Introduction

The emergence of coronavirus disease 2019 (COVID-19) as a pandemic in 2020 spurred many investigations on functional RNA structures in coronaviruses, particularly SARS coronavirus 2 (SARS-CoV-2) [**? ? ? ? ? ? ? ? ? ?**]. Among the more unexpected findings was an RNA:RNA interaction between the frameshifting stimulation element (FSE) and another sequence up to 1,475 nt downstream, which the authors named the FSE-arch [**?**]. The FSE-arch was detected in infected cells using COMRADES [**?**] and proposed to comprise three nested long-range RNA:RNA interactions (Figure **??**a): an outer 38 bp bulged stem spanning coordinates 13,370-14,842 (which encompasses the FSE); a middle 18 bp bulged stem spanning coordinates 13,533-14,673; and an inner 14 bp bulged stem spanning coordinates 13,580-14,552 [**?**]. We had discovered that the FSE folds into at least two alternative structures in infected cells, in roughly equal proportions, and that the predicted structure for one of them resembles the FSE-arch [**?**]. Because computational RNA structure prediction – even guided by chemical probing data – is unreliable for long RNA sequences especially [**?**], we sought stronger, hypothesis-driven evidence for the existence of the FSE-arch.

Chemical probing followed by mutational profiling is a common strategy for inferring secondary structures of RNA molecules [**? ?**].

Here, we present a method to probe RNA–RNA interactions spanning hundreds to thousands of nucleotides, "Structure Ensemble Ablation by Reverse Complement Hybridization with Mutational Profiling" (SEARCH-MaP). To compute, compare, and deconvolute data from mutational profiling experiments (including SEARCH-MaP, DMS-MaPseq, and SHAPE-MaP), we introduce the software "Structure Ensemble Inference by Sequencing, Mutation Identification, and Clustering of RNA" (SEISMIC-RNA).

Results

Strategy of SEARCH-MaP and SEISMIC-RNA

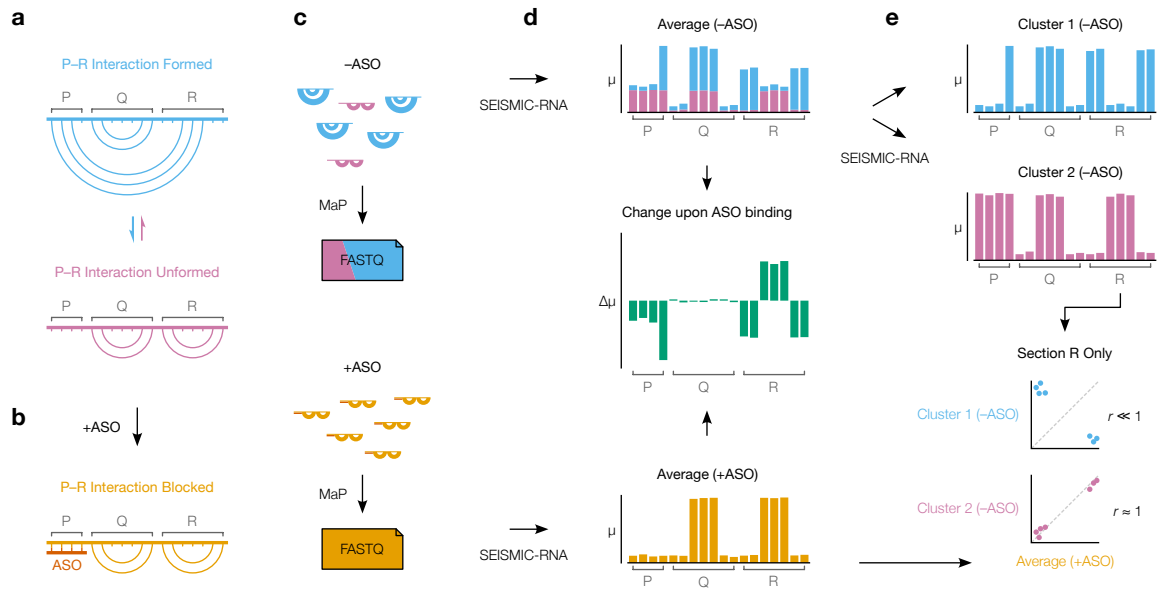

Figure 1: **The strategy of SEARCH-MaP and SEISMIC-RNA. (a)** This toy RNA is partitioned into three sections (P, Q, and R) whose molecules exist in two structural states: one in which an interaction between P and R forms (blue) and one in which it does not (purple). **(b)** Hybridizing an ASO (red) 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 (at the 5' side of the FSE-arch) would be perturbed by – and only by – ASOs targeting the 5' or 3' 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. 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. Every ASO group bound properly, evidenced by suppression of DMS reactivities over its target site (SFIG).

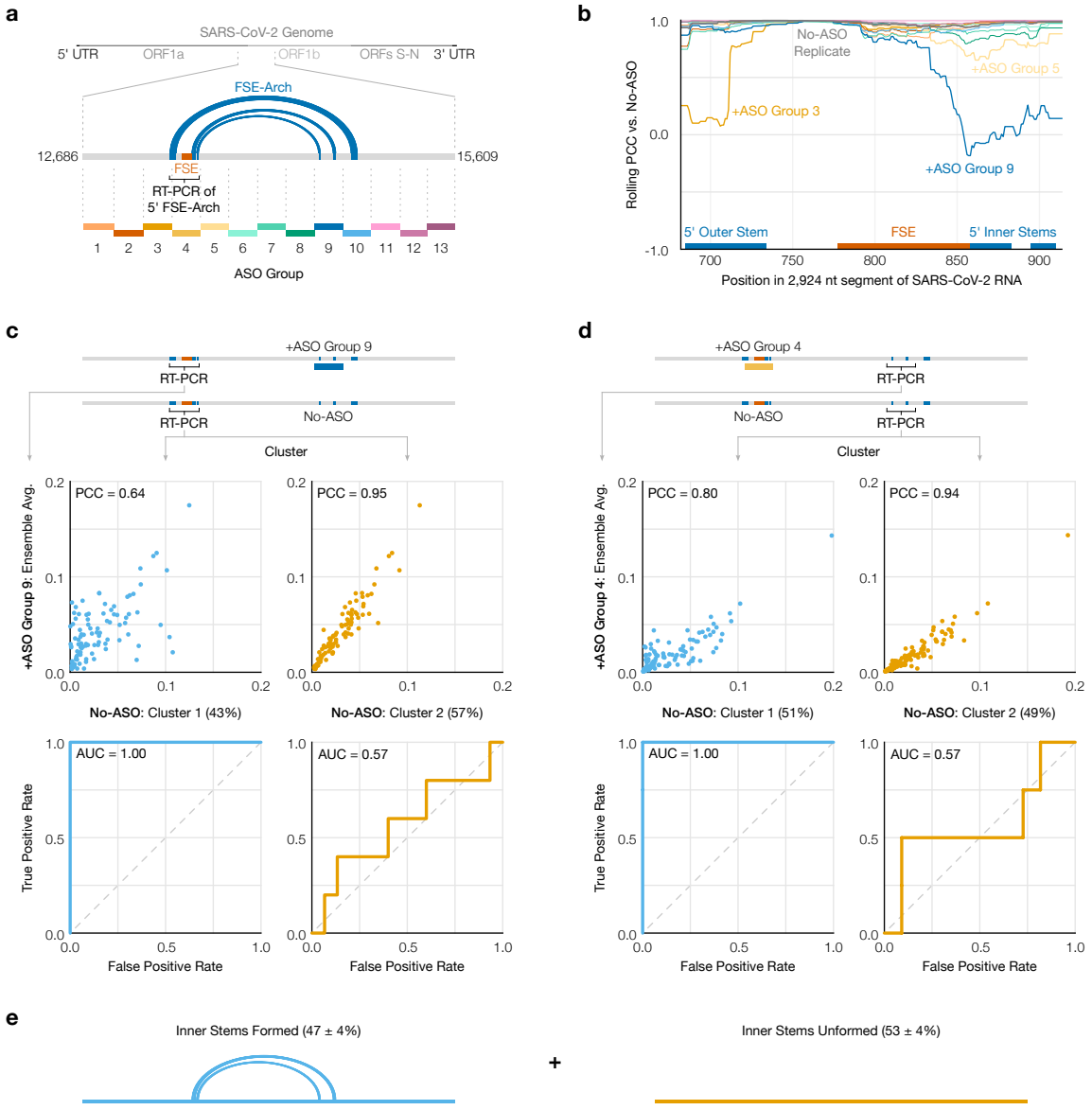

Figure 3: **Characterization of a long-range RNA–RNA interaction in SARS-CoV-2.** **(a)** The 2,924 nt segment of the SARS-CoV-2 genome containing the frameshift stimulation element (FSE) and putative FSE-arch [**?**]. The target of each ASO group is indicated by dotted lines; lengths are to scale. **(b)** Rolling (window = 45 nt) Pearson correlation coefficient (PCC) of the DMS reactivities between each sample and a no-ASO control over the 5' FSE-arch. Each curve represents one sample, colored as in (a); ASO groups 4 and 13 are not shown. The 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. **(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 sample and a no-ASO control via a sliding window of 45 nt (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 near the 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. 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 distinct clusters, with a 43%/57% split. To determine if these clusters 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 verify this correspondence, we leveraged the preexisting model of the FSE-arch [**?**]. If cluster 1 did correspond to the two inner stems formed, then 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 really 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 un-

formed 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 and found – similar to the previous result – that cluster 1 (51%) matched the structure of the inner two stems (AUC = 1.00), while cluster 2 (49%) agreed with the DMS reactivities (PCC = 0.94) after blocking the FSE-arch with ASO group 4 (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). This result is consistent with our previous finding that an inner stem of the FSE-arch constitutes 45% of the ensemble in cells infected with SARS-CoV-2 [**?**].

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

We used the DMS reactivities of the interaction-formed clusters on both sides (Figure 3c, d) to refine the model of 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 participate in 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 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, 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 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 other 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.

To confirm we could detect long-range interactions, we compared the rolling SCC for the SARS-CoV-2 segment to our model of the long-range interaction (Figure 4a, green). The SCC dipped below 0.9 at positions 1,483-1,560 and at 1,611-1,642, which coincide with stems LS2a-LS3b (positions 1,476-1,550 within the 1,799 nt segment) and stem LS4 (positions 1,600-1,622) of the long-range interaction. 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 LS2a-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, suggesting 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 mutational profiles as DMS constraints [**?**]. Predictions based on ensemble averages (of all structural states) generally underperform those based on clustered mutational profiles (of fewer states or one state); the ensemble average prediction for SARS-CoV-2 included LS1 and LS2b but missed the other long-range stems. Nevertheless, we found long-range base pairs consistent with the SEARCH-MaP data for both murine hepatitis virus and transmissible gastroenteritis virus (Figure 5, orange). We conclude that long-range interactions involving the FSE occur more widely than in just SARS-CoV-2, including in *Alphacoronavirus* species.

We To verify that the long-range interaction also forms in live transmissible gastroenteritis virus (TGEV), we treated TGEV-infected ST cells with DMS (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.

Discussion

In this work, we developed SEARCH-MaP and SEISMIC-RNA and applied them jointly to detect structural ensembles involving long-range RNA:RNA interactions in SARS-CoV-2 and other coronaviruses. This study is certainly not the first to perturb RNA structure with ASOs, nor even the first to use DMS-MaPseq to quantify the structural changes upon binding ASOs to SARS-CoV-2 RNA [**?**]. But while this previous study examined local structural perturbations caused by binding an ASO, we show that we can detect changes in the structure at more distant locations in an RNA molecule that interact with the nucleotides bound by an ASO.

A previous study detected two long-range RNA–RNA interactions in the genome of satellite tobacco mosaic virus by binding an ASO (in this case, an LNA 9-mer) to each site, followed by chemical probing [**?**]. However, SEARCH-MaP and SEISMIC-RNA go further by also determining the mutational profile and proportion of the interaction-formed and -unformed states (Figure 3c, d). With a collection of candidate structure models, these methods even reveal how adding an ASO ablates specific structures, collapsing the ensemble into one predominant structure (Figure 4d).

Many methods have been developed to find long-range (and intermolecular) RNA–RNA base pairing using crosslinking (with psoralen or a derivative), proximity ligation, and

deep sequencing [**?** **?** **?** **?**]. These methods require no prior knowledge of RNA–RNA interactions and have no limit to the length of the interactions they can detect. They do, however, suffer from several limitations including inefficient ligation [QUANTIFY, I think I read that less than 5% of molecules actually ligate] necessitating either enrichment or very deep sequencing, as well as bias towards U-rich sequences. They are not single-molecule techniques, either, meaning that although they can detect mutually exclusive base pairs, they cannot determine which specific alternative structures exist or quantify their proportions, as SEARCH-MaP/SEISMIC-RNA can. There is also no straightforward way to focus on one specific RNA–RNA interaction.

Another method based on applying many ASO ”patches” in parallel and reading out the signal with microarray probes has also recently been developed [**?**]. Like proximity ligation, this method has no limitation to the length of the interactions it could find, yet it is also not a single-molecule technique, meaning that it cannot resolve individual structures in an ensemble.

SEARCH-MaP bears conceptual similarity to another method, mutate-and-map read out through next-generation sequencing (M2-seq) [**?**]. Both involve perturbing one region of an RNA molecule (in the case of M2-seq, by pre-installing mutations through error-prone PCR) and measuring the effects on other bases in the RNA using chemical probing. The major differences are the precision and scale of the interactions identified, as well as the throughput. M2-seq can pinpoint interactions down to the resolution of a single base pair, and is thus more precise than SEARCH-MaP. However, DMS-guided RNA structure prediction can propose structure models at single-base-pair resolution, which SEARCH-MaP can validate, and in this way achieve single-base-pair resolution. SEARCH-MaP is also capable of finding interactions over a much longer range because M2-seq requires the interacting bases to be in the same Illumina sequencing read. Within this length limit, one M2-seq experiment can theoretically find all pairwise interactions between bases, while one SEARCH-MaP experiment can find only interactions that involve the region to which the ASOs were hybridized. M2-seq is also limited by the formation of alternative structures. Some methods, such as [CITE something by Rhiju,

maybe REEFIT] and DANCE-MaP [**?**], have been designed to work around this limitation SEARCH-MaP; however, [something by Rhiju] has [this problem], and DANCE-MaP requires extremely high sequencing depth of several million reads [MORE PRECISE]. SEARCH-MaP, by contrast, assumes from the start that the RNA may form alternative structures; for simply detecting long-range interactions, even a 5,000 read depth is sufficient coverage; and for clustering, we have found [SOME LIMIT].

Another limitation of SEARCH-MaP as presented here is that it cannot distinguish between direct and indirect interactions. If RNA segment A interacts with segment B, while B interacts with both segment A and C, then hybridizing an ASO to segment A would perturb the structure of B, which could consequentially perturb the structure of C. Hence, C would appear to interact with A, even though this interaction is indirect, through B. One possible workaround (not shown in this study) would be to mutate or hybridize an ASO to segment B, and then repeat the experiment with hybridizing an ASO to segment A. If the interaction between A and C is direct, then C should still be perturbed even when segment B is incapable of interacting with A or C. But if B mediates an indirect interaction between A and C, then disrupting B should eliminate the apparent interaction between A and C.

Functional long-range interactions up to four kilobases involving an FSE have been been found previously in two plant viruses [**? ?**]. In both cases, frameshifting required the long-range interaction, suggesting that this interaction enables negative feedback on synthesis of viral RNA polymerase [**?**]. When polymerase levels are low, the interaction would form and stimulate frameshifting, which is needed to synthesize RNA polymerase. Once the polymerase had accumulated, it would begin to replicate the genomic RNA; in its passage from the genomic 3' end to the 5' end, it would disrupt the 3' side of the long-range interaction, attenuating frameshifting and reducing synthesis of more polymerase.

However, this strategy cannot be the role, if any, of the long-range interactions in coronaviruses. Unlike in the two plant viruses, a long-range interaction is not required to stimulate frameshifting in coronaviruses: numerous studies have shown that even the isolated FSE can cause 15 - 40% of ribosomes to frameshift [**? ? ? ? ? ? ?**]. In

coronaviruses, the long-range interaction is not only unnecessary for frameshifting but also may even attenuate it, given that in SARS-CoV-2, the FSE-arch and the frameshift-stimulating pseudoknot seem to be mutually exclusive. Moreover, coronaviruses partition translation and RNA synthesis into two different cellular compartments (the cytosol and the double-membrane vesicles, respectively) [**?**], so structural changes induced by RNA polymerases would not be seen by ribosomes.

The functions of these long-range interactions involving the FSE in coronaviruses remain mysterious. However, given that they occur in multiple coronaviruses across at least two genera, it seems reasonable that they could play a role in the viral life cycle, possibly by affecting the rate of frameshifting. Further research may reveal new mechanisms of translational regulation in coronaviruses via long-range RNA:RNA interactions.

Methods

Screening coronavirus long-range interactions computationally

All coronaviruses with reference genomes in the NCBI Reference Sequence Database [**?**] were searched for using the following query:

```
refseq[filter] AND ("Alphacoronavirus"[Organism] OR
                    "Betacoronavirus"[Organism] OR
                    "Gammacoronavirus"[Organism] OR
                    "Deltacoronavirus"[Organism])
```

The complete record of every reference genome was downloaded both in FASTA format (for the reference sequence) and in Feature Table format (for feature locations). The location of the frameshift stimulating element (FSE) in each genome was estimated from the feature table, and the nearest instance of `TTTAAAC` was used as the slippery site, using a custom Python script. The 2,000 nt segment beginning 100 nt upstream of and ending

1,893 nt downstream of the slippery site was used for predicting long-range interactions involving the FSE. Genomes with ambiguous nucleotides (e.g. `N`) in this segment were discarded. For each coronavirus genome, up to 100 secondary structure models of the 2,000 nt segment were generated using Fold version 6.3 from RNAstructure [**?**] with `-M 100` and otherwise default parameters. Then, for each position, the fraction of models for the coronavirus in which the base at the position paired with any other base between positions 101 (the first base of the slippery sequence) and 250 was calculated using a custom Python script. The coronaviruses were clustered by their fraction vectors using the unweighted pair group method with arithmetic mean (UPGMA) and a euclidean distance metric, implemented in Seaborn version 0.11 [**?**] and SciPy version 1.7 [**?**]. The resulting hierarchically-clustered heatmap was examined manually to select coronaviruses based on the prominence of potential long-range interactions with the FSE (relatively large fractions far from positions 101-250).

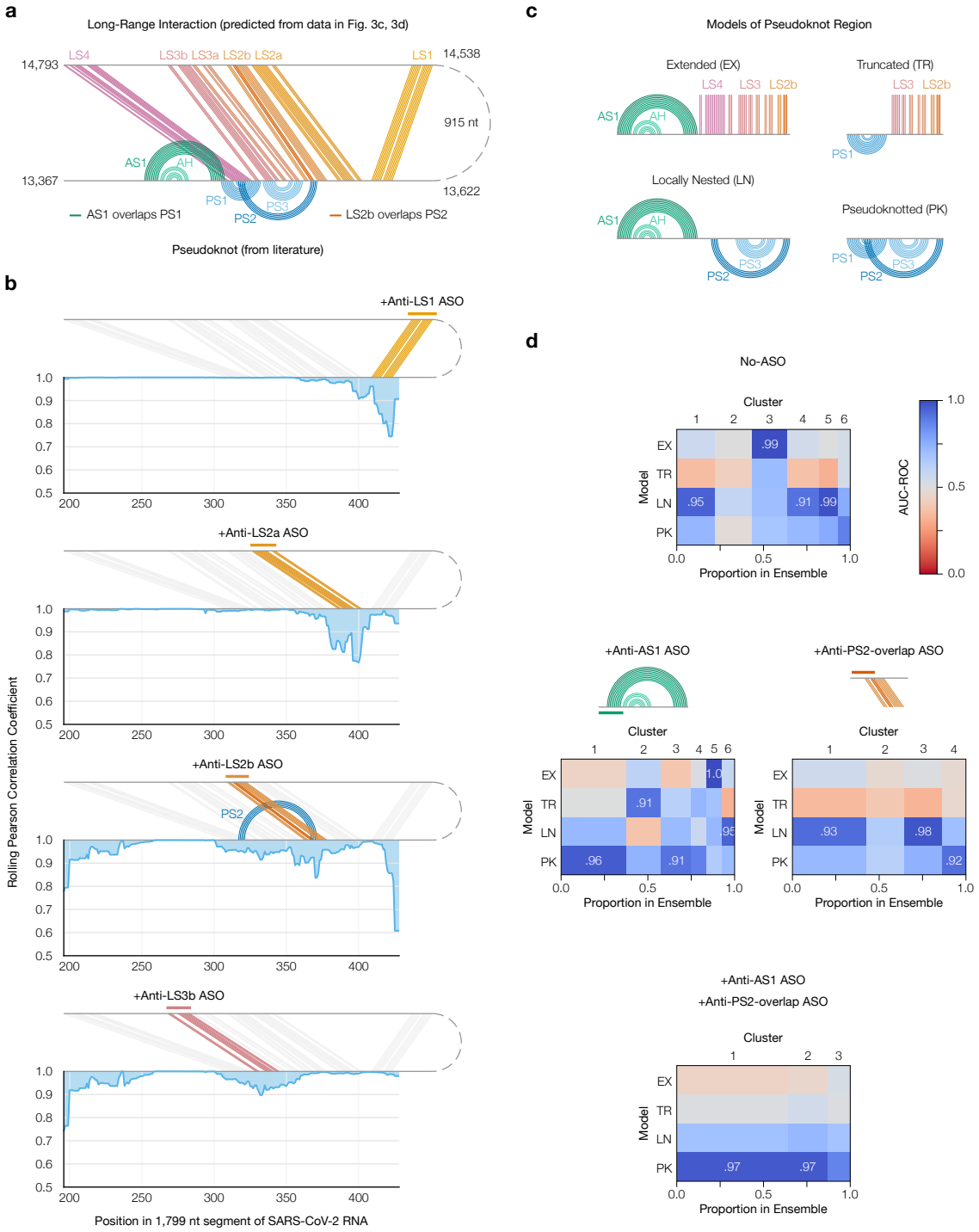
a
Long-Range Interaction (predicted from data in Fig. 3c, 3d)
14,793
LS4
LS3b LS3a LS2b LS2a
LS1
14,538
915 nt
AS1
AH
13,367
PS1
PS3
PS2
13,622
AS1 overlaps PS1
LS2b overlaps PS2
Pseudoknot (from literature)
b
+Anti-LS1 ASO
+Anti-LS2a ASO
+Anti-LS2b ASO
PS2
+Anti-LS3b ASO
Rolling Pearson Correlation Coefficient
Position in 1,799 nt segment of SARS-CoV-2 RNA
c
Models of Pseudoknot Region
Extended (EX)
LS4 LS3 LS2b
AS1
AH
Truncated (TR)
LS3 LS2b
PS1
Locally Nested (LN)
Pseudoknotted (PK)
PS3
PS2
d
No-ASO
Cluster
Model
EX
TR
LN
PK
Proportion in Ensemble
AUC-ROC
+Anti-AS1 ASO
+Anti-PS2-overlap ASO
+Anti-AS1 ASO
+Anti-PS2-overlap ASO

Figure 4:

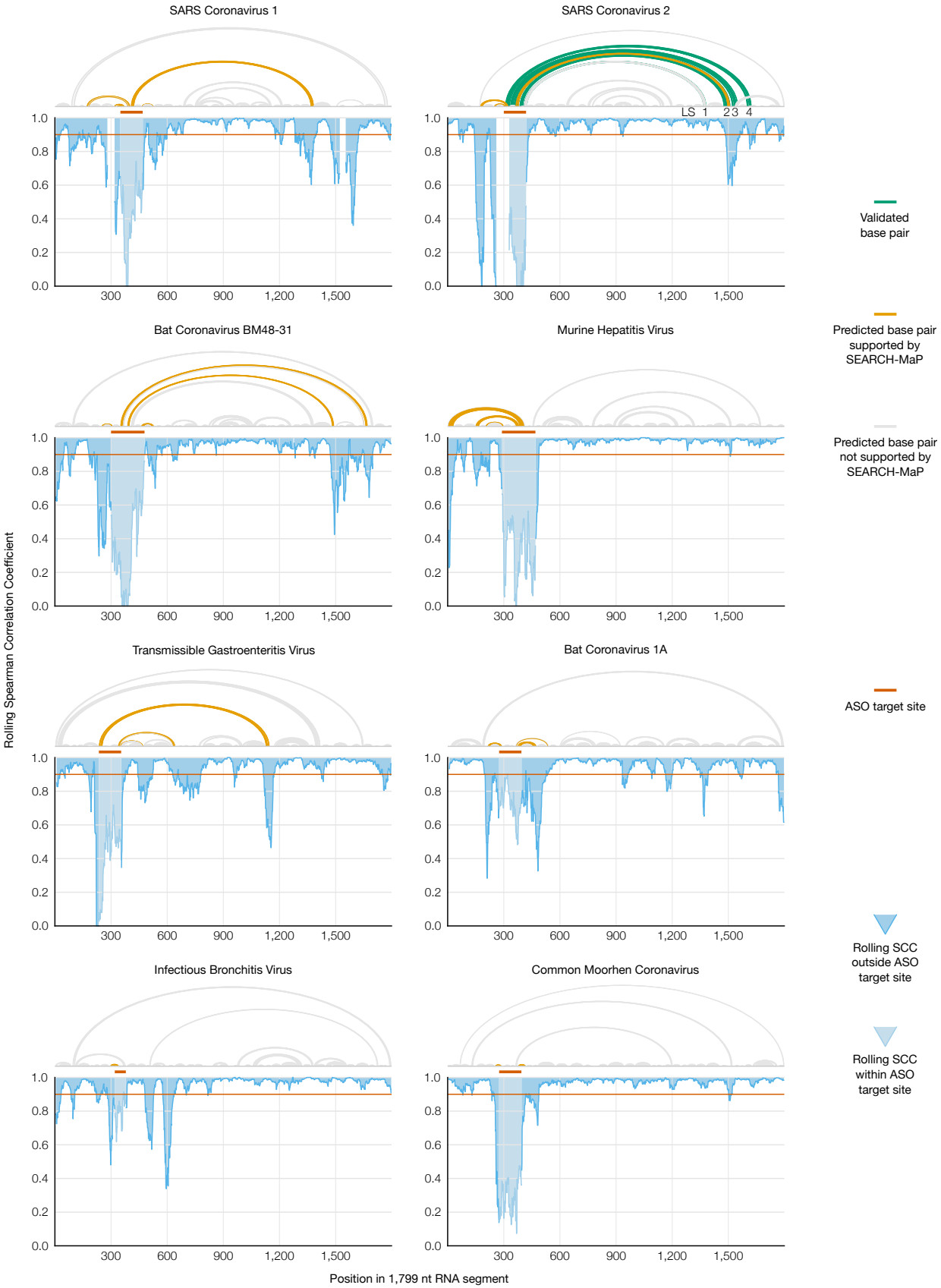
SARS Coronavirus 1
SARS Coronavirus 2
LS 1 2 3 4
Bat Coronavirus BM48-31
Murine Hepatitis Virus
Transmissible Gastroenteritis Virus
Bat Coronavirus 1A
Infectious Bronchitis Virus
Common Moorhen Coronavirus
Rolling Spearman Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
300
600
900
1,200
1,500
Position in 1,799 nt RNA segment
Validated base pair
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: