUC) indicated. **(d)** Like (c) but over the 3' FSE-arch, and comparing to the sample with ASO group 4. One highly reactive outlier was ignored when calculating PCC (which is sensitive to outliers) but included in the ROC (which is robust). **(e)** Model of the inner two stems in the ensemble of structures formed by the 2,924 nt segment.

We next sought to determine in what fraction of molecules the two inner stems of the FSE-arch form. Using SEISMIC-RNA, we clustered reads from the 5' side of the FSE-arch for the no-ASO control and found two clusters with a 43/57% split. To determine if they corresponded to the two inner stems formed and unformed, we compared their DMS reactivities to those after adding ASO group 9, which blocks the two inner stems (Figure **??**c, top). Cluster 2 had similar DMS reactivities (PCC = 0.95), indicating it corresponds to the stems unformed. Meanwhile, the DMS reactivities of cluster 1 differed (PCC = 0.64), suggesting it corresponds to the stems formed.

To further support this result, we leveraged the preexisting model of the FSE-arch **?**. If cluster 1 did correspond to the two inner stems formed, its DMS reactivities would agree well with their structures (i.e. paired and unpaired bases should have low and high reactivities, respectively), while those of cluster 2 would agree less. We quantified this agreement using receiver operating characteristic (ROC) curves (Figure **??**c, bottom). The area under the curve (AUC) for cluster 1 was 1.00, indicating perfect agreement with the two inner stems of the FSE-arch; while that of cluster 2 was 0.57, close to no agreement (0.50). This result further supports that clusters 1 and 2 correspond to the two inner stems formed and unformed, respectively.

If the RNA did exist as an ensemble of the two inner stems formed and unformed, the 3' side of the FSE-arch would also cluster into formed and unformed states. To investigate, we performed RT-PCR with primers flanking the 3' side of the inner two stems – both without ASOs and with ASO group 4 (targeting the 5' side of the FSE-arch). We clustered the no-ASO control into two clusters (51/49% split) and found – similar to the previous result – that the DMS reactivities after blocking the 5' FSE-arch with ASO group 4 resembled those of cluster 2 (PCC = 0.94) but not cluster 1 (PCC = 0.80), while the structure of the two inner stems agreed with cluster 1 (AUC = 1.00) but not cluster 2 (AUC = 0.57) (Figure **??**d). We concluded that the RNA exists as an ensemble of structures in which the two inner stems of the FSE-arch form in 47% ± 4% of molecules (Figure **??**e).

The long-range stems compete with the frameshift pseudoknot in SARS-CoV-2

To determine if the FSE forms other long-range stems, in lieu of the original outer stem of the FSE-arch **?**, we modeled a 1,799 nt segment centered on the FSE-arch. Although computationally predicting long-range base pairs is notoriously unreliable **??**, we speculated that we could increase consistency by incorporating the DMS reactivities of cluster 1 on both sides of the FSE-arch (Figure **??**c and d). In agreement, among the top 20 structures predicted from the cluster 1 DMS reactivities, 100% included the inner two stems, compared to ? using the ensemble average, ? using cluster 2, and ? using no DMS reactivities (Supplementary Figure). Using the DMS reactivities of the long-range cluster thus enabled predicting the long-range stems consistently.

Our refined model based on the long-range cluster (Figure **??**) included not only the two inner stems of the FSE-arch – which we hereafter call long stems 1 (LS1) and 2 (LS2) – but also two stems (LS3 and LS4) that were not in the original FSE-arch model **?**. The structure also contained the alternative stem 1 (AS1) that we had previously discovered **?**. To our surprise, LS2b, LS3, and LS4 of the new model collectively overlapped all three stems of the pseudoknot (PS1, PS2, and PS3) that is thought to stimulate frameshifting **???**. Thus, these long stems – if they exist – would be mutually exclusive with the pseudoknot.

To verify this refined model, we performed SEARCH-MaP on the 1,799 nt segment using 15-20 nt LNA/DNA mixmer ASOs for single-stem precision (Figure **??**b). Each ASO targeted the 3' side of one stem, and we measured the change in DMS reactivities of the FSE. ASOs targeting the 3' sides of LS1 and LS2a perturbed the DMS reactivities in exactly the expected locations on the 5' sides. Binding an ASO to the 3' side of LS2b caused a larger perturbation with more off-target effects, likely because this stem overlaps with pseudoknot stem 2 (PS2). Blocking LS3b also resulted in a main effect around the intended location, with one off-target effect upstream, suggesting that other base pairs between the pseudoknot and this upstream

region may exist. Therefore, stems LS1, LS2a/b, and LS3b do exist – at least in a portion of the ensemble.

We then sought to determine whether the long-range stems compete with the pseudoknot. If so, blocking them with ASOs would increase the proportion of the pseudoknot in the ensemble. To test this hypothesis, we first generated four possible models of the FSE structure by combining mutually compatible stems from the refined model (Figure **??**c). Then, we clustered the 1,799 nt segment without ASOs up to 6 clusters – the maximum number reproducible between replicates – (Supplementary Figure) and compared each cluster to each structure model using the area under the receiver operating characteristic curve (AUC-ROC) over the positions spanned by the pseudoknot, 305-371 (Figure **??**d, top). We considered a cluster and model to be "consistent" if the AUC-ROC was at least 0.90. The locally nested model (AS1 plus PS2 and PS3) was consistent with three clusters totaling 52% of the ensemble, while the extended model (AS1 plus all long-range stems) was consistent with one cluster (20%). No clusters were fully consistent with the pseudoknotted model, though the least-abundant cluster (7%) came close with an AUC-ROC of 0.88. The remaining cluster (21%) was not consistent with any model, suggesting that the ensemble contains structures beyond those in Figure **??**c.

Adding an ASO targeting the 5' side of AS1 reduced the proportion of AS1-containing states (extended and locally nested) from 72% to 16% (Figure **??**d, left). In their absence emerged clusters consistent with the pseudoknotted and truncated models, representing 56% and 20% of the ensemble, respectively. Meanwhile, adding an ASO that blocked the part of LS2b that overlaps PS2 eliminated the extended state (which includes LS2b) and produced one cluster (13%) consistent with the pseudoknotted model (Figure **??**d, right). Adding both ASOs simultaneously collapsed the ensemble into three clusters of which two (87%) were highly consistent with the pseudoknotted model (Figure **??**d, bottom). Since blocking the PS2-overlapping portion of LS2b increased the proportion of clusters consistent (or nearly so) with the pseudoknotted model – both alone and combined with the anti-AS1 ASO – we conclude that the long-range stems do outcompete the pseudoknot.

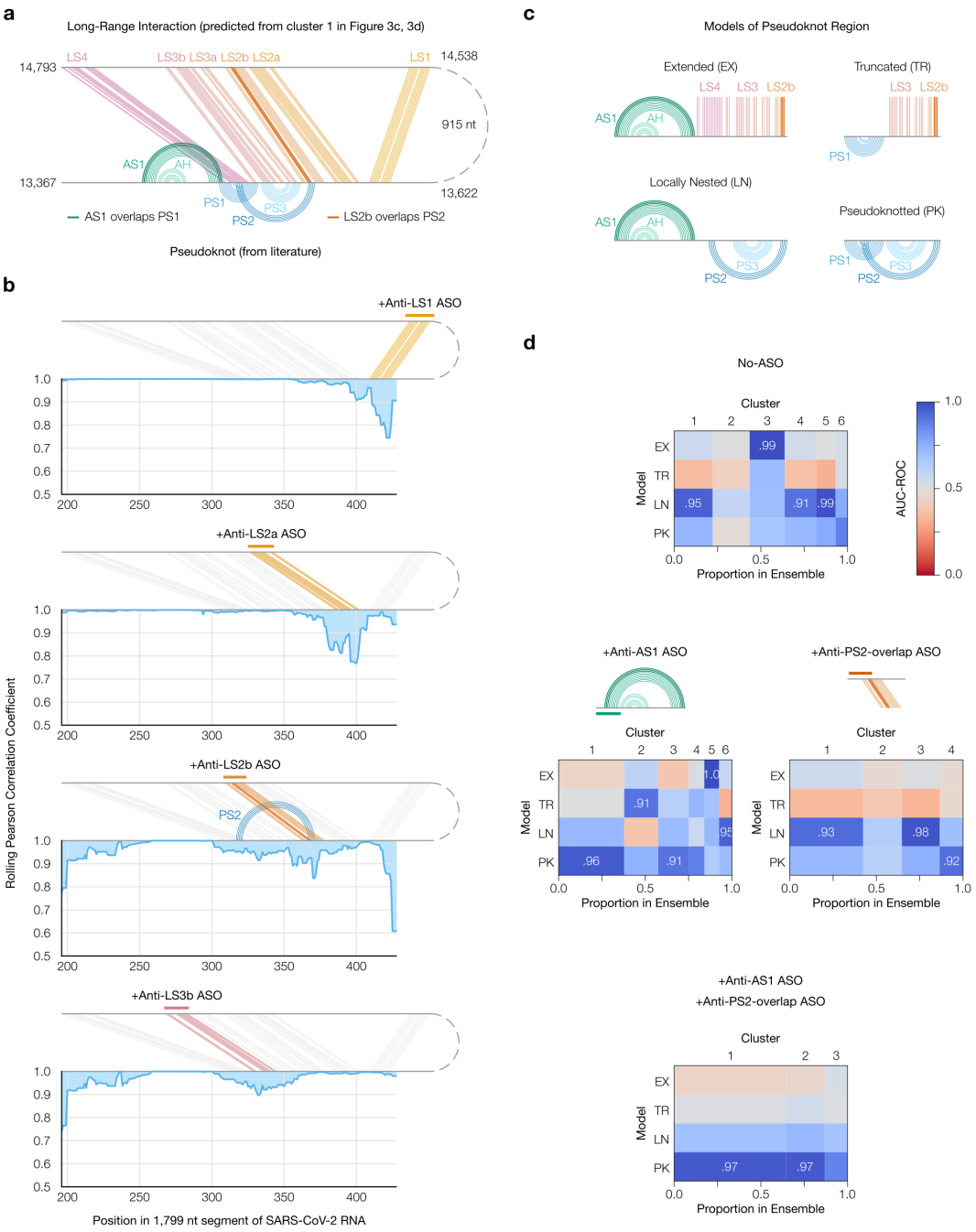


Figure 4: **Refinement of the long-range structure model and competition with the frameshift pseudoknot.** **(a)** Refined model of the long-range stems (minimum free energy prediction based on cluster 1 in Figure **??**c and d) including alternative stem 1 (AS1) **?**; the attenuator hairpin (AH) **?**; and long stems LS1, LS2a/b, LS3a/b, and LS4. Locations of pseudoknot stems PS1, PS2, and PS3 are also shown; as are the base pairs they overlap in AS1 and LS2b. **(b)** Rolling (window = 21 nt) Pearson correlation coefficient of DMS reactivities between each +ASO sample and a no-ASO control; base pairs targeted by each ASO are colored. **(c)** Models of possible structures for the FSE, by combining non-overlapping stems from (a). **(d)** Heatmaps comparing models in (c) to clusters of DMS reactivities over positions 305-371 via the area under the receiver operating characteristic curve (AUC-ROC). AUC-ROCs at least 0.90 are annotated. Cluster widths indicate proportions in the ensemble.

Frameshift stimulating elements of multiple coronaviruses form long-range base pairs

We surmised that other coronaviruses would also feature long-range base pairs involving the FSE. To search for these structures, we performed SEARCH-MaP with FSE-targeted ASOs on 1,799 nt segments from eight coronaviral genomes.

As of December 2021, the NCBI Reference Sequence Database **?** contained 62 complete genomes of coronaviruses. To focus on those likely to have long-range base pairs involving the FSE, we predicted the likelihood that each base in a 2,000 nt section surrounding the FSE would pair with a base in the FSE (Supplementary Figure). Based on these predicted structures, we selected ten coronaviruses – at least one from each genus (Supplementary Figure) – including SARS-CoV-2 as a positive control. Within the genus *Betacoronavirus*, we included all three SARS-related viruses – SARS coronaviruses 1 (`NC_004718.3`) and 2 (`NC_045512.2`) and bat coronavirus BM48-31 (`NC_014470.1`) – because they clustered into their own structural outgroup. The other three strains of *Betacoronavirus* that we selected were MERS coronavirus (`NC_019843.3`) with predicted base pairs at positions 510-530; and human coronavirus OC43 (`NC_006213.1`) and murine hepatitis virus strain A59 (`NC_048217.1`), both with a predicted upstream base pairs at positions 10-20. We selected two strains of *Alphacoronavirus*: transmissible gastroenteritis virus (`NC_038861.1`) and bat coronavirus 1A (`NC_010437.1`), predicted to have base pairs at positions 440-460 and 350-360, respectively. For avian infectious bronchitis virus strain Beaudette (`NC_001451.1`) – a strain of *Gammacoronavirus* – the FSE was predicted to base-pair with positions 330-350; while common moorhen coronavirus HKU21 (`NC_016996.1`) was the species of *Deltacoronavirus* with the most promising long-range base pairs.

