Results

Strategy of SEARCH-MaP and SEISMIC-RNA

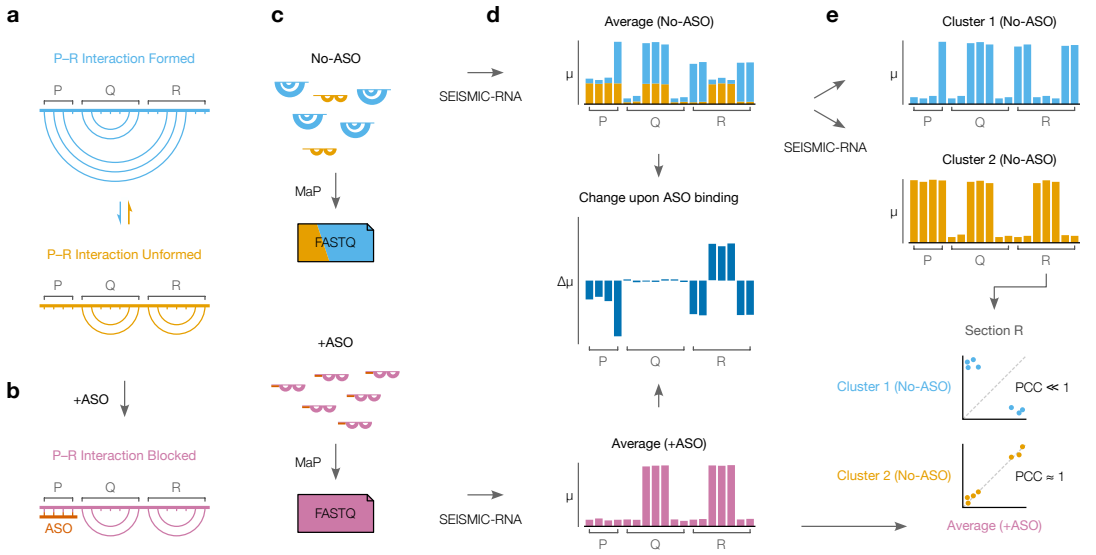

Figure 1: **The strategy of SEARCH-MaP and SEISMIC-RNA. (a)** This toy RNA is partitioned into three sections (P, Q, and R) whose molecules exist in two structural states: one in which an interaction between P and R forms (blue) and one in which it does not (purple). **(b)** Hybridizing an ASO (red) to P blocks it from interacting with R and forces all RNA molecules into the state where the P–R interaction is unformed. **(c)** A SEARCH-MaP experiment entails separate chemical probing and mutational profiling (MaP) with (+ASO) and without (–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)** Ensemble average mutational profiles with (+ASO) and without (–ASO) the ASO, computed with SEISMIC-RNA. The *x*-axis is the position in the RNA sequence; the *y*-axis is the fraction of mutations (μ) at the position. Each bar in the –ASO profile is drawn in two colors merely to illustrate how much each structural state contributes to each position; in a real experiment, states cannot be distinguished before clustering. The change upon ASO binding (green) indicates the difference in the fraction of mutations ($\Delta\mu$) between the +ASO and –ASO conditions. **(e)** Mutational profiles of two clusters (top) obtained by clustering the –ASO ensemble in (d) using SEISMIC-RNA, and the scatter plot of the mutation rates of bases in R (bottom) between the +ASO ensemble average (*x*-axis) and each cluster (*y*-axis). The expected correlation (r) is shown beside each scatter plot.

We illustrate SEARCH-MaP with an RNA comprising three sections (P, Q, and R) that folds into an ensemble of two structural states: one in which a base-pairing

interaction between P and R forms, another in which it does not (Figure 1a). Searching for sections that interact with P begins with hybridizing an antisense oligonucleotide (ASO) to P, which blocks P from base pairing with any other section, ablating the state in which the P–R interaction forms (Figure 1b). The RNA is chemically probed separately with (+ASO) and without (–ASO) the ASO, followed by mutational profiling and sequencing, e.g. using DMS-MaPseq **?** (Figure 1c).

SEISMIC-RNA can detect RNA–RNA interactions by comparing the +ASO and –ASO mutational profiles. Theoretically, each structural state has its own mutational profile **?**, but the mutational profile of a single state is not directly observable because all states are physically mixed during the experiment (Figure 1c, top). Instead, the directly observable mutational profile is the "ensemble average" – the average of the states' (unobserved) mutational profiles, weighted by the states' (unobserved) proportions (Figure 1d, top). Because the structures – and therefore mutational profiles – of R differ between the interaction-formed and -unformed states, the ensemble averages of R also differ between the +ASO and –ASO conditions (Figure 1d, middle). However, this is not the case for element Q, which has the same secondary structure in both states (Figure 1d, 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.

After identifying RNA–RNA interactions, SEISMIC-RNA can also determine the mutational profiles of the states where the P–R interaction is formed and unformed – even if their secondary structures are unknown. Inferring mutational profiles for the interaction-formed and -unformed states requires clustering the –ASO ensemble into two clusters of RNA molecules (Figure 1e, top). Each cluster has its own mutational profile and corresponds to one structural state, but which cluster corresponds to the interaction-formed (or -unformed) state is not yet known. The interaction-unformed state has a mutational profile similar to that of the +ASO ensemble average, since the ASO blocks the interaction and forces the RNA into the interaction-unformed state. Therefore, a cluster that correlates well ($r \approx 1$) with the +ASO ensemble average (here, Cluster 2) corresponds to the interaction-

unformed state; while a cluster that correlates weakly ($r \ll 1$) corresponds to the interaction-formed state (Figure 1e, bottom).

(I hope) SEARCH-MaP detects long-range base-pairing in ribosomal RNA

We first validated SEARCH-MaP using 16S and 23S ribosomal RNA (rRNA) from *E. coli*. For each rRNA, we selected two RNA–RNA interactions spanning ¿ [HOW MANY] nt that had been detected in a cell-free system **?**. For each interaction, we hypothesized that binding an ASO to either side would break the interaction and perturb the structure of the other side (distant from the ASO binding site) and designed two ASOs, one targeting each side. As a negative control, we also designed one ASO targeting a stem loop in each rRNA, which we hypothesized would perturb only the structure near the ASO binding site.

We folded the 16S and 23S rRNAs with each ASO, performed DMS-MaPseq over the entire transcripts, and compared ensemble average mutational profiles with and without ASOs using SEISMIC-RNA. [DESCRIBE THE RESULTS]

Figure 2:

SEARCH-MaP detects, separates, and quantifies a long-range RNA–RNA interaction in SARS-CoV-2

Aside from ribosomes, many of the best-characterized functional long-range RNA–RNA interactions 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 SARS coronavirus 2 (SARS-CoV-2) FSE 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 a long-range interaction **?**.

We used SEARCH-MaP to find the long-range interaction involving the FSE. We hypothesized it would turn out to be the 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 3a). 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. All ASO groups bound properly, evidenced by suppression of DMS reactivities over their target sites (SFIG).

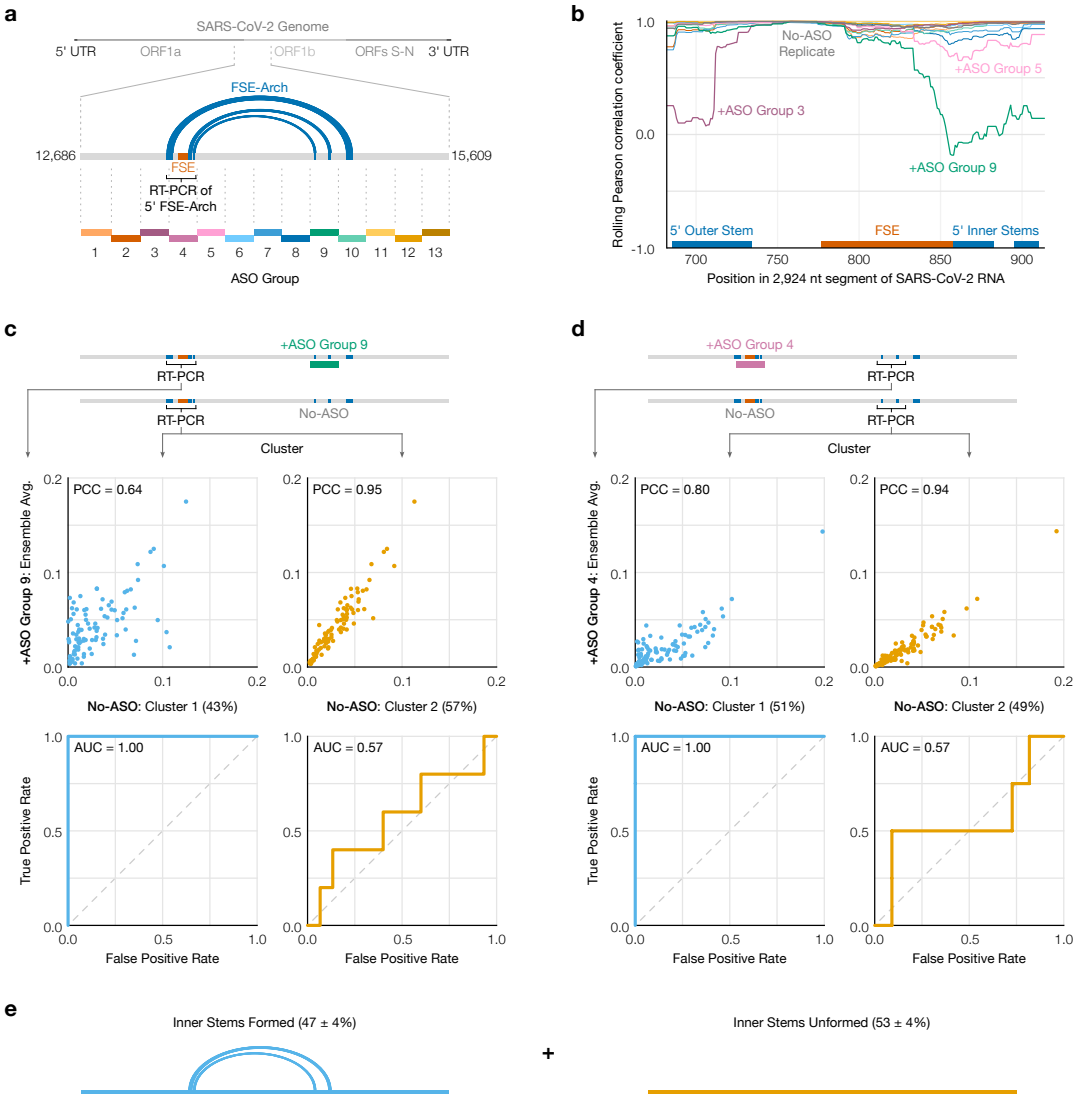

Figure 3: **Search for a long-range RNA–RNA interaction with the SARS-CoV-2 FSE. (a)** The 2,924 nt segment of the SARS-CoV-2 genome containing the frameshift stimulation element (FSE) and putative FSE-arch **?**. The target site of each ASO group is indicated by dotted lines; lengths are to scale. **(b)** Rolling (window = 45 nt) Pearson correlation coefficient 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. **(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. **(e)** Model of the inner two stems in the ensemble of structures formed by the 2,924 nt segment.

To quantify structural changes over the 5' FSE-arch, we calculated the rolling Pearson correlation coefficient (PCC) of the DMS reactivities between each sam-

ple and a no-ASO control (Figure 3b). A no-ASO replicate had a rolling PCC consistently 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 interactions involving the immediate vicinity of the FSE.

