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

Abstract

In every living organism and virus, RNA molecules perform a diversity of essential functions for which their linear sequences must fold into higher-order structures. Techniques including crystallography and cryogenic electron microscopy have revealed 3D structures of ribosomal, transfer, and other well-structured RNAs; while chemical probing with sequencing facilitates secondary structure modeling of arbitrary RNAs, even within cells. Ongoing efforts continue increasing the accuracy, resolution, and ability to distinguish coexisting alternative structures. However, no method can identify and quantify alternative structures with base pairs spanning arbitrarily long distances, which, as mounting evidence indicates, occur abundantly in the genomes of RNA viruses. Here, we develop the method of Structure Ensemble Ablation by Reverse Complement Hybridization with Mutational Profiling (SEARCH-MaP) and the data analysis software Structure Ensemble Inference by Sequencing, Mutation Identification, and Clustering of RNA (SEISMIC-RNA). We use SEARCH-MaP and SEISMIC-RNA to discover that the frameshift stimulating element of SARS coronavirus 2 base-pairs with another element 1 kilobase downstream in nearly half of RNA molecules, and that this structure inhibits the folding of a pseudoknot that stimulates ribosomal frameshifting. Moreover, we identify long-range base pairs involving the frameshift stimulating element in other coronaviruses including SARS coronavirus 1 and transmissible gastroenteritis virus, and model the full genomic secondary structure of the latter. These findings suggest that stable long-range base pairs are common in coronaviruses and may regulate ribosomal frameshifting, which is essential for viral RNA synthesis. We anticipate that SEARCH-MaP and SEISMIC-RNA will facilitate studies on viral, messenger, and long noncoding RNAs – particularly of alternative structures and long-range base pairing. Ultimately, the results of many such studies could be collected in a vast database of long RNA structures for training and benchmarking RNA folding algorithms, making possible a true “AlphaFold for RNA”.

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

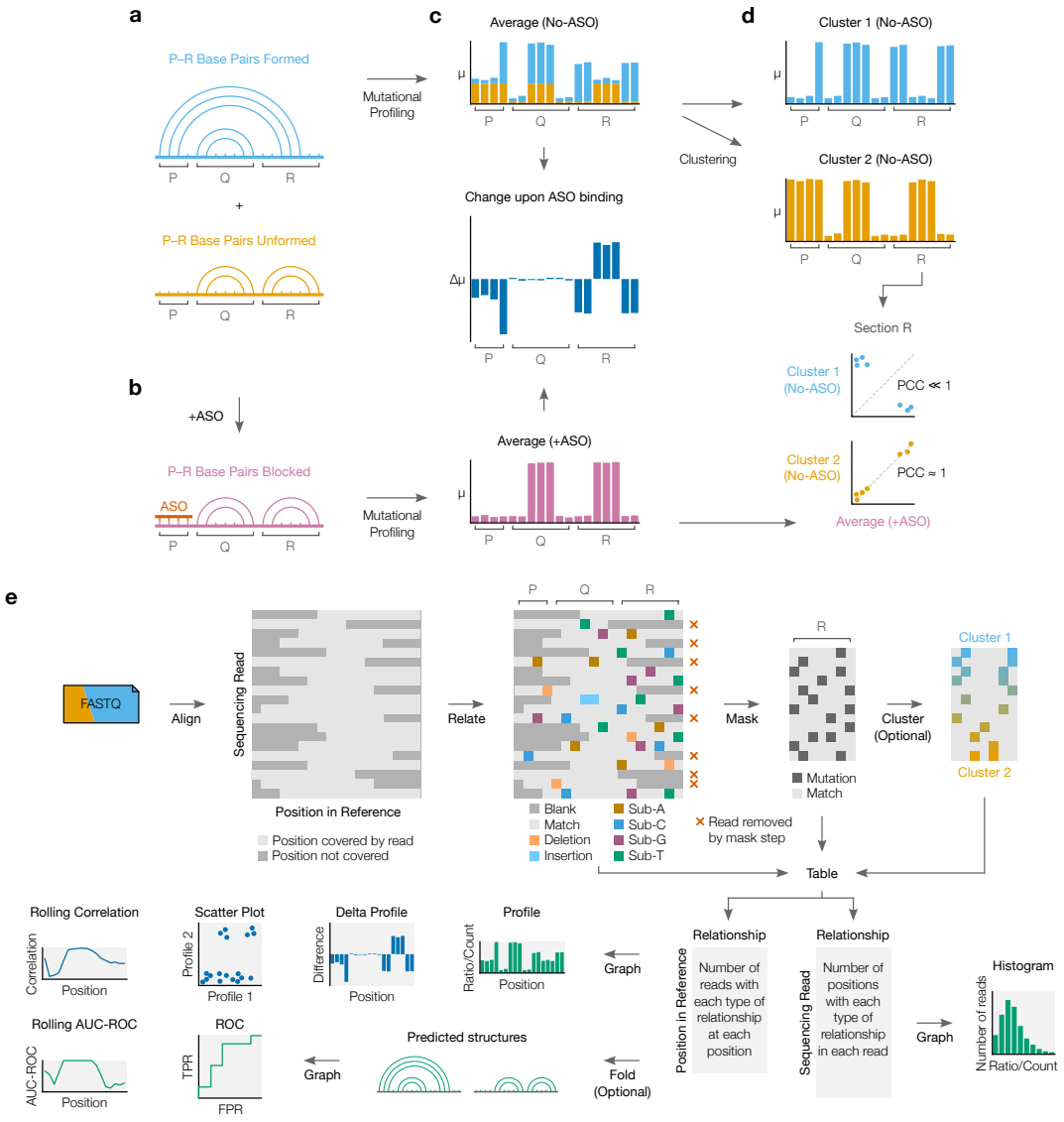


Figure 1: **The workflow of SEARCH-MaP and SEISMIC-RNA.** (Continued on next page.)

Figure 1: (Continued from previous page.) **(a)** This 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 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]. In coronaviruses, the first open reading frame (ORF1) contains a frameshift stimulation element (FSE) that makes a fraction of ribosomes slip into the -1 reading frame, bypass a 0-frame stop codon, and translate to the end of ORF1 [42]. Every coronaviral FSE contains a "slippery site" (UUUAAAC) and a structure characterized as a pseudoknot in multiple species [43, 44, 45]. For SARS coronavirus 2 (SARS-CoV-2), 80-90 nt segments of the core FSE have been shown to fold into a pseudoknot with three stems [39, 46, 47]. However, in intact SARS-CoV-2, the FSE adopts a different structure: one that could be recapitulated with a 2,924 nt segment but not a 283 nt

segment [48]. This finding suggested that the SARS-CoV-2 FSE involves long-range base pairs.

We hypothesized that the long-range base pairs would match a proposed structure named the "FSE-arch" [49]. If so, the structure of the FSE would be perturbed by – and only by – ASOs targeting either side of the FSE-arch. To investigate, we performed DMS-MaPseq [29] on a 2,924 nt segment of SARS-CoV-2 after adding each of thirteen groups of DNA ASOs (Figure 2a). We used RT-PCR primers flanking the ASO target site to confirm binding (Supplementary Figure 1) and flanking the 5' side of the FSE-arch to measure changes in its structure (Supplementary Figure 2), except for ASO group 13, for which we obtained no data.

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 2b). 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, 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 (but not the outer stem) of the proposed FSE-arch [49] exist and are the predominant long-range base pairs involving the immediate vicinity of the FSE.

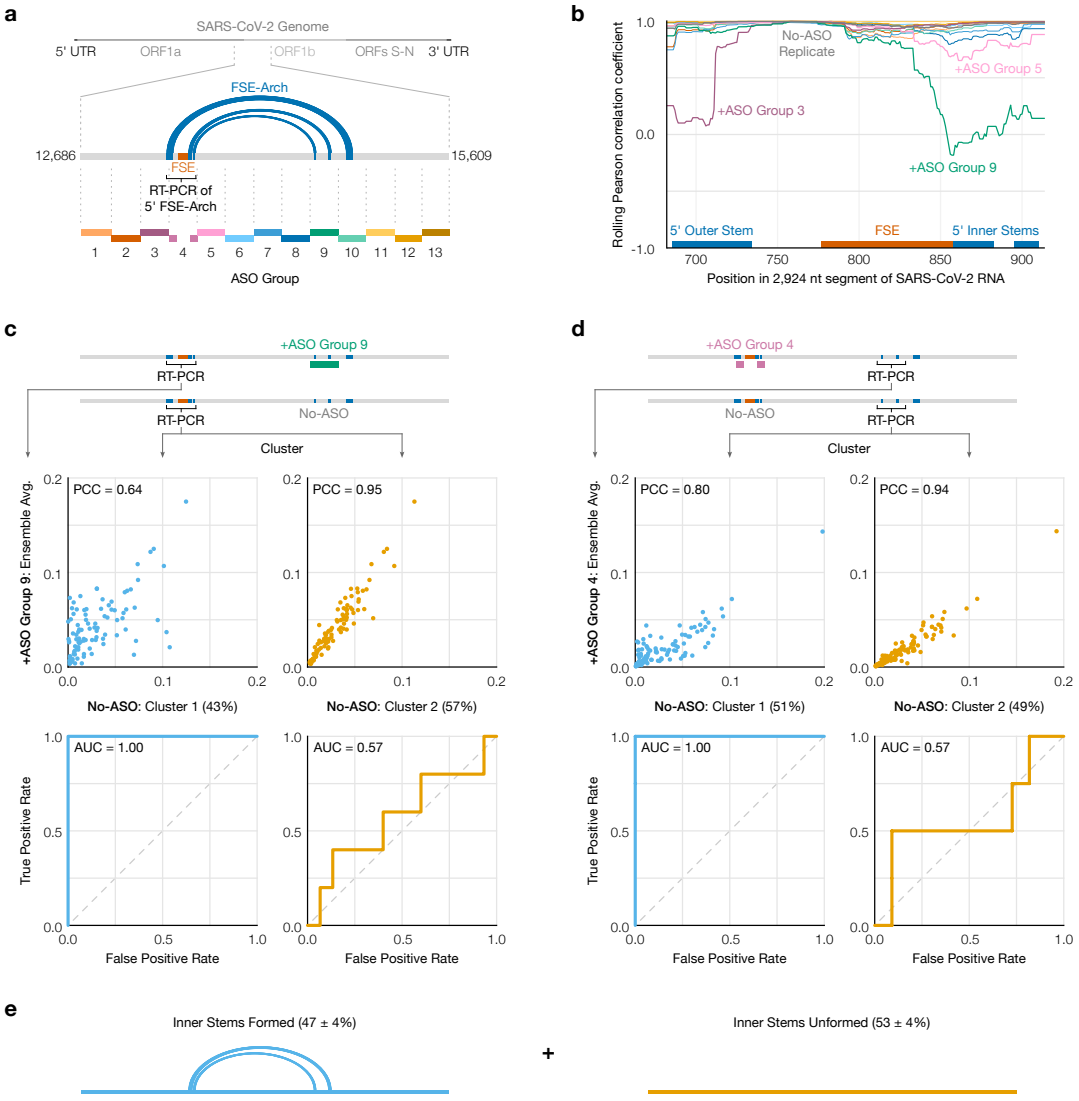


Figure 2: **SEARCH-MaP of long-range base pairs involving the SARS-CoV-2 FSE.** **(a)** The 2,924 nt segment of the SARS-CoV-2 genome containing the frameshift stimulation element (FSE) and putative FSE-arch [49]. The target site for each group of antisense oligonucleotides (ASOs) is indicated by dotted 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, with PCC indicated; each point is one base 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.

To determine in what fraction of molecules the two inner stems of the FSE-arch form, we clustered reads from the 5' side of the FSE-arch for the no-ASO control. We found two clusters with a 43/57% split and – to determine if they corresponded to the two inner stems formed versus unformed – compared their DMS reactivities to those after adding ASO group 9, which would block the two inner stems (Figure 2c, 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 [49]. 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 2c, 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 states with the two inner stems formed and unformed. We thus RT-PCRed and clustered the 3' FSE-arch in the no-ASO control and +ASO group 4 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 2d). 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 2e).

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 [49], 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 3) included not only the two inner stems of the FSE-arch – LS1 and LS2a/b – but also two long stems (LS3a/b and LS4) absent from the original FSE-arch model [49], as well as alternative stem 1 (AS1) [48]. 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 3b, Supplementary Table 3). 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.

LS2b, LS3, and LS4 of the refined model overlap all three stems of the pseudoknot (PS1, PS2, and PS3) that stimulates frameshifting [38, 39, 47]. To test whether these long stems actually compete with the pseudoknot, we first generated four possible models of the FSE structure by combining mutually compatible stems from the refined model (Figure 3c). 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 3d, 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 3c.

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 3d, 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 3d, right; Supplementary Figure 4c). Adding both ASOs simultaneously collapsed the ensemble into three clusters of which two (87%) were highly consistent with the pseudoknotted model (Figure 3d, 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 – the long-range stems did appear to outcompete the pseudoknot.

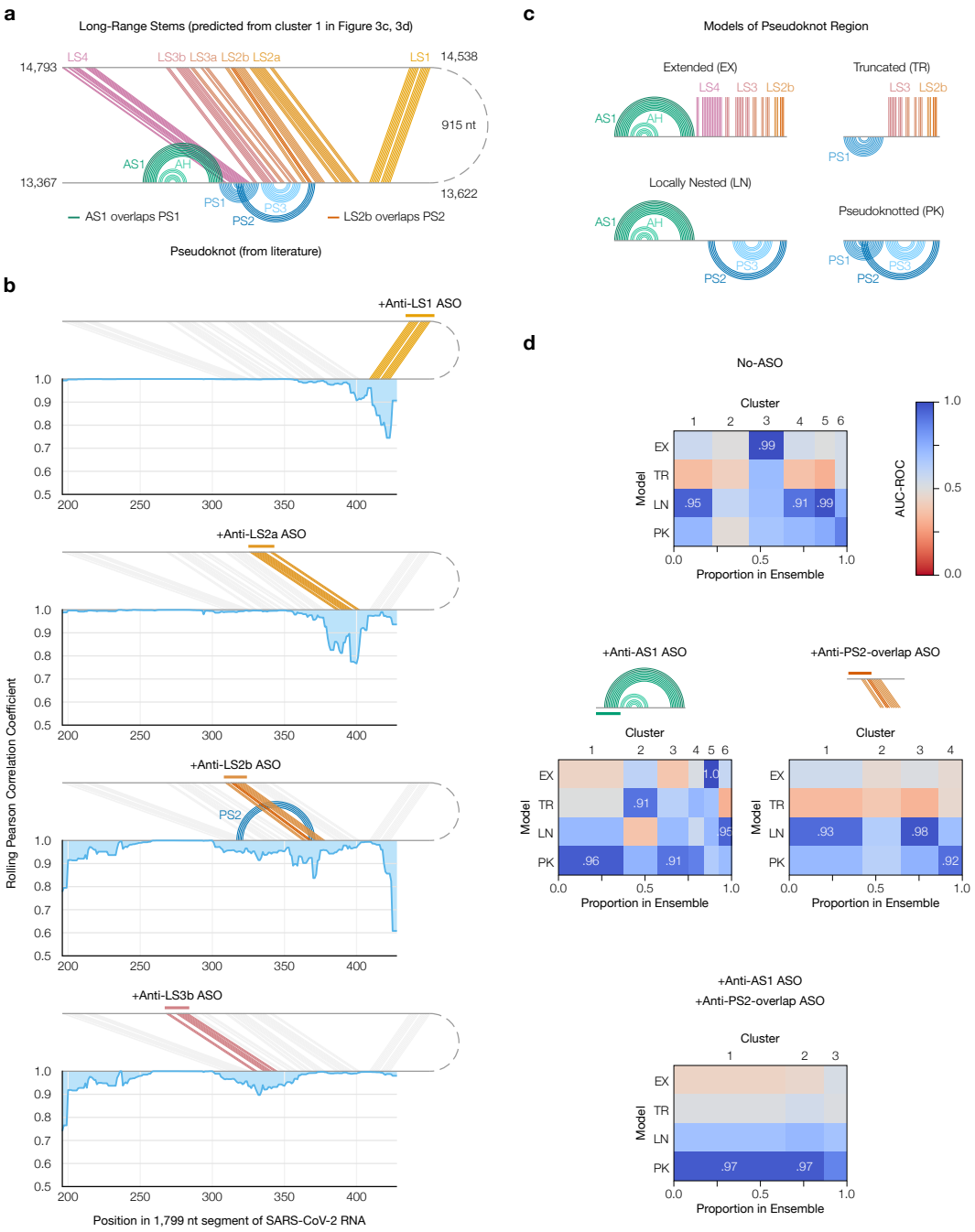


Figure 3: **Refinement of the long-range structure model and competition with the frameshift pseudoknot. (a)** Refined model of the long-range stems (minimum free energy prediction based on cluster 1 in Figure 2c and d) including alternative stem 1 (AS1) [48]; the attenuator hairpin (AH) [50]; and long 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 identify coronaviruses with potential long-range base pairs involving the FSE, we downloaded 62 complete coronavirus genomes from the NCBI Reference Sequence Database [51] and predicted structures of 2,000 nt sections surrounding their FSEs (Supplementary Figure 5). We selected ten coronaviruses, at least one from each genus (Supplementary Figure 6a), based on with which bases the FSE was predicted to base-pair. In *Betacoronavirus*, 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; MERS coronavirus (`NC_019843.3`), predicted to pair with positions 510-530; and human coronavirus OC43 (`NC_006213.1`) and murine hepatitis virus strain A59 (`NC_048217.1`), both predicted to pair with positions 10-20. In *Alphacoronavirus*, transmissible gastroenteritis virus (`NC_038861.1`) and bat coronavirus 1A (`NC_010437.1`), predicted to pair with positions 440-460 and 350-360, respectively. In *Gammacoronavirus*, avian infectious bronchitis virus strain Beaudette (`NC_001451.1`), predicted to pair with positions 330-350. And in *Deltacoronavirus*, common moorhen coronavirus HKU21 (`NC_016996.1`) had the most promising long-range base pairs.

We screened each of these ten coronaviruses for long-range base pairs with the FSE by comparing the DMS reactivities of 239 nt segment comprising the FSE with 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 and 1,799 nt segments (Supplementary Figure 6b), suggesting their FSEs formed 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 tar-

geting the vicinity of the FSE (Figure 4). 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 4, 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.

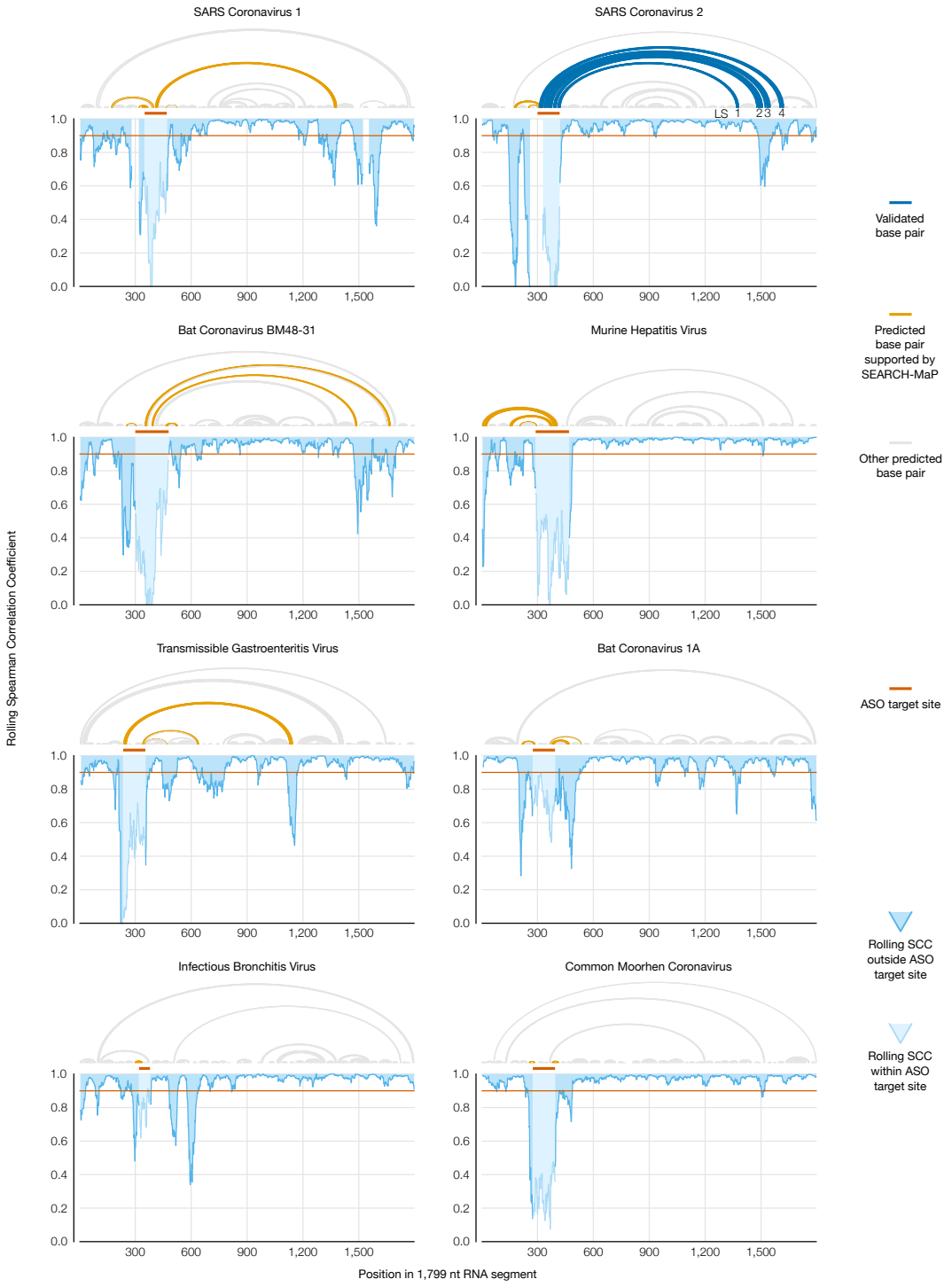


Figure 4: **Evidence for long-range RNA–RNA base pairs involving the FSE in four additional coronaviruses.** Rolling (window = 45 nt) Spearman correlation coefficient (SCC) of DMS reactivities between the +ASO and no-ASO samples for each 1,799 nt segment of a coronaviral genome. The target site of each ASO is highlighted on the SCC data and shown above each graph. Structures predicted with RNAstructure [52] 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 3a) 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 determine what long-range base pairs could have caused those dips, we used RNAstructure Fold [52, 33] guided by the DMS reactivities of the no-ASO ensemble average, as clustered data – while more accurate (Supplementary Figure 3) – were unavailable. Nevertheless, we were able to find long-range base pairs consistent with the SEARCH-MaP data for both murine hepatitis virus and transmissible gastroenteritis virus (Figure 4, 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, often fatally [54]. Due to the impacts of TGEV [54] and our evidence of long-range base pairs, we sought to model the genomic secondary structures of live TGEV. We began by infecting ST cells with TGEV and performing DMS-MaPseq [29] (Figure 5a). The DMS reactivities over the full genome were consistent between technical and biological replicates (PCC = 0.94, Supplementary Figure 7a and b), albeit not with the 1,799 nt segment *in vitro* (PCC = 0.77), which showed that verifying the long-range stem in live TGEV would be necessary (Supplementary Figure 7c).

To quantify the long-range base pairs, we RT-PCRed the 5' and 3' sides, confirmed their DMS reactivities were consistent with the full genome's (Supplementary Figure 7d), and clustered both sides (Figure 5b). Although the clusters were indistinguishable by their correlations with the +ASO sample or AUC-ROC scores (Supplementary Figure 8a and b), bases involved in the predicted long-range pairs were generally less DMS-reactive in cluster 2 (Figure 5b), which we hypothesized corresponded to the long-range base pairs forming. In support, the long-range stem (hereafter, LS3) appeared when using DMS reactivities from cluster 2 on both

sides (Figure 5c), but not cluster 1 (Supplementary Figure 8c). The refined model based on cluster 2 included another long-range stem, LS2, which was also supported by a dip in the rolling SCC (Figure 5c).

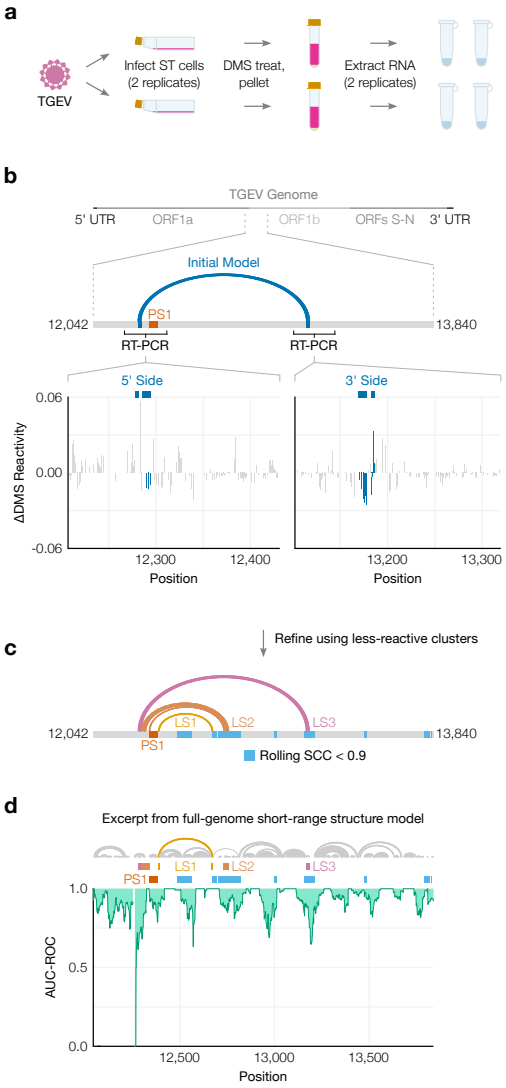


Figure 5: **Genomic secondary structure of live TGEV. (a)** Schematic of the experiment in which two biological replicates of ST cells were infected with TGEV, DMS-treated, and pelleted. Cell pellets were divided into two technical replicates prior to extraction of DMS-modified RNA. **(b)** Differences 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 4), 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 4), 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 short-range (up to 300 nt) base pairs in the full TGEV genome (Supplementary Figure 9). To verify the model quality, we confirmed that the modeled structure of the first 330 nt included the highly conserved stem loops SL1, SL2, SL4, and SL5a/b/c in the 5' UTR [10] (Supplementary Figure 10a) and was consistent with the DMS reactivities (AUC-ROC = 0.97) (Supplementary Figure 10b). The AUC-ROC was lower in many locations throughout the rest of the genome (Supplementary Figure 9), indicating that a single secondary structure consistent with the ensemble average DMS reactivities could not be found – which suggests alternative structures, long-range base pairs, or both [48]. Accordingly, 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 5d). 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 that regions with low AUC-ROC in general are good starting points for investigating alternative structures and long-range base pairing with SEARCH-MaP.

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 3) and TGEV (Figure 5).

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 [49]. 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 5d). 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 [43, 44, 45, 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, 48]. 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 [73].

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) [74]. 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 [75] 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 [76]. 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 2), 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

Development of SEISMIC-RNA

SEISMIC-RNA was written in Python (currently compatible with v3.10 or greater) using PyCharm Community Edition. Its dependencies include Python packages NumPy [77], Numba [78], pandas [79], and SciPy [80]; as well as Samtools [81], Cutadapt [82], Bowtie 2 [83], and RNAstructure [52].

SEARCH-MaP of 2,924 nt SARS-CoV-2 RNA

Synthesis of 2,924 nt SARS-CoV-2 RNA

A DNA template of the 2,924 nt segment of SARS-CoV-2, including a T7 promoter, was amplified from a previously constructed plasmid [48] ([Supplementary Data]) in 50 µl using 2X CloneAmp HiFi PCR Premix (Takara Bio) with 250 nM primers TAATACGACTCACTATAGAATAATGAGCTTAGTCCTGTTGCACTACG and TAAATTGCGGACATACTTATCGGCAATTTTGTTACC (Thermo Fisher Scientific); 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 with 10 µl of 6X Purple Loading Dye (New England Biolabs) was electrophoresed through a 50 ml gel – 1% SeaKem Agarose (Lonza), 1X tris-acetate-EDTA (Boston BioProducts), and 1X SYBR Safe (Invitrogen) – at 60 V for 60 min. The band at roughly 3 kb was extracted using a Zymoclean Gel DNA Recovery Kit (Zymo Research) according to the manufacturer's protocol, eluted in 10 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific). To increase yield, the gel-extracted DNA was fed into a second round of PCR and gel extraction using the same protocol. Due to remaining contaminants, the DNA was further purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 10 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific).

150 ng of DNA template was transcribed using a MEGAscript T7 Transcription Kit (Invitrogen) according to the manufacturer's protocol, incubating at 37°C for 3 hr. DNA template was then degraded by incubating with 1 µl of TURBO DNase (Invitrogen) at 37°C for 15 min. RNA was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 20 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific).

DMS treatment of 2,924 nt SARS-CoV-2 RNA

Antisense oligonucleotides (ASOs) were ordered from Integrated DNA Technologies already resuspended to 10 µM in 1X IDTE buffer (10 mM Tris, 0.1 mM EDTA) in a 96-well PCR plate. Each ASO pool was assembled from 25 pmol of each constituent ASO (Supplementary Table 1); volume was adjusted to 12.5 µl by adding TE Buffer – 10 mM Tris (Invitrogen) with 0.1 mM EDTA (Invitrogen). 450 fmol of 2,924 nt SARS-CoV-2 RNA was added to each ASO pool for a total of 13.5 µl in a PCR tube. The tube was heated to 95°C for 60 s to denature the RNA, placed on ice for several minutes, and transferred to a 1.5 ml tube. To refold the RNA, 35 µl of 1.4X refolding buffer comprising 400 mM sodium cacodylate pH 7.2 (Electron Microscopy Sciences) and 6 mM magnesium chloride (Invitrogen) was added, then incubated at 37°C for 25 min. For no-ASO control 1, 12.5 µl of TE Buffer was used instead of an ASO pool. For no-ASO control 2, 12.5 µl of TE Buffer was added after placing on ice and before refolding to confirm the timing of adding TE Buffer would not alter the RNA structure.

RNA was treated with DMS in 50 µl containing 1.5 µl (320 nM) of DMS (Sigma-Aldrich) while shaking at 500 rpm in a ThermoMixer C (Eppendorf) at 37°C for 5 min. To quench, 30 µl of beta-mercaptoethanol (Sigma-Aldrich) was added and mixed thoroughly. RNA was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 10 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific).

ASOs were removed from 4 µl of DMS-modified RNA in 10 µl containing 1 µl of TURBO DNase (Invitrogen) and 1X TURBO DNase Buffer (Invitrogen), incubated at 37°C for 30 min. To stop the reaction, 2 µl of DNase Inactivation Reagent was added and incubated at room temperature for 10 min, mixing several times throughout by flicking. DNase Inactivation Reagent was precipitated by spinning on a benchtop PCR tube centrifuge for 10 min and transferring 4 µl of supernatant to a new tube.

Library generation of 2,924 nt SARS-CoV-2 RNA

4 µl RNA was reverse transcribed in 20 µl containing 1X First Strand Buffer (Invitrogen), 500 µM dNTPs (Promega), 5 mM dithiothreitol (Invitrogen), 500 nM FSE primer CTTCGTCCTTTTCTTGGAAGCGACA (Integrated DNA Technologies), 500 nM section-specific reverse primer (Integrated DNA Technologies, Supplementary Table 2), 1 µl of RNaseOUT (Invitrogen), and 1 µl of TGIRT-III enzyme (InGex) at 57°C for 90 min, followed by inactivation at 85°C for 15 min. To degrade the RNA, 1 µl of Hybridase Thermostable RNase H (Lucigen) was added to each tube and incubated at 37°C for 20 min. 1 µl of unpurified RT product was amplified in 12.5 µl using the Advantage HF 2 PCR Kit (Takara Bio) with 1X Advantage 2 PCR Buffer, 1X Advantage-HF 2 dNTP Mix, 1X Advantage-HF 2 Polymerase Mix, 250 nM primers (Integrated DNA Technologies) for either the FSE (CCCTGTGGGTTTTACACTTAAAAAC and CTTCGTCCTTTTCTTGGAAGCGACA) or specific section (Supplementary Table 2); 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; and final extension at 68°C for 60 s. 5 µl of every amplicon from the same RT product was pooled and then purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 20 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific).

200 ng of pooled PCR product was 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 size se-

lection after adapter ligation, 27.5 µl and 12.5 µl of NEBNext Sample Purification Beads (New England Biolabs) were used in the first and second steps, respectively, to select inserts of 280-300 bp. Indexing PCR was run at half volume (25 µl) for 3 cycles. In lieu of the final bead cleanup, 420 bp inserts were selected using a 2% E-Gel SizeSelect II Agarose Gel (Invitrogen) according to the manufacturer's protocol. 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 2 x 150 bp paired-end reads according to the manufacturer's protocol.

Data analysis of 2,924 nt SARS-CoV-2 RNA

Sequencing data were processed with SEISMIC-RNA v0.12 and v0.13 to compute mutation rates, clusters, correlations, and secondary structures. Effects of each ASO group (Figure 2b, Supplementary Figures 1 and 2) were computed with the script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-2924/run-tile.sh`. Clustering and structure modeling (Figure 2c and d, Supplementary Figure 3a and b) were performed with the script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-2924/run-deep.sh`. Because some samples contained amplicons that overlapped each other, sequence alignment map (SAM) files were filtered by amplicon using the script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-2924/filter-deep.py`. The fraction of structures containing long-range stems (Supplementary Figure 3c) was determined using the script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-2924/fraction_folded.py`.

SEARCH-MaP of long-range base pairs in multiple coronaviruses

Computational screen for long-range base pairs in coronaviruses

All coronaviruses with reference genomes in the NCBI Reference Sequence Database [51] as of December 2021 were searched for using the following query:

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

The reference sequences (https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/cov_refseq.fasta) and table of features (https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/cov_features.txt) were downloaded and used to locate the slippery site in each genome using a custom Python script (https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/extract_long_fse.py). For each genome, up to 100 secondary structure models of the 2,000 nt segment from 100 nt upstream to 1,893 nt downstream of the slippery site (excluding genomes with ambiguous nucleotides in this segment) were generated using Fold v6.3 from RNAstructure [52] via the script https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/fold_long_fse.py. The fraction of models in which each base paired with any other base between positions 101 and 250 was calculated using the script https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/analyze_interactions.py. Using these fractions, coronaviruses were clustered via the unweighted pair group method with arithmetic mean (UPGMA) and a euclidean distance metric, implemented in Seaborn v0.11 [84] and SciPy v1.7 [80] (Supplementary Figure 5). From each cluster with prominent

potential long-range interactions involving the FSE, coronaviruses were manually selected for experimental study.

Synthesis of 239 and 1,799 nt coronaviral RNAs

For each selected coronavirus, the 1,799 nt segment from 290 to 1,502 nt downstream of the slippery site was ordered from Twist Bioscience as a gene fragment flanked by the standard 5' and 3' adapters CAATCCGCCCTCACTACAACCG and CTACTCTGGCGTCGATGAGGGA, respectively. Gene fragments were resuspended to 10 ng/µl in 10 mM Tris-HCl pH 8 (Invitrogen). Each DNA template for transcription of 1,799 nt RNA segments, including a T7 promoter, was amplified from 0.5 µl (5 ng) of a gene fragment in 20 µl using 2X CloneAmp HiFi PCR Premix (Takara Bio) with 250 nM of each primer TAATACGACTCACTATAGGCAATCCGCCCTCACTACAACCG and TCCCTCATCGACGCCAGAGTAG; initial denaturation at 98°C for 30 s; 30 cycles of 98°C for 10 s, X°C (see Supplementary Table 4) for 10 s, and 72°C for 15 s; and final extension at 72°C for 60 s. DNA templates for transcription of 239 nt RNA segments were amplified using the same procedure but with the forward primers with T7 promoters (F+T7) and reverse primers (R) in Supplementary Table 5. For experiments in which the RNAs were transcribed as a pool of all coronaviruses, all PCR products of the same length (i.e. 239 or 1,799 nt) were pooled, then purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol; concentrations were measured with a NanoDrop (Thermo Fisher Scientific). Otherwise, PCR products were purified individually.

50 ng of DNA template was transcribed using a HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs) according to the manufacturer's protocol but at one-quarter volume (5 µl), supplemented with 0.25 µl RNaseOUT (Invitrogen), for 16 hr. DNA template was degraded by incubating with 0.5 µl of TURBO DNase (Invitrogen) at 37°C for 30 min. RNA was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol,

eluted in 50 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific).

DMS treatment of 239 and 1,799 nt coronaviral RNAs

Antisense oligonucleotides (ASOs) in Supplementary Table 6 were ordered from Integrated DNA Technologies and resuspended to 100 µM in low-EDTA TE buffer: 10 mM Tris pH 7.4 with 0.1 mM EDTA (Integrated DNA Technologies). For each coronavirus, 5 µl of each corresponding ASO (Supplementary Table 6) was pooled; the pool of ASOs was diluted with low-EDTA TE buffer to a final volume of 100 µl, bringing each ASO to 5 µM. 1X refolding buffer comprising 300 mM sodium cacodylate pH 7.2 (Electron Microscopy Sciences) and 6 mM magnesium chloride (Invitrogen) was assembled, then pre-warmed to 37°C.

For already-pooled RNA, 300 ng was diluted in 2.5 µl of nuclease-free water (Fisher Bioreagents) in a PCR tube, heated to 95°C for 1 min to denature, chilled on ice for 3 min, added to 95 µl of pre-warmed refolding buffer, and incubated at 37°C for 20 min to refold. For individually transcribed RNA, 1 pmol was mixed with 10 µl of either low-EDTA TE buffer (for probing without ASOs) or the ASO pool for the corresponding coronavirus (for probing with ASOs) in a PCR tube, heated to 95°C for 1 min to denature the RNA, chilled on ice for 3 min, added to pre-warmed refolding buffer for a total volume of 100 µl, and incubated at 37°C for 20 min to refold the RNA (possibly with ASOs). Subsequently, equimolar amounts of all refolded RNAs were combined into one 97 µl pool in a 1.5 ml tube.

RNA was treated with DMS (Sigma-Aldrich) – 2.5 µl (260 mM) for RNAs transcribed as pools or 3 µl (320 mM) for RNAs pooled after transcription – in 100 µl while shaking at 800 rpm in a ThermoMixer C (Eppendorf) at 37°C for 5 min. To quench, 60 µl of beta-mercaptoethanol (Sigma-Aldrich) was added and mixed thoroughly. DMS-modified RNA was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 16 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific). If added, ASOs were then degraded in 50 µl

containing 1X TURBO DNase Buffer (Invitrogen) and 1 µl of TURBO DNase Enzyme (Invitrogen) at 37°C for 30 min; RNA was purified with an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 16 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop (Thermo Fisher Scientific).

Sequencing library generation of 239 and 1,799 nt coronaviral RNAs

100 ng of DMS-modified RNA was prepared for sequencing using the xGen Broad-Range RNA Library Preparation Kit (Integrated DNA Technologies) according to the manufacturer's protocol, with the following modifications. During fragmentation, 8 µl of RNA was combined with 1 µl of Reagent F1, 4 µl of Reagent F3, and 2 µl of Reagent F2. For reverse transcription, 1 µl of Enzyme R1, 2 µl of TGIRT-III enzyme (InGex), and 1 µl of 100 mM dithiothreitol (Invitrogen) was used instead of the reaction mix, then incubated at room temperature for 30 minutes before adding 2 µl of Reagent F2. Reverse transcription was stopped by adding 1 µl of 4 M sodium hydroxide (Fluka), heating to 95°C for 3 min, chilling at 4°C, then neutralizing with 1µl of 4 M hydrochloric acid. Instead of a bead cleanup after the final PCR, unpurified PCR products with 6X DNA loading dye (Invitrogen) were elecrophoresed through an 8% polyacrylamide Tris-borate-EDTA (TBE) gel (Invitrogen) at 180 V for 55 min. The gel was stained with SYBR Gold (Invitrogen); the section between 250 and 500 bp was excised and placed in a 0.5 ml tube with a hole punctured in the bottom by an 18-gauge needle (BD Biosciences), which was nested inside a 1.5 ml tube and centrifuged at 21,300 x g for 1 min to crush the gel slice into the latter. Crushed gel pieces were suspended in 500 µl of 300 mM sodium chloride (Boston Bioproducts), shaken at 1,500 rpm in a ThermoMixer C (Eppendorf) at 70°C for 20 min, and centrifuged at 21,300 x g through a 0.22 µm Costar Spin-X filter column to remove the gel pieces. Filtrate was mixed with 600 µl isopropanol (Sigma-Aldrich) and 3 µl GlycoBlue Coprecipitant (Invitrogen), vortexed briefly, and stored at -20°C overnight. DNA was then pelleted by centrifugation at 4°C at 18,200 x g

for 45 min. The supernatant was aspirated, and the pellet was washed with 1 ml of ice-cold 70% ethanol (Sigma-Aldritch), resuspended in 15 µl nuclease-free water (Fisher Bioreagents), and quantified using the 1X dsDNA High Sensitivity Assay Kit for the 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 2 x 150 bp paired-end reads according to the manufacturer's protocol.

Data analysis of 239 and 1,799 nt coronaviral RNAs

Sequencing data were processed with SEISMIC-RNA v0.11 and v0.12 to compute mutation rates, correlations between samples, and secondary structure models using the commands in the shell script `https://github.com/rouskinlab/search-map/tree/main/Compute/covs-1799/run.sh`. For the 239 and 1,799 nt RNAs that had been pooled during transcription, the two replicates for each coronavirus for each length were confirmed to give similar results, then merged before comparing the 239 and 1,799 nt RNAs to each other. For the comparison of RNAs with and without ASOs, the no-ASO samples that had been transcribed individually were confirmed to give similar results to those transcribed as a pool; then, all no-ASO samples were pooled before comparing to samples with ASOs. For each coronavirus, the DMS reactivities of the combined no-ASO samples were used to model up to 20 secondary structures of the 1,799 nt segment using Fold from RNAstructure v6.3 [52]. Structure models were checked manually for correspondence with the rolling correlation between the +ASO and no-ASO conditions; the minimum free energy structure was chosen for every coronavirus except for transmissible gastroenteritis virus, in which the first sub-optimal structure – but not the minimum free energy structure – contained long-range base pairs supported by the rolling correlation. Rolling correlations between +ASO and no-ASO conditions superimposed on secondary structure models (Figure 4) were graphed using the Python script `https://github.com/rouskinlab/search-map/tree/main/Compute/util/pairs_vs_correl.py`.

SEARCH-MaP of 1,799 nt SARS-CoV-2 RNA

RNA synthesis of 1,799 nt SARS-CoV-2 RNA

A DNA template for transcription, including a T7 promoter, was amplified from the 1,799 bp gene fragment of SARS-CoV-2 as described above but with primers TAATACGACTCACTATAGGTACTGGTCAGGCAATAACAGTTACAC and GACCCCATTTATTAAATGGAAAACCAGCTG, an annealing temperature of 65°C, and an extension time of 10 s; eluted in 18 µl of 10 mM Tris-HCl pH 8 (Invitrogen); and measured with a NanoDrop One (Thermo Fisher Scientific). 100 ng of DNA template was transcribed using a HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs) according to the manufacturer's protocol for 11 hr. DNA template was degraded by incubating with 1 µl of TURBO DNase (Invitrogen) at 37°C for 30 min. RNA was purified using an RNA Clean & Concentrator-25 kit (Zymo Research) according to the manufacturer's protocol, eluted in 50 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop One (Thermo Fisher Scientific).

DMS treatment of 1,799 nt SARS-CoV-2 RNA

1.15X refolding buffer comprising 345 mM sodium cacodylate pH 7.2 (Electron Microscopy Sciences) and 7 mM magnesium chloride (Invitrogen) was assembled and pre-warmed to 37°C. 1 pmol of RNA was mixed with 100 pmol of each ASO (Integrated DNA Technologies, Supplementary Table 3) in 10 µl total, heated to 95°C for 60 s to denature, chilled on ice for 5-10 min, and added to 87.1 µl of pre-warmed refolding buffer. If no ASO would be added during refolding, then 1 µl of nuclease-free water (Fisher Bioreagents) was added. RNA was incubated at 37°C for 15-20 min to refold. If an ASO would be added during refolding, then 100 pmol (1 µl) of ASO was added. RNA was incubated for another 15 min to allow any newly added ASOs to bind.

RNA was probed in 100 µl containing 1.9 µl (300 mM) DMS (Sigma-Aldrich) while shaking at 500 rpm in a ThermoMixer C (Eppendorf) at 37°C for 5 min. To

quench, 20 µl of beta-mercaptoethanol (Sigma-Aldrich) was added and mixed thoroughly. DMS-modified RNA was purified using an RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 15 µl of nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop One (Thermo Fisher Scientific).

Library generation 1,799 nt SARS-CoV-2 RNA

1 µl of DMS-modified RNA was reverse transcribed in 20 µl using Induro Reverse Transcriptase (New England Biolabs) according to the manufacturer's protocol with 500 nM of primer CTTCGTCCTTTTCTTGGAAGCGACA (Integrated DNA Technologies) at 57°C for 30 min, followed by inactivation at 95°C for 1 min. 1 µl of unpurified RT product was amplified in 20 µl using Q5 High-Fidelity 2X Master Mix (New England Biolabs) with 500 nM of each primer CCCTGTGGGTTTTACACTTAAAAAC and CTTCGTCCTTTTCTTGGAAGCGACA (Integrated DNA Technologies); initial denaturation at 98°C for 30 s; 30 cycles of 98°C for 10 s, 65°C for 20 s, and 72°C for 20 s; and final extension at 72°C for 120 s. The PCR product was purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 20 µl of 10 mM Tris-HCl pH 8 (Invitrogen), and measured with a NanoDrop One (Thermo Fisher Scientific).

50-100 ng of purified PCR product was 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. All steps were performed at half of the volume specified in the protocol, including reactions, bead cleanups, and washes. During size selection after adapter ligation, 14 µl and 7 µl of SPRIselect Beads (Beckman Coulter) were used in the first and second steps, respectively, to select inserts of 283 bp. Indexing PCR was run with 400 nM of each primer for 4 cycles. After indexing, PCR products were pooled in pairs; in lieu of the final bead cleanup, 405 bp products were selected using a 2% E-Gel SizeSelect II Agarose Gel (Invitrogen) according to the manufacturer's protocol. DNA concentrations were measured using a Qubit 4 Fluorometer (Thermo Fisher Scientific)

according to the manufacturer's protocol. Samples were pooled and sequenced using a NextSeq 1000 Sequencing System (Illumina) with 2 x 150 bp paired-end reads according to the manufacturer's protocol.

Data analysis of 1,799 nt SARS-CoV-2 RNA

Sequencing data were processed with SEISMIC-RNA v0.11 and v0.12 to compute mutation rates, clusters, and correlations between samples using the commands in the shell script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/run.sh`. Heatmaps of the reproducibility of clustering between replicates (Supplementary Figure 4) were generated using the Python script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/compare-clusters.py`. After the two replicates were confirmed to give similar clusters, they were pooled for subsequent analyses. Secondary structures with rolling correlations (Figure 3b) were drawn using the Python script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/draw-structure.py`. Alternative structure models (Figure 3c) were selected and created with the help of the Python scripts `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/choose-model-parts.py` and `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/make-models.py`. Heatmaps of areas under the curve (Figure 3d) were generated using the Python script `https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/atlas-plot.py`.

DMS-MaPseq of transmissible gastroenteritis virus in ST cells

Cells and Viruses

Transmissible gastroenteritis virus (TGEV, TC-adapted Miller strain, ATCC VR-1740) and ST cells (ATCC CRL-1746) were ordered from American Type

Culture Collection (ATCC). ST cells were maintained in Eagle's Minimum Essential Medium (EMEM, Gibco) supplemented with 10% fetal bovine serum (Gibco), 1% sodium pyruvate (Gibco), and 1% Pen Strep (Gibco) at 37°C with 5% carbon dioxide. For TGEV, the infection medium (IM) comprised EMEM (Gibgo) supplemented with 10% fetal bovine serum (Gibco), 1% sodium pyruvate, and 1 μg/μl of TPCK trypsin (Thermo Fisher Scientific).

Production and titering of TGEV

A 150 mm dish was seeded with 1 x 10^7 ST cells, grown overnight, and washed twice with phosphate-buffered saline (PBS, Gibco). Cells were inoculated with 8 ml of TGEV in IM at a multiplicity of infection (MOI) of 0.1, which was kept on for 60 min with rocking every 15 min. The inoculum was removed, cells were washed twice with PBS, and 26 ml of IM was added. Cells were checked daily for cytopathic effects (CPE) and were harvested after 5 days upon development of significant (80%) CPE.

Harvested TGEV was titered via tissue culture infectious dose ($TCID_{50}$). Briefly, ST cells were seeded in a poly-L-lysine coated 96-well plate at 4 x 10^4 cells per well and grown overnight. TGEV was thawed on ice and serially diluted in a 12-well plate from 10_{-1} to 10_{-10} in IM. Cells were washed once with PBS, and each well was inoculated with one serial dilution of TGEV (8 replicates per dilution level). The plate was wrapped in parafilm and incubated until CPE appeared. Then, media was aspirated and cells were fixed with 4% paraformaldehyde for 30 min and decanted. 0.5% crystal violet was then added to each well; the plate was rocked for 10 min, submerged in water to remove excess crystal violet, and dried. Wells with CPE were counted and the titer determined using the Spearman-Kärber method.

TGEV infection and DMS treatment

Four 150 mm dishes were each seeded with 1 x 10^7 ST cells, grown overnight, and washed twice with phosphate-buffered saline. Cells were inoculated with 8 ml of TGEV in IM at a multiplicity of infection (MOI) of 2, which was kept on for 60 min

with rocking every 15 min. The inoculum was removed, cells were washed twice with PBS, and 26 ml of IM was added.

After 48 hr, media was aspirated. 250 µl of DMS was mixed with 10 ml of IM and immediately added to two plates; the other two received 10 ml IM without DMS. Plates were incubated at 37°C for 5 min. The media was aspirated and replaced with stop solution (30% beta-mercaptoethanol in 1X PBS). Cells were scraped off using a cell scraper, spun down at 3000 x g for 3 min. The pellet was washed with stop solution, spun down again, washed with 10 ml PBS, dissolved in 3 ml of TRIzol, and split into 1 ml techincal replicates.

RNA purification

200 µl of chloroform was added to each 1 ml techincal replicate, vortexed for 20 s, and rested until the phases separated. Samples were then spun at 18,200 x g for 15 min at 4°C; the aqueous phase transferred to a new tube and mixed with an equal volume of 100% ethanol. RNA was purified using a 50 µg Monarch RNA Cleanup Column (New England Biolabs), eluted in nuclease-free water, and quantified with a NanoDrop.

To remove rRNA, 10 µg of total RNA was diluted in 6 µl of nuclease-free water and mixed with 1 µl of anti-rRNA ASOs (Integrated DNA Technologies) and 3 µl of HYBE buffer (200 mM sodium chloride, 100 mM Tris-HCl pH 7.5). The mixture was incubated at 95°C for 2 min and cooled by 0.1°C/s until reaching 45°C. A preheated mixture of 10 µl of RNase H and 2 µl of RNase H Buffer was added and incubated at 45°C for 30 min. RNA was purified using a 10 µg Monarch RNA Cleanup Column (New England Biolabs) and eluted in 42 µl of nuclease-free water.

To remove DNA (including anti-rRNA ASOs), 5 µl of 10X Turbo DNase Buffer (Thermo Fisher) and 3 µl of TURBO RNase (Thermo Fisher) were added and incubated at 37°C for 20 min. RNA was purified using a 10 µg Monarch RNA Cleanup Column (New England Biolabs) and eluted in 10 µl of nuclease-free water.

Library generation for the full TGEV genome

RNA was prepared for sequencing using the xGen Broad-Range RNA Library Preparation Kit (Integrated DNA Technologies) according to the manufacturer's protocol, with the same modifications as described above (5.3.4), notably the substitution of TGIRT-III (InGex) for the kit reverse transcriptase. Samples were pooled and sequenced using a NextSeq 1000 Sequencing System (Illumina) with 2 x 150 bp paired-end reads according to the manufacturer's protocol.

Library generation for amplicons

1 µl of rRNA-depleted, DNased RNA was reverse transcribed in 20 µl using Induro Reverse Transcriptase (New England Biolabs) according to the manufacturer's protocol with 500 nM of primer ACAATTCGTCTTAAGGAATTTACCAATACACGCAA (Integrated DNA Technologies) at 57°C for 30 min, followed by inactivation at 95°C for 1 min. 1 µl of unpurified RT product was amplified in 10 µl using Q5 High-Fidelity 2X Master Mix (New England Biolabs) with 1 µM of each primer, either GCCGCTACAAAGGTAAGTTCGTGCAAATACCAACT and ACAATTCGTCTTAAGGAATTTACCAATACACGCAA or GTGAAAAGTGACATCTATGGTTCTGATTATAAGCAGTA and CTATACCAAGTTGTTTGAAATGGTAACCTGCAGTAACA (Integrated DNA Technologies); initial denaturation at 98°C for 30 s; 30 cycles of 98°C for 5 s, 69°C for 20 s, and 72°C for 15 s; and final extension at 72°C for 120 s. Amplification was confirmed by electrophoresing 1 µl of each PCR product. PCR products for both pairs of primers were pooled and then purified using a DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 18 µl of 10 mM Tris-HCl pH 8 (Invitrogen), and measured with a NanoDrop (Thermo Fisher Scientific).

175-225 ng of purified PCR product was 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. All steps were

performed at half of the volume specified in the protocol, including reactions, bead cleanups, and washes. During size selection after adapter ligation, 14 µl and 7 µl of SPRIselect Beads (Beckman Coulter) were used in the first and second steps, respectively, to select inserts of 295 bp. Indexing PCR was run with 400 nM of each primer for 4 cycles. In lieu of the final bead cleanup, 415 bp products were selected using a 2% E-Gel SizeSelect II Agarose Gel (Invitrogen) according to the manufacturer's protocol. DNA concentrations were measured using a Qubit 4 Fluorometer (Thermo Fisher Scientific) according to the manufacturer's protocol. Samples were pooled and sequenced using a NextSeq 1000 Sequencing System (Illumina) with 2 x 150 bp paired-end reads according to the manufacturer's protocol.

Data analysis of transmissible gastroenteritis virus in ST cells

The genomic sequence of this TGEV strain was determined using the script `https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/consensus.sh`: reads from the untreated sample were aligned to the TGEV reference genome (`NC_038861.1`) using Bowtie 2 [83] and the consensus sequence was determined using Samtools [81]. All reads were processed with SEISMIC-RNA v0.15 to compute mutation rates, correlations between samples, and secondary structure models using the commands in the shell script `https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/run.sh`. Positions in the untreated sample with mutation rates greater than 0.01 were masked. Replicates were checked for reproducibility and pooled for clustering and structure modeling. A model of short-range base pairs (maximum distance 300 nt) in the TGEV genome was generated from the DMS reactivities using Fold-smp from RNAstructure [52] in five overlapping 10 kb segments, which were merged using the script `https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/assemble-tgev-ss.py`. Rolling area under the curve superimposed on secondary structure models in Figure 5d was graphed using the script `https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/make-figure-6d.py`, and in Supplementary Figure 9 using

the script https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/plot_genome.py.

Acknowledgements

We thank Miriam L. Rittenberg for assistance with experiments. This research was supported by the National Institute of Allergy and Infectious Diseases grant DP2 AI175475 to S.R. and a National Science Foundation Graduate Research Fellowship to M.F.A.

Author Contributions

S.R. and M.F.A. conceived the project. M.F.A. performed the experiments with SARS-CoV-2 segments. J.A. performed the experiments with other coronavirus segments. J.P. performed the experiments with TGEV-infected ST cells. M.F.A. wrote SEISMIC-RNA with contributions from S.G., Y.M., A.L., and J.A. M.F.A. and J.A. analyzed the data. M.F.A. drafted the manuscript. All authors reviewed the manuscript and provided comments.

Ethics Declarations

The authors declare no competing interests.

Data Availability

All sequencing data generated in this study have been deposited into the NCBI Short Read Archive under accession code PRJNA1103196.

Code Availability

Documentation for SEISMIC-RNA, including instructions for installation, is hosted on GitHub Pages: `https://rouskinlab.github.io/seismic-rna`. Source code for SEISMIC-RNA is available from GitHub: `https://github.com/rouskinlab/`

`seismic-rna`. Shell scripts for running SEISMIC-RNA, auxiliary scripts for data analysis, supplementary files, and LaTeX source code for this manuscript are also available from GitHub: `https://github.com/rouskinlab/search-map`.

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 non-coding 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. Chapter four - improving RNA secondary structure prediction with structure mapping data. In Shi-Jie Chen and Donald H. Burke-Aguero, editors, *Computational Methods for Understanding Riboswitches*, volume 553 of *Methods in Enzymology*, pages 91–114. Academic Press, 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 by mutate-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. *Frontiers in Bioscience*, 13:4873–4881, 2008.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Jessica S. Reuter and David H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129, 2010.

[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 Alphacoronavirus 1 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 musD 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] Matthew F. Allan, Amir Brivanlou, and Silvi Rouskin. RNA levers and switches controlling viral gene expression. *Trends in Biochemical Sciences*, 48, 2023.

[74] 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.

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

[76] 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.

[77] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 9 2020.

[78] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: a LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, New York, NY, USA, 2015. Association for Computing Machinery.

[79] Wes McKinney. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.

[80] 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.

[81] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009.

[82] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[83] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 2012.

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

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

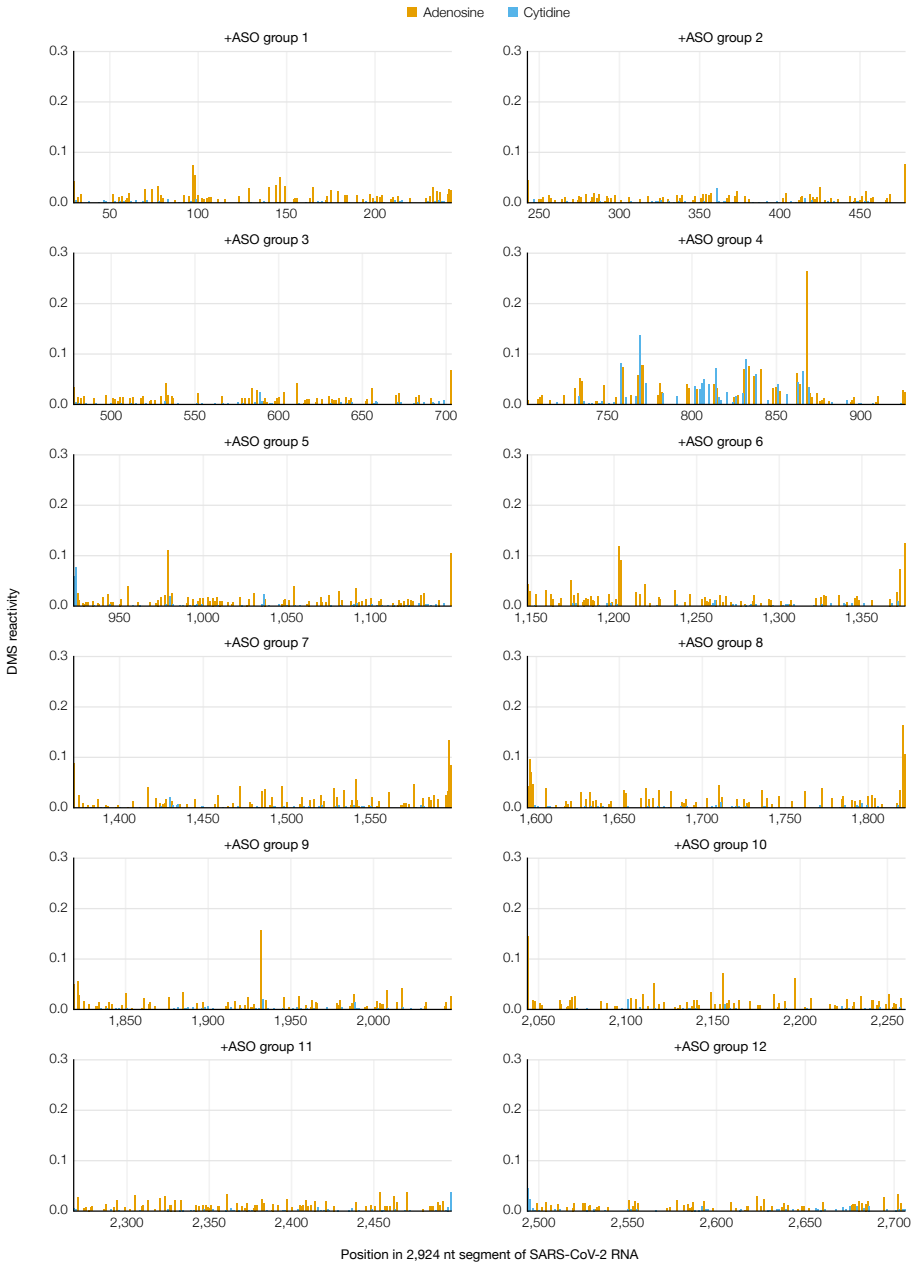
[86] Sara Alonso, Ander Izeta, Isabel Sola, and Luis Enjuanes. Transcription regulatory sequences and mRNA expression levels in the coronavirus transmissible gastroenteritis virus. *Journal of Virology*, 76(3):1293–1308, 2002.

[87] 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.

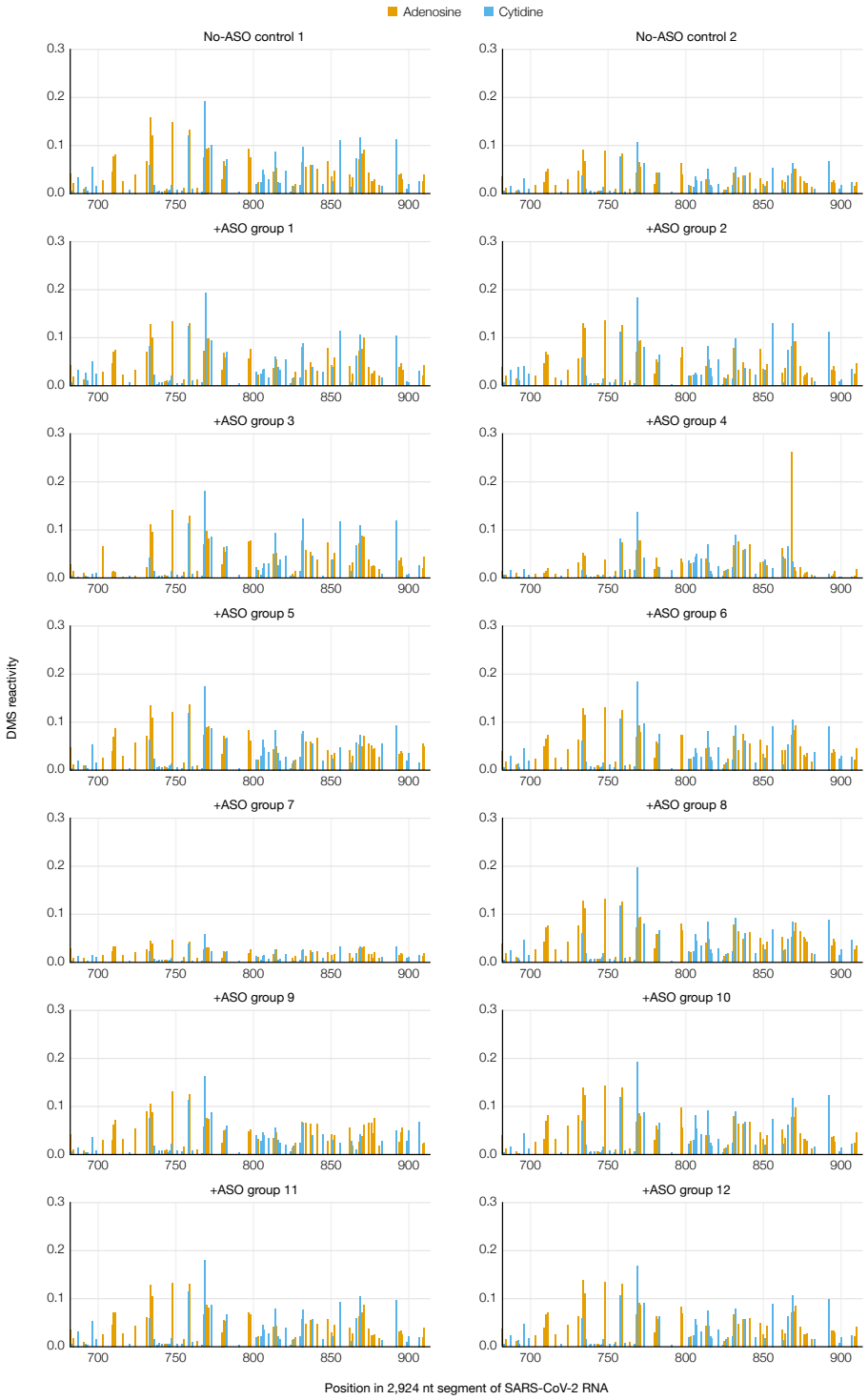
[88] D.A. Knoll and D.E. Keyes. Jacobian-free Newton-Krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, 193(2):357–397, 2004.

Supplementary Information

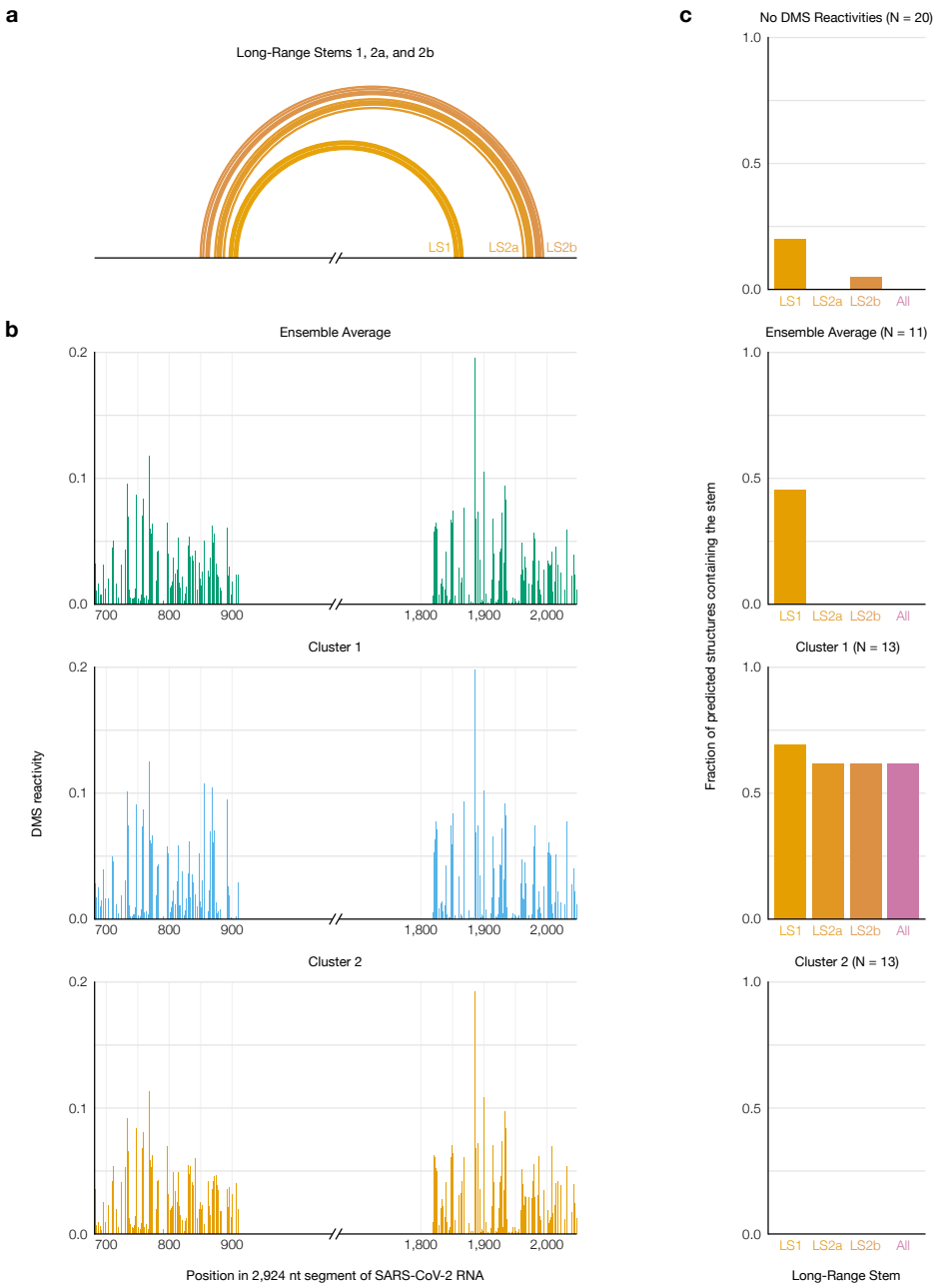
Supplementary Figures



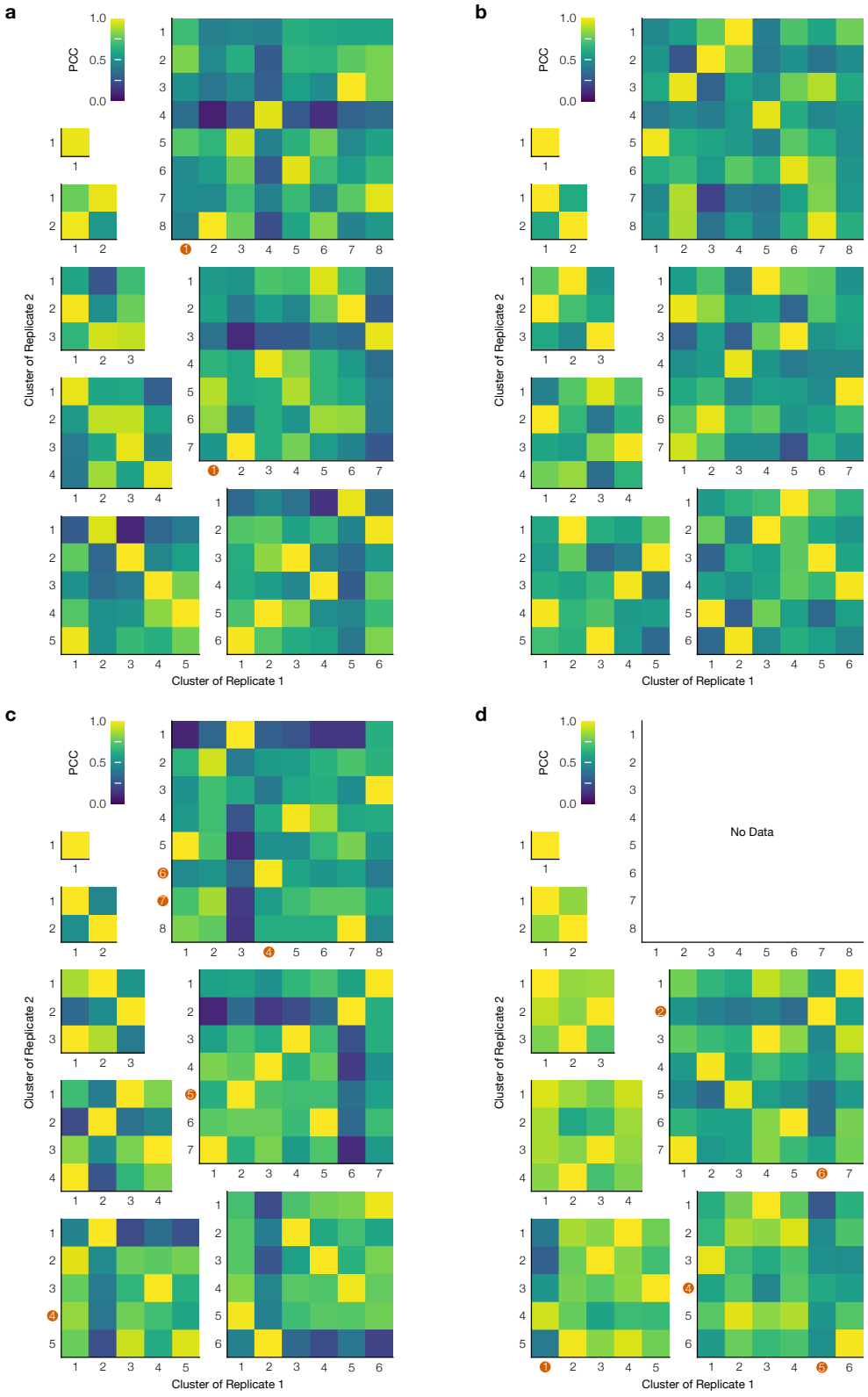
Supplementary Figure 1: **Mutational profile of each ASO target section upon adding the corresponding group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.** Positions are colored based on the RNA sequence.



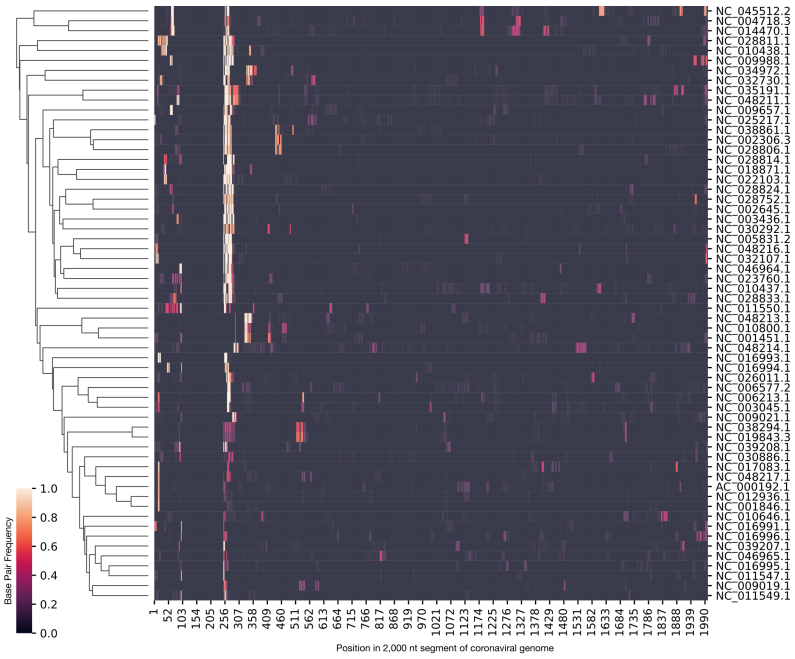
Supplementary Figure 2: **Mutational profiles of the FSE section upon adding each group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.** Positions are colored based on the RNA sequence.



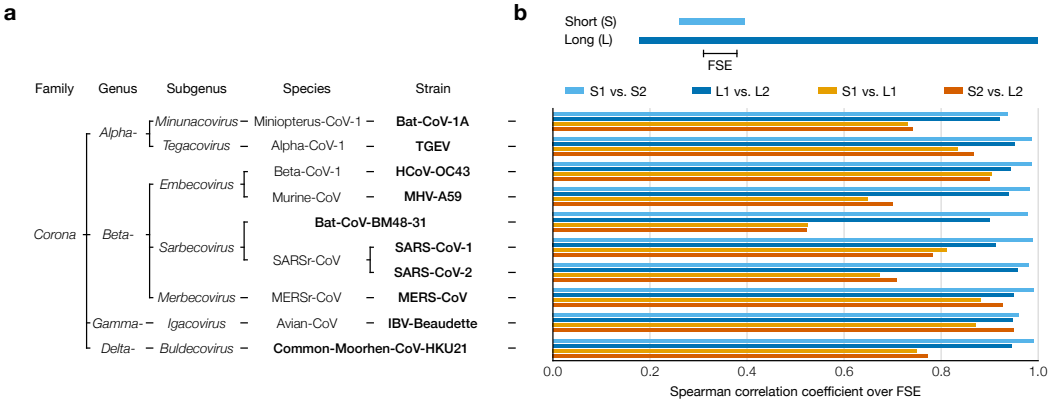
Supplementary Figure 3: **Improved prediction of long-range stems in SARS-CoV-2 using clustered DMS reactivities. (a)** Model of the two inner stems of the FSE-arch [49], denoted long stems (LS) 1 and 2a/b. **(b)** Mutational profiles of the ensemble average and of clusters 1 and 2 on both sides of the 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.



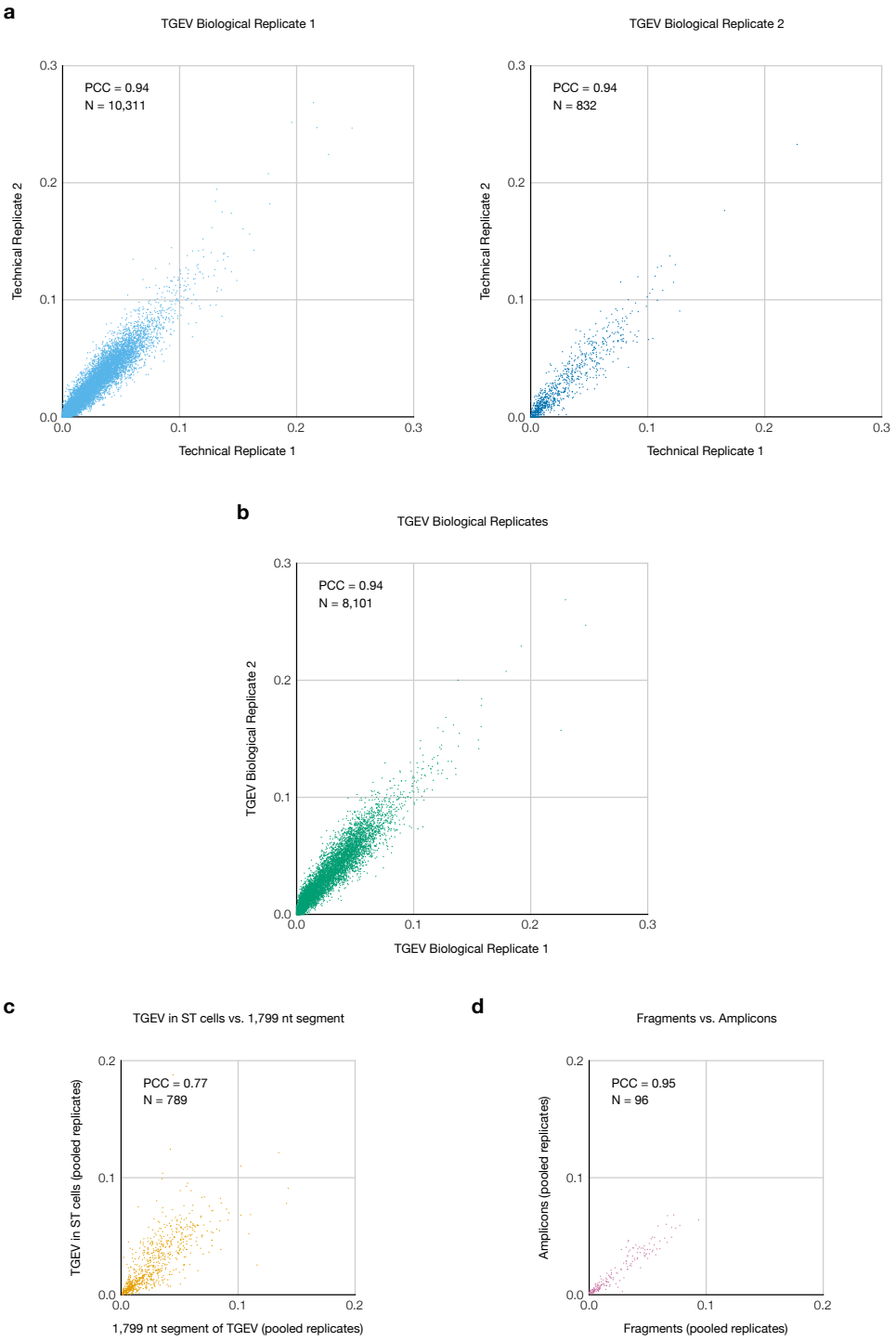
Supplementary Figure 4: **Reproducibility of clustering the SARS-CoV-2 FSE after adding ASOs. (a)** Heatmaps of the Pearson correlation coefficient (PCC) between each pair of clusters from two replicates of the 1,799 nt segment of SARS-CoV-2. Each heatmap corresponds to one order (i.e. number of clusters). 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.



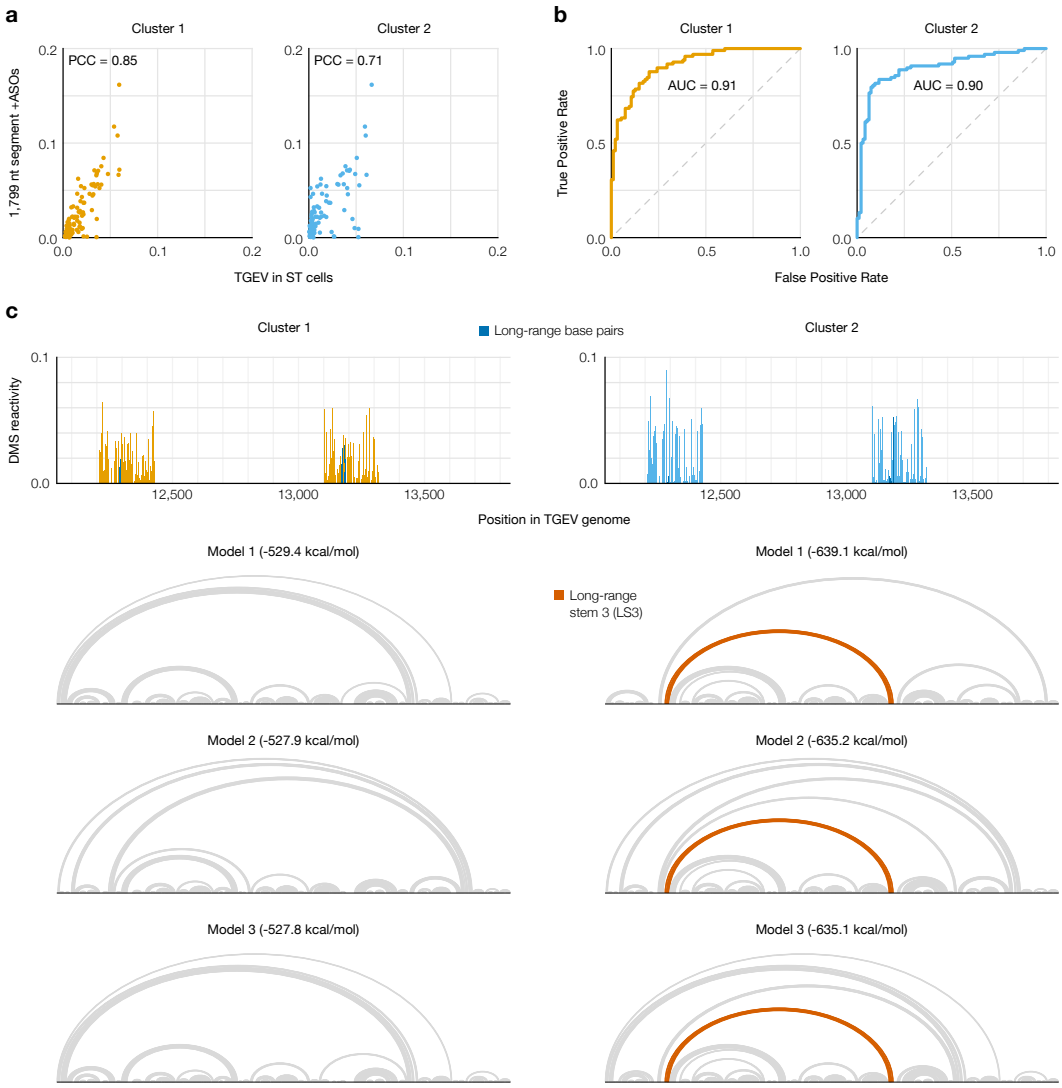
Supplementary Figure 5: **Computational screen of long-range base pairing near the FSE in 60 coronaviruses.** For each 2,000 nt segment of each coronaviral genome, the fraction of predicted structures in which each position outside the range 101-250 base-paired with any position in the range 101-250 is indicated. Genomes are clustered by their base-pairing frequencies. For each genome, the accession number for NCBI [51] is indicated.



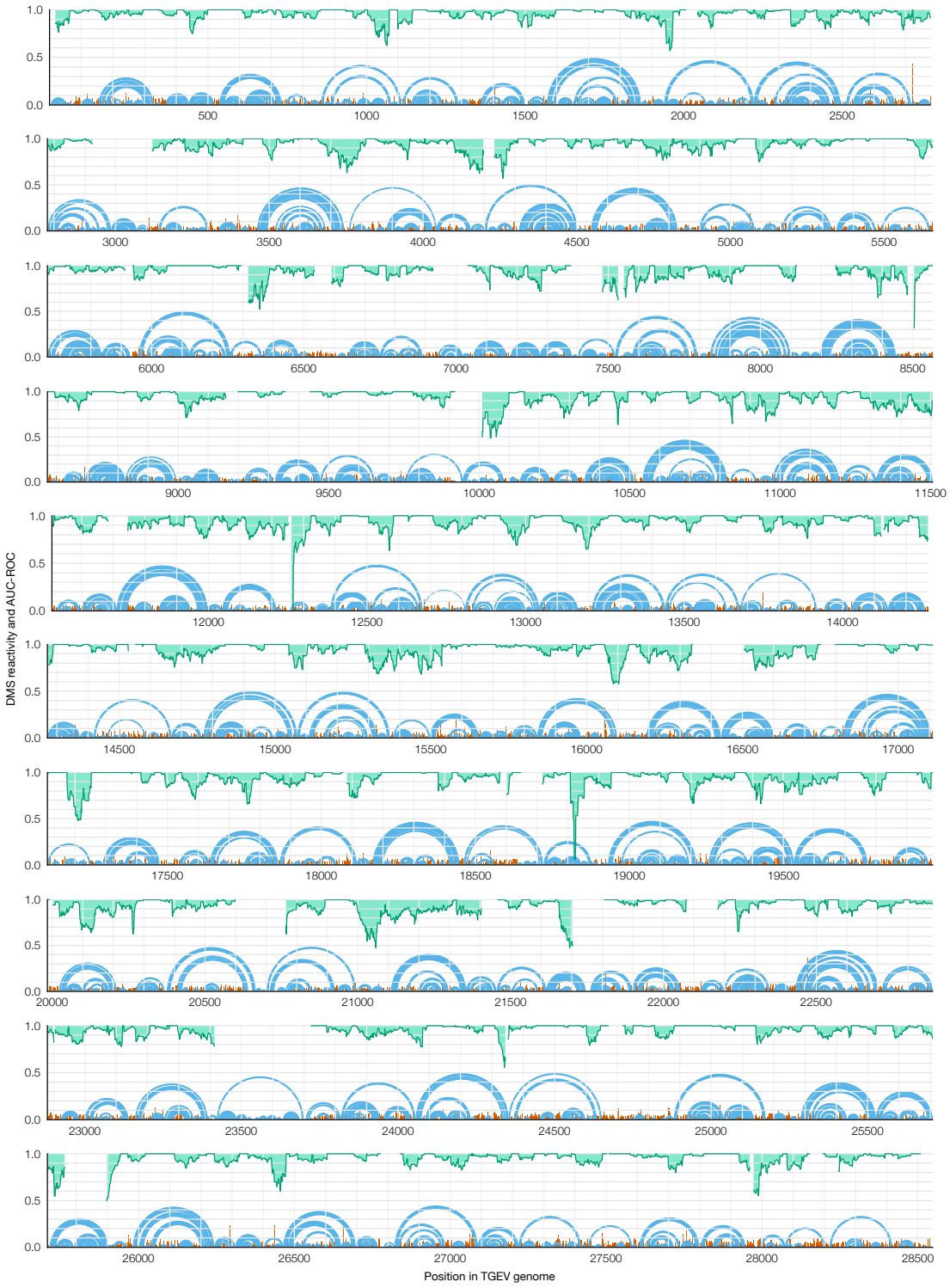
Supplementary Figure 6: **Experimental screen of long-range base pairing near the FSE in 10 coronaviruses. (a)** Taxonomy of the ten coronavirus species/strains in this screen; the lowest-level group for each virus is bolded. Bat-CoV-1A: bat coronavirus 1A (NC_010437.1), TGEV: transmissible gastroenteritis 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.



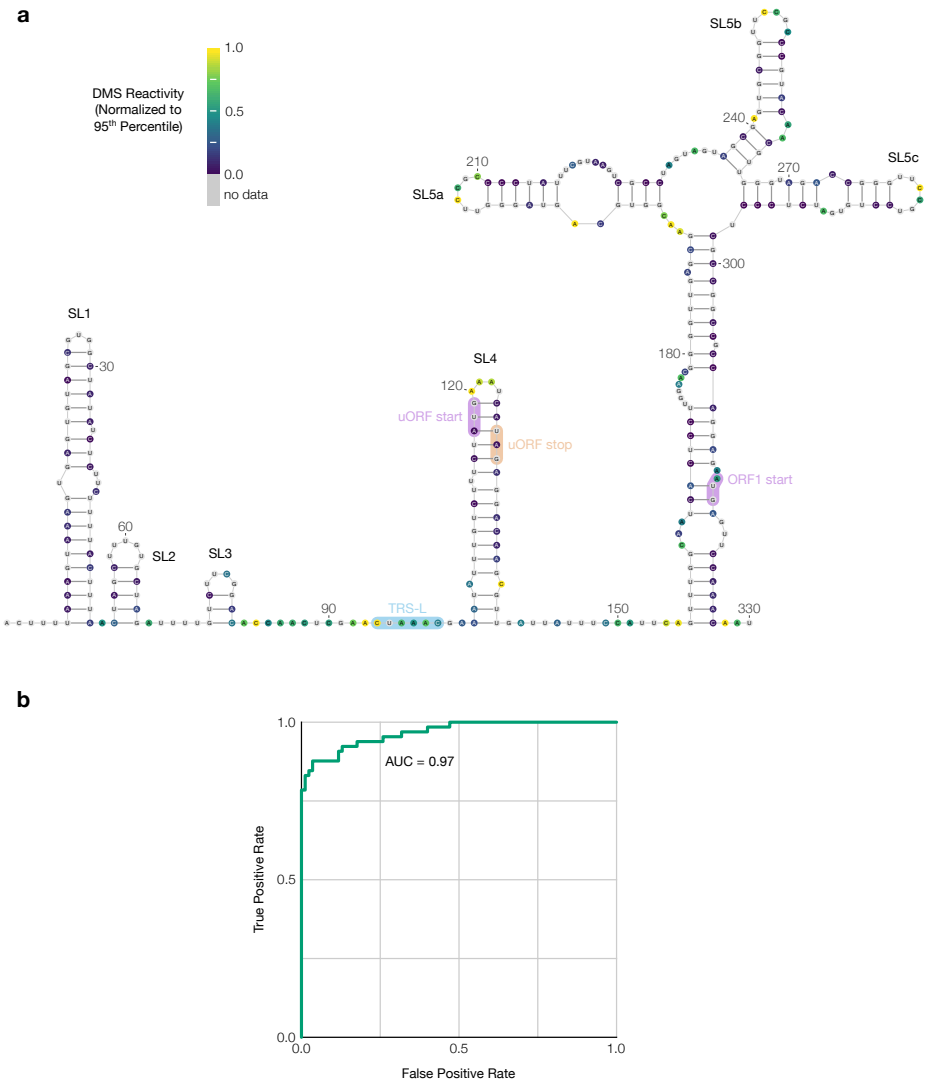
Supplementary Figure 7: **Replicates of TGEV in ST cells and comparison to the 1,799 nt segment. (a)** Comparison of DMS reactivities of the two technical replicates for each biological replicate of TGEV in ST cells. Each point represents one base in the sequence. The number of points (N) and Pearson correlation coefficient (PCC) are indicated for each plot. One point with DMS reactivity exceeding 0.3 in both technical replicates of biological replicate 1 is not shown. **(b)** Comparison of DMS reactivities of the two biological replicates (pooled technical replicates). **(c)** DMS reactivities of TGEV in ST cells using random fragmentation versus amplicons (pooled biological replicates). **(d)** DMS reactivities of TGEV in ST cells (pooled biological replicates) versus the 1,799 nt segment.



Supplementary Figure 8: **Alternative structures on both sides of the long-range base pairs in TGEV. (a)** Scatter plots of DMS reactivities over the 3' side of the predicted long-range stem in TGEV (Figure 4) comparing each cluster from amplicons in ST cells to the 1,799 nt segment with ASOs targeting the 5' side of the long-range stem, with Pearson correlation coefficient (PCC) indicated; each point is one base. **(b)** Receiver operating characteristic (ROC) curves comparing each cluster from amplicons in ST cells to the structure model of the 1,799 nt segment of TGEV including the long-range base pairs (Figure 4), with area under the curve (AUC) indicated. **(c)** DMS reactivities of clusters 1 and 2 and the three lowest-energy structure models of the 1,799 nt segment (positions 12,042-13,840) based on each cluster. Long-range stem 3 (LS3) is highlighted when it appears in a model. Structures were drawn with VARNA [85].



Supplementary Figure 9: **Short-range base pairs across the full TGEV genome.** Model of the secondary structure of the entire TGEV genome with a maximum distance of 300 nt between paired bases (blue). DMS reactivities used to generate the model are shown in red. Rolling (45 nt) area under the receiver operating characteristic curve (AUC-ROC), measuring how well the secondary structure model fits the DMS reactivities, is shown in green.



Supplementary Figure 10: **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) [86], upstream open reading frame (uORF) [87], and start codon of ORF1 are also labeled. The model was drawn using VARNA [85]. **(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 [88] implemented in SciPy [80].

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 10.1.4.

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 10.1.2. Likewise, the third part, $P(C_k|G_g) = p_k$, can also be calculated from the parameters, as explained in 10.1.3. 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	GGAAGCGACAACAATTAGTTTTTAGGAATTTAGCAAAACCAGCTACTTTATCATTG
5	1	TGTCTCTTAACTACAAAGTAAGAATCAATTAAATTGTCATCTTCGTCCTTTTCTT
	2	GACAATCCTTAAGTAAATTATAAATTGTTTCTTCATGTTGGTAGTTAGAGAAAGTG
	3	GGTACCATGTCACCGTCTATTCTAAACTTAAAGAAGTCATGTTTAGCAACAGCTG
	4	AAGCATAGACGAGGTCTGCCATTGTGTATTTAGTAAGACGTTGACGTGATATATGT
6	1	TGTATGTGACAAGTATTTCTTTTAATGTGTCACAATTACCTTCATCAAAATGCCTTA
	2	GGTTTTCTACAAAATCATACCAGTCCTTTTTATTGAAATAATCATCATCACAACAAT
	3	TTAACAAAGCTTGGCGTACACGTTCACCTAAGTTGGCGTATACGCGTAATATATCTG
	4	ATGTCAGTACACCAACAATACCAGCATTTCGCATGGCATCACAGAATTGTACTGTTT

Continued on next page

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

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

Continued on next page

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

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
	3	CATAATTCTAAGCATGTTAGGCATGGCTCTATCACATTTAG-GATAATCCCAAC
	4	ACGGTGTGACAAGCTACAACACGTTGTATGTTTGCGAGCA-AGAACAAGTGAGGC

Continued on next page

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

13	1	ACACATGACCATTTCACTCAATACTTGAGCACACTCATTAGCTAATCTATAGAA
	2	AGTTGTGGCATCTCCTGATGAGGTTCCACCTGGTTTAACATATAGTGAACCGCC
	3	ATTAACATTGGCCGTGACAGCTTGACAAATGTTAAAAACACTATTAGCATAAGC
	4	TAAATTGCGGACATACTTATCGGCAATTTTGTTACCATCAGTAGATAAAAGTGC

Table 2: **Sequences of the forward (F) and reverse (R) primers for amplifying the target 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 target 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 antisense oligonucleotides (ASOs) targeting the 1,799 nt segment of SARS-CoV-2 genomic RNA.** A plus sign (+) indicates that the following nucleotide is locked nucleic acid (LNA).

ASO	Sequence
Anti-LS1	GTAATTC+CTTAAAA+CTAAG
Anti-LS2a	TGAAA+AGCAA+CATTGTT
Anti-LS2b	TA+CCGGGTTTGACAG
Anti-LS3b	A+CCCTTAGACACAGCA
Anti-AS1	TGGGTTCGCG+GAGTTG
Anti-PS2-overlap	GT+TAAAATTA+CCG+GG

Table 4: **PCR primer annealing temperatures for coronavirus gene fragments.**

Coronavirus	**Annealing Temperature (°C)**
Bat Coronavirus 1A	55
Bat Coronavirus BM48-31	60
Common Moorhen Coronavirus	55
Human Coronavirus OC43	55
Infectious Bronchitis Virus	60
MERS Coronavirus	60
Murine Hepaitis Virus	60
SARS Coronavirus 1	60
SARS Coronavirus 2	55
Transmissible Gastroenteritis Virus	55

Table 5: **Sequences of the forward (F), forward with T7 promoter (F+T7), and reverse (R) primers for amplifying the 239 nt 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 5: **Sequences of the forward (F), forward with T7 promoter (F+T7), and reverse (R) primers for amplifying the 239 nt 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 5: **Sequences of the forward (F), forward with T7 promoter (F+T7), and reverse (R) primers for amplifying the 239 nt 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 6: **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 6: **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 6: **Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs.** (Continued)

2	GGTCTGGATCAGTACCATTGCA-GGGTTCTAGTCGAGCTGCACTA-GAACCCCGCACTCGTT