We reasoned that if an FSE does interact with a distant RNA element, removing that element by truncating the RNA would change the structure of the FSE, which we could detect with DMS-MaPseq **?**. For each of the ten coronaviruses that passed the computational screen, we *in vitro* transcribed and performed DMS-MaPseq on

both a 239 nt segment comprising the FSE and minimal flanking sequences and a 1,799 nt segment encompassing the FSE and all sites with which it was predicted to interact. All coronaviruses except for human coronavirus OC43 and MERS coronavirus showed differences in their DMS reactivity profiles between the 239 nt and 1,799 nt segments (Supplementary Figure), suggesting the FSE forms long-range base pairs.

To determine which RNA elements the FSE base-pairs with in each coronavirus, we performed SEARCH-MaP on the 1,799 nt RNA segment using DNA ASOs targeting the vicinity of the FSE (Figure **??**). The rolling Spearman correlation coefficient (SCC) between the +ASO and no-ASO mutational profiles dipped below 0.9 at the ASO target site in every coronavirus segment, confirming the ASOs bound and altered the structure.

To confirm we could detect long-range base pairs, we compared the rolling SCC for the SARS-CoV-2 segment to our refined model of the FSE structure (Figure **??**, blue). The SCC dipped below 0.9 at positions 1,483-1,560 and at 1,611-1,642, which coincide with stems LS2-LS3 (positions 1,476-1,550 within the 1,799 nt segment) and stem LS4 (positions 1,600-1,622). These dips were the two largest downstream of the FSE; although others (corresponding to no known base pairs) existed, they were barely below 0.9 and could have resulted from base pairing between these regions and other (non-FSE) regions. Near LS1 (positions 1,367-1,381), the SCC dipped only slightly to a minimum of 0.95, presumably because LS1 is the smallest (15 nt) and most isolated long-range stem. Therefore, this method was sensitive enough to detect all but the smallest long-range stem, and specific enough that the two largest dips corresponded to validated long-range stems.

We found similar long-range stems in SARS-CoV-1 and another SARS-related virus, bat coronavirus BM48-31. Both viruses showed dips in SCC at roughly the same positions as LS2-LS4 in SARS-CoV-2, indicating that they have homologous structures. SARS-CoV-1 also had a wide dip below 0.9 at positions 1,284-1,394, corresponding to a homologous LS1. Thus, three SARS-related viruses share these long-range stems involving the FSE, hinting that these structures are functional.

