# Results

## Strategy of SEARCH-MaP and SEISMIC-RNA

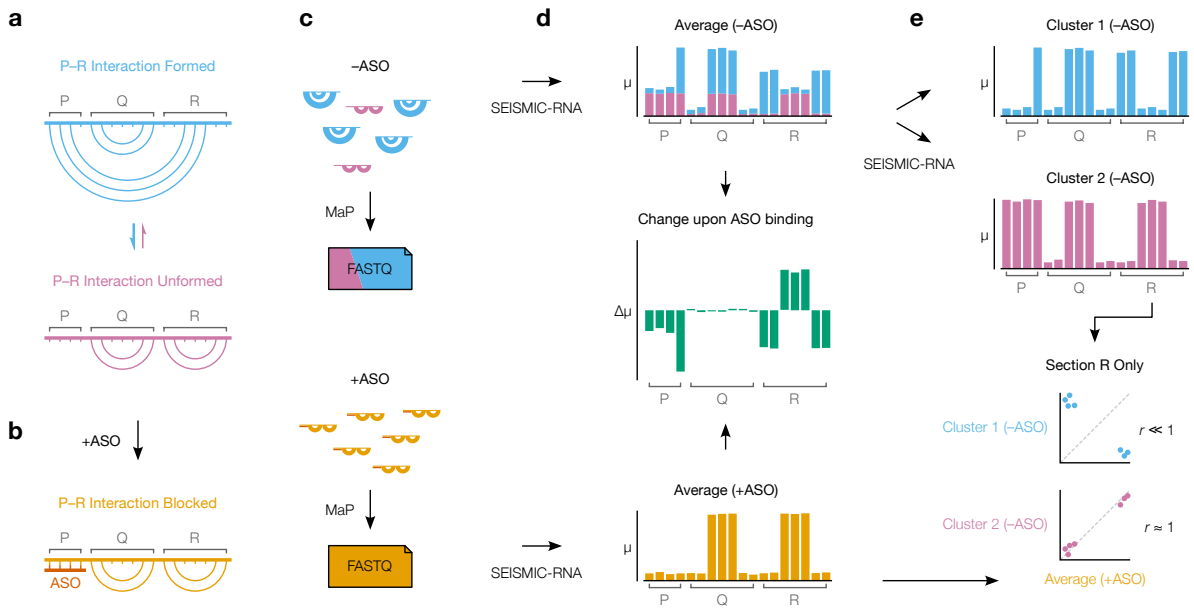


Figure 1: **The strategy of SEARCH-MaP and SEISMIC-RNA. (a)** This toy RNA is partitioned into three sections (P, Q, and R) whose molecules exist in two structural states: one in which an interaction between P and R forms (blue) and one in which it does not (purple). **(b)** Hybridizing an ASO (red) 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 ($\mu$) 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:

## SARS-CoV-2 tiling

Long-range RNA–RNA interactions in many species of virus regulate core processes such as viral protein synthesis **?**.

In SARS coronavirus 2 (SARS-CoV-2), the frameshift stimulating element (FSE) was shown to base pair with another genomic element over 1,000 nt downstream, a structure the

authors named the ”FSE-arch” **?**. We had found that about 45% of the genomic RNA molecules within infected cells have DMS mutational profiles consistent with the FSE-arch **?**, which had surprised us given the length of this RNA–RNA interaction. Therefore, we sought to investigate and potentially improve the model of the FSE-arch using SEARCH-MaP.

We *in vitro* transcribed a 2,924 nt RNA segment of the SARS-CoV-2 genome centered on the long-range interaction (Figure 3a). We added groups of DNA ASOs; each group targeted a different section of the RNA (Figure 3a). Groups 9 and 10 targeted the 3’ side of the FSE-arch; we expected that if this structure exists, then adding either group should block part of the FSE-arch and change the structure near the FSE. We confirmed the ASOs bound using DMS-MaPseq (SFIG). We then assessed the structure near the FSE via DMS-MaPseq with RT-PCR primers flanking the FSE, including the entire 5’ side of the FSE-arch.