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 3c, 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, we would expect its DMS reactivities to agree well with their structures (i.e. paired and unpaired bases should have low and high reactivities, respectively) and those of cluster 2 to agree much less. We quantified this agreement using receiver operating characteristic (ROC) curves (Figure 3c, 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 null (0.50). This result further supports that cluster 1 (43%) corresponds to the two inner stems formed, and cluster 2 (57%) to these stems unformed.

If the RNA exists as an ensemble of the two inner stems formed and unformed, then we would also expect the 3' side of the FSE-arch to 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 3d). 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 3e).

The long-range interaction competes with the frameshift pseudoknot in SARS-CoV-2

Having clustered out the DMS reactivities of the interaction-formed state on both sides of the FSE-arch (cluster 1 in Figure 3c and d), we used them as DMS constraints **?** in RNAstructure **?** to fold a 1,799 nt segment centered on the long-range interaction. This refined model (Figure 4) 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 generally thought to stimulate frameshifting **???**. Thus, these long stems – if they exist – and the pseudoknot would be mutually exclusive.

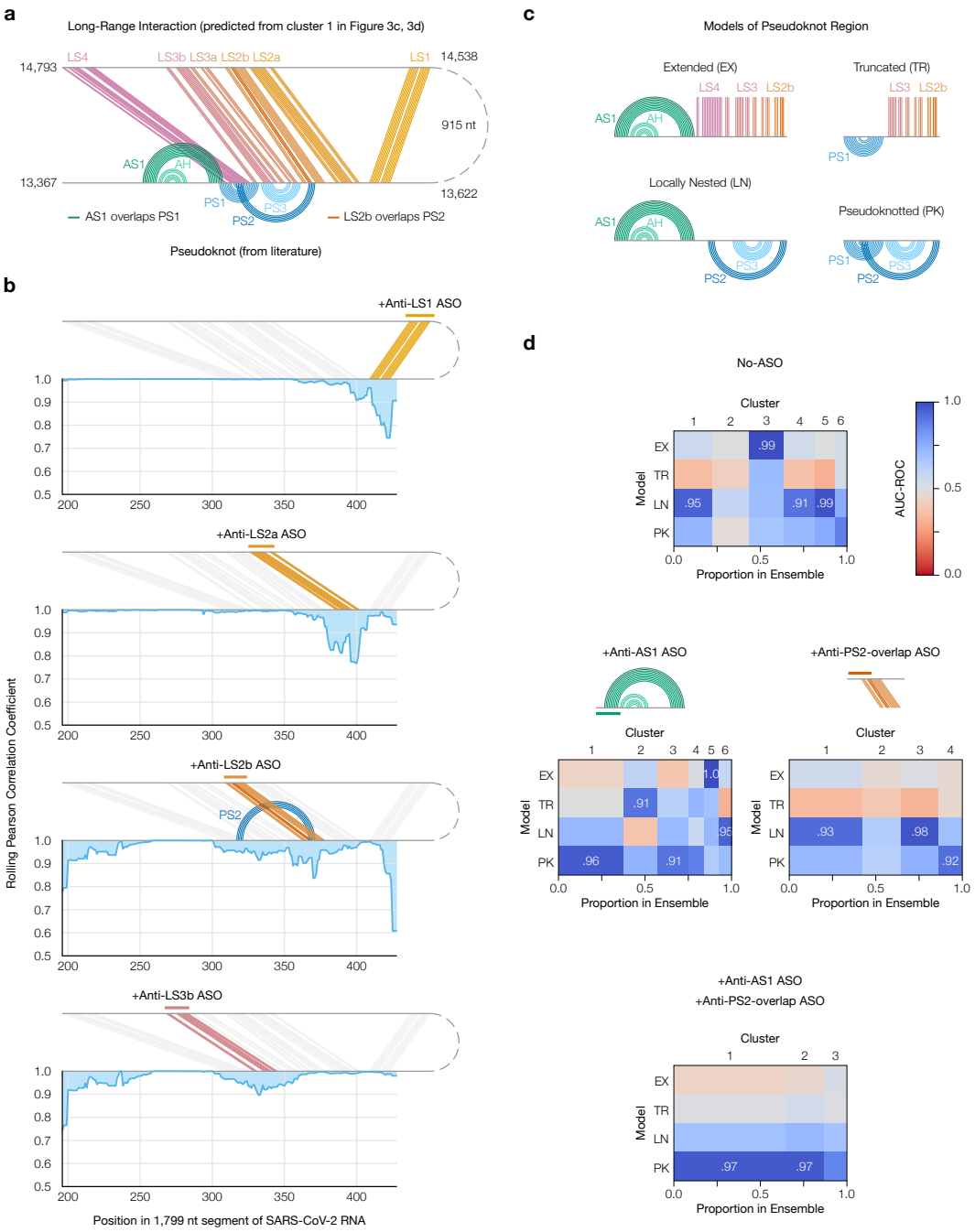

Figure 4: **Refinement of the long-range interaction and competition with the frameshift pseudoknot. (a)** Refined model of the long-range interaction (minimum free energy prediction based on cluster 1 in Figure 3c and d) including alternative stem 1 (AS1) **?**; the attenuator hairpin (AH) **?**; and 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.

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 4b). Each ASO targeted one stem in the downstream portion of the interaction, 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 there may be another RNA–RNA interaction with the pseudoknot and this upstream region. 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 4c). Then, we clustered the 1,799 nt segment without ASOs up to 6 clusters (the maximum number reproducible between replicates) 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 4d, 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 4c.

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 4d, left). In their absence emerged clusters consistent with the pseudoknotted and trun-

cated 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 4d, right). Adding both ASOs simultaneously collapsed the ensemble into three clusters of which two (87%) were highly consistent with the pseudoknotted model (Figure 4d, 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 interaction does outcompete the pseudoknot.

Frameshift stimulating elements of multiple coronaviruses participate in long-range RNA–RNA interactions

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

Computational and experimental screening identifies eight coronaviruses with potential long-range interactions

As of December 2021, the NCBI Reference Sequence Database **?** contained 62 complete genomes of coronaviruses. To focus on those likely to have long-range interactions 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 (SFIG). Based on these predicted interactions, we selected ten coronaviruses – at least one from each genus (SFIG) – 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 a predicted interaction 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 interaction 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 interactions at positions 440-460 and 350-360, respectively. Avian infectious bronchitis virus strain Beaudette (NC_001451.1) – a strain of *Gammacoronavirus* – was predicted to have a strong interaction at positions 330-350, while common moorhen coronavirus HKU21 (NC_016996.1) was the species of *Deltacoronavirus* with the most promising FSE interactions.

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 (SFIG), suggesting long-range interactions involving the FSE.

SEARCH-MaP reveals long-range interactions involving the FSE in four additional coronaviruses

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 5). 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.

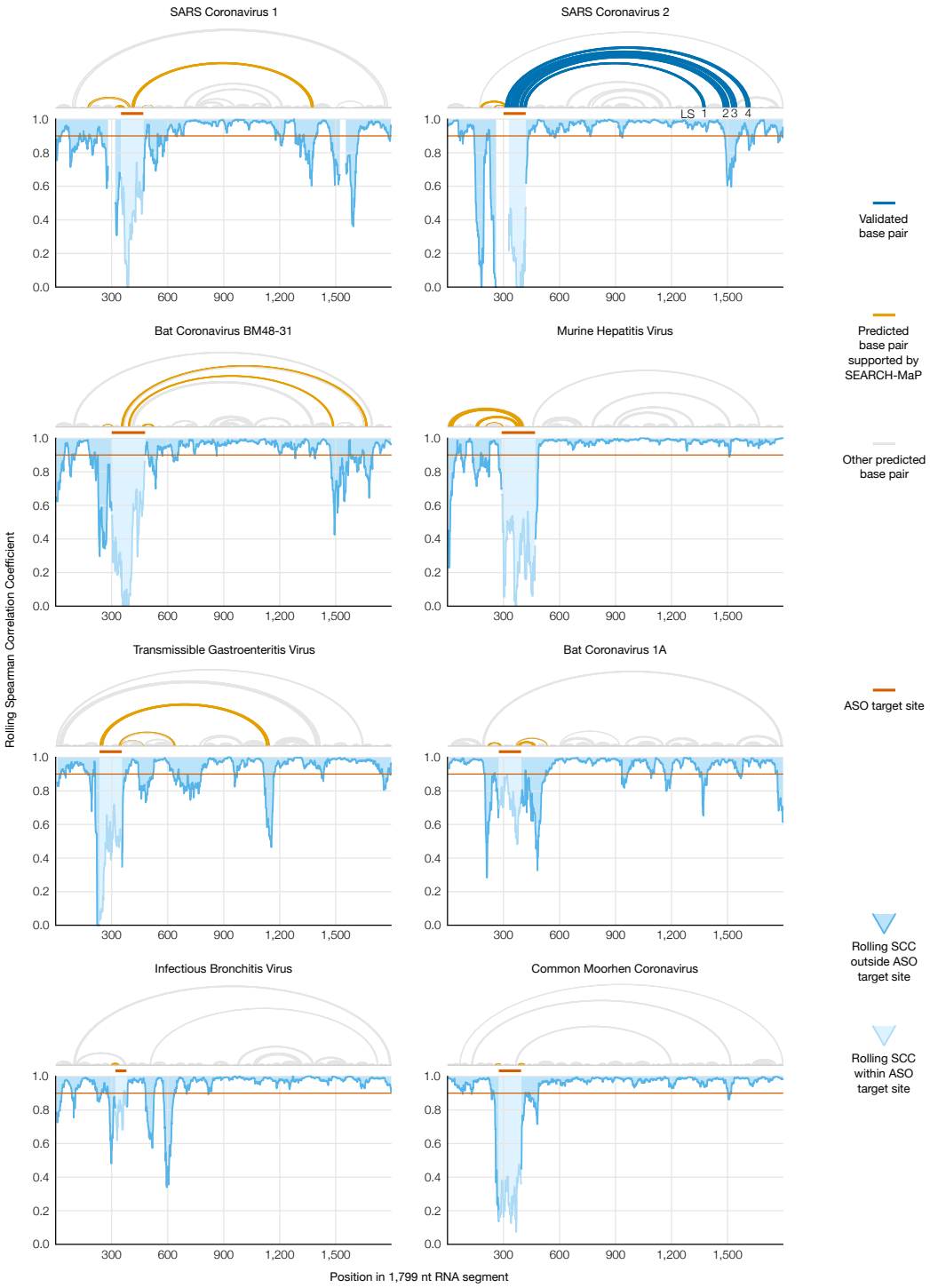

Figure 5: **Evidence for long-range RNA–RNA interactions involving the FSE in five coronaviruses.** Rolling (window = 45 nt) Spearman correlation coefficient (SCC) of DMS reactivities between the +ASO and no-ASO samples for each 1,799 nt segment of a coronaviral genome. The target site of each ASO is 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 4a) is also drawn, with LS1-LS4 labeled.

To confirm we could detect long-range interactions, we compared the rolling SCC for the SARS-CoV-2 segment to our refined model of the long-range interaction (Figure 5, 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 base pairs.

We found similar long-range interactions 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 this multi-stemmed long-range interaction involving the FSE, hinting that this structure is functional.

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 DMS reactivities of clusters corresponding to long-range interactions would generally yield more accurate predictions of long-range interactions than would ensemble averages (over all structural states). 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 gas-

troenteritis virus (Figure 5, orange). We conclude that long-range interactions involving the FSE occur more widely than in just SARS-CoV-2, including in the genus *Alphacoronavirus*.

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 RNA–RNA interaction, 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) (Figure 6a). 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 interaction in live virus would be necessary (Supplementary Figure **??**).

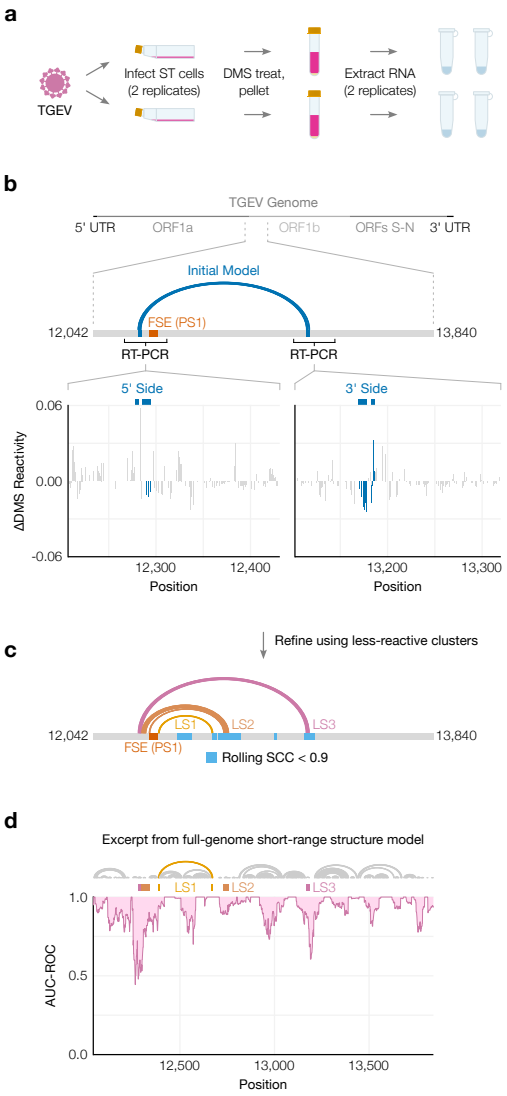

Figure 6: **Structure of the TGEV genome. (a)** Schematic of the experiment in which two biological replicates of ST cells were infected with TGEV, DMS-treated, and pelleted. Cell pellets were divided into two technical replicates prior to extraction of DMS-modified RNA.

To determine the structure ensembles, we clustered the reads from the amplicons on both sides of the long-range interaction into two clusters. The two 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 consistently lower DMS reactivities in one cluster, and hypothesized that this cluster corresponded to the long-range interaction (Figure). To investigate, we refined the structure of the 1,799 nt segment using the DMS reactivities from both less-reactive clusters. Consistent with our hypoth-

esis, the minimum free energy (MFE) model included the long-range interaction, while the MFE model re-predicted using both more-reactive clusters did not (Figure). Both models based also featured prominent new interaction connecting 20 nt upstream of the FSE with 400 nt downstream – which likely exists because it coincides with a prominent dip in SCC at positions 661-779 in the 1,799 nt segment of TGEV (Figure 5).

We used the ensemble average DMS reactivities to produce one "ensemble average" model of short-range base pairs (spanning up to 300 nt) in the TGEV genome (SUPPLEMENTARY FIGURE). 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). This structure was also consistent with the DMS reactivities: the area under the receiver operating characteristic curve (AUC-ROC) was 0.94 (Supplementary Figure **??**b). We had found the first roughly 500 nt of the SARS-CoV-2 genome in Vero cells was also consistent with a single secondary structure (AUC-ROC = 0.98) **?**. However, for many other regions – including the FSE – a single structure consistent with the DMS reactivities could not be found, evidenced by lower AUC-ROC scores (below 0.6 for the FSE) **?**. Likewise for TGEV, we found regions with low AUC-ROC scores between the ensemble average model and DMS reactivities (SUPPLEMENTARY FIGURE). We surmised that such regions likely formed alternative structures or long-range interactions – or both – that the ensemble average model could not capture.