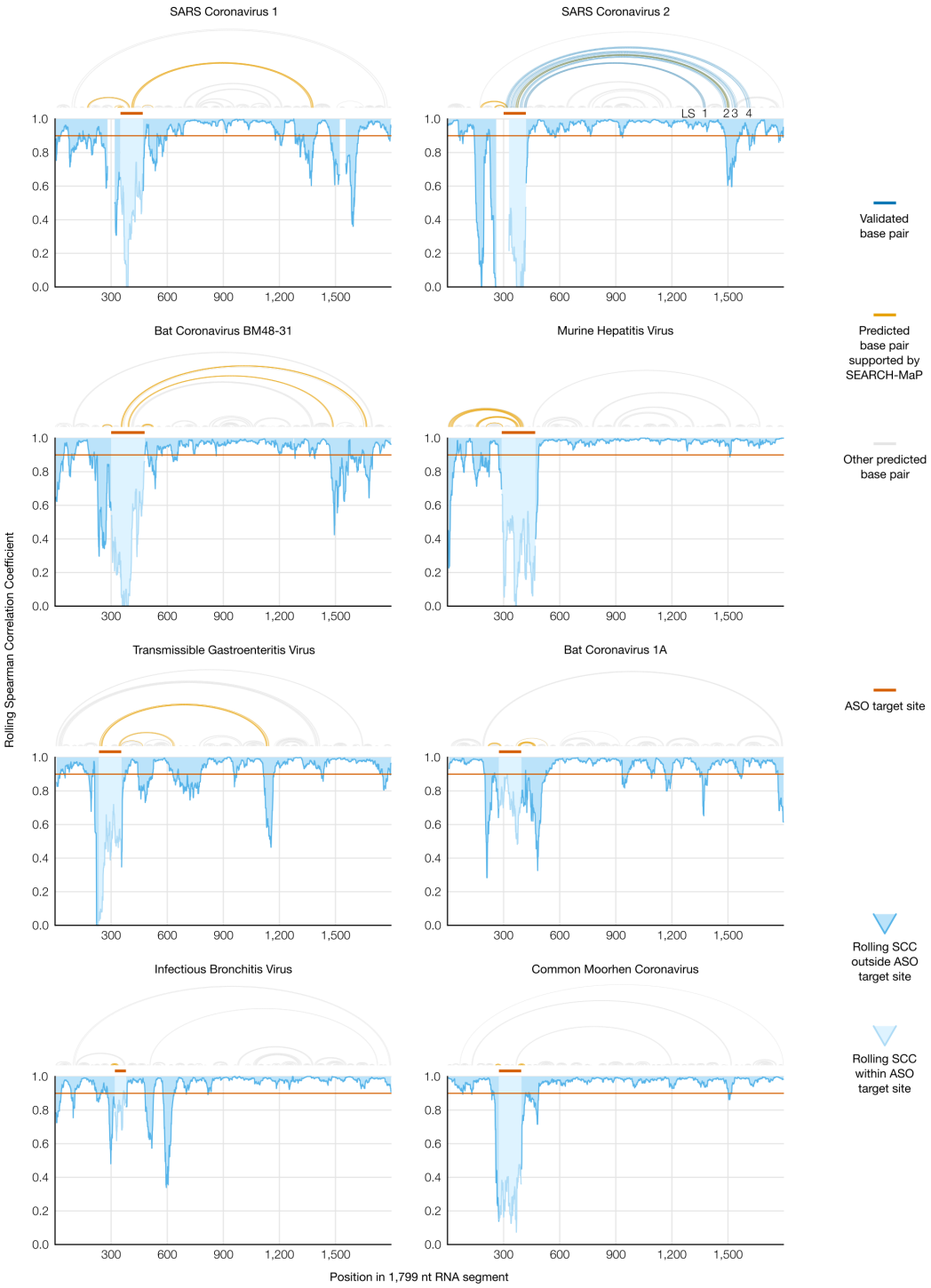


Figure 5: **Evidence for long-range RNA–RNA base pairs involving the FSE in four additional coronaviruses.** Rolling (window = 45 nt) Spearman correlation coefficient (SCC) of DMS reactivities between the +ASO and no-ASO samples for each 1,799 nt segment of a coronaviral genome. The target site of each ASO is highlighted on the SCC data and shown above each graph. Structures predicted with RNAstructure **?** using no-ASO ensemble average DMS reactivities as constraints **?** are drawn above each graph; base pairs connecting the ASO target site to an off-target position with SCC less than 0.9 are colored. For SARS-CoV-2, the refined model (Figure **??**a) is also drawn, with LS1-LS4 labeled.

In every other species except common moorhen coronavirus, we found prominent dips in SCC at least 200 nt from the ASO target site. To model potential base pairing between these dip positions and the FSE, we used the Fold program from RNAstructure **?** with the no-ASO ensemble average DMS reactivities as constraints **?**. We surmised that using the DMS reactivities corresponding to the long-range base pairs formed would generally yield more accurate predictions of the long-range structure than would using the ensemble average DMS reactivities (a mixture of all structures). For instance, the prediction for SARS-CoV-2 based on the ensemble average included LS1 and LS2b but missed the other long-range stems. Although clustered data were unavailable in this case, we were still able to find long-range base pairs consistent with the SEARCH-MaP data for both murine hepatitis virus and transmissible gastroenteritis virus (Figure **??**, orange). We conclude that long-range base pairing involving the FSE occurs more widely than in just SARS-CoV-2, including in the genus *Alphacoronavirus*.

Structure of the full TGEV genome in ST cells supports long-range base pairing involving the FSE

Transmissible gastroenteritis virus (TGEV) is a strain of *Alphacoronavirus 1* **?** that infects pigs and causes vomiting and diarrhea – almost always fatally in baby piglets **?**. Due to the impacts of TGEV on animal health and economics **?** and our evidence of a long-range stem, we sought to model the genomic secondary structures of live TGEV. We began by treating TGEV-infected ST cells with DMS (two biological replicates) and performing DMS-MaPseq **?** (two technical replicates per biological replicate) on the extracted RNA (Figure **??**a). The DMS reactivities over the full genome were consistent between biological replicates (PCC = 0.97), albeit not with the 1,799 nt segment *in vitro* (PCC = 0.82), which showed that verifying the long-range stem in live TGEV would be necessary (Supplementary Figure **??**).

To determine the structure ensembles, we performed RT-PCR on the extracted RNA using primers targeting both sides of the long-range stems. The DMS reactivities from RT-PCR were consistent with those over the full genome (Supplementary Figure). For each side of the long-range stem, we clustered the reads into two clusters (Figure **??**b). These clusters had similar correlations with the +ASO sample and similar AUC-ROC scores (Supplementary Figure), making it more difficult to identify them for TGEV than for SARS-CoV-2 (Figure **??**). Nevertheless, we realized that on each side, the bases that were predicted to interact had generally lower DMS reactivities in one cluster compared to the other cluster, and hypothesized that this cluster corresponded to the long-range stem formed (Figure **??**b). On the 5' side, the less-reactive cluster constituted 52% of the ensemble; on the 3' side, 60%. To investigate, we refined the structure of the 1,799 nt segment using the DMS reactivities from both of these clusters. Consistent with our hypothesis, the minimum free energy (MFE) model included the long-range stem, which we hereafter call long stem 3 (LS3) (Figure **??**c); predicting the structure using both more-reactive clusters did not produce LS3 (Supplementary Figure). The refined model also featured a prominent new stem connecting 20 nt upstream of the FSE with 400 nt downstream, which we call LS2. We suspect that LS2 exists because it coincides with a broad region perturbed by adding an ASO to the FSE in the 1,799 nt segment of TGEV (Figure **??**c). Another stem spanning just under 300 nt, which we call LS1, was also predicted in the same location as in the 1,799 nt segment.

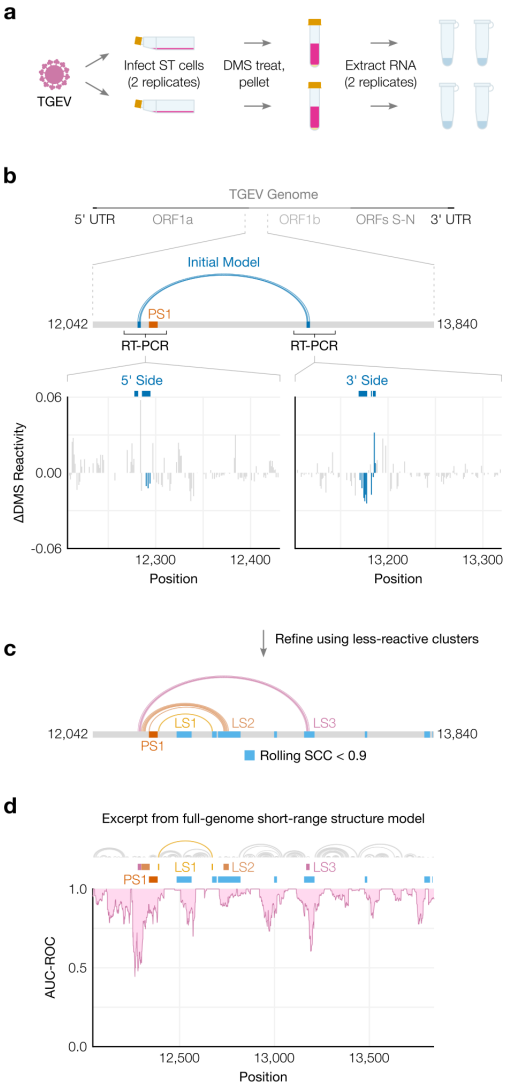


Figure 6: **Genomic secondary structure of live TGEV. (a)** Schematic of the experiment in which two biological replicates of ST cells were infected with TGEV, DMS-treated, and pelleted. Cell pellets were divided into two technical replicates prior to extraction of DMS-modified RNA. **(b)** Differences in DMS reactivities between the two clusters on each side of the long-range stem. Each bar represents one base. Bases are shaded dark blue if they pair in the initial model of the long-range stem (from Figure **??**), shown above along with its location in the full genome. The locations of FSE pseudoknot stem 1 (PS1) and the regions amplified for clustering are also indicated. **(c)** Refined model of the long-range stem in TGEV based on the DMS reactivities of the less-reactive cluster from both sides. Long stems 1 (LS1), 2 (LS2), and 3 (LS3) are labeled. For comparison with the regions of the 1,799 nt segment perturbed by the ASO (Figure **??**), positions after the FSE where the Spearman correlation coefficient (SCC) dipped below 0.9 are shaded light blue. **(d)** Rolling AUC-ROC (window = 45 nt) between the full-genome DMS reactivities and full-genome secondary structure modeled from the DMS reactivities (maximum 300 nt between paired bases). The structure model is drawn above the graph. Only positions 12,042-13,840 are shown here. For comparison, the locations of PS1, LS1, LS2, LS3, and dips in SCC after the FSE are also indicated.

We used the ensemble average DMS reactivities to produce one "ensemble average" model of the secondary structure of the full TGEV genome (Supplementary Figure). We restricted base pairs to a maximum distance of 300 nt to make the computation tractable and avoid over-predicting spurious long-range base pairs. To verify the model quality, we confirmed that the predicted structure of the first 520 nt included the highly conserved stem loops SL1, SL2, SL4, and SL5a/b/c in the 5' UTR **?** (Supplementary Figure **??**a) and was consistent with the DMS reactivities (AUC-ROC = 0.94) (Supplementary Figure **??**b).

The AUC-ROC was lower in many locations throughout the rest of the genome (Supplementary Figure), indicating that a single secondary structure consistent with the ensemble average DMS reactivities could not be found. We had noticed a similar phenomenon in SARS-CoV-2 – in particular, at the FSE **?**. Thus, we surmised that regions with low AUC-ROC scores likely form alternative structures or long-range base pairs – or both – that a single secondary structure model could not capture. Checking if this relationship also held for TGEV, we found a large dip in AUC-ROC just upstream of the FSE, centered on the 5' ends of LS2 and LS3, as well as smaller dips at the 3' ends of both stems (Figure **??**d). In fact, at or near every location that SEARCH-MaP had evidenced to interact with the FSE – where the rolling SCC had dipped – the AUC-ROC also dipped. This finding supports the hypothesis that long-range base pairs and/or alternative structures are often the reason why predicted structures are not locally consistent with the DMS reactivities on which they were based.

Discussion

In this work, we developed SEARCH-MaP and SEISMIC-RNA and applied them to detect structural ensembles involving long-range base pairs in SARS-CoV-2 and other coronaviruses. A previous study demonstrated that binding an ASO to one side of a long-range stem would perturb the chemical probing reactivities of the other side **?**. Here, we separated and identified the reactivities corresponding to long-range stems formed and unformed. This advance enables isolating the reactivities of the long-range stem formed – on not just one but both sides of the stem, linking corresponding alternative structures over distances much greater than the length of a read, which has not been possible in previous studies **??**. Using the linked reactivities from both sides of a long-range stem, its secondary structure can be modeled more accurately than would be possible using the ensemble average reactivities, as we have done for SARS-CoV-2 (Figure **??**) and TGEV (Figure **??**).

SEISMIC-RNA builds upon our previous work, the DREEM algorithm **?**. Here, we have optimized the algorithm to run approximately 10-30 times faster and built an entirely new workflow around it for aligning reads, calling mutations, masking data, and outputting a variety of graphs. SEISMIC-RNA can process data from any mutational profiling experiment, including DMS-MaPseq **?** and SHAPE-MaP **?**, not just SEARCH-MaP. The software is available from the Python Package Index (pypi.org/project/seismic-rna) or GitHub (github.com/rouskinlab/seismic-rna) and can be used as a command line executable program (`seismic`) or via its Python application programming interface (`import seismicrna`).

We envision SEARCH-MaP and SEISMIC-RNA bridging the gap between broad and detailed investigations of RNA structure. Other methods such as proximity ligation **?????** provide broad, transcriptome-wide information on RNA structure and could be used as a starting point to find structures of interest for deeper investigation with SEARCH-MaP/SEISMIC-RNA. Indeed, the first evidence of the FSE-arch in SARS-CoV-2 came from such a study **?**. To investigate RNA structures in detail, M2-seq **?** and related methods **?** can pinpoint base pairs with up to single-nucleotide

resolution and minimal need for structure prediction. However, base pairs are detectible only if the paired bases occur on the same sequencing read, which restricts their spans to at most the read length (typically 300 nt). Because the capabilities of M2-seq and SEARCH-MaP complement each other, they could be integrated: first SEARCH-MaP/SEISMIC-RNA to discover, quantify, and model long-range base pairs; then M2-seq for short-range base pairs. By providing the missing link – structure ensembles involving long-range base pairs – SEARCH-MaP and SEISMIC-RNA could combine broad and detailed views of RNA structure into one coherent model.

To understand structures of long RNA molecules, SEARCH-MaP and SEISMIC-RNA could also be used to validate predicted secondary structures and benchmark structure prediction algorithms. Algorithms that predict secondary structures achieve lower accuracies for longer sequences **??**, hence long-range base pairs in particular must be confirmed independently. We envision a workflow to determine the structure ensembles of an arbitrarily long RNA molecule that begins with DMS-MaPseq **?**. The DMS reactivities would be used **?** to predict two initial models of the structure: one with a limit to the base pair length (for short-range pairs), the other without (for long-range pairs). Sections of the RNA with potential long-range pairs would be flagged from the long-range model and from regions of the short-range model that disagreed with the DMS reactivities (as in Figure **??**d). Then, SEARCH-MaP/SEISMIC-RNA could be used to validate, quantify, and refine the potential long-range base pairs; and other methods such as M2-seq **?** to do likewise for short-range base pairs. This integrated workflow could characterize the secondary structures of RNA molecules that have evaded existing methods (e.g. messenger RNAs **?**) as well provide much-needed benchmarks for secondary structure prediction algorithms **?**.

In this study, we focused on the genomes of coronaviruses, specifically long-range base pairs involving the frameshift stimulating element (FSE). Long-range base pairs implicated in frameshifting also occur in several plant viruses of the family *Tombusviridae* **???**. However, in *Tombusviridae* species, the frameshift pseudoknots themselves are made of long-range base pairs; in coronaviruses, the pseudoknots are

local structures **????** and (at least in SARS-CoV-2) compete with long-range base pairs. Consequently, the long-range base pairs are necessary for frameshifting in *Tombusviridae* species **???** but dispensable in coronaviruses: even the 80-90 nt core FSE of SARS-CoV-2 has stimulated 15-40% of ribosomes to frameshift in dual luciferase constructs **??????**. Surprisingly, frameshifting has appeared to be nearly twice as frequent (50-70%) in live SARS-CoV-2 **???**; whether this discrepancy is due to long-range base-pairing, methodological artifacts, or *trans* factors **?** is unknown **?**.

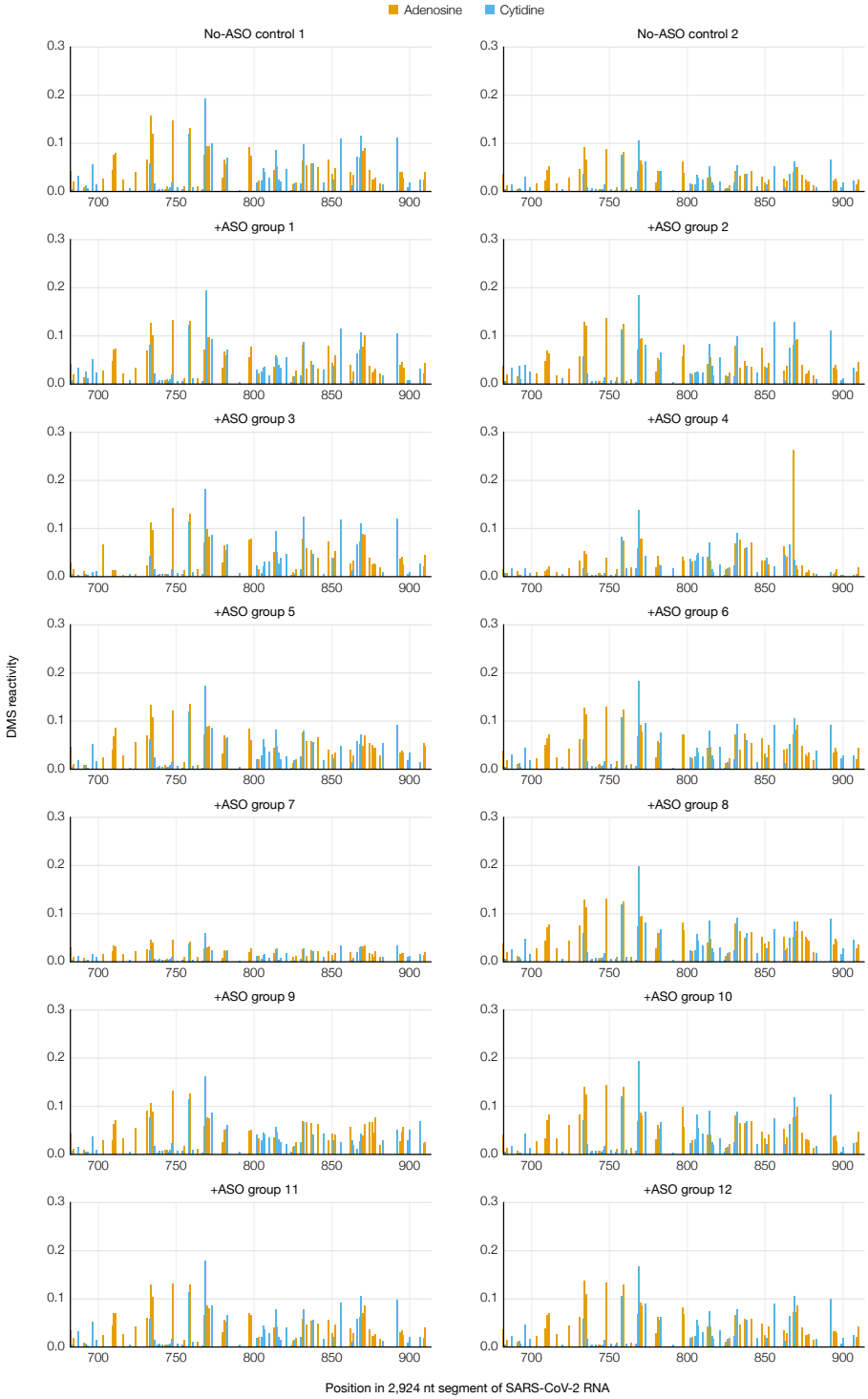
If, how, and why the long-range base pairs affect frameshifting in coronaviruses are open questions. For *Tombusviridae*, one study **?** suggested that the long-range stem regulates viral RNA synthesis by negative feedback: without RNA polymerase, the long-range stem would form and stimulate frameshifting to produce polymerase, which would then unwind the long-range stem while replicating the genome. However, this mechanism seems implausible in coronaviruses, where RNA synthesis and translation occur in separate subcellular compartments (the double-membrane vesicles and the cytosol, respectively) **?**. Another study on *Tombusviridae* **?** hypothesized that after the ribosome has frameshifted, long-range stems destabilize the FSE so the ribosome can unwind it and continue translating. As the long-range base pairs in SARS-CoV-2 do compete with the pseudoknot, they might also have this role, which – for coronaviruses – could not be strictly necessary for frameshifting. One study **?** of translation in SARS-CoV-2 at different time points measured frameshifting around 20% at 4 hours post infection but 60-80% at 12-36 hours. This result is consistent with a previous hypothesis **?** that coronaviruses use frameshifting to time protein synthesis: first translating ORF1a to suppress the immune system, then translating ORF1b containing the RNA polymerase. We surmise the long-range base pairs would form in virions and persist when the virus released its genome into a host cell, where they would initially suppress frameshifting. Once host protein synthesis had been inhibited and the double-membrane vesicles formed, a signal specific to the cytosol would disassemble the long-range base pairs so that frameshifting could occur efficiently and produce the replication machinery from ORF1b. The long-range base pairs would form in viral progeny but not in genomic RNA released into

the cytosol for translation, so that more ORF1b could be translated. This possible role of long-range base pairs in the coronaviral life cycle could be tested by probing the RNA structure in subcellular compartments and virions, identifying cytosolic factors that could disassemble the long-range base pairs, and quantifying how they affect frameshifting in the context of a live coronavirus.

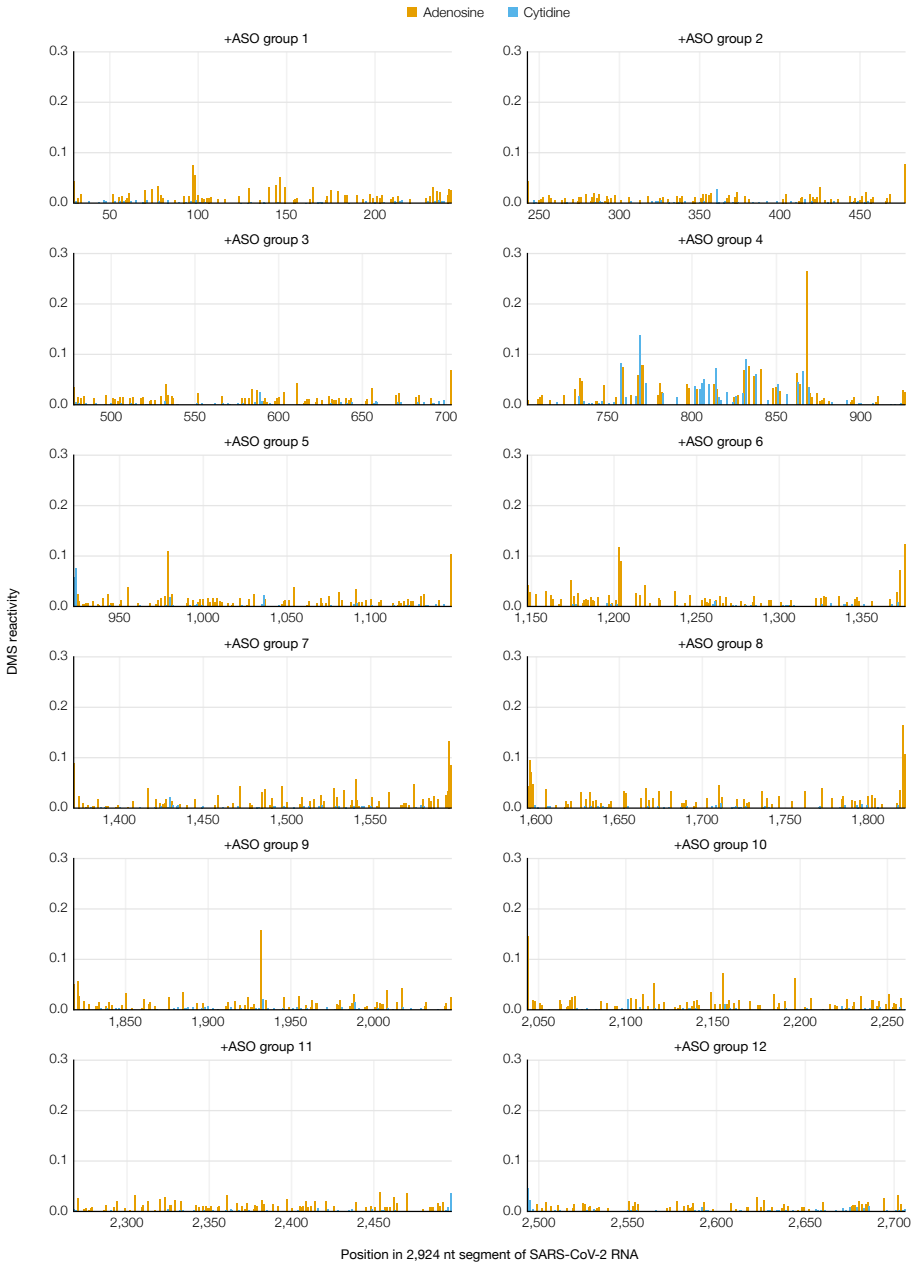
Future studies could also expand the scope of SEARCH-MaP and SEISMIC-RNA. While all SEARCH-MaP experiments in this study were performed *in vitro*, the method would likely also be feasible *in cellulo*: DMS-MaPseq can detect ASOs binding to RNAs within cells **?**. The main challenges would likely involve optimizing the ASO probes and transfection protocols to maximize the signal while minimizing unwanted side effects such as immunogenicity. SEARCH-MaP can screen an entire transcript (as in Figure **??**), but scaling up to an entire transcriptome could prove challenging. One strategy for probing many RNAs simultaneously could involve adding a pool of ASOs – with no more than one ASO capable of binding each RNA – rather than one ASO at a time. In this manner, a similar number of samples would be needed to search all RNAs as would be needed for the longest RNA. Distinguishing direct from indirect base pairing is another area for development: if segment Q could base-pair with either P or R, then blocking P could perturb R (and vice versa) as a consequence of perturbing Q, even though P and R could not base-pair directly. A solution could be to first block Q with one ASO; then, if blocking P with another ASO caused no change in R (and vice versa), it would suggest that they could only interact indirectly (through Q).

We imagine that SEARCH-MaP and SEISMIC-RNA will make it practical to determine accurate secondary structure ensembles of entire messenger, long noncoding, and viral RNAs. Collected in a database of long RNA structures, these results would facilitate subsequent efforts to predict RNA structures and benchmark algorithms, culminating in a real "AlphaFold for RNA" **?** in the hands of every biologist.

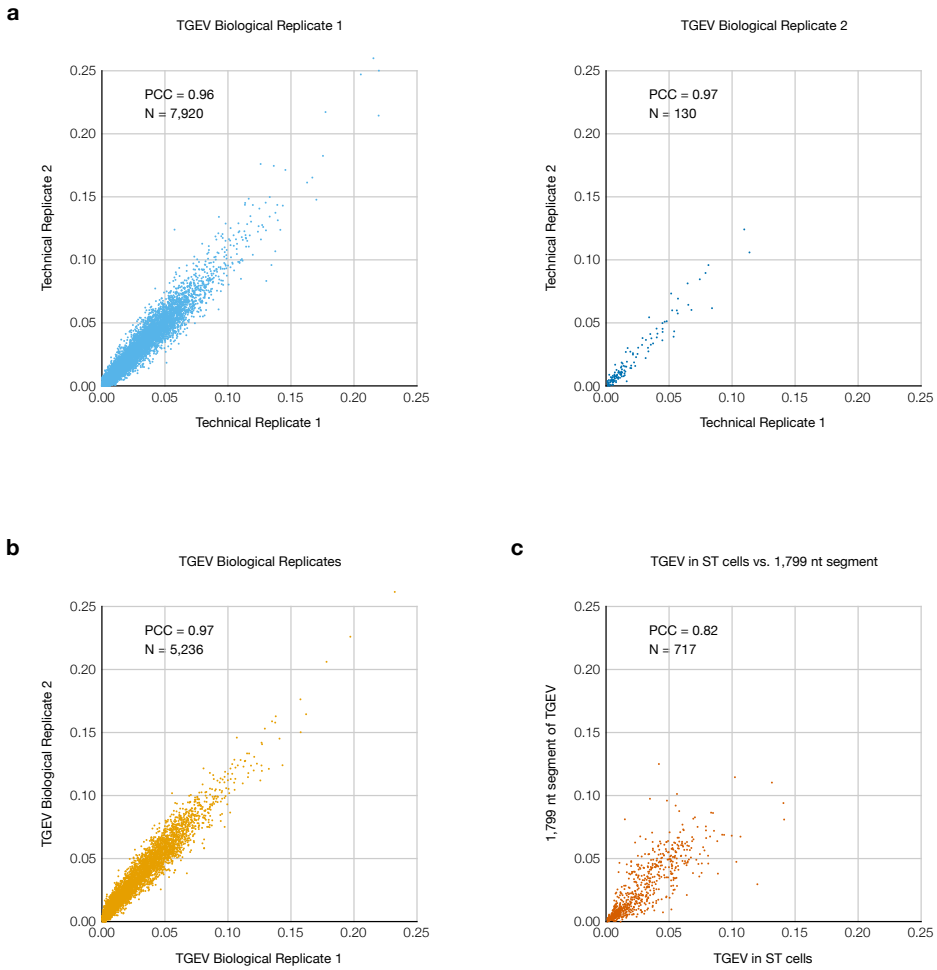
Supplementary Figures



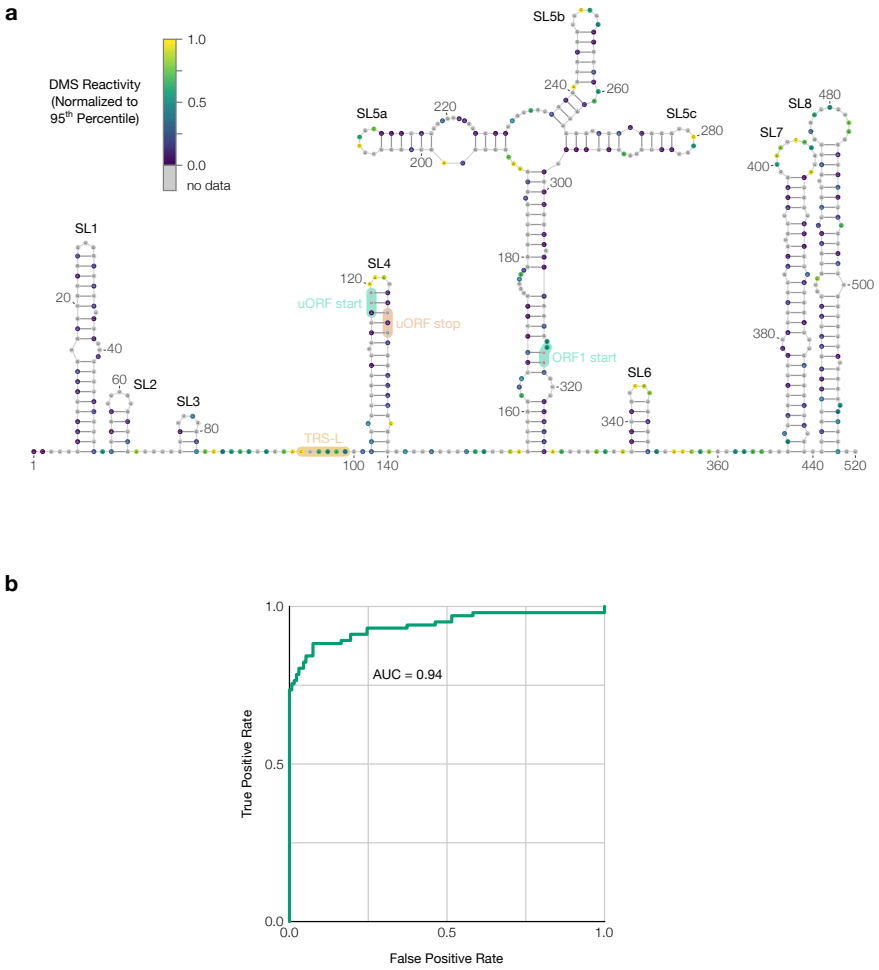
Supplementary Figure 1: **Mutational profiles of the FSE section upon adding each group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.** Positions are colored based on the RNA sequence.



Supplementary Figure 2: **Mutational profile of each ASO target section upon adding the corresponding group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.** Positions are colored based on the RNA sequence.



Supplementary Figure 3: **Replicates of TGEV in ST cells and comparison to the 1,799 nt segment.** **(a)** Scatter plots comparing the DMS reactivities of the two technical replicates for each biological replicate of TGEV in ST cells. Each point represents one base in the sequence. The number of points (N) and Pearson correlation coefficient (PCC) are indicated for each plot. **(b)** Scatter plot comparing the DMS reactivities of the two biological replicates (each biological replicate comprises the reads for both of its technical replicates pooled together). **(c)** Scatter plot comparing the DMS reactivities of TGEV in ST cells (the reads for both biological replicates pooled together) and for the 1,799 nt segment *in vitro*.



Supplementary Figure 4: **Secondary structure of the 5' UTR of TGEV. (a)** Model of the secondary structure of the first 520 nt of the TGEV genome, based on DMS reactivities in infected ST cells normalized to the 95th percentile. Bases are colored by DMS reactivity. The model includes the highly conserved stem loops SL1, SL2, SL4, SL5a, SL5b, and SL5c, as well as the more variable stem loops SL3, SL6, SL7, and SL8 **?**. The leader transcription regulatory sequence (TRS-L) **?**, upstream open reading frame (uORF) **?**, and start codon of ORF1 are also labeled. The model was drawn using VARNA **?**. **(b)** Receiver operating characteristic curve showing agreement between the DMS reactivities and the secondary structure model; the area under the curve (AUC) is indicated.