The mutational profiles of the FSE region with no ASOs were highly reproducible: the Pearson correlation coefficient (PCC) between two replicates was 0.98 (Figure 3b, light gray). Binding ASO group F (targeting the FSE itself) plunged the correlation with the no-ASO control to 0.55 (Figure 3b, dark gray), confirming that we could detect ASO-induced structural changes around the FSE. Of the other ASO groups, only the addition of group 9 dropped the correlation with the no-ASO control below 0.90 (Figure 3b). This drop in correlation localized to the stems predicted to be part of the FSE-arch (SFIG, on rolling correlation). This result supports the inner two stems of the FSE-arch. That adding ASO group 10 had no effect on the FSE (PCC = 0.97) suggests that the outer stem either does not exist or forms less often or under more specific conditions than do the inner two stems.

We next sought to determine in what fraction of molecules the inner two stems of the FSE-arch fold. We clustered the reads for the no-ASO control using SEISMIC-RNA and found that they form at least two distinct clusters. These clusters were consistent with our previous data in Vero and Huh-7 cells **?** (SFIG), showing that this RNA segment adequately models of the RNA structure in the full-length virus. We compared the mutational profile

of each cluster to that of the ensemble average after adding ASO group 9 (Figure 3c, top). Cluster 2 (57% of the ensemble) was very similar (PCC = 0.95), suggesting that this cluster corresponds to the FSE-arch unformed. Cluster 1 (43% of the ensemble) was distinct (PCC = 0.64), suggesting that it corresponds to the the FSE-arch formed.

To gain further support for these assignments, we took advantage of having a preexisting model of the FSE-arch **?**. If these assignments were true, then the mutational profile of Cluster 1 should agree well with the structure of the FSE-arch (i.e. paired and unpaired bases should have low and high mutation rates, respectively); and Cluster 2 should agree less. We assessed the agreement by constructing a receiver operating characteristic (ROC) curve with respect to the inner two stems of the preexisting model (Figure 3c, bottom). The area under the curve (AUC) for Cluster 1 was 1.0, indicating perfect agreement with the inner two stems of the FSE-arch, while that of Cluster 2 (AUC = 0.57) was only marginally better than the null expectation of 0.50. This result supports that Cluster 1 is the mutational profile when the inner two stems of the FSE-arch form, and Cluster 2 when they do not.

If the FSE-arch exists, then blocking the 5' side of the FSE-arch should also alter the structure of the 3' side. We investigated by RT-PCRing the region surrounding the inner two stems of the FSE-arch on the 3' side with and without adding ASO group F, which targets the FSE. Similar to the previous result, we found that the 3' side of the FSE-arch also forms two clusters of roughly even proportions (Figure 3d). The mutational profile (ignoring one outlier) of Cluster 2 resembled blocking the FSE-arch (PCC = 0.95), while Cluster 1 did not (PCC = 0.80); and Cluster 1 agreed with the FSE-arch model (AUC = 1.0), while Cluster 2 did not (AUC = 0.57). Thus, the 3' side of the FSE-arch also appears to form two structural states, one of which corresponds to the long-range interaction forming.

We conclude that this 2,924 nt segment mimics the FSE-arch in cells and exists as a structure ensemble in which the inner two stems of the FSE-arch fold in 47% ± 4% of the molecules (Figure 3e). As this structure folds *in vitro*, it depends on the RNA sequence itself,

not on proteins or other cellular/viral factors. Moreover, we generated a mutational profile of the formed and the unformed states on both sides of the FSE-arch, which could assist with structure modeling.

## The frameshifting pseudoknot competes with the long-range interaction

We refined the model of the long-range interaction using the mutational profiling data from both clusters that corresponded to this interaction forming (Figure 3c, d). We predicted the structure of a 1,799 nt segment of the genome centered on the long-range interaction using RNAstructure Fold **?**. The minimum free energy structure (Figure 4) contained not only the two inner stems of the FSE-arch (LS1 and LS2a/b) but also two additional stems that were not part of the original FSE-arch model **?** (LS3a/b and LS4). The structure also contained the alternative stem 1 (AS1) that we had previously discovered **?**, which encircles the attenuator hairpin (AH) **?**. Because we had collected DMS-MaPseq data on both ends of LS3a/b, but only the 5' end of LS4, we focused on LS3a/b.

To our surprise, LS2b, LS3a/b, and LS4 of the new model penetrate into structures within the FSE pseudoknot that is widely thought to stimulate frameshifting – LS2b overlaps with PS2, LS3 with PS3, and LS4 with PS1. We had previously shown that AS1 also overlaps with and seems to outcompete PS1 **?**. If these stems existed, then they would be mutually exclusive with the FSE pseudoknot, which suggests that the long-range interaction could inhibit the formation of the pseudoknot and possibly regulate the rate of frameshifting.

