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, and Silvi Rouskin

April 17, 2024

Introduction

Across all domains of life, RNA molecules perform myriad functions in development [1], immunity [2], translation [3], sensing [4, 5], epigenetics [6], cancer [7], and more. RNA also constitutes the genomes of many threatening viruses [8], including influenza viruses [9] and coronaviruses [10]. 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) [11].

Although high-quality tertiary structures provide the most information, resolving them often proves difficult or impossible with mainstay methods used for proteins [12]. Consequently, the world's largest database of tertiary structures – the Protein Data Bank [13] – 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 [14].

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 [15]. A formalization known as the covariance model [16] underlies the widely-used Rfam database [17] 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 [18, 19]). 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 [20], messenger [21], and viral genomic [22] 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) [23, 24]. Second, without known secondary structures of many diverse RNAs, the accuracy of any prediction method cannot be properly benchmarked [21, 25]. For these reasons, and because thermodynamic-based models also tend to be less accurate for longer RNAs [22] and base pairs spanning longer distances [26], 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 [27]. 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) [28, 29]. A key advantage of MaP is that the sequencing reads can be clustered to detect multiple secondary structures in an ensemble [30, 31]. Determining the base pairs in those structures still requires structure prediction [32], although incorporating chemical probing data does improve accuracy [33, 34].

Several experimental methods have been developed to find base pairs directly, with minimal reliance on structure prediction. M2-seq [35] 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 [36], 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 [37]. 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 [38, 39], 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" [14].

Results

Workflow of SEARCH-MaP and SEISMIC-RNA

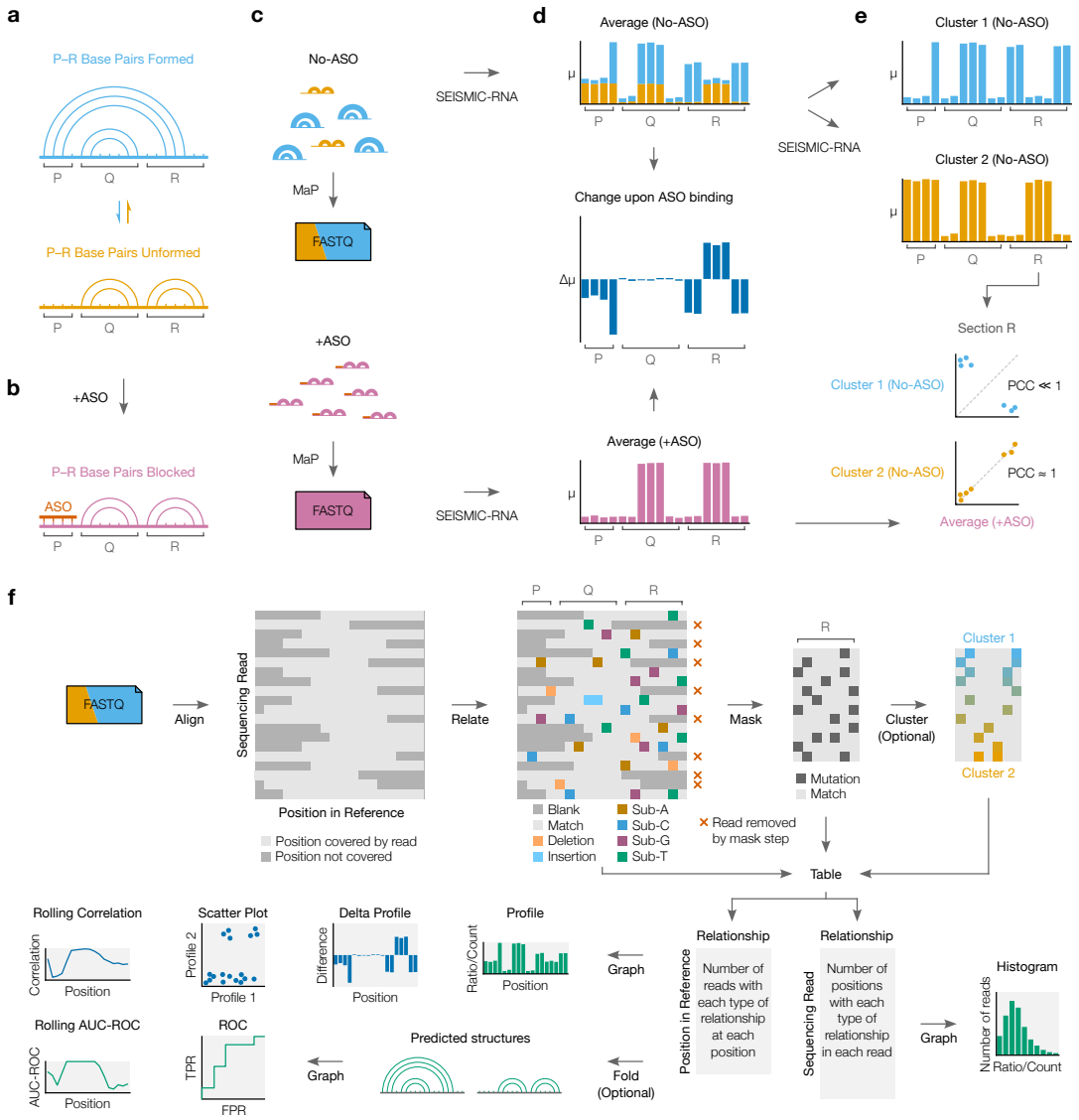


Figure 1: **The workflow of SEARCH-MaP and SEISMIC-RNA.** (Continued on next page.)

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 1a). 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 1b). 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 [29] (Figure 1c).

SEISMIC-RNA can detect base pairs by comparing the +ASO and no-ASO mutational profiles. Theoretically, each structure has its own mutational profile [40], but the mutational profile of a single structure is not directly observable because all structures are physically mixed during the experiment (Figure 1c, 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 1d, 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 1d, middle). However, the ASO has little effect on section Q because this section does not base-pair with P (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.

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 1e). 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 [15] and selected two known stems. For each stem, we designed two ASOs, one targeting each side.

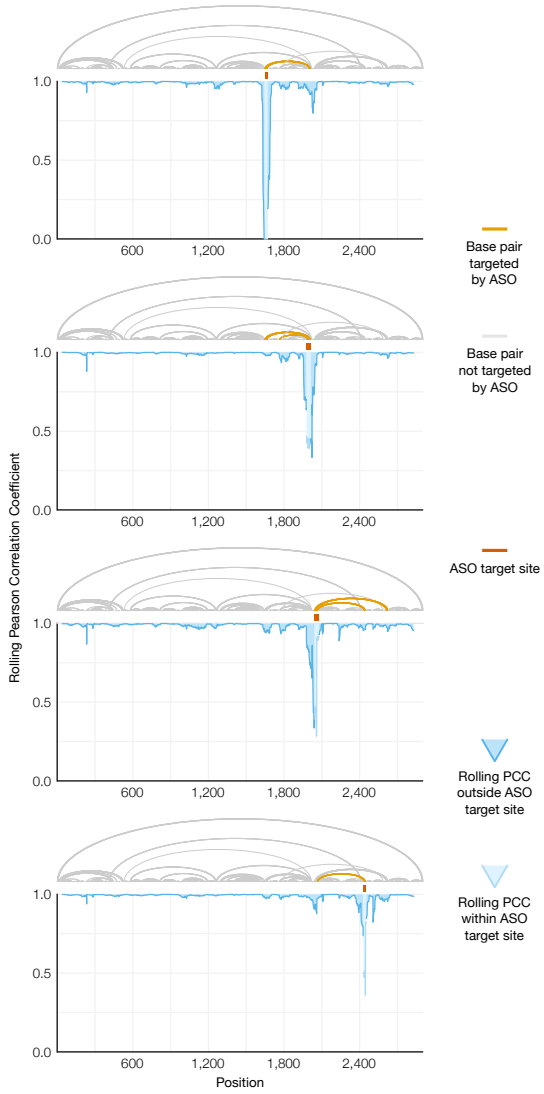


Figure 2: **Validation of SEARCH-MaP on 23S ribosomal RNA from *E. coli*.** Each graph shows the rolling (window = 45 nt) Pearson correlation coefficient (PCC) between the rRNA to which one ASO was added and a no-ASO control. The known secondary structure of the 23S rRNA [15] is drawn above the graph; 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 2). 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. Con-

versely, 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 [41]. Coronaviruses regulate translation of their first open reading frame (ORF1) using programmed ribosomal frameshifting [42]. 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 [43], yet all have FSEs [42].

Every coronaviral FSE contains a "slippery site" (UUUAAAC) and a structure characterized as a pseudoknot in multiple species [44, 45, 46]. Indeed, the isolated core of the FSE in SARS-CoV-2 was shown to fold into a pseudoknot with three stems [39, 47, 48]. 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 [49]. 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 [49].

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" [50]. 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 up to five ASOs targeting a contiguous 213-244 nt section of the RNA; target sites of adjacent groups abutted without overlapping (Supplementary Table 1). After adding each group of ASOs, we performed DMS-MaPseq [29] with two pairs of RT-PCR primers. With the first pair of primers, flanking the ASO target site (Supplementary Table 2), we confirmed that the DMS reactivities were suppressed – hence the ASO groups bound – except for group 13, for which we obtained no data (Supplementary Figure 1). With the second pair of primers, flanking the 5' side of the FSE-arch, we investigated how its structure was perturbed by each ASO group (Supplementary Figure 2).

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 3b). 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.

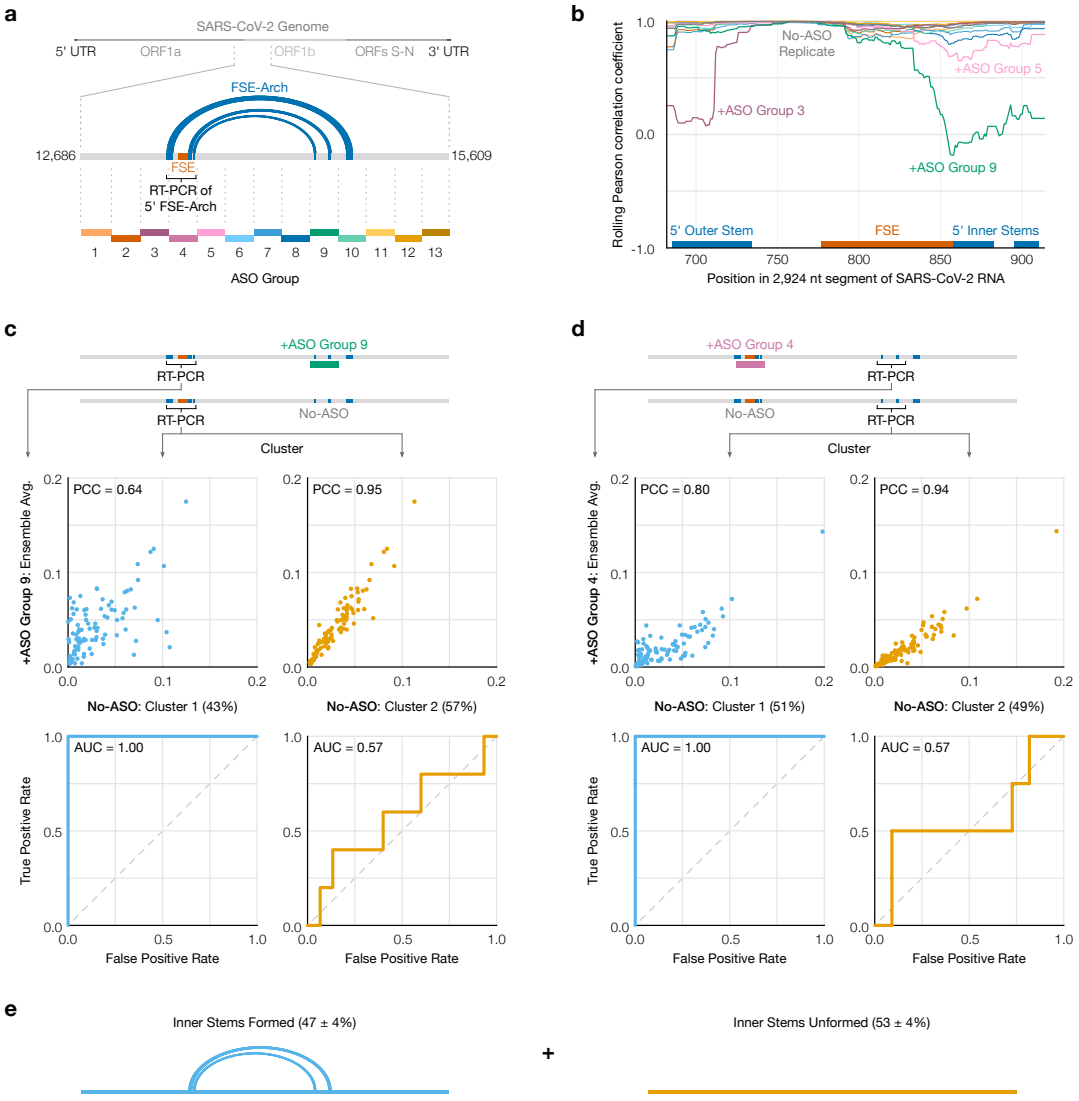


Figure 3: **Search for a long-range base pairs involving the SARS-CoV-2 FSE. (a)** The 2,924 nt segment of the SARS-CoV-2 genome containing the frameshift stimulation element (FSE) and putative FSE-arch [50]. The target site for each group of antisense oligonucleotides (ASOs) is indicated by dotted lines; 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 the fraction of molecules in which 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 [50]. 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 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 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 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 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 [50], we modeled a 1,799 nt segment centered on the FSE-arch. Although computationally predicting long-range base pairs is notoriously unreliable [26, 22], we speculated that we could improve accuracy by incorporating the DMS reactivities of cluster 1 on both sides of the FSE-arch (Supplementary Figure 3). For the innermost stem – which we call long stem 1 (LS1) – nine of thirteen structures (69%) predicted using the cluster 1 DMS reactivities contained LS1, compared to five of eleven (45%) using the ensemble average and four of twenty (20%) using no DMS reactivities. For the second-most inner stem (LS2), eight structures (62%) predicted using cluster 1 contained LS2, while none did using average or no DMS reactivities. Thus, the DMS reactivities corresponding to the long-range cluster enabled predicting the long-range stems more consistently, allowing us to refine our model of the long-range stems.

Our refined model based on the long-range cluster (Figure 4) included not only the two inner stems of the FSE-arch – LS1 and LS2a/b – but also two long stems (LS3a/b and LS4) that were not in the original FSE-arch model [50]. The structure also contained the alternative stem 1 (AS1) that we had previously discovered [49]. To our surprise, LS2b, LS3, and LS4 of the refined model collectively overlapped all three stems of the pseudoknot (PS1, PS2, and PS3) that is thought to stimulate frameshifting [38, 39, 48]. 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 4b). 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 investigated whether the long-range stems compete with the pseudoknot. If they did, 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 – (Supplementary Figure 4a) 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; Supplementary Figure 4b). In their absence emerged clusters consistent with the pseudoknotted and truncated models, constituting 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; Supplementary Figure 4c). Adding both ASOs simultaneously collapsed the ensemble into three clusters of which two (87%) were highly consistent with the pseudo-

knotted model (Figure 4d, bottom; Supplementary Figure 4d). 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.

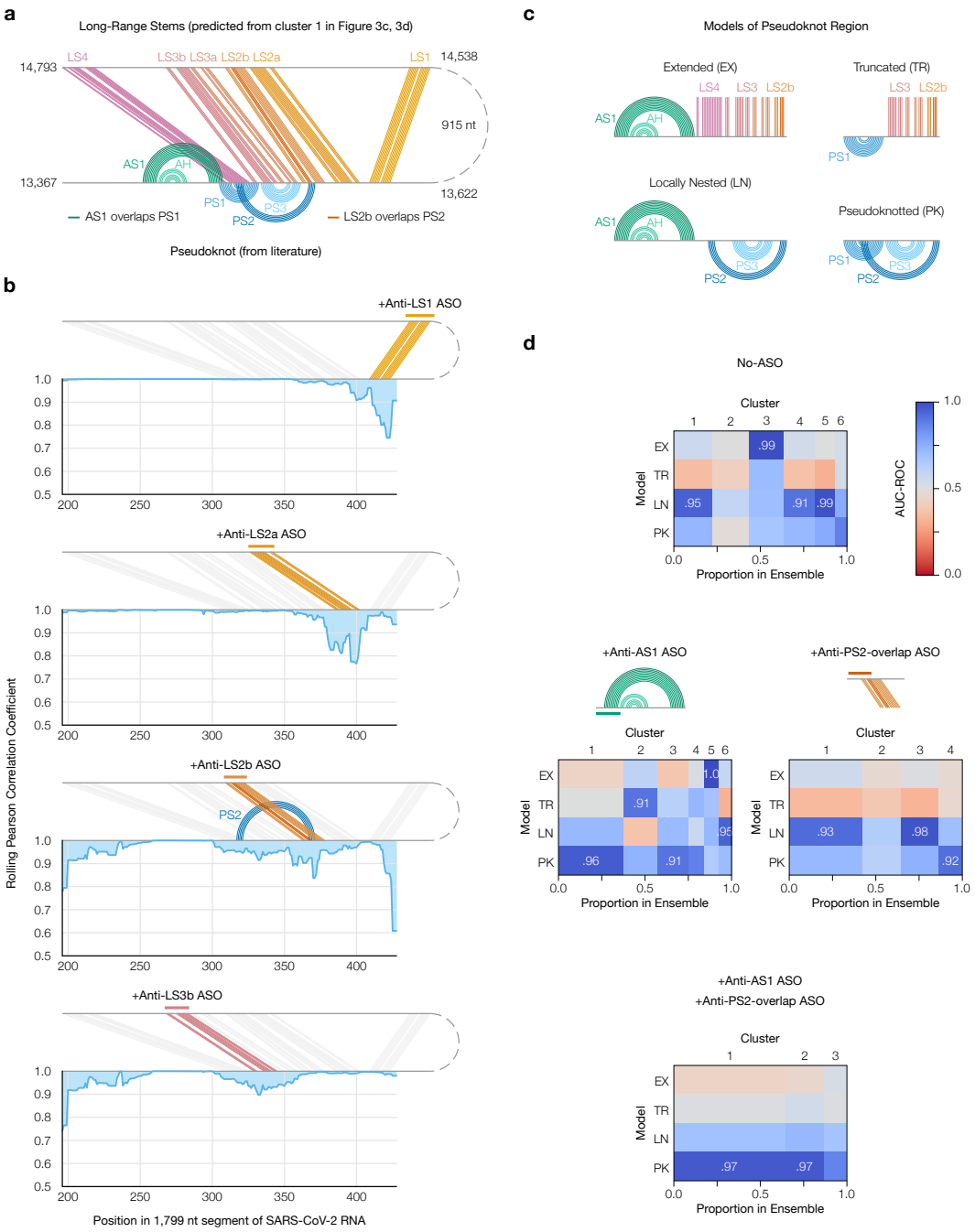


Figure 4: **Refinement of the long-range structure model and competition with the frameshift pseudoknot. (a)** Refined model of the long-range stems (minimum free energy prediction based on cluster 1 in Figure 3c and d) including alternative stem 1 (AS1) [49]; the attenuator hairpin (AH) [51]; and long 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 hypothesized 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 [52] 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 5). Based on these predicted structures, we selected ten coronaviruses – at least one from each genus (Supplementary Figure 6a) – 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 [29]. 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 6b), 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 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 base pairs, we compared the rolling SCC for the SARS-CoV-2 segment to our refined model of the FSE structure (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 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.

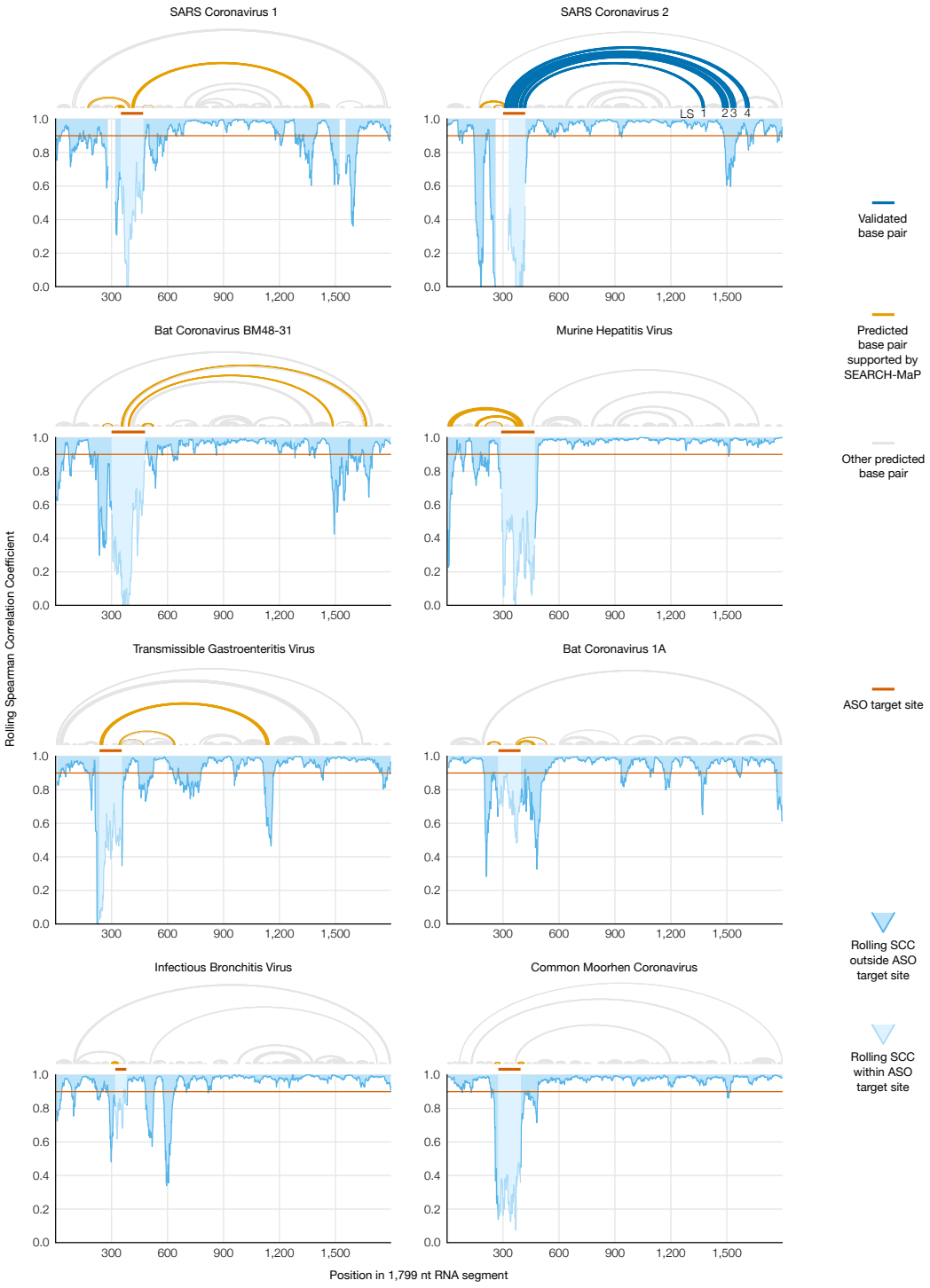


Figure 5: **Evidence for long-range RNA–RNA base pairs involving the FSE in four additional coronaviruses.** Rolling (window = 45 nt) Spearman correlation coefficient (SCC) of DMS reactivities between the +ASO and no-ASO samples for each 1,799 nt segment of a coronaviral genome. The target site of each ASO is highlighted on the SCC data and shown above each graph. Structures predicted with RNAstructure [32] using no-ASO ensemble average DMS reactivities as constraints [33] 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.

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 [32] with the no-ASO ensemble average DMS reactivities as constraints [33]. 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 5, 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* [53] that infects pigs and causes vomiting and diarrhea – almost always fatally in baby piglets [54]. Due to the impacts of TGEV on animal health and economics [54] 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 [29] (two technical replicates per biological replicate) on the extracted RNA (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 stem in live TGEV would be necessary (Supplementary Figure 7).

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 6b). 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 3). 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 6b). 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 6c); 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 6c). 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.

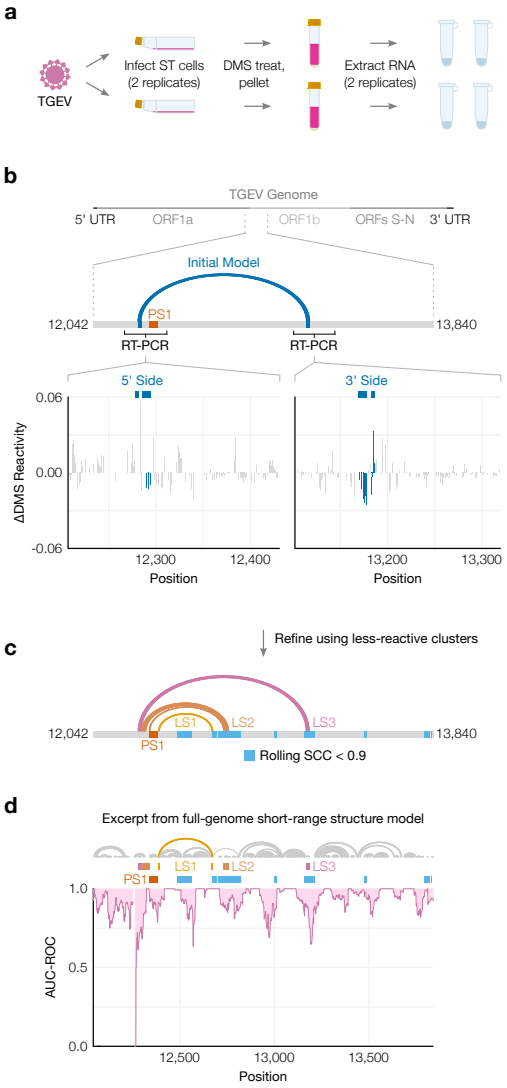


Figure 6: **Genomic secondary structure of live TGEV. (a)** Schematic of the experiment in which two biological replicates of ST cells were infected with TGEV, DMS-treated, and pelleted. Cell pellets were divided into two technical replicates prior to extraction of DMS-modified RNA. **(b)** Differences 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 5), 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 5), 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 [10] (Supplementary Figure 8a) and was consistent with the DMS reactivities (AUC-ROC = 0.94) (Supplementary Figure 8b).

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 [49]. 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 6d). 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. Previous studies have demonstrated that binding an ASO to one side of a long-range stem would perturb the chemical probing reactivities of the other side [55, 56, 57]. 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 [30, 31]. 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 4) and TGEV (Figure 6).

SEISMIC-RNA builds upon our previous work, the DREEM algorithm [30]. 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 [29] and SHAPE-MaP [28], 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 [58, 59, 60, 61, 62] 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 [50]. To investigate RNA structures in detail, M2-seq [35] and related methods [36] 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 [26, 22], 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 [29]. The DMS reactivities would be used [33] 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 6d). 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 [35] 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 [21]) as well provide much-needed benchmarks for secondary structure prediction algorithms [25].

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* [63, 64, 65]. However, in *Tombusviridae* species, the frameshift pseudoknots themselves are made of long-range base pairs; in coronaviruses, the pseudoknots are local structures [44, 45, 46, 39] 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 [63, 64, 65] 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 [38, 66, 39, 67, 68, 49]. Surprisingly, frameshifting has appeared to be nearly twice as frequent (50-70%) in live SARS-CoV-2 [69, 70, 71]; whether this discrepancy is due to long-range base-pairing, methodological artifacts, or *trans* factors [72] is unknown [43].

If, how, and why the long-range base pairs affect frameshifting in coronaviruses are open questions. For *Tombusviridae*, one study [63] 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) [73]. Another study on *Tombusviridae* [65] 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 [70] 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 [74] 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 [75]. 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 3), 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 re-

sults would facilitate subsequent efforts to predict RNA structures and benchmark algorithms, culminating in a real “AlphaFold for RNA” [14] in the hands of every biologist.

Methods

SEARCH-MaP of the 2,924 nt segment of SARS-CoV-2 genomic RNA

A DNA template of the 2,924 nt segment of the SARS-CoV-2 genome, plus an upstream T7 promoter, was amplified from our previously constructed pmirGLO plasmid [49] with 250 nM primers TAATACGACTCACTATAGAATAATGAGCTTAGTCCTGTTGCACTACG and TAAATTGCGGACATACTTATCGGCAATTTTGTTACC (Thermo Fisher Scientific) using 2X CloneAmp HiFi PCR Premix (Takara Bio) in a 50 µl volume with initial denaturation at 98°C for 60 s; 35 cycles of 98°C for 10 s, 65°C for 10 s, and 72°C for 15 s; and final extension at 72°C for 60 s. The 50 µl PCR product was mixed with 10 µl of 6X Purple Loading Dye (New England Biolabs) alongside 10 µl of 0.1X 1 kb DNA Ladder (New England Biolabs) and electrophoresed through a 1% agarose gel – 50 ml of 1X tris-acetate-EDTA buffer (Boston BioProducts), 0.5 g of agarose powder, and 5 µl of 10,000X SYBR Safe (Invitrogen) – in 1X tris-acetate-EDTA buffer (Boston BioProducts) within a Mini-Sub cell GT (Bio-Rad) at 60 V for 60 min. The band at roughly 3 kb was excised and the DNA purified using a Zymoclean Gel DNA Recovery Kit (Zymo Research) according to the manufacturer's protocol; samples were eluted in 10 µl of nuclease-free water (Fisher Bioreagents) and measured with a NanoDrop (Thermo Fisher Scientific). To increase the DNA yield, the gel-extracted DNA was amplified by a second round of PCR followed by gel extraction, using the same protocol as above. To remove contaminants after the second gel extraction, the DNA was further purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol; samples were eluted in 10 µl of nuclease-free water (Fisher Bioreagents) and measured with a NanoDrop (Thermo Fisher Scientific).

RNA was transcribed using a MEGAscript T7 Transcription Kit (Invitrogen) according to the manufacturer's protocol. Specifically, 1 µl (150 ng) of DNA template

from the previous step was diluted in 7 ul of nuclease-free water (Fisher Bioreagents), mixed with 2 µl of each 10X ribonucleotide solution followed by 2 µl of the 10X reaction buffer and 2 µl of the 10X enzyme mix, and then incubated at 37°C for 3 hr. The DNA template was then degraded by adding 1 µl of TURBO DNase (Invitrogen) and incubating at 37°C for 15 min. The RNA transcript was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol; samples were eluted in 20 µl of nuclease-free water (Fisher Bioreagents) and measured with a NanoDrop (Thermo Fisher Scientific). DID YOU CONFIRM THE RNA HAS NO OFF-TARGET BANDS?

Antisense oligonucleotides (ASOs) were ordered from Integrated DNA Technologies in a 96-well PCR plate, each ASO already resuspended to 10 µM in 1X IDTE buffer (10 mM Tris, 0.1 mM EDTA). Each ASO pool (of up to 5 ASOs) was made by mixing 2.5 µl (25 pmol) of each constituent ASO (Supplementary Table 1); the total volume of each pool was adjusted to 12.5 µl by adding TE Buffer: 10 mM Tris (Invitrogen) with 0.1 mM EDTA (Invitrogen). Each 12.5 µl ASO pool was then mixed with 1 µl (425 ng, 453 fmol) of RNA (55X molar excess of each ASO) in a PCR tube. The tubes were heated to 95°C for 60 seconds to denature the RNA, then placed on ice for several minutes. The RNA was transferred to 1.5 ml tubes; to each, 35 µl of 1.4X refolding buffer comprising 400 mM sodium cacodylate (Electron Microscopy Sciences) and 6 mM magnesium chloride (Invitrogen) was added, followed by incubation at 37°C for 25 min to allow the RNA to refold and bind the ASOs. No-ASO control 1 was handled in the same manner but with 12.5 µl of TE Buffer in lieu of an ASO pool. For no-ASO control 2, 12.5 µl of TE Buffer was added after placing on ice and before adding refolding buffer, to confirm that the timing of adding TE buffer would not alter the RNA structure.

For chemical probing, 1.5 µl of neat DMS (Sigma-Aldrich) was added to each tube for a total volume of 50 µl including 320 mM DMS, 280 mM cacodylate, and 9.1 nM RNA. DMS was initially mixed by swirling the pipette tip and kept resuspended by shaking at 500 rpm in a thermomixer (Eppendorf) throughout the treatment at 37°C for 5 min. Reactions were quenched by adding 30 µl neat beta-

mercaptoethanol (Sigma-Aldrich) and mixing thoroughly by pipetting. Each sample of DMS-modified RNA was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol; samples were eluted in 10 µl of nuclease-free water (Fisher Bioreagents) and measured with a NanoDrop (Thermo Fisher Scientific).

ASOs were removed from each sample using TURBO DNase (Invitrogen) according to the manufacturer's protocol. Briefly, 4 µl of each DMS-modified RNA was mixed with 4 µl of nuclease-free water (Fisher Bioreagents), 1 µl of 10X TURBO DNase Buffer, and 1 µl of TURBO DNase in a PCR tube; and then incubated at 37°C for 30 min. To stop each reaction, 2 µl of DNase Inactivation Reagent was mixed in and incubated at room temperature for 10 min, and mixed throughout by flicking several times. The DNase Inactivation Reagent was precipitated by spinning the tubes on a benchtop PCR tube centrifuge for 10 min, then transferring 4 µl of each supernatant to a new tube.

For reverse transcription, each 4 µl sample of DNased, DMS-modified RNA was mixed with 6 µl of nuclease-free water (Fisher Bioreagents), 4 µl of 5X First Strand Buffer (Invitrogen), 1 µl of 10 mM dNTPs (Promega), 1 µl of 100 mM dithiothreitol (Invitrogen), 1 µl of RNaseOUT (Invitrogen), 1 µl of TGIRT-III (InGex), 1 µl of 10 mM FSE reverse primer CTTCGTCCTTTTCTTGGAAGCGACA (Integrated DNA Technologies), and 1 µl of 10 mM section-specific reverse primer (Integrated DNA Technologies, Supplementary Table 2), for a total volume of 20 µl. Synthesis of cDNA proceeded at 57°C for 90 min, then inactivated at 85°C for 15 min. To remove the RNA template from the cDNA, 1 µl of Hybridase Thermostable RNase H (Lucigen) was added to each tube and incubated at 37°C for 20 min.

The cDNA products were amplified using the Advantage HF 2 PCR Kit (Takara Bio) according to the manufacturer's protocol. Specifically, 1 µl of unpurified cDNA was mixed with 8.25 µl of nuclease-free water (Fisher Bioreagents), 1.25 µl of 10X Advantage 2 PCR Buffer, 1.25 µl of 10X Advantage-HF 2 dNTP Mix, 0.25 µl of 50X Advantage-HF 2 Polymerase Mix, 0.25 µl of 10 mM forward primer (Integrated DNA Technologies, Supplementary Table 2), and 0.25 µl of 10 mM reverse primer

(Integrated DNA Technologies, Supplementary Table 2), for a total volume of 12.5 µl. After an initial denaturation at 94°C for 60 s, 25 cycles of 94°C for 30 s, 60°C for 30 s, and 68°C for 60 s were run, followed by a final extension at 68°C for 60 s. For each cDNA, the PCR products were pooled (5 µl each) and then purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol; samples were eluted in 20 µl of nuclease-free water (Fisher Bioreagents) and measured with a NanoDrop (Thermo Fisher Scientific).

A 200 ng aliquot of each pool of PCR products was diluted in 10 mM Tris-HCl, pH 8 (Invitrogen) to a total of 50 µl. Aliquots were prepared for sequencing using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer's protocol with the following modifications. During two-step size selection after adapter ligation, 27.5 µl and 12.5 µl of NEBNext Sample Purification Beads were used in the first and second steps, respectively, to select inserts of 280-290 bp. Indexing PCR was run at half scale (25 µl total volume) for 3 cycles. Each PCR product was mixed with 2.5 µl of E-Gel Sample Loading Buffer (Invitrogen) and electrophoresed through a 2% E-Gel SizeSelect II Agarose Gel (Invitrogen) according to the manufacturer's instructions. Samples were extracted in 50 µl nuclease-free water (Fisher Bioreagents). DNA concentrations were measured using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific) according to the manufacturer's protocol. Samples were pooled and sequenced using an iSeq 100 Sequencing System (Illumina) with a 2 x 150 bp paired-end read length according to the manufacturer's protocol.

DMS-MaPseq of live transmissible gastroenteritis virus

Screening coronavirus long-range interactions computationally

All coronaviruses with reference genomes in the NCBI Reference Sequence Database [52] 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 [32] 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 [76] and SciPy version 1.7 [77]. 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).

References

[1] Carla A. Klattenhoff, Johanna C. Scheuermann, Lauren E. Surface, Robert K. Bradley, Paul A. Fields, Matthew L. Steinhauser, Huiming Ding, Vincent L. Butty, Lillian Torrey, Simon Haas, Ryan Abo, Mohammadsharif Tabebordbar, Richard T. Lee, Christopher B. Burge, and Laurie A. Boyer. Braveheart, a long noncoding rna required for cardiovascular lineage commitment. *Cell*, 152:570–583, 2013.

[2] Blake Wiedenheft, Samuel H. Sternberg, and Jennifer A. Doudna. Rna-guided genetic silencing systems in bacteria and archaea. *Nature*, 482(7385):331–338, 2012.

[3] Harry F Noller. Evolution of protein synthesis from an rna world. *Cold Spring Harb Perspect Biol*, 4(4):a003681, Apr 2012.

[4] Jens Kortmann and Franz Narberhaus. Bacterial rna thermometers: molecular zippers and switches. *Nature Reviews Microbiology*, 10(4):255–265, 2012.

[5] Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152:17–24, 2013.

[6] Arunoday Bhan and Subhrangsu S. Mandal. Lncrna hotair: A master regulator of chromatin dynamics and cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1856(1):151–164, 2015.

[7] Mohammadreza Hajjari and Adrian Salavaty. Hotair: an oncogenic long noncoding rna in different cancers. *Cancer Biol Med*, 12(1):1–9, Mar 2015.

[8] Mark E. J. Woolhouse and Liam Brierley. Epidemiological characteristics of human-infective rna viruses. *Scientific Data*, 5(1):180017, 2018.

[9] Nicole M. Bouvier and Peter Palese. The biology of influenza viruses. *Vaccine*, 26:D49–D53, 2008. Influenza Vaccines: Research, Development and Public Health Challenges.

[10] Dong Yang and Julian L. Leibowitz. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Research*, 206:120–133, 2015.

[11] Stefanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. Insights into rna structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7):469–479, 2014.

[12] Kalli Kappel, Kaiming Zhang, Zhaoming Su, Andrew M. Watkins, Wipapat Kladwang, Shanshan Li, Grigore Pintilie, Ved V. Topkar, Ramya Rangan, Ivan N. Zheludev, Joseph D. Yesselman, Wah Chiu, and Rhiju Das. Accelerated cryo-em-guided determination of three-dimensional rna-only structures. *Nature Methods*, 17:699–707, 2020.

[13] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 1 2000.

[14] Bohdan Schneider, Blake Alexander Sweeney, Alex Bateman, Jiri Cerny, Tomasz Zok, and Marta Szachniuk. When will RNA get its AlphaFold moment? *Nucleic Acids Research*, 51(18):9522–9532, 09 2023.

[15] Jamie J Cannone, Sankar Subramanian, Murray N Schnare, James R Collett, Lisa M D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V Madabusi, Kirsten M Müller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin R Gutell. The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC Bioinformatics*, 3:2, 2002.

[16] Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 06 1994.

[17] Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, Elena Rivas, Sean R Eddy, Robert D Finn, Alex Bateman, and Anton I Petrov. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1):D192–D200, 11 2020.

[18] Anthony M. Mustoe, Charles L. Brooks, and Hashim M. Al-Hashimi. Hierarchy of rna functional dynamics. *Annual Review of Biochemistry*, 83(1):441–466, 2014. PMID: 24606137.

[19] Robert C. Spitale and Danny Incarnato. Probing the dynamic rna structurome and its functions. *Nature Reviews Genetics*, 24(3):178–196, 2023.

[20] Jeffrey J. Quinn and Howard Y. Chang. Unique features of long non-coding rna biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016.

[21] Sita J. Lange, Daniel Maticzka, Mathias Mohl, Joshua N. Gagnon, Chris M. Brown, and Rolf Backofen. Global or local? predicting secondary structure and accessibility in mrnas. *Nucleic Acids Research*, 2012.

[22] Beth L Nicholson and K Andrew White. Exploring the architecture of viral rna genomes. *Current Opinion in Virology*, 12:66–74, 2015. Antiviral strategies • Virus structure and expression.

[23] Christoph Flamm, Julia Wielach, Michael T. Wolfinger, Stefan Badelt, Ronny Lorenz, and Ivo L. Hofacker. Caveats to deep learning approaches to rna secondary structure prediction. *Frontiers in Bioinformatics*, 2, 2022.

[24] Kengo Sato and Michiaki Hamada. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. *Briefings in Bioinformatics*, 24(4):bbad186, 05 2023.

[25] David H. Mathews. How to benchmark rna secondary structure prediction accuracy. *Methods*, 162-163:60–67, 2019. Experimental and Computational Techniques for Studying Structural Dynamics and Function of RNA.

[26] Kishore J. Doshi, Jamie J. Cannone, Christian W. Cobaugh, and Robin R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction. *BMC Bioinformatics*, 5(1):105, 2004.

[27] Miles Kubota, Catherine Tran, and Robert C Spitale. Progress and challenges for chemical probing of rna structure inside living cells. *Nature Chemical Biology*, 11(12):933–941, 2015.

[28] Nathan A. Siegfried, Steven Busan, Greggory M. Rice, Julie A.E. Nelson, and Kevin M. Weeks. Rna motif discovery by shape and mutational profiling (shape-map). *Nature methods*, 2014.

[29] Meghan Zubradt, Paromita Gupta, Sitara Persad, Alan M. Lambowitz, Jonathan S. Weissman, and Silvi Rouskin. Dms-mapseq for genome-wide or targeted rna structure probing in vivo. *Nature Methods*, 2254:219–238, 2016.

[30] Phillip J. Tomezsko, Vincent D.A. Corbin, Paromita Gupta, Harish Swaminathan, Margalit Glasgow, Sitara Persad, Matthew D. Edwards, Lachlan Mcintosh, Anthony T. Papenfuss, Ann Emery, Ronald Swanstrom, Trinity Zang, Tammy C.T. Lan, Paul Bieniasz, Daniel R. Kuritzkes, Athe Tsibris, and Silvi Rouskin. Determination of rna structural diversity and its role in hiv-1 rna splicing. *Nature*, 582:438–442, 2020.

[31] Edoardo Morandi, Ilaria Manfredonia, Lisa M. Simon, Francesca Anselmi, Martijn J. van Hemert, Salvatore Oliviero, and Danny Incarnato. Genome-scale deconvolution of rna structure ensembles. *Nature Methods*, 18:249–252, 2 2021.

[32] David H. Mathews, Matthew D. Disney, Jessica L. Childs, Susan J. Schroeder, Michael Zuker, and Douglas H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proceedings of the National Academy of Sciences*, 101:7287–7292, 5 2004.

[33] Pablo Cordero, Wipapat Kladwang, Christopher C. Vanlang, and Rhiju Das. Quantitative dimethyl sulfate mapping for automated rna secondary structure inference. *Biochemistry*, 51:7037–7039, 9 2012.

[34] Michael F. Sloma and David H. Mathews. Improving rna secondary structure prediction with structure mapping data, 2015.

[35] Clarence Y. Cheng, Wipapat Kladwang, Joseph D. Yesselman, and Rhiju Das. Rna structure inference through chemical mapping after accidental or intentional mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 114:9876–9881, 9 2017.

[36] Pablo Cordero and Rhiju Das. Rich rna structure landscapes revealed bymutate-and-map analysis. *PLOS Computational Biology*, 11:e1004473, 11 2015.

[37] Grzegorz Kudla, Yue Wan, and Aleksandra Helwak. Rna conformation capture by proximity ligation. *Annual Review of Genomics and Human Genetics*, 21(1):81–100, 2020. PMID: 32320281.

[38] Jamie A. Kelly, Alexandra N. Olson, Krishna Neupane, Sneha Munshi, Josue San Emeterio, Lois Pollack, Michael T. Woodside, and Jonathan D. Dinman. Structural and functional conservation of the programmed -1 ribosomal frameshift signal of sars coronavirus 2 (sars-cov-2). *Journal of Biological Chemistry*, 295:10741–10748, 7 2020.

[39] Kaiming Zhang, Ivan N. Zheludev, Rachel J. Hagey, Raphael Haslecker, Yixuan J. Hou, Rachael Kretsch, Grigore D. Pintilie, Ramya Rangan, Wipapat Kladwang, Shanshan Li, Marie Teng Pei Wu, Edward A. Pham, Claire Bernardin-Souibgui, Ralph S. Baric, Timothy P. Sheahan, Victoria D'Souza, Jeffrey S. Glenn, Wah Chiu, and Rhiju Das. Cryo-em and antisense targeting of the 28-kda frameshift stimulation element from the sars-cov-2 rna genome. *Nature Structural & Molecular Biology*, 28:747–754, 8 2021.

[40] Chringma Sherpa, Jason W. Rausch, Stuart F.J. Le Grice, Marie Louise Hammarskjold, and David Rekosh. The hiv-1 rev response element (rre) adopts alternative conformations that promote different rates of virus replication. *Nucleic Acids Research*, 43:4676–4686, 3 2015.

[41] Beth L. Nicholson and K. Andrew White. Functional long-range rna–rna interactions in positive-strand rna viruses. *Nature Reviews Microbiology*, 12:493–504, 6 2014.

[42] Ewan P. Plant and Jonathan D. Dinman. The role of programmed-1 ribosomal frameshifting in coronavirus propagation, 2008.

[43] Matthew F. Allan, Amir Brivanlou, and Silvi Rouskin. Rna levers and switches controlling viral gene expression. *Trends in Biochemical Sciences*, 48, 2023.

[44] Ian Brierley, Paul Digard, and Stephen C. Inglis. Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an rna pseudoknot. *Cell*, 1989.

[45] J. Herald and S. G. Siddell. An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of hcv 229e polymerase mrna. *Nucleic Acids Research*, 21:5838–5842, 1993.

[46] Ewan P. Plant, Gabriela C. Pérez-Alvarado, Jonathan L. Jacobs, Bani Mukhopadhyay, Mirko Hennig, and Jonathan D. Dinman. A three-stemmed mrna pseudoknot in the sars coronavirus frameshift signal. *PLoS Biology*, 3:e172, 2005.

[47] Christina Roman, Anna Lewicka, Deepak Koirala, Nan-Sheng Li, and Joseph A. Piccirilli. The sars-cov-2 programmed -1 ribosomal frameshifting element crystal structure solved to 2.09 åusing chaperone-assisted rna crystallography. *ACS Chemical Biology*, 16(8):1469–1481, 08 2021.

[48] Christopher P. Jones and Adrian R. Ferré-D'Amaré. Crystal structure of the severe acute respiratory syndrome coronavirus 2 (sars-cov-2) frameshifting pseudoknot. *RNA*, 28:239–249, 2022.

[49] Tammy C.T. Lan, Matty F. Allan, Lauren E. Malsick, Jia Z. Woo, Chi Zhu, Fengrui Zhang, Stuti Khandwala, Sherry S.Y. Nyeo, Yu Sun, Junjie U. Guo, Mark Bathe, Anders Näär, Anthony Griffiths, and Silvi Rouskin. Secondary structural

ensembles of the sars-cov-2 rna genome in infected cells. *Nature Communications*, 13:1128, 3 2022.

[50] Omer Ziv, Jonathan Price, Lyudmila Shalamova, Tsveta Kamenova, Ian Goodfellow, Friedemann Weber, and Eric A. Miska. The short- and long-range rna-rna interactome of sars-cov-2. *Molecular Cell*, 80:1067–1077.e5, 12 2020.

[51] Mei Chi Su, Chung Te Chang, Chiu Hui Chu, Ching Hsiu Tsai, and Kung Yao Chang. An atypical rna pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal frameshifting of sars coronavirus. *Nucleic Acids Research*, 33:4265–4275, 2005.

[52] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, Kim D. Pruitt, O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad Haft D, McVeigh R, Robbertse Rajput B, Robbertse Rajput B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb Gupta T, Goldfarb Gupta T, Haddad Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, and Pruitt KD. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44:D733–D745, 2016.

[53] Gary R. Whittaker, Nicole M. André, and Jean Kaoru Millet. Improving virus taxonomy by recontextualizing sequence-based classification with biologically relevant data: the case of the ¡i¿alphacoronavirus 1¡/i¿ species. *mSphere*, 3(1):10.1128/mspheredirect.00463–17, 2018.

[54] Qiang Liu and Huai-Yu Wang. Porcine enteric coronaviruses: an updated overview of the pathogenesis, prevalence, and diagnosis. *Veterinary Research Communications*, 45(2):75–86, 2021.

[55] Michal Legiewicz, Andrei S. Zolotukhin, Guy R. Pilkington, Katarzyna J. Purzycka, Michelle Mitchell, Hiroaki Uranishi, Jenifer Bear, George N. Pavlakis, Stuart F. J. Le Grice, and Barbara K. Felber. The rna transport element of the murine ¡em¿musd¡/em¿ retrotransposon requires long-range intramolecular interactions for function *. *Journal of Biological Chemistry*, 285(53):42097–42104, 2024/03/11 2010.

[56] Eva J. Archer, Mark A. Simpson, Nicholas J. Watts, Rory O'Kane, Bangchen Wang, Dorothy A. Erie, Alex McPherson, and Kevin M. Weeks. Long-range architecture in a viral rna genome. *Biochemistry*, 52(18):3182–3190, 2013. PMID: 23614526.

[57] Yun Bai, Akshay Tambe, Kaihong Zhou, and Jennifer A Doudna. Rna-guided assembly of rev-rre nuclear export complexes. *eLife*, 3:e03656, aug 2014.

[58] Jong Ghut Ashley Aw, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim, Kum Loong Boon, Sidika Tapsin, Yun Shen Chan, Cheng Peow Tan, Adelene Y.L. Sim, Tong Zhang, Teodorus Theo Susanto, Zhiyan Fu, Niranjan Nagarajan, and Yue Wan. In vivo mapping of eukaryotic rna interactomes reveals principles of higher-order organization and regulation. *Molecular Cell*, 62:603–617, 2016.

[59] Zhipeng Lu, Qiangfeng Cliff Zhang, Byron Lee, Ryan A. Flynn, Martin A. Smith, James T. Robinson, Chen Davidovich, Anne R. Gooding, Karen J. Goodrich, John S. Mattick, Jill P. Mesirov, Thomas R. Cech, and Howard Y. Chang. Rna duplex map in living cells reveals higher-order transcriptome structure. *Cell*, 165:1267–1279, 2016.

[60] Eesha Sharma, Tim Sterne-Weiler, Dave O'Hanlon, and Benjamin J. Blencowe. Global mapping of human rna-rna interactions. *Molecular Cell*, 62:618–626, 2016.

[61] Omer Ziv, Marta M. Gabryelska, Aaron T.L. Lun, Luca F.R. Gebert, Jessica Sheu-Gruttadauria, Luke W. Meredith, Zhong Yu Liu, Chun Kit Kwok, Cheng Feng Qin, Ian J. MacRae, Ian Goodfellow, John C. Marioni, Grzegorz Kudla, and Eric A. Miska. Comrades determines in vivo rna structures and interactions. *Nature Methods*, 15:785–788, 9 2018.

[62] Ryan Van Damme, Kongpan Li, Minjie Zhang, Jianhui Bai, Wilson H. Lee, Joseph D. Yesselman, Zhipeng Lu, and Willem A. Velema. Chemical reversible crosslinking enables measurement of rna 3d distances and alternative conformations in cells. *Nature Communications*, 13(1):911, 2022.

[63] Jennifer K. Barry and W. Allen Miller. A -1 ribosomal frameshift element that requires base pairing across four kilobases suggests a mechanism of regulating ribosome and replicase traffic on a viral rna. *Proceedings of the National Academy of Sciences of the United States of America*, 99:11133–11138, 8 2002.

[64] Yuri Tajima, Hiro oki Iwakawa, Masanori Kaido, Kazuyuki Mise, and Tetsuro Okuno. A long-distance rna–rna interaction plays an important role in programmed - 1 ribosomal frameshifting in the translation of p88 replicase protein of red clover necrotic mosaic virus. *Virology*, 417:169–178, 8 2011.

[65] Anna A Mikkelsen, Feng Gao, Elizabeth Carino, Sayanta Bera, and Anne E Simon. -1 programmed ribosomal frameshifting in class 2 umbravirus-like rnas uses multiple long-distance interactions to shift between active and inactive structures and destabilize the frameshift stimulating element. *Nucleic Acids Research*, 51(19):10700–10718, 09 2023.

[66] Hafeez S. Haniff, Yuquan Tong, Xiaohui Liu, Jonathan L. Chen, Blessy M. Suresh, Ryan J. Andrews, Jake M. Peterson, Collin A. O'Leary, Raphael I. Benhamou, Walter N. Moss, and Matthew D. Disney. Targeting the sars-cov-2 rna genome with small molecule binders and ribonuclease targeting chimera (ribotac) degraders. *ACS Central Science*, 6:1713–1721, 2020.

[67] Pramod R. Bhatt, Alain Scaiola, Gary Loughran, Marc Leibundgut, Annika Kratzel, Romane Meurs, René Dreos, Kate M. O'Connor, Angus McMillan, Jeffrey W. Bode, Volker Thiel, David Gatfield, John F. Atkins, and Nenad Ban. Structural basis of ribosomal frameshifting during translation of the sars-cov-2 rna genome. *Science*, 372:1306–1313, 5 2021.

[68] Yu Sun, Laura Abriola, Rachel O. Niederer, Savannah F. Pedersen, Mia M. Alfajaro, Valter Silva Monteiro, Craig B. Wilen, Ya-Chi Ho, Wendy V. Gilbert, Yulia V. Surovtseva, Brett D. Lindenbach, and Junjie U. Guo. Restriction of sars-cov-2 replication by targeting programmed -1 ribosomal frameshifting. *Proceedings of the National Academy of Sciences of the United States of America*, 118:e2023051118, 6 2021.

[69] Yaara Finkel, Orel Mizrahi, Aharon Nachshon, Shira Weingarten-Gabbay, David Morgenstern, Yfat Yahalom-Ronen, Hadas Tamir, Hagit Achdout, Dana Stein, Ofir Israeli, Adi Beth-Din, Sharon Melamed, Shay Weiss, Tomer Israely, Nir Paran, Michal Schwartz, and Noam Stern-Ginossar. The coding capacity of sars-cov-2. *Nature*, 589:125–130, 1 2021.

[70] Doyeon Kim, Sukjun Kim, Joori Park, Hee Ryung Chang, Jeeyoon Chang, Junhak Ahn, Heedo Park, Junehee Park, Narae Son, Gihyeon Kang, Jeonghun Kim, Kisoon Kim, Man Seong Park, Yoon Ki Kim, and Daehyun Baek. A high-resolution temporal atlas of the sars-cov-2 translatome and transcriptome. *Nature Communications*, 12:5120, 8 2021.

[71] Maritza Puray-Chavez, Nakyung Lee, Kasyap Tenneti, Yiqing Wang, Hung R Vuong, Yating Liu, Amjad Horani, Tao Huang, Sean P Gunsten, James B Case, Wei Yang, Michael S Diamond, Steven L Brody, Joseph Dougherty, Sebla B Kutluay, and Kellie Jurado. The translational landscape of sars-cov-2-infected cells reveals suppression of innate immune genes. *mBio*, 13:e00815–22, 6 2022.

[72] Ricarda J Riegger and Neva Caliskan. Thinking outside the frame: Impacting genomes capacity by programmed ribosomal frameshifting. *Frontiers in Molecular Biosciences*, 9:842261, 2022.

[73] Georg Wolff, Charlotte E. Melia, Eric J. Snijder, and Montserrat Bárcena. Double-membrane vesicles as platforms for viral replication. *Trends in Microbiology*, 28:1022–1033, 12 2020.

[74] Jamie A. Kelly, Michael T. Woodside, and Jonathan D. Dinman. Programmed -1 ribosomal frameshifting in coronaviruses: A therapeutic target. *Virology*, 554:75–82, 2021.

[75] Chi Zhu, Justin Y. Lee, Jia Z. Woo, Lei Xu, Xammy Nguyenla, Livia H. Yamashiro, Fei Ji, Scott B. Biering, Erik Van Dis, Federico Gonzalez, Douglas

Fox, Eddie Wehri, Arjun Rustagi, Benjamin A. Pinsky, Julia Schaletzky, Catherine A. Blish, Charles Chiu, Eva Harris, Ruslan I. Sadreyev, Sarah Stanley, Sakari Kauppinen, Silvi Rouskin, and Anders M. Näär. An intranasal aso therapeutic targeting sars-cov-2. *Nature Communications*, 13:4503, 12 2022.

[76] Michael Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6, 2021.

[77] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[78] Sara Alonso, Ander Izeta, Isabel Sola, and Luis Enjuanes. Transcription regulatory sequences and mrna expression levels in the coronavirus transmissible gastroenteritis virus. *Journal of Virology*, 76(3):1293–1308, 2002.

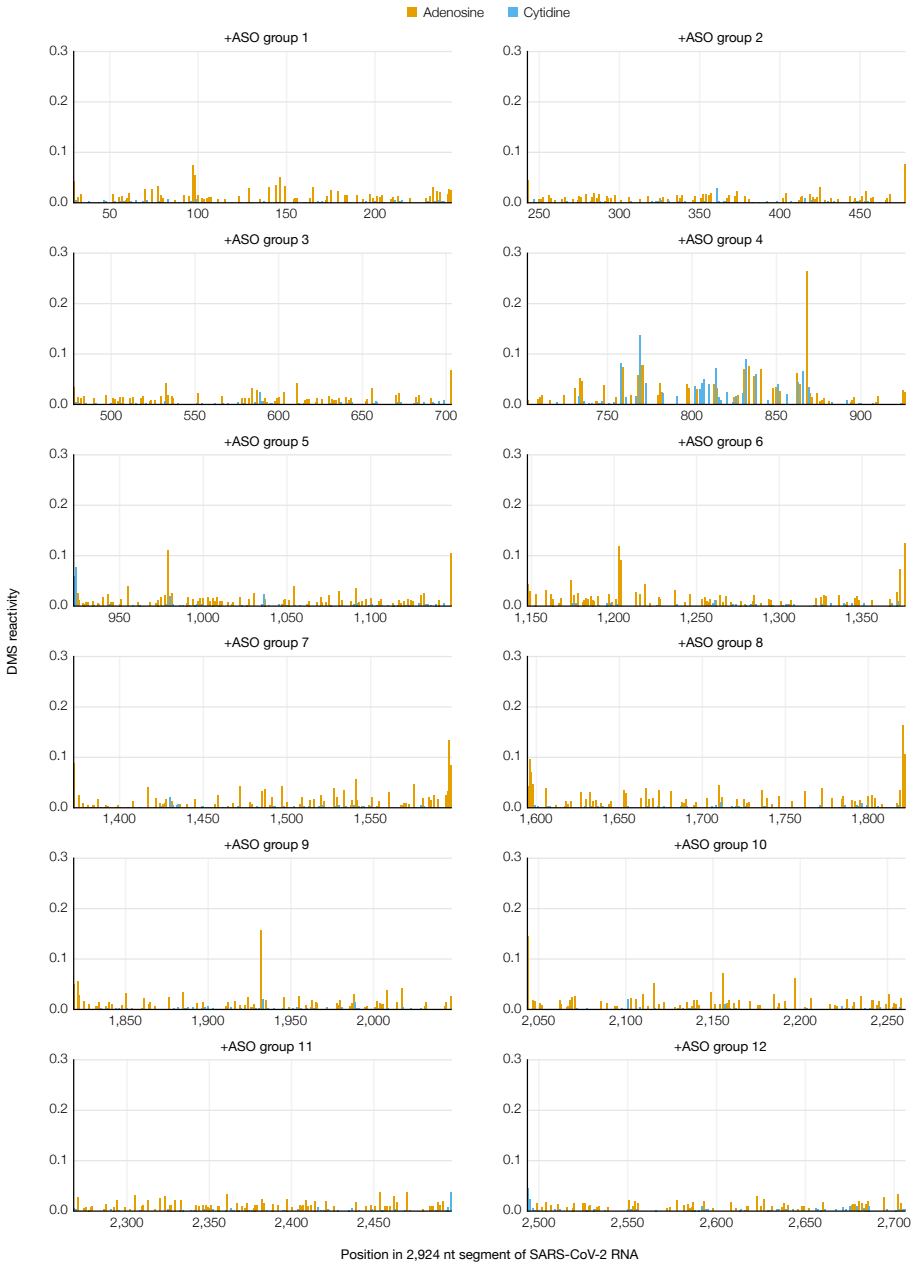
[79] K. Nakagawa, K.G. Lokugamage, and S. Makino. Viral and cellular mrna translation in coronavirus-infected cells. *Advances in Virus Research*, 96:165, 12 2016.

[80] Kévin Darty, Alain Denise, and Yann Ponty. Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, 25:1974–1975, 2009.

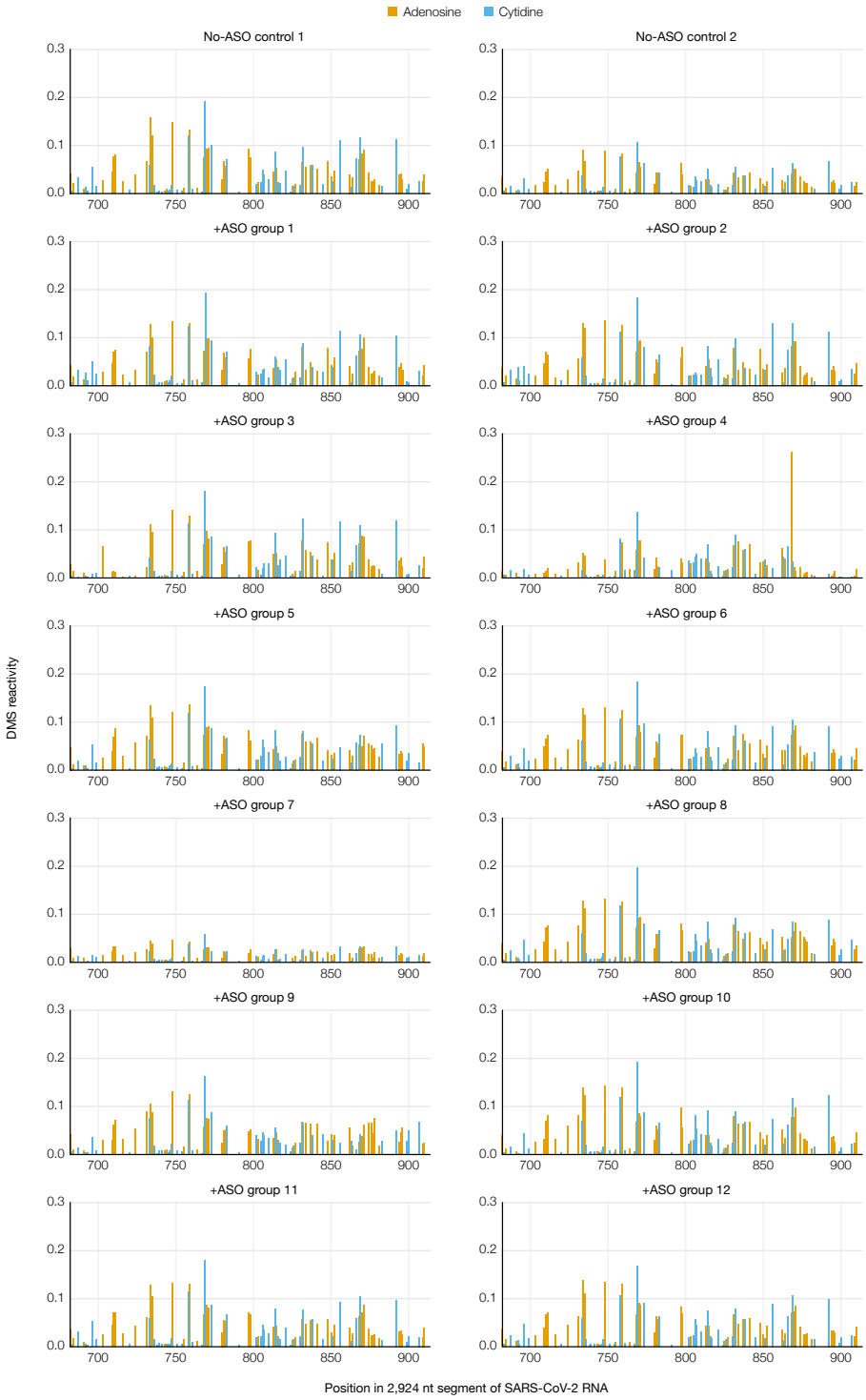
[81] D.A. Knoll and D.E. Keyes. Jacobian-free newton–krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, 193(2):357–397, 2004.

Supplementary Information

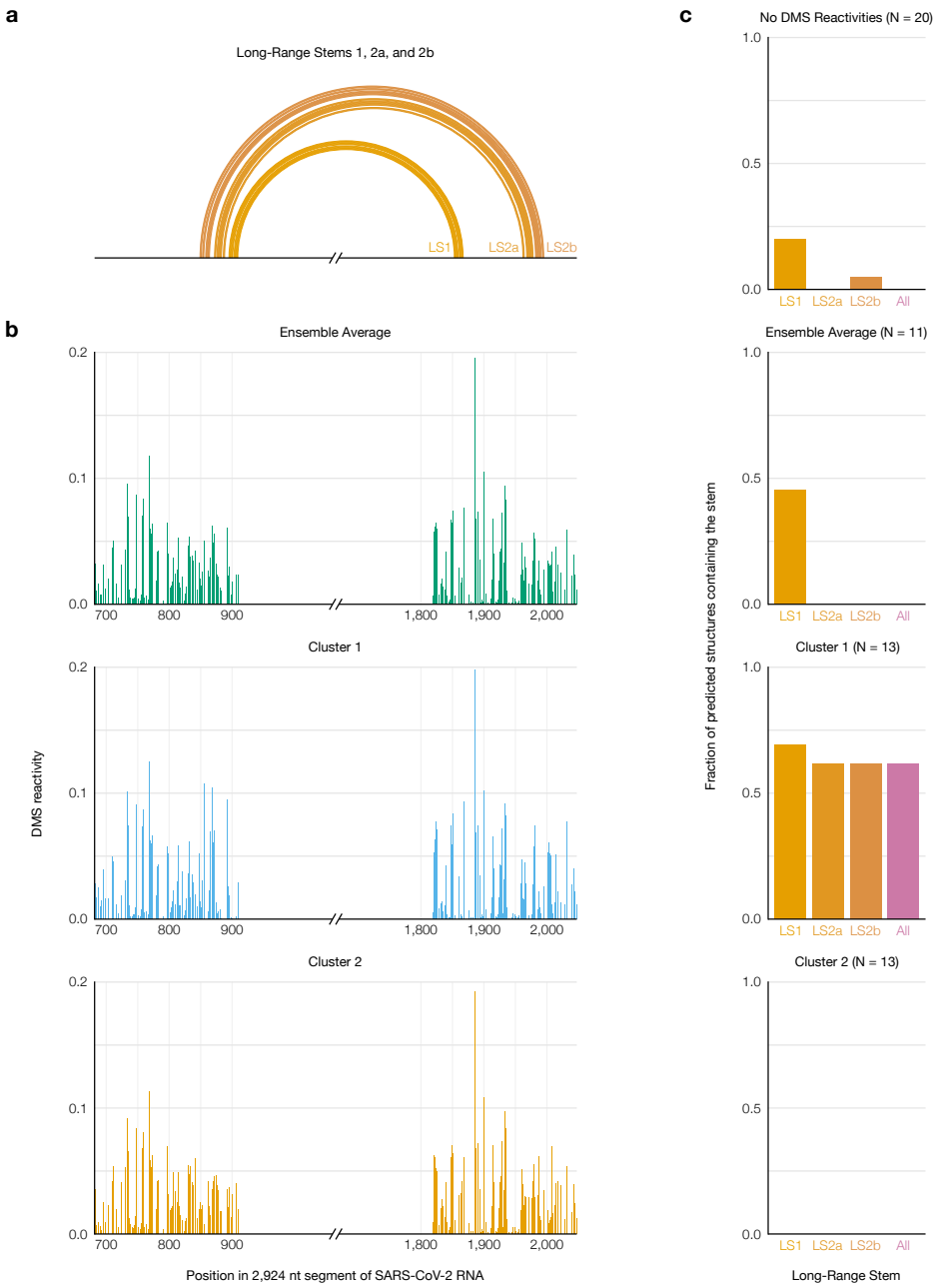
Supplementary Figures



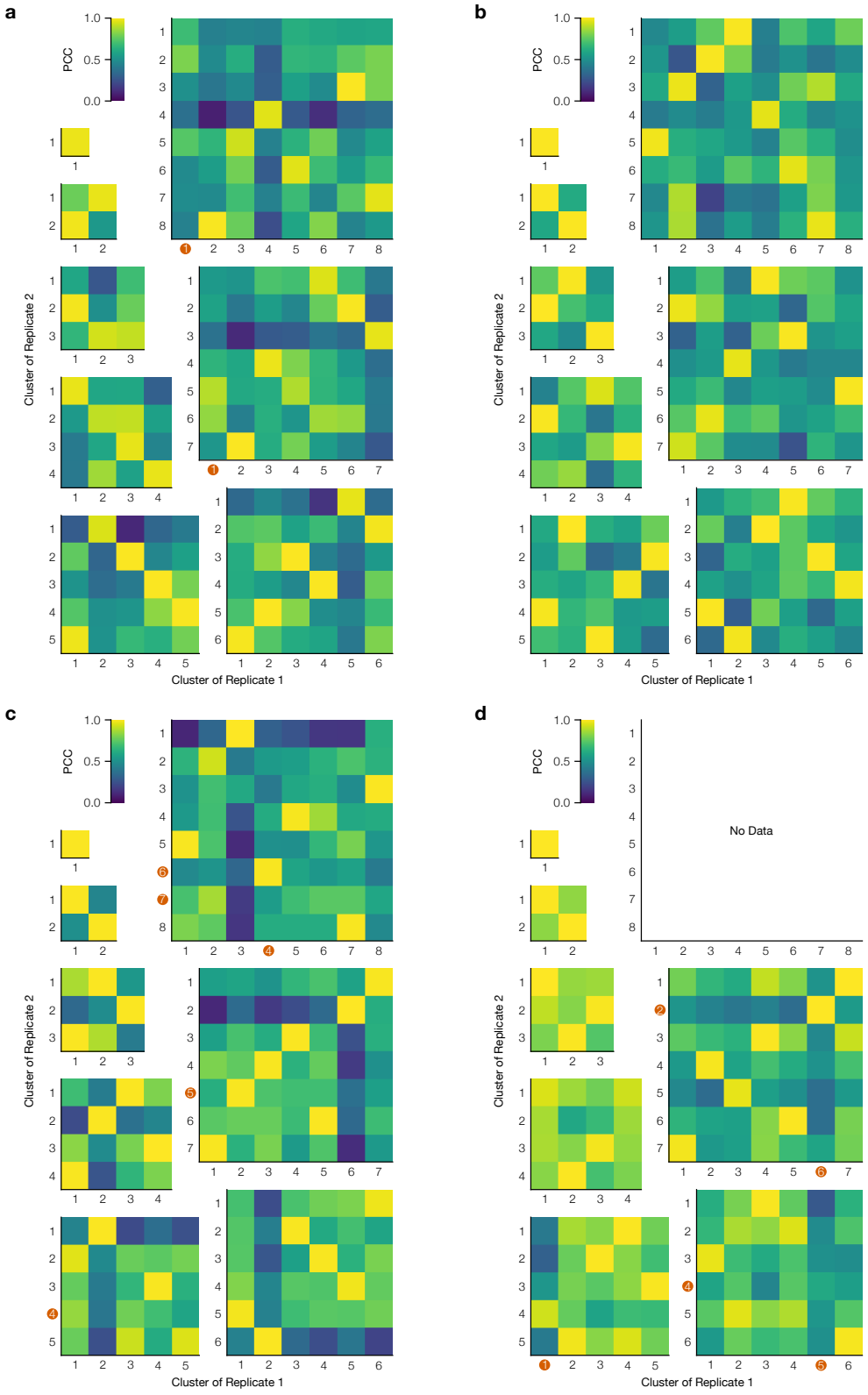
Supplementary Figure 1: **Mutational profile of each ASO target section upon adding the corresponding group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.** Positions are colored based on the RNA sequence.



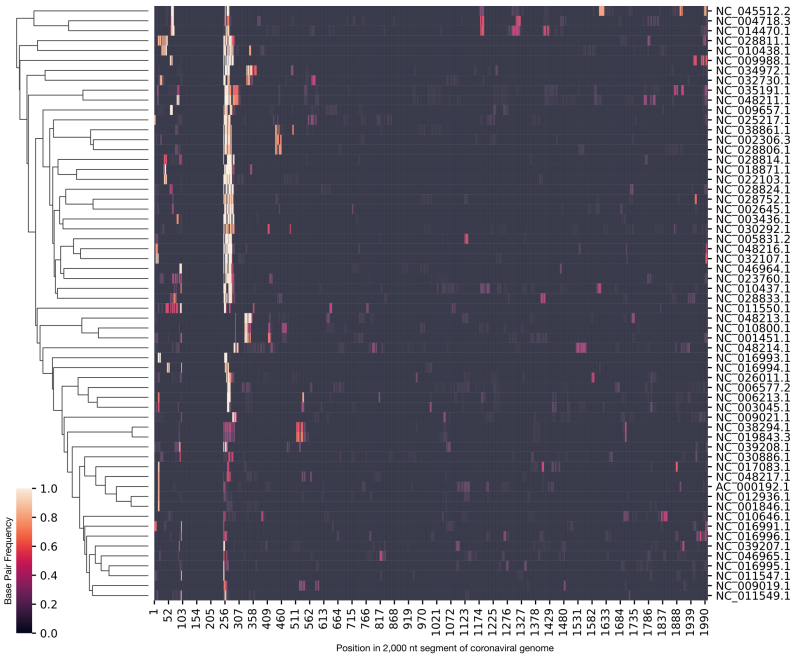
Supplementary Figure 2: **Mutational profiles of the FSE section upon adding each group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.** Positions are colored based on the RNA sequence.



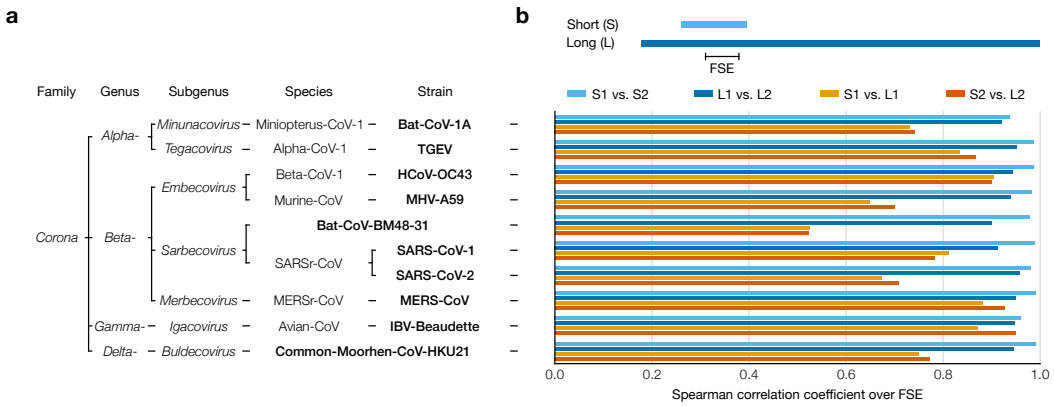
Supplementary Figure 3: **Improved prediction of long-range stems in SARS-CoV-2 using clustered DMS reactivities. (a)** Model of the two inner stems of the FSE-arch [50], denoted long stems (LS) 1 and 2a/b. **(b)** Mutational profiles of the ensemble average and of clusters 1 and 2 on both sides of the FSE-arch. **(c)** For each mutational profile (as well as a purely thermodynamic prediction with no DMS reactivities), the fraction of predicted structures in which each long stem was predicted perfectly (i.e. all base pairs were present). The numbers of predicted structures (N) are indicated.



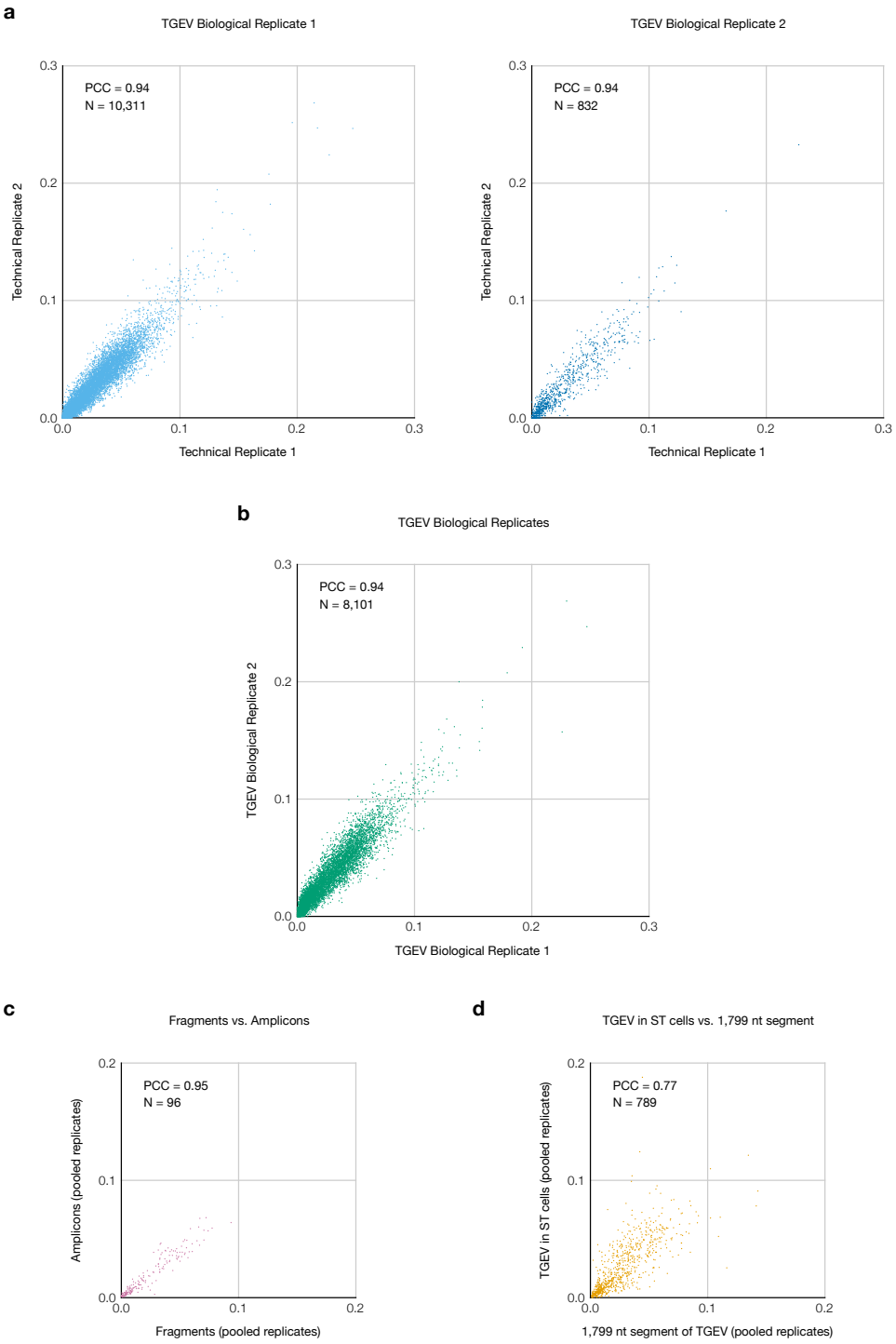
Supplementary Figure 4: **Reproducibility of clustering the SARS-CoV-2 FSE after adding ASOs. (a)** Heatmaps of the Pearson correlation coefficient (PCC) between each pair of clusters from two replicates of the 1,799 nt segment of SARS-CoV-2. Each heatmap corresponds to one order (i.e. number of clusters). Clusters are marked with red circles if at least one DMS reactivity exceeded 0.3. **(b)** Same as (a) plus Anti-AS1 ASO. **(c)** Same as (a) plus Anti-PS2-overlap ASO. **(d)** Same as (a) plus Anti-AS1 and Anti-PS2-overlap ASOs.



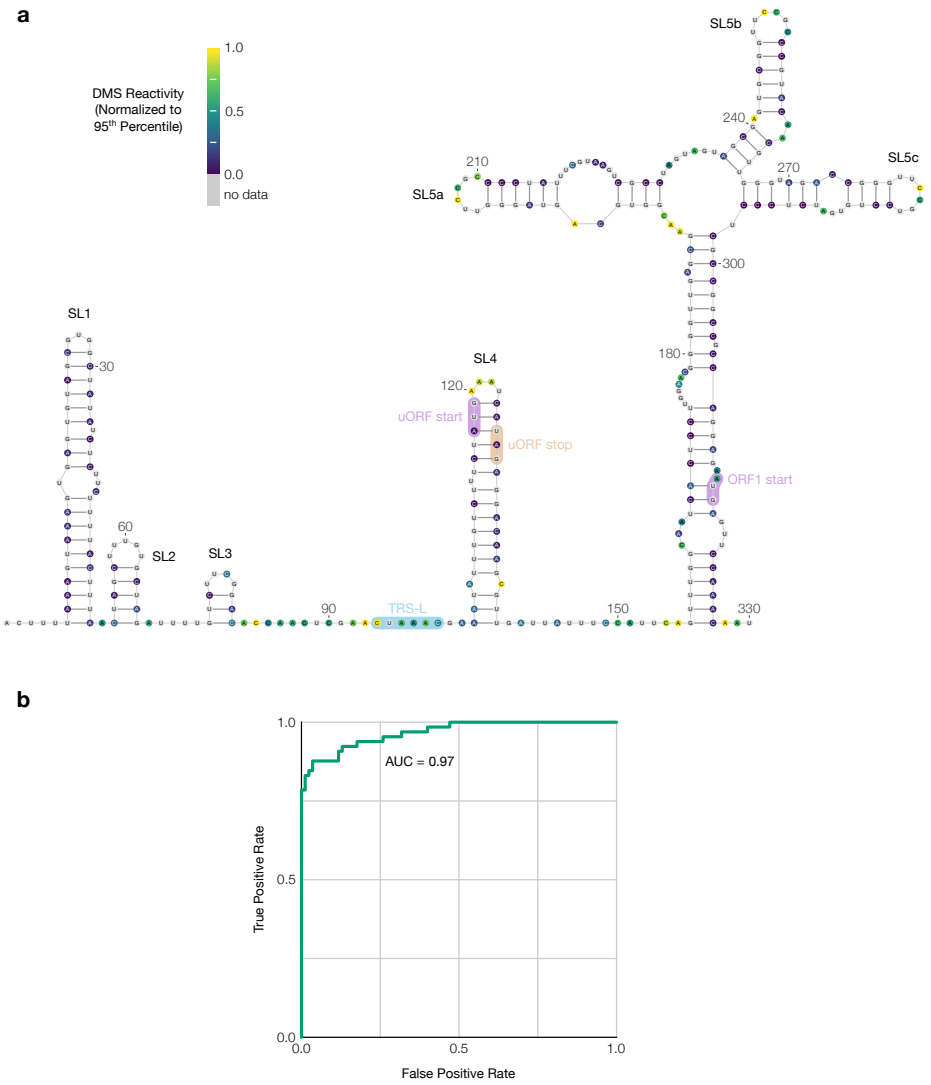
Supplementary Figure 5: **Computational screen of long-range base pairing near the FSE in 60 coronaviruses.** For each 2,000 nt segment of each coronaviral genome, the fraction of predicted structures in which each position outside the range 101-250 base-paired with any position in the range 101-250 is indicated. Genomes are clustered by their base-pairing frequencies. For each genome, the accession number for NCBI [52] is indicated.



Supplementary Figure 6: **Experimental screen of long-range base pairing near the FSE in 10 coronaviruses. (a)** Taxonomy of the ten coronavirus species/strains in this screen; the lowest-level group for each virus is bolded. Bat-CoV-1A: bat coronavirus 1A (NC_010437.1), TGEV: transmissible gastroenteritis virus (NC_038861.1), HCoV-OC43: human coronavirus OC43 (NC_006213.1), MHV-A59: murine hepatitis virus strain A59 (NC_048217.1), Bat-CoV-BM48-31: bat coronavirus BM48-31 (NC_014470.1), SARS-CoV-1: severe acute respiratory syndrome coronavirus 1 (NC_004718.3), SARS-CoV-2: severe acute respiratory syndrome coronavirus 2 (NC_045512.2), MERS-CoV: Middle East respiratory syndrome coronavirus (NC_019843.3), IBV-Beaudette: avian infectious bronchitis virus strain Beaudette (NC_001451.1), Common-Moorhen-CoV-HKU21: common moorhen coronavirus HKU21 (NC_016996.1). **(b)** Spearman correlation coefficients of DMS reactivities over the FSE between replicates 1 and 2 of short (239 nt) and long (1,799 nt) segments of each coronaviral genome.



Supplementary Figure 7: **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 8: **Secondary structure of the TGEV 5' UTR. (a)** Model of the secondary structure of the first 330 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 conserved stem loops SL1, SL2, SL3, SL4, SL5a, SL5b, and SL5c [10]. The leader transcription regulatory sequence (TRS-L) [78], upstream open reading frame (uORF) [79], and start codon of ORF1 are also labeled. The model was drawn using VARNA [80]. **(b)** Receiver operating characteristic curve showing agreement between the DMS reactivities and the secondary structure model; the area under the curve (AUC) is indicated.

Supplementary Methods

Correcting observer bias due to drop-out of reads

Let N reads from K clusters align to a reference sequence of length L. Let the proportion of reads whose 5' and 3' ends align, respectively, to coordinates a and b ($1 \leq a \leq b \leq L$) be η_{ab} (assuming these proportions are equal for all clusters). Let the mutation rate of base j ($1 \leq j \leq L$) in cluster k ($1 \leq k \leq K$) be μ_{jk}. Let the proportion of cluster k in the ensemble be π_k. To express these quantities as probabilities, let C_k be the event that a read comes from cluster k; let E_{ab} be the event that a read aligns with 5' and 3' coordinates a and b, respectively; let S_j be the event that a read contains position j (i.e. its alignment coordinates a and b satisfy $1 \leq a \leq j \leq b \leq L$); let M_j be the event that a read has a mutation at position j; and let G_g be the event that a read has no two mutations separated by fewer than g non-mutated bases.

Deriving mutation rates of reads with no two mutations too close

In terms of these events, the total mutation rates (μ_{jk}) are $P(M_j|S_jC_k)$, i.e. the probability that a read would have a mutation at position j given that it contained position j and came from cluster k; and the observable mutation rates (m_{jk}) are $P(M_j|S_jC_kG_g)$, i.e. the probability that a read would have a mutation at position j given that it contained position j, came from cluster k, and had no two mutations closer than g bases. Using these definitions and Bayes' theorem yields a probabilistic formula for m_{jk}:

$$m_{jk} = P(M_j|S_jC_kG_g) = P(M_j|S_jC_k)\frac{P(G_g|S_jM_jC_k)}{P(G_g|S_jC_k)} = \mu_{jk}\frac{P(G_g|S_jM_jC_k)}{P(G_g|S_jC_k)}$$

The term $P(G_g|S_jC_k)$ is the probability that a read would have no two mutations closer than g bases given that it contained position j and came from cluster k. It can be computed using $P(G_g|E_{ab}C_k)$ (abbreviated d_{abk}): the probability that a

read would contain no two mutations closer than g bases given that its 5' and 3' coordinates are a and b, repectively ($1 \leq a \leq b \leq L$), and that it came from cluster k. If position b were mutated (probability μ_{bk}), then the read would contain no two mutations closer than g bases if and only if none of the g bases preceding b (i.e. positions $b-g$ to $b-1$, inclusive) were mutated (probability $\prod_{j'=\max(b-g,a)}^{b-1}(1-\mu_{j'k})$, abbreviated $w_{\max(b-g,a),b-1,k}$) and two no mutations between positions a and $b-(g+1)$, inclusive, were too close (probability $d_{a,\max(b-(g+1),a),k}$)). If position b were not mutated (probability $1-\mu_{bk}$), then the read would contain no two mutations closer than g bases if and only if no mutations between positions a and $b-1$, inclusive, were too close (probability $d_{a,\max(b-1,a),k}$). These two possibilities generate a recurrence relation:

$$d_{abk} = \mu_{bk} w_{\max(b-g,a),b-1,k} d_{a,\max(b-(g+1),a),k} + (1-\mu_{bk}) d_{a,\max(b-1,a),k}$$

The base case is $d_{abk} = 1$ when $a = b$ because such a read would contain one position and thus be guaranteed to have no two mutations too close. Then, $P(G_g|S_jC_k)$ is the average of d_{abk} over every read that contains position j, weighted by the proportions η_{ab}:

$$P(G_g|S_jC_k) = \frac{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}d_{abk}}{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}}$$

The term $P(G_g|S_jM_jC_k)$ is the probability that a read would have no two mutations too close given that it contained a mutation at position j and came from cluster k. It can be computed using $P(G_g|M_jE_{ab}C_k)$ (abbreviated f_{abjk}): the probability that a read would contain no two mutations too close given that position j is mutated ($1 \leq a \leq j \leq b \leq L$), that its 5' and 3' coordinates are a and b (respectively), and that it came from cluster k. Because position j is mutated, having no two mutations too close requires that none of the g bases on both sides of position j be mutated. The probability that none of the preceding g positions ($j-g$ to $j-1$) is mutated is $w_{\max(j-g,a),j-1,k}$, while that of the following g positions ($j+1$ to $j+g$) is $w_{j+1,\min(j+g,b),k}$. Upstream of the g bases flanking position j (i.e. positions a to $j-(g+1)$), the probability that no two mutations are too close is $d_{a,\max(j-(g+1),a),k}$;

downstream (i.e. positions $j+(g+1)$ to b), the probability is $d_{\min(j+(g+1),b),b,k}$. Since mutations in these four sections are independent, the probability that the read contains no two mutations too close is the product:

$$f_{abjk} = d_{a,\max(j-(g+1),a),k} w_{\max(j-g,a),j-1,k} w_{j+1,\min(j+g,b),k} d_{\min(j+(g+1),b),b,k}$$

Then, $P(G_g|S_jM_jC_k)$ is the average of f_{abjk} over every read that contains position j, weighted by the proportions η_{ab}.

$$P(G_g|S_jM_jC_k) = \frac{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}f_{abjk}}{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}}$$

Combining the above results yields an explicit formula for m_{jk}:

$$m_{jk} = \mu_{jk}\frac{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}f_{abjk}}{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}d_{abk}}$$

Deriving end coordinate proportions of reads with no two mutations too close

The total proportions (η_{ab}) of reads aligned to 5' and 3' coordinates a and b, respectively, are $P(E_{ab})$; and the proportions of reads with no two mutations too close that align with coordinates a and b (e_{abk}) are $P(E_{ab}|G_gC_k)$. Note that, while reads are assumed to come from the same distribution of coordinates (η_{ab}) regardless of their cluster k, the observable distribution of coordinates (e_{abk}) varies by cluster because P(G_gC_k) depends on k. Using these definitions and Bayes' theorem yields a probabilistic formula for e_{abk}:

$$e_{abk} = P(E_{ab}|G_gC_k) = P(G_g|E_{ab}C_k)\frac{P(E_{ab}|C_k)}{P(G_g|C_k)} = d_{abk}\frac{\eta_{ab}}{P(G_g|C_k)}$$

The term $P(G_g|C_k)$ is the probability that a read would have no two mutations too close given that it came from cluster k. It can be computed as an average of $P(G_g|E_{ab}C_k)$ (i.e. d_{abk}) over all coordinates a and b (such that $1 \leq a \leq b \leq L$),

weighted by the proportion of each coordinate, $P(E_{ab})$ (i.e. η_{ab}):

$$P(G_g|C_k) = \frac{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}} = \sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}$$

This expression is already normalized because $\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab} = 1$, by definition.

Combining the above results yields an explicit formula for e_{abk}:

$$e_{abk} = \frac{\eta_{ab}d_{abk}}{\sum_{a'=1}^{L}\sum_{b'=a'}^{L}\eta_{a'b'}d_{a'b'k}}$$

Deriving cluster proportions of reads with no two mutations too close

The proportion of total reads in cluster k is $\pi_k = P(C_k)$. The proportion among only reads with no two mutations closer than g bases is

$$p_k = P(C_k|G_g) = P(G_g|C_k)\frac{P(C_k)}{P(G_g)} = \pi_k\frac{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{P(G_g)}$$

The term $P(G_g)$ is the probability that a read from any cluster would have no two mutations closer than g bases and can be solved for by leveraging that the cluster proportions (p_k) must sum to 1:

$$1 = \sum_{k=1}^{K}p_k = \sum_{k=1}^{K}\pi_k\frac{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{P(G_g)} = \frac{1}{P(G_g)}\sum_{k=1}^{K}\pi_k\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}$$

$$P(G_g) = \sum_{k=1}^{K}\pi_k\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}$$

The result is an explicit formula for p_k:

$$p_k = \frac{\pi_k\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{\sum_{k'=1}^{K}\pi_{k'}\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk'}}$$

Solving total mutation rates and cluster and coordinate proportions

The observed mutation rates (m_{jk}), end coordinate proportions (e_{abk}), and cluster proportions (p_k) can be calculated as weighted averages over the N reads with no

two mutations too close:

$$m_{jk} = \frac{\sum_{i=1}^{N} z_{ik} x_{ij}}{\sum_{i=1}^{N} z_{ik}}$$

$$e_{abk} = \frac{\sum_{i=1}^{N} z_{ik} y_{abi}}{\sum_{i=1}^{N} z_{ik}}$$

$$p_k = \frac{\sum_{i=1}^{N} z_{ik}}{N}$$

where x_{ij} is 1 if read i has a mutation at position j, otherwise 0; y_{abi} is 1 if read i aligns to coordinates a and b, otherwise 0; and z_{ik} is the probability that read i came from cluster k.

The original parameters μ_{jk}, η_{abk}, and π_k can be solved by setting the two formula each for m_{jk}, e_{abk}, and p_k equal to each other, creating a system of equations:

$$\mu_{jk} \frac{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} f_{abjk}}{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} d_{abk}} = m_{jk} = \frac{\sum_{i=1}^{N} z_{ik} x_{ij}}{\sum_{i=1}^{N} z_{ik}}$$

$$\eta_{ab} \frac{d_{abk}}{\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k}} = e_{ab} = \frac{\sum_{i=1}^{N} z_{ik} y_{abi}}{\sum_{i=1}^{N} z_{ik}}$$

$$\pi_k \frac{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}{\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}} = p_k = \frac{\sum_{i=1}^{N} z_{ik}}{N}$$

Solving this entire system at once has proven computationally impractical for all but extremely short sequences. A more feasible approach is to first solve for μ_{jk} given an initial guess for η_{ab}, next solve for η_{ab} given the updated μ_{jk}, then solve for π_k given the updated μ_{jk} and η_{ab}, and iterate until all three sets of parameters converge.

Even assuming every η_{ab} is a constant, these equations are still too complex to solve for μ_{jk} analytically because d_{abk} and f_{abjk} also depend on μ_{jk} (as well as on other μ variables). Thus, every μ_{jk} is solved for numerically by rearranging each equation to

$$\mu_{jk} \frac{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} f_{abjk}}{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} d_{abk}} - m_{jk} = 0$$

and applying the Netwon-Krylov method [81] implemented in SciPy [77].

Once every μ_{jk} has been solved for, every η_{ab} can be updated. Because d_{abk} does not depend on η_{ab} (except indirectly through the μ_{jk} parameters, which are

now assumed to be constants), each equation can be rearranged to

$$\eta_{ab} = \frac{e_{ab}}{d_{abk}} \sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k}$$

Leveraging that $\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} = 1$, by definition, leads to

$$\sum_{a=1}^{L} \sum_{b=a}^{L} \frac{e_{ab}}{d_{abk}} \sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k} = 1$$

$$\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k} = \frac{1}{\sum_{a=1}^{L} \sum_{b=a}^{L} \frac{e_{ab}}{d_{abk}}}$$

and finally a closed-form expression for each η_{ab} given μ_{jk} (and hence d_{abk}) and e_{abk}:

$$\eta_{ab} = \frac{\frac{e_{ab}}{d_{abk}}}{\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \frac{e_{a'b'}}{d_{a'b'k}}}$$

This equation should theoretically yield the same value of η_{ab} for every k. In practice, the values will differ due to inexactness in floating-point arithmetic. Thus, the consensus value of η_{ab} is taken to be the average η_{ab} over every k, weighted by π_k:

$$\eta_{ab} = \sum_{k=1}^{K} \pi_k \frac{\frac{e_{ab}}{d_{abk}}}{\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \frac{e_{a'b'}}{d_{a'b'k}}}$$

With updated values of μ_{jk} and η_{ab}, π_k can also be solved. The above equations can be rearranged to

$$\pi_k = p_k \frac{\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}$$

Given that $\sum_{k=1}^{K} \pi_k = 1$, by definition:

$$\sum_{k=1}^{K} p_k \frac{\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}} = 1$$

$$\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'} = \frac{1}{\sum_{k=1}^{K} \frac{p_k}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}}$$

which leads to a closed-form expression for each π_k given μ_{jk} (and hence d_{abk}), η_{ab}, and p_k:

$$\pi_k = \frac{\frac{p_k}{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}}{\sum_{k'=1}^{K}\frac{p_{k'}}{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk'}}}$$

Clustering reads with the expectation-maximization algorithm

Let N reads from K clusters align to a reference sequence of length L. Let the proportion of reads whose 5' and 3' ends align, respectively, to coordinates a and b ($1 \leq a \leq b \leq L$) be η_{ab} (assuming these proportions are equal for all clusters). Let the mutation rate of base j ($1 \leq j \leq L$) in cluster k ($1 \leq k \leq K$) be μ_{jk}. Let the proportion of cluster k in the ensemble be π_k.

Maximization step

The maximization step updates the parameters (μ_{jk}, η_{ab}, and π_k) using the current cluster memberships (z_{ik}). The observed estimates of the parameters m_{jk}, e_{ab}, and p_k are first computed; then, the underlying parameters μ_{jk}, η_{ab}, and π_k are solved for as described in 4.4.4.

Expectation step

The expectation step updates the cluster memberships (z_{ik}) and the likelihood function (L) using the current parameters (μ_{jk}, η_{ab}, and π_k). Each cluster membership is defined as the probability that read i came from cluster k given its 5'/3' end coordinates (E_{ab}) and mutations (M) and given that no two mutations are too close (G_g): $z_{ik} = P(C_k|E_{ab}MG_g)$. The likelihood of the model (L) is the product of the marginal probability (L_i) of observing each read i from any cluster: $L_i = P(E_{ab}M|G_g)$. Both L_i and z_{ik} can be expressed in terms of the joint probability ($L_{ik} = P(E_{ab}MC_k|G_g)$)

of observing each read i from each cluster k:

$$L_i = P(E_{ab}M|G_g) = \sum_{k=1}^{K} P(E_{ab}MC_k|G_g) = \sum_{k=1}^{K} L_{ik}$$

$$z_{ik} = P(C_k|E_{ab}MG_g) = \frac{P(E_{ab}MC_kG_g)}{P(E_{ab}MG_g)} = \frac{P(E_{ab}MC_k|G_g)}{P(E_{ab}M|G_g)} = \frac{L_{ik}}{L_i}$$

To derive a formula for L_{ik}, it can be factored into three parts using the chain rule for probability:

$$L_{ik} = P(E_{ab}MC_k|G_g) = \frac{P(E_{ab}MC_kG_g)}{P(G_g)} = P(M|E_{ab}C_kG_g)P(E_{ab}|C_kG_g)P(C_k|G_g)$$

The first part – the probability that a read would have the specific mutations x_{ij} given that its 5'/3' end coordinates are a and b (respectively), it comes from cluster k, and no two mutations are too close – is the product over every position j from a to b of the probability of a mutation (μ_{jk}) if read i is mutated at position j ($x_{ij} = 1$), otherwise ($x_{ij} = 0$) the probability of no mutation ($1 - \mu_{jk}$), normalized by the probability that no two mutations would be too close (d_{abk}):

$$P(M|E_{ab}C_kG_g) = \frac{1}{d_{abk}} \prod_{j=a}^{b} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{(1-x_{ij})}$$

The second part, $P(E_{ab}|C_kG_g) = e_{abk}$, can be calculated from the parameters μ_{jk}, η_{ab}, and π_k, as explained in 4.4.2. Likewise, the third part, $P(C_k|G_g) = p_k$, can also be calculated from the parameters, as explained in 4.4.3. Combining all parts yields a formula for L_{ik} in terms of the parameters μ_{jk}, η_{ab}, and π_k and of their derived values d_{abk}, e_{abk}, and p_k:

$$L_{ik} = p_k \frac{e_{abk}}{d_{abk}} \prod_{j=a}^{b} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{(1-x_{ij})}$$

The formula for the total likelihood of the model and its parameters follows:

$$L(\mu, \eta, \pi) = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k \frac{e_{abk}}{d_{abk}} \prod_{j=a}^{b} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{(1-x_{ij})}$$

Supplementary Tables

Table 1: **Sequences of the antisense oligonucleotides (ASOs) targeting the 2,924 nt segment of SARS-CoV-2 RNA.**

Group	ASO	Sequence
1	1	GGCAGCACAAGACATCTGTCGTAGTGCAACAGGACTAAGC-TCATTATT
	2	TGTAGTAAGCTAACGCATTGTCATCAGTGCAAGCAGTTTGT-GTAGTACC
	3	TGTAAATCGGATAACAGTGCAAGTACAAACCTACCTCCCTT-TGTTGTGT
	4	GATAGTACCAGTTCCATCACTCTTAGGGAATCTAGCCCATT-TCAAATCC
	5	CTTTAGGTGTGTCTGTAACAAACCTACAAGGTGGTTCCAGT-TCTGTATA
2	1	ATACCTCTATTTAGGTTGTTTAATCCTTTAATAAAGTATAAA-TACTTCACTTTAGGAC
	2	CACTTCTGTTGCATTACCAGCTTGTAGACGTACTGTGGCAG-CTAAACTACCAAGTACC
	3	AAGCTTTAGCAGCATCTACAGCAAAAGCACAGAAAGATAA-TACAGTTGAATTGGCAGG
	4	CACAACATCTTAACACAATTAGTGATTGGTTGTCCCCCACT-AGCTAGATAATCTTTGT
3	1	GATCCATATTGGCTTCCGGTGTAACTGTTATTGCCTGACCA-GTACCAGTGTGTGTA
	2	ATGATCTATGTGGCAACGGCAGTACAGACAACACGATGCA-CCACCAAAGGATTCTT

Continued on next page

Table 1: **Sequences of the antisense oligonucleotides (ASOs) targeting the 2,924 nt segment of SARS-CoV-2 RNA.** (Continued)

	3	GTTGTAGGTATTTGTACATACTTACCTTTTAAGTCACAAAATCCTTTAGGATTTGG
	4	CCGCAGACGGTACAGACTGTGTTTTTAAGTGTAAAACCCACAGGGTCATTAGCACAA
4	1	CTGAAGCATGGGTTCGCGGAGTTGATCACAACTACAGCCATAACCTTTCCACATA
	2	AAGACGGGCTGCACTTACACCGCAAACCCGTTTAAAAACGATTGTGCATCAGCTGA
	3	TAGATGTCAAAAGCCCTGTATACGACATCAGTACTAGTGCCTGTGCCGCACGGTGT
	4	GGAAGCGACAACAATTAGTTTTTAGGAATTTAGCAAAACCAGCTACTTTATCATTG
5	1	TGTCTCTTAACTACAAAGTAAGAATCAATTAAATTGTCATCTTCGTCCTTTTCTT
	2	GACAATCCTTAAGTAAATTATAAATTGTTTCTTCATGTTGGTAGTTAGAGAAAGTG
	3	GGTACCATGTCACCGTCTATTCTAAACTTAAAGAAGTCATGTTTAGCAACAGCTG
	4	AAGCATAGACGAGGTCTGCCATTGTGTATTTAGTAAGACGTTGACGTGATATATGT
6	1	TGTATGTGACAAGTATTTCTTTTAATGTGTCACAATTACCTTCATCAAAATGCCTTA
	2	GGTTTTCTACAAAATCATACCAGTCCTTTTTATTGAAATAATCATCATCACAACAAT

Continued on next page

Table 1: **Sequences of the antisense oligonucleotides (ASOs) targeting the 2,924 nt segment of SARS-CoV-2 RNA.** (Continued)

	3	TTAACAAAGCTTGGCGTACACGTTCACCTAAGTTGGCGTATACGCGTAATATATCTG
	4	ATGTCAGTACACCAACAATACCAGCATTTCGCATGGCATCACAGAATTGTACTGTTT
7	1	GTTTGTATGAAATCACCGAAATCATACCAGTTACCATTGAGATCTTGATTATCTA
	2	TAGGCATTAACAATGAATAATAAGAATCTACAACAGGAACTCCACTACCTGGCGTG
	3	GTTAAGTCAGTGTCAACATGTGACTCTGCAGTTAAAGCCCTGGTCAAGGTTAATA
	4	TTAACCTCTCTTCCGTGAAGTCATATTTTAACAAATCCCACTTAATGTAAGGCTTT
8	1	AACACAATTTGGGTGGTATGTCTGATCCCAATATTTAAAATAACGGTCAAAGAGTT
	2	GAGAATAAAACATTAAAGTTTGCACAATGCAGAATGCATCTGTCATCCAAACAGTT
	3	CATCAACAAATATTTTTCTCACTAGTGGTCCAAAACTTGTAGGTGGGAACACTGTA
	4	ATGTACAACACCTAGCTCTCTGAAGTGGTATCCAGTTGAAACTACAAATGGAACAC
9	1	TACACAAGTAATTCCTTAAAACTAAGTCTAGAGCTATGTAAGTTTACATCCTGATT
	2	TGCGTTTATCTAGTAATAGATTACCAGAAGCAGCGTGCATAGCAGGGTCAGCAGCA

Continued on next page

Table 1: **Sequences of the antisense oligonucleotides (ASOs) targeting the 2,924 nt segment of SARS-CoV-2 RNA.** (Continued)

	3	TTTGACAGTTTGAAAAGCAACATTGTTAGTAAGTGCAGCTA-CTGAAAAGCACGTAG
	4	CTTAAAGAAACCCTTAGACACAGCAAAGTCATAGAAGTCTT-TGTTAAAATTACCGGG
10	1	CAGCATTACCATCCTGAGCAAAGAAGAAGTGTTTTAATTCA-ACAGAACTTCCTTC
	2	CTGATATCACACATTGTTGGTAGATTATAACGATAGTAGTC-ATAATCGCTGATAG
	3	ACCATCGTAACAATCAAAGTACTTATCAACAACTTCAACTA-CAAATAGTAGTTGT
	4	AACCAGCTGATTTGTCTAGGTTGTTGACGATGACTTGGTTA-GCATTAATACAGCC
11	1	CCTCATAACTCATTGAATCATAATAAAGTCTAGCCTTACCC-CATTTATTAAATGGAA
	2	ATTTGAGTTATAGTAGGGATGACATTACGTTTTGTATATGC-GAAAAGTGCATCTTGAT
	3	GAGACACCAGCTACGGTGCGAGCTCTATTCTTTGCACTAAT-GGCATACTTAAGATTC
	4	GGCTATTGATTTCAATAATTTTTGATGAAACTGTCTATTGGT-CATAGTACTACAGATA
12	1	CAACCACCATAGAATTTGCTTGTTCCAATTACTACAGTAGC-TCCTCTAGTGGC
	2	CCATAAGGTGAGGGTTTTCTACATCACTATAAACAGTTTTT-AACATGTTGTGC

Continued on next page

Table 1: **Sequences of the antisense oligonucleotides (ASOs) targeting the 2,924 nt segment of SARS-CoV-2 RNA.** (Continued)

	3	CATAATTCTAAGCATGTTAGGCATGGCTCTATCACATTTAG-GATAATCCCAAC
	4	ACGGTGTGACAAGCTACAACACGTTGTATGTTTGCGAGCA-AGAACAAGTGAGGC
13	1	ACACATGACCATTTCACTCAATACTTGAGCACACTCATTAG-CTAATCTATAGAA
	2	AGTTGTGGCATCTCCTGATGAGGTTCCACCTGGTTTAACAT-ATAGTGAACCGCC
	3	ATTAACATTGGCCGTGACAGCTTGACAAATGTTAAAAACAC-TATTAGCATAAGC
	4	TAAATTGCGGACATACTTATCGGCAATTTTGTTACCATCAG-TAGATAAAAGTGC

Table 2: **Sequences of the forward (F) and reverse (R) primers for amplifying the binding site of each ASO group in the 2,924 nt segment of SARS-CoV-2 RNA.**

Group	Primer	Sequence
1	F	AATAATGAGCTTAGTCCTGTTGCACTACG
	R	AGGTTGTTTAATCCTTTAATAAAGTATAAATACTTCACTTT-AGG
2	F	ACCTTGTAGGTTTGTTACAGACACACCTAA
	R	TTGCCTGACCAGTACCAGTGTGTG
3	F	GGACAACCAATCACTAATTGTGTTAAGATGTTG
	R	TCACAACTACAGCCATAACCTTTCCACA
4	F	CTTAAAAACACAGTCTGTACCGTCTGC
	R	GTAAGAATCAATTAAATTGTCATCTTCGTCCTTTTC
5	F	TGCTAAATTCCTAAAAACTAATTGTTGTCGCTT
	R	ATGTGTCACAATTACCTTCATCAAAATGCCT
6	F	CAATGGCAGACCTCGTCTATGC
	R	GAAATCATACCAGTTACCATTGAGATCTTGATTATC
7	F	CGAAATGCTGGTATTGTTGGTGTACTGAC
	R	GTCTGATCCCAATATTTAAAATAACGGTCAAAGAG
8	F	TGTTAAAATATGACTTCACGGAAGAGAGGTT
	R	AAGTCTAGAGCTATGTAAGTTTACATCCTGA
9	F	CCACTTCAGAGAGCTAGGTGTTGTAC
	R	CAAAGAAGAAGTGTTTTAATTCAACAGAACTTCCT
10	F	TGACTTTGCTGTGTCTAAGGGTTTCTTTA
	R	CATAATAAAGTCTAGCCTTACCCCATTTATTAAATGG
11	F	CGTCAACAACCTAGACAAATCAGCTGG

Continued on next page

Table 2: **Sequences of the forward (F) and reverse (R) primers for amplifying the binding site of each ASO group in the 2,924 nt segment of SARS-CoV-2 RNA.** (Continued)

	R	TTCCAATTACTACAGTAGCTCCTCTAGTG
12	F	GACCAATAGACAGTTTCATCAAAAATTATTGAAATCAATA-G
	R	ATACTTGAGCACACTCATTAGCTAATCTATAG
13	F	ACAACGTGTTGTAGCTTGTCACACC
	R	TAAATTGCGGACATACTTATCGGCAATTTTG

Table 3: **Sequences of the forward (F), forward with T7 promoter (F+T7), and reverse (R) primers for amplifying the short segment of each 1,799 nt segment of coronaviral RNAs.**

Coronavirus	Primer	Sequence
Bat Coronavirus 1A	F	GGACCCTATACGGTTTTGTCTTGAAAA
	F+T7	TAATACGACTCACTATAGGACCCTATACGGTTTTGTCTTGAAAA
	R	TTTTACAATAAAGAAAGCATCATGCTT
Bat Coronavirus BM48-31	F	GGGTTTTATTCTTAGAAACACAGTCTG
	F+T7	TAATACGACTCACTATAGGGTTTTATTCTTAGAAACACAGTCTG
	R	GGAGTCTAATAAGTTGCCCTCTTCATC
Common Moorhen Coronavirus	F	GGATAAAGATAAGGAACCTGTTTCTTT
	F+T7	TAATACGACTCACTATAGGATAAAGATAAGGAACCTGTTTCTTT
	R	ACTATTAGGTATTGGCAAATTAATGCG
Human Coronavirus OC43	F	GGCTGTGTCTTATGTTTTGACACATGA

Continued on next page

Table 3: **Sequences of the forward (F), forward with T7 promoter (F+T7), and reverse (R) primers for amplifying the short segment of each 1,799 nt segment of coronaviral RNAs.** (Continued)

	F+T7	TAATACGACTCACTATAGGCTGTGTCTTATGTTTTGACACATGA
	R	ATCTAATTTATCACCGTTCTCATCAAC
Infectious Bronchitis Virus	F	GGTTTGCACTGTTTGCCAGTGTTGGAT
	F+T7	TAATACGACTCACTATAGGTTTGCACTGTTTGCCAGTGTTGGAT
	R	CTCAAGATTTCCATCTTCAGTATCGCG
MERS Coronavirus	F	GGGATTTTGTTTGTCAAATACCCCCTG
	F+T7	TAATACGACTCACTATAGGGATTTTGTTTGTCAAATACCCCCTG
	R	ATGATGCCCTTGGTCATCTAATTCTAC
Murine HepATitis Virus	F	GGCTGTGTCATATGTGTTGACGCATGA
	F+T7	TAATACGACTCACTATAGGCTGTGTCATATGTGTTGACGCATGA

Continued on next page

Table 3: **Sequences of the forward (F), forward with T7 promoter (F+T7), and reverse (R) primers for amplifying the short segment of each 1,799 nt segment of coronaviral RNAs.** (Continued)

	R	ATCCAACTTGTTGCCGTCCTCATCTAC
SARS Coronavirus 1	F	GGGTTTTACACTTAGAAACACAGTCTG
	F+T7	TAATACGACTCACTATAGGGTTTTACACTTAGAAACACAGTCTG
	R	AGAGTCTAATAAATTGCCTTCCTCATC
SARS Coronavirus 2	F	GGGTTTTACACTTAAAAACACAGTCTG
	F+T7	TAATACGACTCACTATAGGGTTTTACACTTAAAAACACAGTCTG
	R	AGAATCAATTAAATTGTCATCTTCGTC
Transmissible Gastroenteritis Virus	F	GGCAATTCGGTTCTGTATTGAAAATGA
	F+T7	TAATACGACTCACTATAGGCAATTCGGTTCTGTATTGAAAATGA
	R	TTTGACAATGTAGTAGGCATCATGTTT

Table 4: **Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs.**

Coronavirus	ASO	Sequence
Bat Coronavirus 1A	1	CAGGGCTCTAGTCGAGCTGCAC-TAGAGCCCCTTGCTCGTTTAAA-TAAGCCTGATCAACAG
	2	GCAACTTCTTTATTGTAAATATC-AAAGGCGCGTACAACATGCTCC-GGTTCAGTACCATTA
Bat Coronavirus BM48-31	1	GACATCAGTGCTTGTGCCTGTG-CCGCACGGTGTAAGACGGGCC-GCACTTACACCGCAAAC
	2	TTTTAGGAACTTTGCAAAACCA-GCAACTTTCTCATTATAAATATC-AAAAGCCCTGTAAAC
	3	AAAATAGGAGTCTAATAAGTTG-CCCTCTTCATCAACTTCCTGGAA-ACGGCAACAATTTGT
Common Moorhen Coronavirus	1	TGGGGTTCTAGACGGGCATCAC-TAGAACCCTTTACTCGTTTAAAT-AAGCTGTATTTTGCA
	2	GTTATATTATTATGTACATGAAA-CGCCCTTTTTACAATATCCGGCT-GAGTGCCAGACTGT
Infectious Bronchitis Virus	1	ACATCAAAGGCTCGCTTTACAA-CATCAGGATCACATCCACTAGC-AAGGGGTATCAGCCGA

Continued on next page

Table 4: **Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs.** (Continued)

Murine Hepatitis Virus	1	AAGCCACTGGCACAGGGTACAA-GACGGGCATTTACACTTGTACC-CCGAATCCGTTTAAAA
	2	CCAATGCCAGCTCGATTAGCAT-TACAAATGTCAAATGCCCTTAA-TTGAACATCAGTGTCC
	3	AACTTGTTGCCGTCCTCATCTAC-ACGCTGGAAGCGGCAGCAATTC-ACTTTATAATACAAA
SARS Coronavirus 1	1	TTTGCAAAACCAGCAACTTTTTC-GTTGTAAATATCAAAAGCCCTG-TAGACGACATCAGTA
	2	TCTAATAAATTGCCTTCCTCATC-CTTCTCCTGGAAGCGACAGCAA-TTAGTTTTTAGGAAC
SARS Coronavirus 2	1	GACATCAGTACTAGTGCCTGTG-CCGCACGGTGTAAGACGGGCT-GCACTTACACCGCAAAC
	2	TTTTAGGAATTTAGCAAAACCA-GCTACTTTATCATTGTAGATGTC-AAAAGCCCTGTATAC
Transmissible Gastroenteritis Virus	1	TAAATAACTTTGATCAACAGTA-AAACTCTGCATAGAAGTACGAT-CGCACATGCAACCATT

Continued on next page

Table 4: **Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs.** (Continued)

2	GGTCTGGATCAGTACCATTGCA-GGGTTCTAGTCGAGCTGCACTA-GAACCCCGCACTCGTT