Thus, we verified these structures of the long-range interaction using SEARCH-MaP. We performed SEARCH-MaP on the same 1,799 nt segment for which we had predicted the structure, this time with shorter LNA/DNA mixmer ASOs (15-20 nt) to reach single-stem precision. Each ASO targeted a single 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 (Figure 4b). Binding an ASO to the 3' side of LS2b caused a larger perturbation with more off-target effects, likely because this stem overlaps with stem 2 of the pseudoknot (PS2), so blocking it with an ASO could promote pseudoknot formation. 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. These results show that stems LS1, LS2a/b, and LS3b in the refined model of the long-range interaction do exist and can be detected with SEARCH-MaP – cleanly if the stem does not interact with anything else, otherwise with off-target effects.

Assuming that all stems in Figure 4a do exist, we found that there are six possible structure models resulting from all possible combinations of non-overlapping stems (Figure 4c). We ordered these models from most long-range character (A) to most pseudoknot-like character (F). We then determined whether each model actually exists in the ensemble and estimated its proportion.

We found that our no-ASO control clustered reproducibly up to 6 clusters (SFIG). For each cluster, we found how well it agreed with each structure model by calculating the area under the receiver operating characteristic curve (AUC-ROC). We graph the results as a heatmap with clusters on the x-axis, in order of and their widths indicating their proportions in the ensemble, and the models on the y-axis. We consider a cluster and model to be consistent with each other if the AUC-ROC is at least 0.90, and explicitly print all such AUC-ROC values. We found that model C – in which AS1 folds along with stems PS2 and PS3 of the pseudoknot – was consistent with three clusters representing 52% of the ensemble (Figure 4d, top). Model A – where the full long-range interaction forms – was consistent with one cluster (20%). No clusters were consistent with the pseudoknot (Model F); the least-abundant cluster (7%) came close with an AUC-ROC of 0.88. The remaining cluster

(21%) was not even close to being consistent with any model, suggesting that there are still other structures in the ensemble besides those in Figure 4c.

These results suggested to us that the primary competitors of the pseudoknot are AS1 and LS2b, since both are present in the most abundant model, C (52%), while LS3 and LS4 fold in model A, which is only 20% of the ensemble. We thus reasoned that if the long-range interaction does actually compete with the pseudoknot, then blocking AS1 and LS2b simultaneously should allow the pseudoknot to fold, while blocking either would have a much smaller effect on the pseudoknot. To test this hypothesis, we repeated the above experiment while adding an ASO targeting either AS1 or just the part of LS2b that overlaps with the pseudoknot, or both ASOs together. Blocking AS1 (Figure 4d, left) reduced the proportion of clusters consistent with AS1 (Models A, B, and C) from 72% to 16%, as expected; it also resulted in two clusters consistent with the pseudoknot (56% total) and one cluster (20%) consistent with Model D, which includes PS1. Blocking the part of LS2b that seems to overlap with PS2 (Figure 4d, right) eliminated Model A but not Model C (as expected, since Model A includes LS2b, while Model C does not), and also produced one cluster (13%) that was consistent with the pseudoknot. Blocking both AS1 and LS2b together (Figure 4d, bottom) forced the entire ensemble into the pseudoknot state, with 87% of the ensemble highly consistent with Model F (AUC-ROC = 0.97) and the remaining 13% still somewhat consistent (AUC-ROC = 0.87). Thus, we conclude that the long-range interaction – particularly LS2b – along with AS1, does compete with the pseudoknot.

## Frameshift stimulating elements of multiple coronaviruses form long-range RNA–RNA interactions

We hypothesized that similar long-range interactions could exist in other coronaviruses – particularly other SARS-related viruses. To test this hypothesis, we performed SEARCH-MaP with FSE-targeted ASOs on 1,799 nt segments from eight selected coronaviruses.

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

As of December 2021, the NCBI Reference Sequence Database **?** contained sixty-two 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 (including SARS-CoV-2) for further study – at least one from each genus (SFIG). Within the genus *Betacoronavirus*, we included all three of the 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, distinct from all other coronaviruses. 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, then removing that element by truncating the RNA would break the interaction, causing a structural change in the FSE that could be detected though chemical probing. 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 se-

quences and a 1.8 kb 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 1.8 kb segments (SFIG), suggesting long-range interactions between the FSE and another element within the 1.8 kb segment.

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

To determine whether the FSE interacts with another RNA element – and if so, which – in each coronavirus, we performed SEARCH-MaP on the 1.8 kb RNA segment using ASOs targeting the bases that changed the most between the 239 nt and 1.8 kb segments. Using SEISMIC-RNA, we computed the mutational profiles with and without ASOs and calculated the Spearman correlation coefficient (SCC) between them via a sliding window (Figure 5). In every coronavirus, the ASO target site (light green) showed a large dip in SCC, confirming that the ASOs bound and altered the structure.

The long-range interaction in SARS-CoV-2 was suggested to comprise three stems **?**. SEARCH-MaP found two (dips in SCC around positions 1,500 and 1,615); the third stem (around position 1,375) may have been missed because it is not only the shortest (15 nt) but also has just one unpaired base on each side. Similar long-range interactions appeared in the other SARS-related viruses: SARS-CoV-1 and bat coronavirus BM48-31. Like SARS-CoV-2, both had a dip in SCC near position 1,500 and another dip downstream, between positions 1,575 and 1,675. The SCC also dipped around position 1,375 for SARS-CoV-1, matching the location of the missed stem in SARS-CoV-2. Therefore, this three-stemmed long-range interaction involving the FSE appears to be conserved among three strains of SARS-related viruses, suggesting it is functional.

Examining the other species, we found prominent dips in SCC below 0.9 further than 200 nt from the ASO target site in all except common moorhen coronavirus. We modeled

potential structures for these long-range interactions using the Fold program from RNAstructure **?** with the no-ASO mutational profiles as DMS constraints **?**. The computationally predicted structures for murine hepatitis virus and transmissible gastroenteritis virus were consistent with the SEARCH-MaP data (Figure 5). We subsequently focused on transmissible gastroenteritis virus because its long-range interaction lay downstream of the FSE, as in the SARS-related viruses.

To verify that the long-range interaction also forms in live transmissible gastroenteritis virus (TGEV), we infected ST cells with TGEV (two biological replicates) and performed DMS-MaPseq (two technical replicates per biological replicate). The DMS reactivities were highly reproducible over the whole TGEV genome ($r$ = 0.96-0.97, SFIG). As expected, they differed from those of the 1.8 kb segment *in vitro* ($r$ = 0.82, SFIG), showing why it is necessary to verify the long-range interaction in TGEV-infected cells.

First, to determine whether the FSE and the region with which it may interact form alternative structures, we amplified and deeply sequenced these two regions from each sample. Clustering the reads using SEISMIC-RNA revealed that both regions adopt at least two alternative structures. The two clusters of the downstream region differed most around positions 1,120-1,140 – the site of the 3' end of the predicted long-range interaction. In cluster 1 (63% of the ensemble), bases 1,129-1,136 (all part of the predicted interaction) had DMS reactivities less than 0.01; while in cluster 2, the DMS reactivities were all greater than 0.01. This result suggests that cluster 1 corresponds to the state in which the long-range interaction forms.

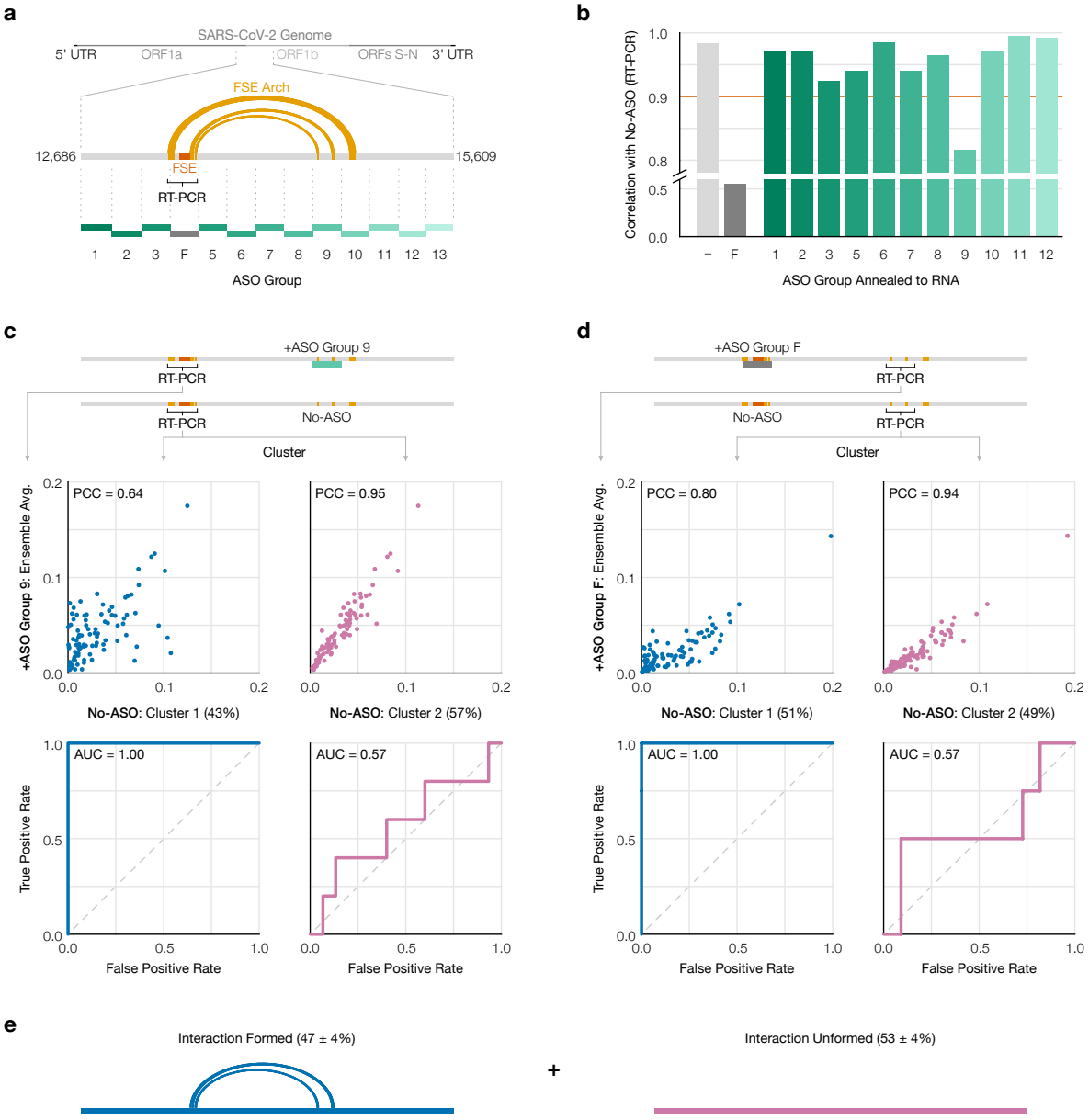
a
SARS-CoV-2 Genome
5' UTR
ORF1a
ORF1b
ORFs S-N
3' UTR
FSE Arch
12,686
15,609
FSE
RT-PCR
1 2 3 F 5 6 7 8 9 10 11 12 13
ASO Group
b
Correlation with No-ASO (RT-PCR)
1.0
0.9
0.8
0.5
0.0
– F 1 2 3 5 6 7 8 9 10 11 12
ASO Group Annealed to RNA
c
+ASO Group 9
RT-PCR
RT-PCR
No-ASO
Cluster
+ASO Group 9: Ensemble Avg.
PCC = 0.64
PCC = 0.95
No-ASO: Cluster 1 (43%)
No-ASO: Cluster 2 (57%)
AUC = 1.00
AUC = 0.57
True Positive Rate
False Positive Rate
d
+ASO Group F
RT-PCR
No-ASO
RT-PCR
Cluster
+ASO Group F: Ensemble Avg.
PCC = 0.80
PCC = 0.94
No-ASO: Cluster 1 (51%)
No-ASO: Cluster 2 (49%)
AUC = 1.00
AUC = 0.57
True Positive Rate
False Positive Rate
e
Interaction Formed (47 ± 4%)
+
Interaction Unformed (53 ± 4%)

Figure 3:

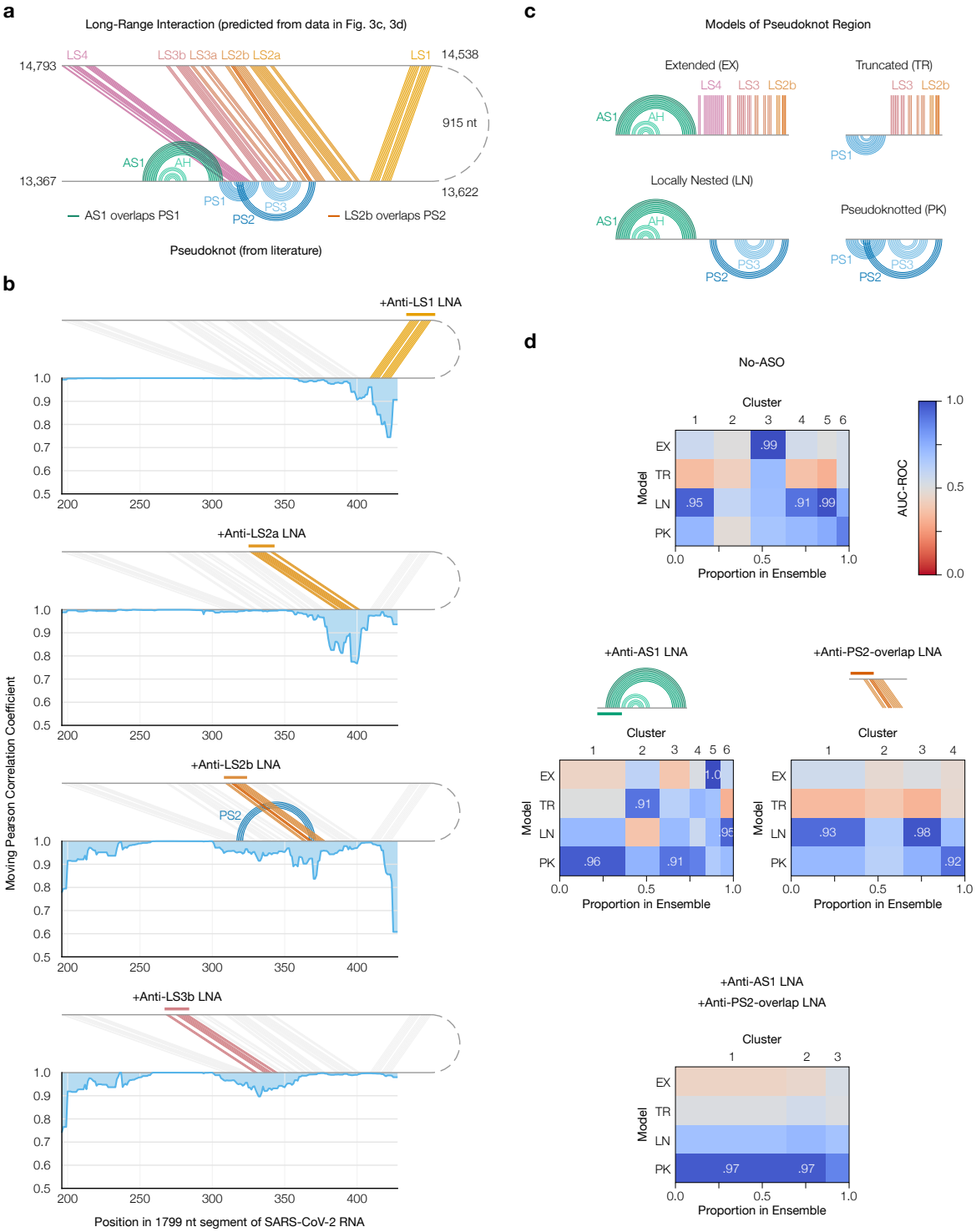
a
Long-Range Interaction (predicted from data in Fig. 3c, 3d)
14,793
LS4
LS3b LS3a LS2b LS2a
LS1
14,538
915 nt
AS1
AH
13,367
PS1
PS3
PS2
13,622
AS1 overlaps PS1
LS2b overlaps PS2
Pseudoknot (from literature)
b
+Anti-LS1 LNA
+Anti-LS2a LNA
+Anti-LS2b LNA
PS2
+Anti-LS3b LNA
Moving Pearson Correlation Coefficient
Position in 1799 nt segment of SARS-CoV-2 RNA
c
Models of Pseudoknot Region
Extended (EX)
Truncated (TR)
Locally Nested (LN)
Pseudoknotted (PK)
d
No-ASO
Cluster
Model
EX
TR
LN
PK
Proportion in Ensemble
AUC-ROC
+Anti-AS1 LNA
+Anti-PS2-overlap LNA
+Anti-AS1 LNA
+Anti-PS2-overlap LNA

Figure 4:

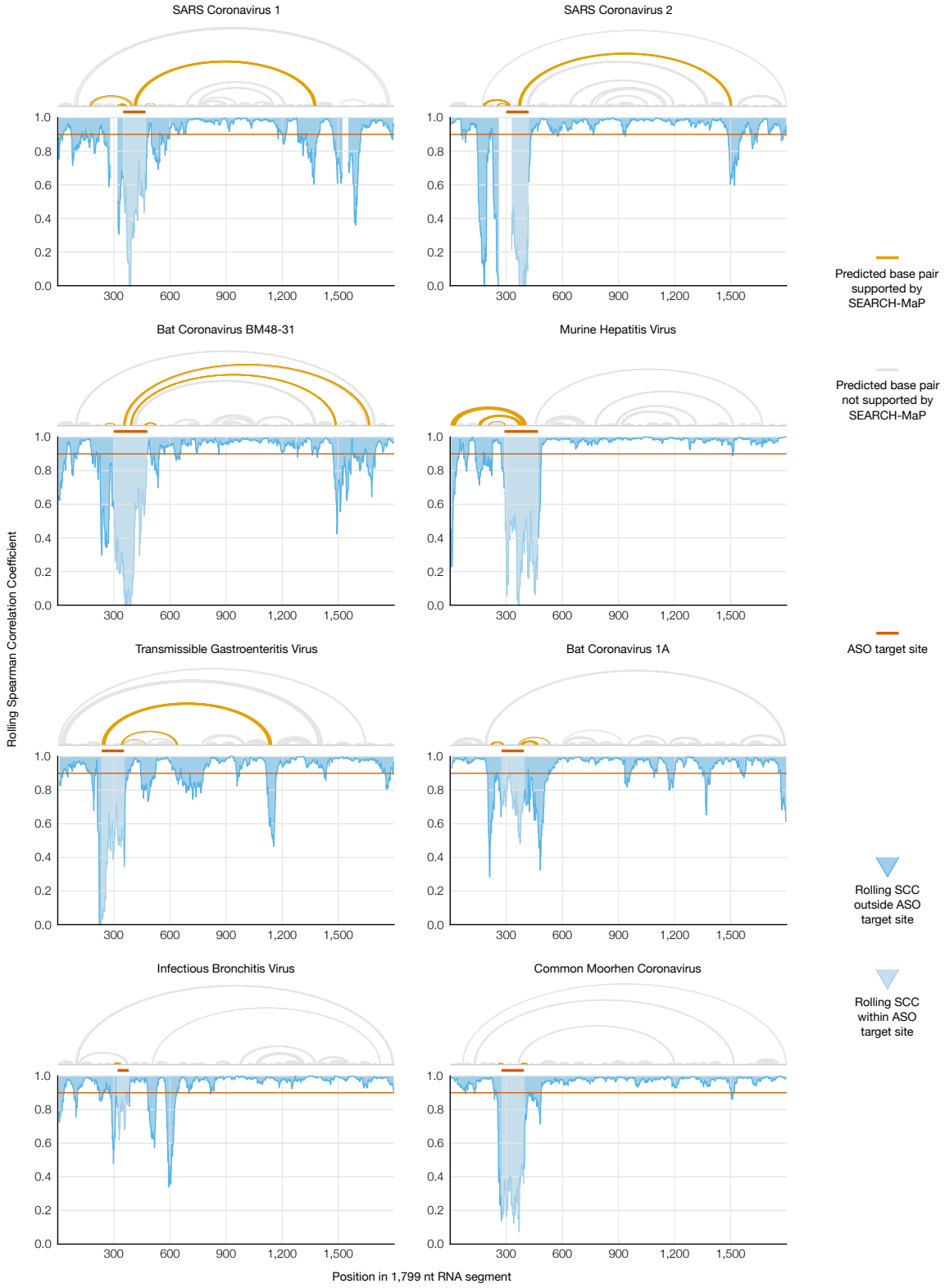
SARS Coronavirus 1
SARS Coronavirus 2
Bat Coronavirus BM48-31
Murine Hepatitis Virus
Transmissible Gastroenteritis Virus
Bat Coronavirus 1A
Infectious Bronchitis Virus
Common Moorhen Coronavirus
Rolling Spearman Correlation Coefficient
Position in 1,799 nt RNA segment
Predicted base pair supported by SEARCH-MaP
Predicted base pair not supported by SEARCH-MaP
ASO target site
Rolling SCC outside ASO target site
Rolling SCC within ASO target site

Figure 5: