

¹ Discovery and Quantification of Long-Range
² RNA Base Pairs in Coronavirus Genomes with
³ SEARCH-MaP and SEISMIC-RNA

⁴ Matthew F. Allan¹²³, Justin Aruda¹⁴, Jesse Plung¹⁵,
Scott Grote¹, Yves J. Martin des Taillades⁶, Alberic de Lajarte¹,
Mark Bathe², and Silvi Rouskin^{1†}

⁵ ¹ Department of Microbiology, Harvard Medical School, Boston,
⁶ Massachusetts, USA 02115

⁷ ² Department of Biological Engineering, Massachusetts Institute of
⁸ Technology, Cambridge, Massachusetts, USA 02139

⁹ ³ Computational and Systems Biology, Massachusetts Institute of
¹⁰ Technology, Cambridge, Massachusetts, USA 02139

¹¹ ⁴ Harvard Program in Biological and Biomedical Sciences, Division
¹² of Medical Sciences, Harvard Medical School, Boston, MA, USA
02115

¹³ ⁵ Harvard Program in Virology, Division of Medical Sciences,
Harvard Medical School, Boston, MA, USA 02115

¹⁴ ⁶ Department of Biochemistry, Stanford University, Stanford,
California, USA 94305

¹⁵ [†] To whom correspondence should be addressed:
silvi@hms.harvard.edu

20 Abstract

21 RNA molecules perform a diversity of essential functions for which their linear se-
22 quences must fold into higher-order structures. Techniques including crystallogra-
23 phy and cryogenic electron microscopy have revealed 3D structures of ribosomal,
24 transfer, and other well-structured RNAs; while chemical probing with sequenc-
25 ing facilitates secondary structure modeling of arbitrary RNAs, even within cells.
26 Ongoing efforts continue increasing the accuracy, resolution, and ability to dis-
27 tinguish coexisting alternative structures. However, no method can discover and
28 quantify alternative structures with base pairs spanning arbitrarily long distances –
29 an obstacle for studying viral, messenger, and long noncoding RNAs, which may
30 form long-range base pairs.

31 Here, we introduce the method of Structure Ensemble Ablation by Reverse
32 Complement Hybridization with Mutational Profiling (SEARCH-MaP) and
33 software for Structure Ensemble Inference by Sequencing, Mutation
34 Identification, and Clustering of RNA (SEISMIC-RNA). We use SEARCH-MaP
35 and SEISMIC-RNA to discover that the frameshift stimulating element of SARS
36 coronavirus 2 base-pairs with another element 1 kilobase downstream in nearly
37 half of RNA molecules, and that this structure competes with a pseudoknot that
38 stimulates ribosomal frameshifting. Moreover, we identify long-range base pairs
39 involving the frameshift stimulating element in other coronaviruses including
40 SARS coronavirus 1 and transmissible gastroenteritis virus, and model the full
41 genomic secondary structure of the latter. These findings suggest that
42 long-range base pairs are common in coronaviruses and may regulate ribosomal
43 frameshifting, which is essential for viral RNA synthesis. We anticipate that
44 SEARCH-MaP will enable solving many RNA structure ensembles that have
45 eluded characterization, thereby enhancing our general understanding of RNA
46 structures and their functions. SEISMIC-RNA, software for analyzing mutational
47 profiling data at any scale, could power future studies on RNA structure and is
48 available on GitHub and the Python Package Index.

⁴⁹ Introduction

⁵⁰ Across all domains of life, RNA molecules perform myriad functions in development [1], immunity [2], translation [3], sensing [4, 5], epigenetics [6], cancer [7],
⁵¹ and more. RNA also constitutes the genomes of many threatening viruses [8],
⁵² including influenza viruses [9] and coronaviruses [10]. The capabilities of an RNA
⁵³ molecule depend not only on its sequence (primary structure) but also on its base
⁵⁴ pairs (secondary structure) and three-dimensional shape (tertiary structure) [11].

⁵⁵ Although high-quality tertiary structures provide the most information, resolving
⁵⁶ them often proves difficult or impossible with mainstay methods used for proteins [12]. Consequently, the world's largest database of tertiary structures – the
⁵⁷ Protein Data Bank [13] – has accumulated only 1,839 structures of RNAs (compared
⁵⁸ to 198,506 of proteins) as of February 2024. Worse, most of those RNAs
⁵⁹ are short: only 119 are longer than 200 nt; of those, only 24 are not ribosomal
⁶⁰ RNAs or group I/II introns. Due partly to the paucity of non-redundant long RNA
⁶¹ structures, methods of predicting tertiary structures for RNAs lag far behind those
⁶² for proteins [14].

⁶³ The situation is only marginally better for RNA secondary structures. If a diverse
⁶⁴ set of homologous RNA sequences is available, a consensus secondary
⁶⁵ structure can often be predicted using comparative sequence analysis, which
⁶⁶ has accurately modeled ribosomal and transfer RNAs, among others [15]. A formalization
⁶⁷ known as the covariance model [16] underlies the widely-used Rfam
⁶⁸ database [17] of consensus secondary structures for 4,170 RNA families (as of
⁶⁹ version 14.10). Although extensive, Rfam contains no protein-coding sequences
⁷⁰ (with some exceptions such as frameshift stimulating elements) and provides only
⁷¹ one secondary structure for each family, even though many RNAs fold into multiple
⁷² functional structures [18, 19]). Each family also models only a short segment
⁷³ of a full RNA sequence; for coronaviruses, existing families encompass the 5' and
⁷⁴ 3' untranslated regions, the frameshift stimulating element, and the packaging signal,
⁷⁵ which collectively constitute only 3% of the genomic RNA.

78 Predicting secondary structures faces two major obstacles due to the scarcity
79 of high-quality RNA structures, particularly for RNAs longer than 200 nt (includ-
80 ing long non-coding [20], messenger [21], and viral genomic [22] RNAs). First,
81 prediction methods trained on known RNA structures are limited to small, low-
82 diversity training datasets (generally of short sequences), which causes overfit-
83 ting and hence inaccurate predictions for dissimilar RNAs (including longer se-
84 quences) [23, 24]. Second, without known secondary structures of many di-
85 verse RNAs, the accuracy of any prediction method cannot be properly bench-
86 marked [21, 25]. For these reasons, and because thermodynamic-based models
87 also tend to be less accurate for longer RNAs [22] and base pairs spanning longer
88 distances [26], predicting secondary structures of long RNAs remains unreliable.

89 The most promising methods for determining the structures of long RNAs
90 use experimental data. Chemical probing experiments involve treating RNA with
91 reagents that modify nucleotides depending on the local secondary structure;
92 for instance, dimethyl sulfate (DMS) methylates adenosine (A) and cytidine (C)
93 residues only if they are not base-paired [27]. Modern methods use reverse
94 transcription to encode modifications of the RNA as mutations in the cDNA, fol-
95 lowed by next-generation sequencing – a strategy known as mutational profiling
96 (MaP) [28, 29]. A key advantage of MaP is that the sequencing reads can be
97 clustered to detect multiple secondary structures in an ensemble [30, 31]. De-
98 termining the base pairs in those structures still requires structure prediction [32],
99 although incorporating chemical probing data does improve accuracy [33, 34].

100 Several experimental methods have been developed to find base pairs directly,
101 with minimal reliance on structure prediction. M2-seq [35] introduces random mu-
102 tations before chemical probing to detect correlated mutations between pairs of
103 bases, which indicates the bases interact. However, alternative structures com-
104 plicate the data analysis [36], and detectable base pairs can be no longer than the
105 sequencing reads (typically 300 nt). For long-range base pairs, many methods in-
106 volving crosslinking, proximity ligation, and sequencing have been developed [37].
107 These methods can find base pairs spanning arbitrarily long distances – as well

108 as between different RNA molecules – but cannot resolve single base pairs or
109 alternative structures. Detecting, resolving, and quantifying alternative structures
110 with base pairs that span arbitrarily long distances remains an open challenge.

111 Here, we introduce “Structure Ensemble Ablation by Reverse Complement Hy-
112 bridization with Mutational Profiling” (SEARCH-MaP), an experimental method to
113 discover RNA base pairs spanning arbitrarily long distances. We also develop the
114 software “Structure Ensemble Inference by Sequencing, Mutation Identification,
115 and Clustering of RNA” (SEISMIC-RNA) to analyze MaP data and resolve alter-
116 native structures. Using SEARCH-MaP and SEISMIC-RNA, we discover an RNA
117 structure in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that
118 comprises dozens of long-range base pairs and folds in nearly half of genomic
119 RNA molecules. We show that it inhibits the folding of a pseudoknot that stim-
120 ulates ribosomal frameshifting [38, 39], hinting a role in regulating viral protein
121 synthesis. We find similar structures in other SARS-related viruses and transmis-
122 sible gastroenteritis virus (TGEV), suggesting that long-range base pairs involving
123 the frameshift stimulation element are a general feature of coronaviruses. In ad-
124 dition to revealing new structures in coronaviral genomes, our findings show how
125 SEARCH-MaP and SEISMIC-RNA can resolve secondary structure ensembles of
126 long RNA molecules – a necessary step towards a true “AlphaFold for RNA” [14].

127 **Results**

128 **Workflows of SEISMIC-RNA and SEARCH-MaP**

129 SEISMIC-RNA is all-in-one software for processing and graphing mutational pro-
130 filing data (e.g. DMS-MaPseq [29] and SHAPE-MaP [28]), inferring alternative
131 structures, and modeling secondary structures (Figure 1a). Designed for mod-
132 ern experiments comprising many samples and reference sequences, SEISMIC-
133 RNA can process as many samples in parallel as there are CPUs, at high speed
134 with low memory and storage footprints. Input files can be in (gzipped) FASTQ
135 format (which SEISMIC-RNA preprocesses with Cutadapt [40] and aligns with
136 Bowtie 2 [41]) or SAM/BAM/CRAM format [42]. SEISMIC-RNA ensures data qual-
137 ity with new algorithms that flag ambiguous base calls (e.g. deletions within repet-
138 itive sequences) and mask unusable reads and positions (e.g. with insufficient
139 coverage). Masking can also be applied to specific types of mutations, such as A-
140 to-G mismatches caused by ADAR editing [43]. To identify alternative structures,
141 SEISMIC-RNA introduces a new clustering algorithm – similar to DREEM [?]
142 yet able to cluster RNAs longer than the reads themselves – enabling analyses
143 of structure ensembles in long transcripts. After computing chemical reactivities,
144 SEISMIC-RNA can use them to model secondary structures with RNAstructure
145 Fold [44, 33] or to generate a variety of graphs – including correlations between
146 samples and receiver operating characteristic (ROC) curves. This entire workflow
147 can be run via the command line, while an additional Python interface enables
148 highly customized analyses, making SEISMIC-RNA accessible yet applicable to
149 many kinds of mutational profiling experiments.

150 In a SEARCH-MaP experiment, an RNA is assumed to fold into an ensemble of
151 one or more structures (Figure 1b). Searching for base pairs involving part of the
152 RNA – in this example, section P – begins by binding an antisense oligonucleotide
153 (ASO) to that part, which prevents it from base pairing (Figure 1c). Separately, the
154 RNA is chemically probed with (+ASO) and without (no-ASO) the ASO, followed
155 by mutational profiling (MaP) and sequencing (e.g. DMS-MaPseq [29]).

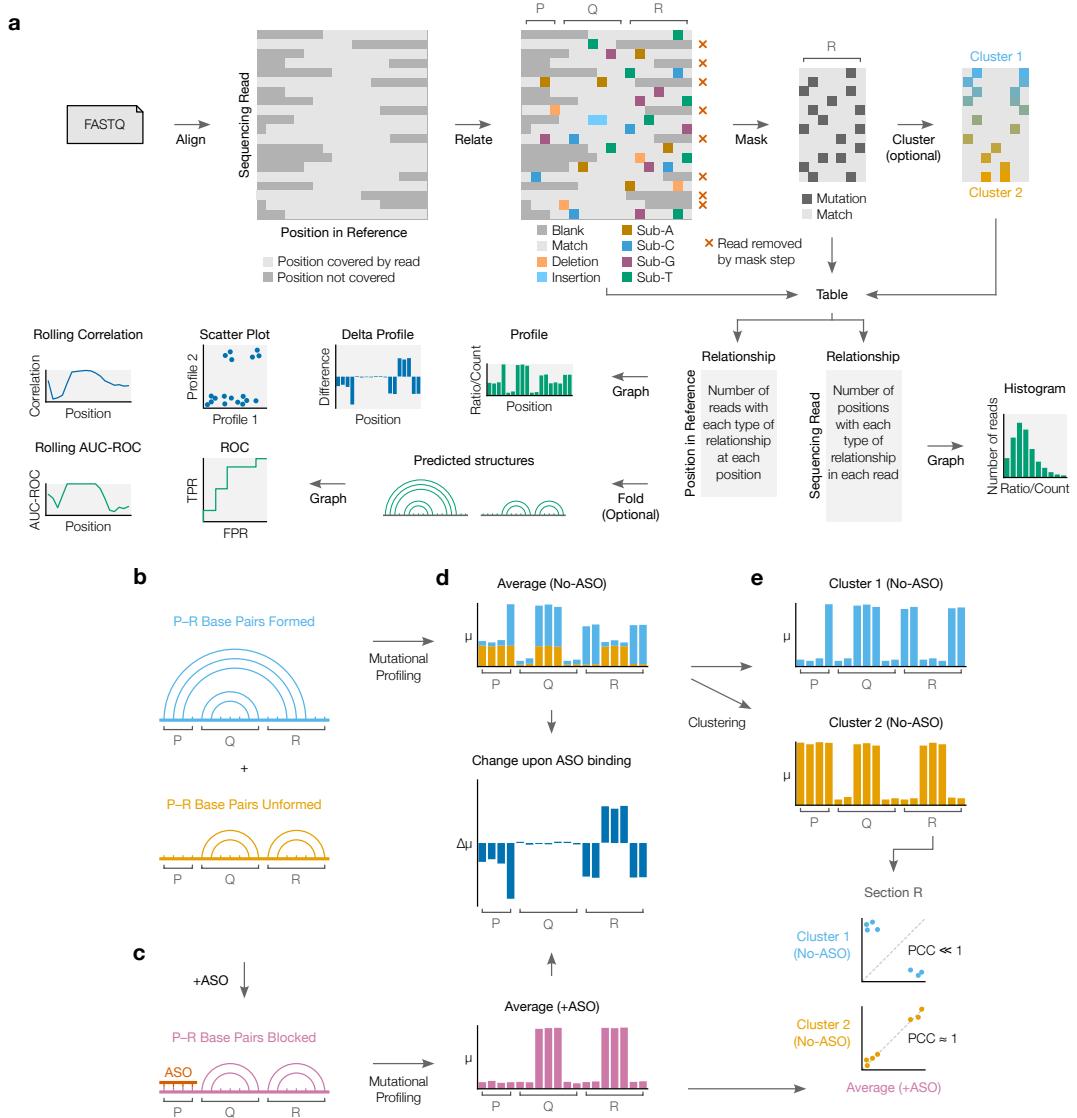


Figure 1: Workflows of SEISMIC-RNA and SEARCH-MaP. (Continued on next page.)

Figure 1: (Continued from previous page.) **(a)** 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 are masked by user-specified filters. Reads can then be clustered to reveal alternative structures. The types of relationships at each position and in each read are 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. **(b)** This RNA comprises 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. **(c)** Hybridizing an ASO to P blocks it from base-pairing with R. **(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.

156 Each structure theoretically has its own mutational profile, which is not directly
157 observable because all structures are physically mixed in an experiment [45]. The
158 only directly observable mutational profile is of the “ensemble average” – the aver-
159 age of the structures’ (unobserved) mutational profiles, weighted by the their (un-
160 observed) proportions (Figure 1d, top). The structure – and thus mutational profile
161 – of section R changes when it base-pairs with P; by preventing such base-pairing,
162 the ASO changes the ensemble average of R (Figure 1d, middle). However, the
163 ASO has no effect on section Q because Q never base-pairs with P. Therefore,
164 one can deduce that P base-pairs with R (but not with Q) because hybridizing an
165 ASO to P alters the mutational profile of R (but not of Q).

166 Furthermore, the mutational profile of the structure in which P and R base-pair
167 can be determined, even without knowing the exact base pairs. This step involves
168 clustering the no-ASO ensemble into two mutational profiles over section R – each
169 corresponding to one structure – and comparing them to the +ASO ensemble

170 average (Figure 1e). Because the ASO blocks the P–R base pairs, the +ASO
171 mutational profile will correlate better with that of the structure where P and R do
172 not base-pair; in this case, cluster 2 correlates better. Therefore, the mutational
173 profile of cluster 1 corresponds to the structure where P and R base-pair.

174 **SEARCH-MaP detects and quantifies long-range
175 base pairing in SARS-CoV-2**

176 Aside from ribosomes, many of the best-characterized functional long-range RNA
177 base pairs occur in the genomes of RNA viruses [46]. In coronaviruses, the first
178 open reading frame (ORF1) contains a frameshift stimulation element (FSE) that
179 makes a fraction of ribosomes slip into the -1 reading frame, bypass a 0-frame stop
180 codon, and translate to the end of ORF1 [47]. Every coronaviral FSE contains
181 a “slippery site” (UUUAAAC) and a structure characterized as a pseudoknot in
182 multiple species [48, 49, 50]. For SARS coronavirus 2 (SARS-CoV-2), 80-90 nt
183 segments of the core FSE have been shown to fold into a pseudoknot with three
184 stems [39, 51, 52]. However, in intact SARS-CoV-2, the FSE adopts a different
185 structure: one that could be recapitulated with a 2,924 nt segment but not a 283 nt
186 segment [53]. This finding suggested that the SARS-CoV-2 FSE involves long-
187 range base pairs.

188 We hypothesized that the long-range base pairs would match a proposed
189 structure named the “FSE-arch” [54]. If so, the structure of the FSE would be
190 perturbed by – and only by – ASOs targeting either side of the FSE-arch. To
191 investigate, we performed DMS-MaPseq [29] on a 2,924 nt segment of SARS-
192 CoV-2 after adding each of thirteen groups of DNA ASOs (Figure 2a). We used
193 RT-PCR primers flanking the ASO target site to confirm binding (Supplementary
194 Figure 1) and flanking the 5' side of the FSE-arch to measure changes in its struc-
195 ture (Supplementary Figure 2), except for ASO group 13, for which we obtained
196 no data.

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

199 ple and a no-ASO control (Figure 2b). The rolling PCC of a no-ASO replicate
200 remained between 0.93 and 1.00 (mean = 0.97), confirming the DMS reactivities
201 were reproducible. ASO group 9 – targeting both 3' inner stems of the FSE-arch –
202 caused the rolling PCC to dip below 0.5 over both 5' inner stems, as expected if the
203 inner stems of the FSE-arch existed. The only other ASO groups with substantial
204 effects were 3, 4, and 5, which overlapped or abutted the FSE and presumably
205 perturbed short-range base pairs; the outer stem of the FSE-arch (targeted by
206 ASO group 10) did not apparently form. These results suggest both inner stems
207 (but not the outer stem) of the proposed FSE-arch [54] exist and are the predom-
208 inant long-range base pairs involving the immediate vicinity of the FSE.

209 To determine in what fraction of molecules the two inner stems of the FSE-
210 arch form, we clustered reads from the 5' side of the FSE-arch for the no-ASO
211 control. We found two clusters with a 43/57% split and – to determine if they
212 corresponded to the two inner stems formed versus unformed – compared their
213 DMS reactivities to those after adding ASO group 9, which would block the two
214 inner stems (Figure 2c, top). Cluster 2 had similar DMS reactivities (PCC = 0.95),
215 indicating it corresponds to the stems unformed. Meanwhile, the DMS reactivities
216 of cluster 1 differed (PCC = 0.64), suggesting it corresponds to the stems formed.

217 To further support this result, we leveraged the preexisting model of the FSE-
218 arch [54]. If cluster 1 did correspond to the two inner stems formed, its DMS
219 reactivities would agree well with their structures (i.e. paired and unpaired bases
220 should have low and high reactivities, respectively), while those of cluster 2 would
221 agree less. We quantified this agreement using receiver operating characteristic
222 (ROC) curves (Figure 2c, bottom). The area under the curve (AUC) for cluster
223 1 was 1.00, indicating perfect agreement with the two inner stems of the FSE-
224 arch; while that of cluster 2 was 0.57, close to no agreement (0.50). This result
225 further supports that clusters 1 and 2 correspond to the two inner stems formed
226 and unformed, respectively.

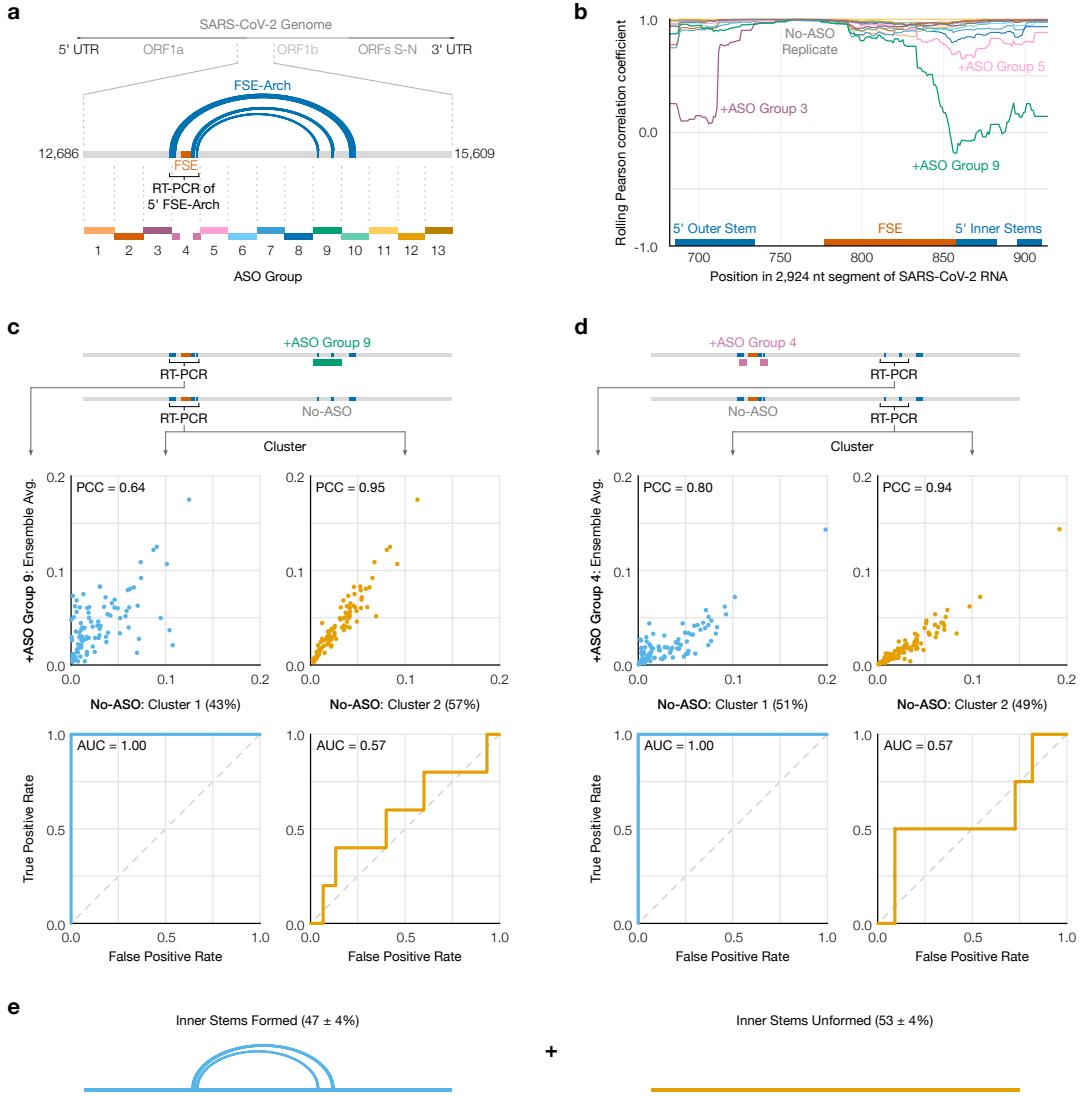


Figure 2: SEARCH-MaP of long-range base pairs involving the SARS-CoV-2 FSE. **(a)** The 2,924 nt segment of the SARS-CoV-2 genome containing the frameshift stimulation element (FSE) and putative FSE-arch [54]. The target site for each group of antisense oligonucleotides (ASOs) is indicated by dotted 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 excluded 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.

227 If the RNA did exist as an ensemble of the two inner stems formed and un-
228 formed, the 3' side of the FSE-arch would also cluster into states with the two
229 inner stems formed and unformed. We thus RT-PCR'd and clustered the 3' FSE-
230 arch in the no-ASO control and +ASO group 4 and found – similar to the previous
231 result – that the DMS reactivities after blocking the 5' FSE-arch with ASO group
232 4 resembled those of cluster 2 ($PCC = 0.94$) but not cluster 1 ($PCC = 0.80$), while
233 the structure of the two inner stems agreed with cluster 1 ($AUC = 1.00$) but not
234 cluster 2 ($AUC = 0.57$) (Figure 2d). We concluded that the RNA exists as an en-
235 semble of structures in which the two inner stems of the FSE-arch form in $47\% \pm$
236 4% of molecules (Figure 2e).

237 **The long-range stems compete with the frameshift 238 pseudoknot in SARS-CoV-2**

239 To determine if the FSE forms other long-range stems, in lieu of the original outer
240 stem of the FSE-arch [54], we modeled a 1,799 nt segment centered on the FSE-
241 arch. Although computationally predicting long-range base pairs is notoriously
242 unreliable [26, 22], we speculated that we could improve accuracy by incorporat-
243 ing the DMS reactivities of cluster 1 on both sides of the FSE-arch (Supplementary
244 Figure 3). For the innermost stem – which we call long stem 1 (LS1) – nine of thir-
245 teen structures (69%) predicted using the cluster 1 DMS reactivities contained
246 LS1, compared to five of eleven (45%) using the ensemble average and four of
247 twenty (20%) using no DMS reactivities. For the second-most inner stem (LS2),
248 eight structures (62%) predicted using cluster 1 contained LS2, while none did us-
249 ing average or no DMS reactivities. Thus, the DMS reactivities corresponding to
250 the long-range cluster enabled predicting the long-range stems more consistently,
251 allowing us to refine our model of the long-range stems.

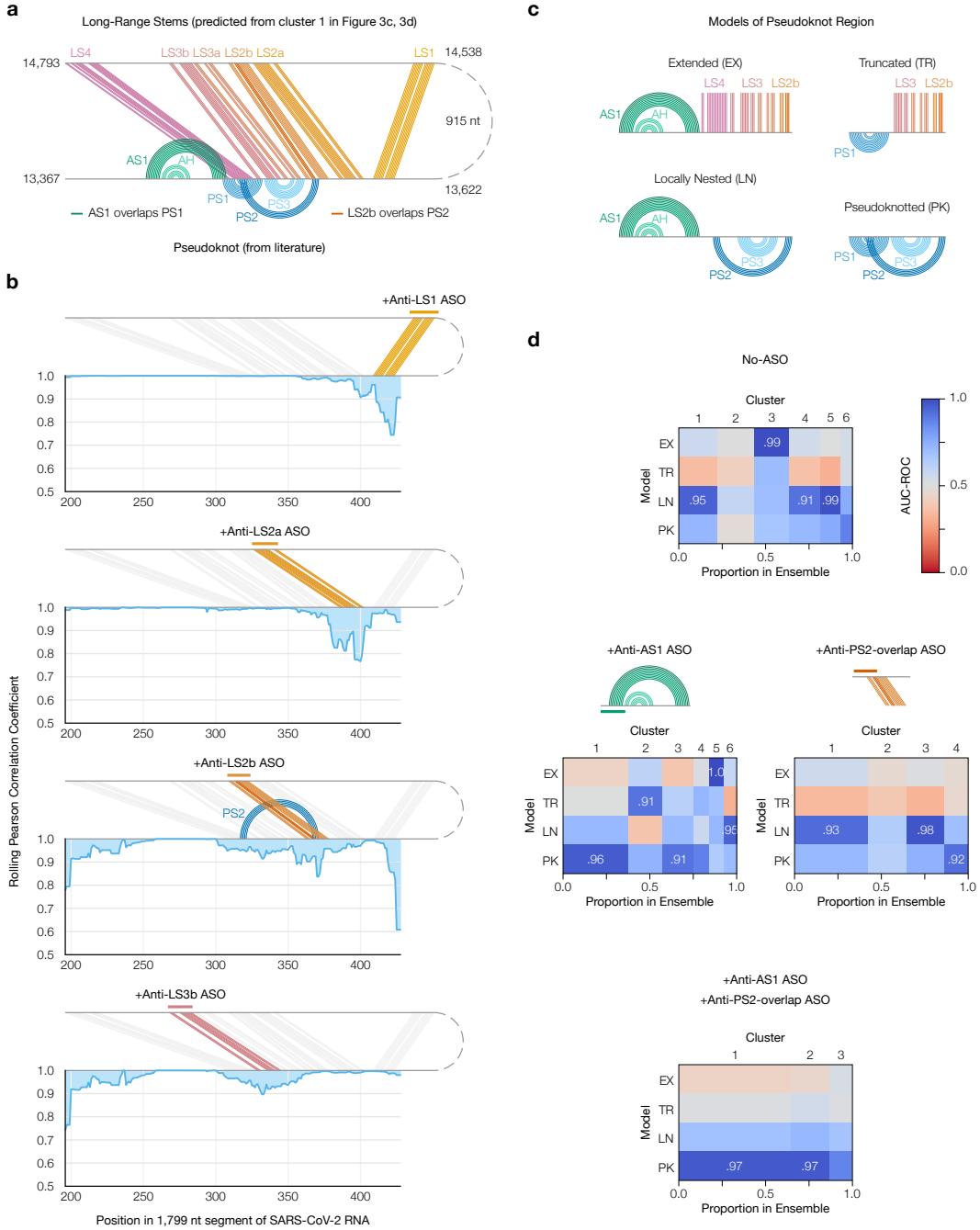


Figure 3: Refinement of the long-range structure model and competition with the frameshift pseudoknot. (a) Refined model of the long-range stems (minimum free energy prediction based on cluster 1 in Figure 2c and d) including alternative stem 1 (AS1) [53]; the attenuator hairpin (AH) [55]; and long stems 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.

252 Our refined model based on the long-range cluster (Figure 3a) included not
253 only the two inner stems of the FSE-arch – LS1 and LS2a/b – but also two long
254 stems (LS3a/b and LS4) absent from the original FSE-arch model [54], as well as
255 alternative stem 1 (AS1) [53]. To verify this model, we performed SEARCH-MaP
256 on the 1,799 nt segment using 15-20 nt LNA/DNA mixmer ASOs for single-stem
257 precision (Supplementary Table 3). Each ASO targeted the 3' side of one stem,
258 and we measured the change in DMS reactivities of the FSE (Figure 3b). ASOs
259 targeting the 3' sides of LS1 and LS2a perturbed the DMS reactivities in exactly
260 the expected locations on the 5' sides. An ASO for the 3' side of LS2b perturbed
261 the FSE with more off-target effects, likely because this stem overlaps with pseu-
262 doknot stem 2 (PS2). Blocking LS3b also resulted in a main effect around the
263 intended location, with one off-target effect upstream, suggesting that other base
264 pairs between the pseudoknot and this upstream region may exist. Therefore,
265 stems LS1, LS2a/b, and LS3b do exist – at least in a portion of the ensemble.

266 LS2b, LS3, and LS4 of the refined model overlap all three stems of the pseu-
267 doknot (PS1, PS2, and PS3) that stimulates frameshifting [39]. To test whether
268 these long stems actually compete with the pseudoknot, we made four possible
269 models of the FSE structure by combining mutually compatible stems (Figure 3c).
270 Then, we clustered the 1,799 nt segment without ASOs up to 6 clusters – the
271 maximum number reproducible between replicates (Supplementary Figure 4a) –
272 and compared each cluster to each structure model using the area under the re-
273 ceiver operating characteristic curve (AUC-ROC) over the positions spanned by
274 the pseudoknot: 305-371 (Figure 3d, top). We considered a cluster and model
275 to be “consistent” if the AUC-ROC was at least 0.90. The locally nested model
276 (AS1 plus PS2 and PS3) was consistent with three clusters totaling 52% of the en-
277 semble, while the extended model (AS1 plus all long-range stems) was consistent
278 with one cluster (20%). No clusters were fully consistent with the pseudoknotted
279 model, though the least-abundant cluster (7%) came close with an AUC-ROC of
280 0.88. The remaining cluster (21%) was not consistent with any model, suggesting
281 that the ensemble contains structures beyond those in Figure 3c.

282 Adding an ASO targeting the 5' side of AS1 reduced the proportion of AS1-
283 containing states (extended and locally nested) from 72% to 16% (Figure 3d, left;
284 Supplementary Figure 4b). In their absence emerged clusters consistent with the
285 pseudoknotted and truncated models, constituting 56% and 20% of the ensemble,
286 respectively. Meanwhile, adding an ASO that blocked the part of LS2b that over-
287 laps PS2 eliminated the extended state (which includes LS2b) and produced one
288 cluster (13%) consistent with the pseudoknotted model (Figure 3d, right; Supple-
289 mentary Figure 4c). Adding both ASOs simultaneously collapsed the ensemble
290 into three clusters of which two (87%) were highly consistent with the pseudo-
291 knotted model (Figure 3d, bottom; Supplementary Figure 4d). Since blocking the
292 PS2-overlapping portion of LS2b increased the proportion of clusters consistent
293 (or nearly so) with the pseudoknotted model – both alone and combined with the
294 anti-AS1 ASO – the long-range stems did appear to outcompete the pseudoknot.

295 **Frameshift stimulating elements of multiple 296 coronaviruses form long-range base pairs**

297 We hypothesized that other coronaviruses would also feature long-range base
298 pairs involving their FSEs. To search for such coronaviruses computationally, we
299 predicted structures of 2,000 nt sections surrounding the FSEs of all 62
300 complete coronavirus genomes in the NCBI Reference Sequence Database [56]
301 (Supplementary Figure 5). We selected ten coronaviruses, at least one from
302 each genus (Supplementary Figure 6a), based on with which bases the FSE
303 was predicted to base-pair. In *Betacoronavirus*, SARS coronaviruses 1
304 (NC_004718.3) and 2 (NC_045512.2) and bat coronavirus BM48-31
305 (NC_014470.1) because they clustered into their own structural outgroup;
306 MERS coronavirus (NC_019843.3), predicted to pair with positions 510-530;
307 and human coronavirus OC43 (NC_006213.1) and murine hepatitis virus strain
308 A59 (NC_048217.1), both predicted to pair with positions 10-20. In
309 *Alphacoronavirus*, transmissible gastroenteritis virus (NC_038861.1) and bat
310 coronavirus 1A (NC_010437.1), predicted to pair with positions 440-460 and

311 350-360, respectively. In *Gammacoronavirus*, avian infectious bronchitis virus
312 strain Beaudette (NC_001451.1), predicted to pair with positions 330-350. And
313 in *Deltacoronavirus*, common moorhen coronavirus HKU21 (NC_016996.1) had
314 the most promising long-range base pairs.

315 We screened each of these ten coronaviruses for long-range base pairs with
316 the FSE by comparing the DMS reactivities of 239 nt segment comprising the
317 FSE with minimal flanking sequences and a 1,799 nt segment encompassing the
318 FSE and all sites with which it was predicted to interact. All coronaviruses except
319 for human coronavirus OC43 and MERS coronavirus showed differences in their
320 DMS reactivity profiles between the 239 and 1,799 nt segments (Supplementary
321 Figure 6b), suggesting their FSEs formed long-range base pairs.

322 To determine which RNA elements the FSE base-pairs with in each corona-
323 virus, we performed SEARCH-MaP on the 1,799 nt RNA segment using DNA
324 ASOs targeting the vicinity of the FSE (Figure 4). The rolling Spearman correla-
325 tion coefficient (SCC) between the +ASO and no-ASO mutational profiles dipped
326 below 0.9 at the ASO target site in every coronavirus segment, confirming the
327 ASOs bound and altered the structure. To confirm we could detect long-range
328 base pairs, we compared the rolling SCC for the SARS-CoV-2 segment to our
329 refined model of the FSE structure (Figure 4, blue). The SCC dipped below 0.9 at
330 positions 1,483-1,560 and at 1,611-1,642, which coincide with stems LS2-LS3 (po-
331 sitions 1,476-1,550 within the 1,799 nt segment) and stem LS4 (positions 1,600-
332 1,622). These dips were the two largest downstream of the FSE; although others
333 (corresponding to no known base pairs) existed, they were barely below 0.9 and
334 could have resulted from base pairing between these regions and other (non-
335 FSE) regions. Near LS1 (positions 1,367-1,381), the SCC dipped only slightly to
336 a minimum of 0.95, presumably because LS1 is the smallest (15 nt) and most
337 isolated long-range stem. Therefore, this method was sensitive enough to detect
338 all but the smallest long-range stem, and specific enough that the two largest dips
339 corresponded to validated long-range stems.

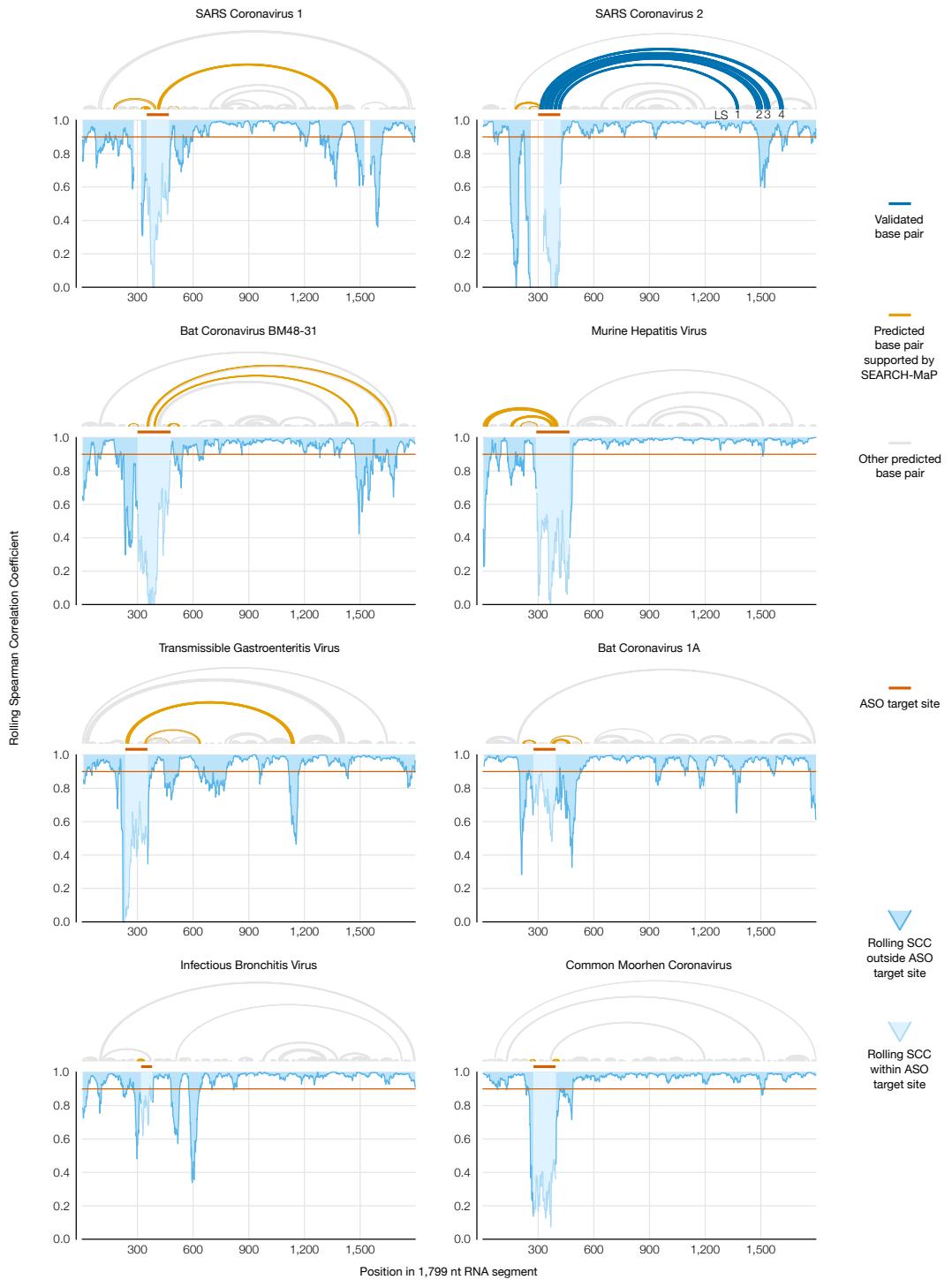


Figure 4: Evidence for long-range RNA–RNA base pairs involving the FSE in four additional coronaviruses. Rolling (window = 45 nt) Spearman correlation coefficient (SCC) of DMS reactivities between the +ASO and no-ASO samples for each 1,799 nt segment of a coronaviral genome. The target site of each ASO is highlighted on the SCC data and shown above each graph. Structures predicted with RNAstructure [44] 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.

340 We found similar long-range stems in SARS-CoV-1 and another SARS-related
341 virus, bat coronavirus BM48-31. Both viruses showed dips in SCC at roughly the
342 same positions as LS2-LS4 in SARS-CoV-2, indicating that they have homologous
343 structures. SARS-CoV-1 also had a wide dip below 0.9 at positions 1,284-1,394,
344 corresponding to a homologous LS1. Thus, three SARS-related viruses share
345 these long-range stems involving the FSE, hinting at a conserved function.

346 In every other species except common moorhen coronavirus, we found promi-
347 nent dips in SCC at least 200 nt from the ASO target site. To determine what long-
348 range base pairs could have caused those dips, we used RNAstructure Fold [44,
349 33] guided by the DMS reactivities of the no-ASO ensemble average; clustered
350 DMS reactivities, while superior for structure modeling (Supplementary Figure 3),
351 were unavailable. Nevertheless, we managed to find long-range base pairs con-
352 sistent with the SEARCH-MaP data for both murine hepatitis virus and transmis-
353 sible gastroenteritis virus (Figure 4, orange). We conclude that long-range base
354 pairing involving the FSE occurs more widely than in just SARS-CoV-2, including
355 in the genus *Alphacoronavirus*.

356 **Structure of the full TGEV genome in ST cells 357 supports long-range base pairing involving the FSE**

358 Transmissible gastroenteritis virus (TGEV) is a strain of *Alphacoronavirus 1* [57]
359 that infects pigs and causes vomiting and diarrhea, often fatally [58]. Due to the
360 impacts of TGEV [58] and our evidence of long-range base pairs, we sought to
361 model the genomic secondary structures of live TGEV. We began by infecting ST
362 cells with TGEV and performing DMS-MaPseq [29] (Figure 5a). The DMS reactiv-
363 ities over the full genome were consistent between technical and biological repli-
364 cates (PCC = 0.94, Supplementary Figure 7a and b), albeit not with the 1,799 nt
365 segment *in vitro* (PCC = 0.77), which showed that verifying the long-range stem
366 in live TGEV would be necessary (Supplementary Figure 7c).

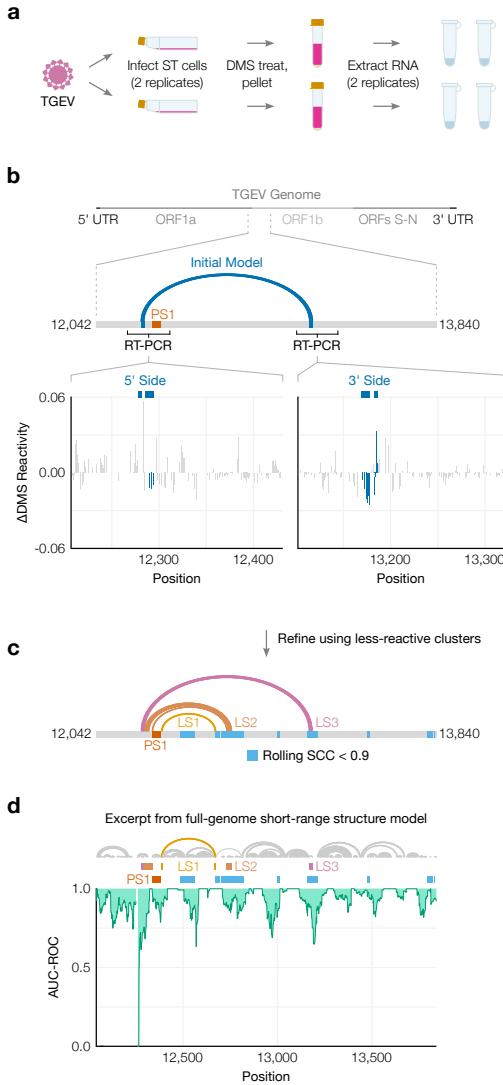


Figure 5: Genomic secondary structure of live TGEV. **(a)** Schematic of the experiment in which two biological replicates of ST cells were infected with TGEV, DMS-treated, and pelleted. Cell pellets were divided into two technical replicates prior to extraction of DMS-modified RNA. **(b)** Differences 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.

367 To quantify the long-range base pairs, we RT-PCRed the 5' and 3' sides, con-
368 firmed their DMS reactivities were consistent with the full genome's (Supplemen-
369 tary Figure 7d), and clustered both sides (Figure 5b). Although the clusters were
370 indistinguishable by their correlations with the +ASO sample or AUC-ROC scores
371 (Supplementary Figure 8a and b), bases involved in the predicted long-range pairs
372 were generally less DMS-reactive in cluster 2 (Figure 5b), which we hypothesized
373 corresponded to the long-range base pairs forming. In support, the long-range
374 stem (hereafter, LS3) appeared when using DMS reactivities from cluster 2 on
375 both sides (Figure 5c), but not cluster 1 (Supplementary Figure 8c). The refined
376 model based on cluster 2 included another long-range stem, LS2, which was also
377 supported by a dip in the rolling SCC (Figure 5c).

378 We used the ensemble average DMS reactivities to produce one "ensemble
379 average" model of short-range (up to 300 nt) base pairs in the full TGEV genome
380 (Supplementary Figure 9). To verify the model quality, we confirmed that the
381 modeled structure of the first 330 nt included the highly conserved stem loops
382 SL1, SL2, SL4, and SL5a/b/c in the 5' UTR [10] (Supplementary Figure 10a) and
383 was consistent with the DMS reactivities (AUC-ROC = 0.97) (Supplementary Fig-
384 ure 10b). The AUC-ROC was lower in many locations throughout the rest of the
385 genome (Supplementary Figure 9), indicating that a single secondary structure
386 consistent with the ensemble average DMS reactivities could not be found – which
387 suggests alternative structures, long-range base pairs, or both [53]. Accordingly,
388 we found a large dip in AUC-ROC just upstream of the FSE, centered on the 5'
389 ends of LS2 and LS3, as well as smaller dips at the 3' ends of both stems (Fig-
390 ure 5d). In fact, at or near every location that SEARCH-MaP had evidenced to
391 interact with the FSE – where the rolling SCC had dipped – the AUC-ROC also
392 dipped. This finding supports that regions with low AUC-ROC in general are good
393 starting points for investigating alternative structures and long-range base pairing
394 with SEARCH-MaP.

395 Discussion

396 In this work, we developed SEARCH-MaP and SEISMIC-RNA and applied them
397 to detect structural ensembles involving long-range base pairs in SARS-CoV-2
398 and other coronaviruses. Previous studies have demonstrated that binding an
399 ASO to one side of a long-range stem would perturb the chemical probing re-
400 activities of the other side [59, 60, 61]. Here, we separated and identified the
401 reactivities corresponding to long-range stems formed and unformed. This ad-
402 vance enables isolating the reactivities of the long-range stem formed – on not
403 just one but both sides of the stem, linking corresponding alternative structures
404 over distances much greater than the length of a read, which has not been pos-
405 sible in previous studies [30, 31]. Using the linked reactivities from both sides of
406 a long-range stem, its secondary structure can be modeled more accurately than
407 would be possible using the ensemble average reactivities, as we have done for
408 SARS-CoV-2 (Figure 3) and TGEV (Figure 5).

409 SEISMIC-RNA builds upon our previous work, the DREEM algorithm [30].
410 Here, we have optimized the algorithm to run approximately 10-30 times faster
411 and built an entirely new workflow around it for aligning reads, calling mutations,
412 masking data, and outputting a variety of graphs. SEISMIC-RNA can process
413 data from any mutational profiling experiment, including DMS-MaPseq [29] and
414 SHAPE-MaP [28], not just SEARCH-MaP. The software is available from
415 the Python Package Index (pypi.org/project/seismic-rna) or GitHub
416 (github.com/rouskinlab/seismic-rna) and can be used as a command line
417 executable program (`seismic`) or via its Python application programming
418 interface (`import seismicrna`).

419 We envision SEARCH-MaP and SEISMIC-RNA bridging the gap
420 between broad and detailed investigations of RNA structure. Other
421 methods such as proximity ligation [62, 63, 64, 65, 66] provide broad,
422 transcriptome-wide information on RNA structure and could be used as a
423 starting point to find structures of interest for deeper investigation with
424 SEARCH-MaP/SEISMIC-RNA. Indeed, the first evidence of the FSE-arch in

425 SARS-CoV-2 came from such a study [54]. To investigate RNA structures in
426 detail, M2-seq [35] and related methods [36] can pinpoint base pairs with up to
427 single-nucleotide resolution and minimal need for structure prediction. However,
428 base pairs are detectable only if the paired bases occur on the same sequencing
429 read, which restricts their spans to at most the read length (typically 300 nt).
430 Because the capabilities of M2-seq and SEARCH-MaP complement each other,
431 they could be integrated: first SEARCH-MaP/SEISMIC-RNA to discover,
432 quantify, and model long-range base pairs; then M2-seq for short-range base
433 pairs. By providing the missing link – structure ensembles involving long-range
434 base pairs – SEARCH-MaP and SEISMIC-RNA could combine broad and
435 detailed views of RNA structure into one coherent model.

436 To understand structures of long RNA molecules, SEARCH-MaP and
437 SEISMIC-RNA could also be used to validate predicted secondary structures
438 and benchmark structure prediction algorithms. Algorithms that predict
439 secondary structures achieve lower accuracies for longer sequences [26, 22],
440 hence long-range base pairs in particular must be confirmed independently. We
441 envision a workflow to determine the structure ensembles of an arbitrarily long
442 RNA molecule that begins with DMS-MaPseq [29]. The DMS reactivities would
443 be used [33] to predict two initial models of the structure: one with a limit to the
444 base pair length (for short-range pairs), the other without (for long-range pairs).
445 Sections of the RNA with potential long-range pairs would be flagged from the
446 long-range model and from regions of the short-range model that disagreed with
447 the DMS reactivities (as in Figure 5d). Then, SEARCH-MaP/SEISMIC-RNA
448 could be used to validate, quantify, and refine the potential long-range base
449 pairs; and other methods such as M2-seq [35] to do likewise for short-range
450 base pairs. This integrated workflow could characterize the secondary structures
451 of RNA molecules that have evaded existing methods (e.g. messenger
452 RNAs [21]) as well provide much-needed benchmarks for secondary structure
453 prediction algorithms [25].

454 In this study, we focused on the genomes of coronaviruses, specifically
455 long-range base pairs involving the frameshift stimulating element (FSE).
456 Long-range base pairs implicated in frameshifting also occur in several plant
457 viruses of the family *Tombusviridae* [67, 68, 69]. However, in *Tombusviridae*
458 species, the frameshift pseudoknots themselves are made of long-range base
459 pairs; in coronaviruses, the pseudoknots are local structures [48, 49, 50, 39] and
460 (at least in SARS-CoV-2) compete with long-range base pairs. Consequently,
461 the long-range base pairs are necessary for frameshifting in *Tombusviridae*
462 species [67, 68, 69] but dispensable in coronaviruses: even the 80-90
463 nt core FSE of SARS-CoV-2 has stimulated 15-40% of ribosomes to
464 frameshift in dual luciferase constructs [38, 70, 39, 71, 72, 53]. Surprisingly,
465 frameshifting has appeared to be nearly twice as frequent (50-70%) in live
466 SARS-CoV-2 [73, 74, 75]; whether this discrepancy is due to long-range
467 base-pairing, methodological artifacts, or *trans* factors [76] is unknown [77].

468 If, how, and why the long-range base pairs affect frameshifting in
469 coronaviruses are open questions. For *Tombusviridae*, one study [67] suggested
470 that the long-range stem regulates viral RNA synthesis by negative feedback:
471 without RNA polymerase, the long-range stem would form and stimulate
472 frameshifting to produce polymerase, which would then unwind the long-range
473 stem while replicating the genome. However, this mechanism seems
474 implausible in coronaviruses, where RNA synthesis and translation occur in
475 separate subcellular compartments (the double-membrane vesicles and the
476 cytosol, respectively) [78]. Another study on *Tombusviridae* [69] hypothesized
477 that after the ribosome has frameshifted, long-range stems destabilize the FSE
478 so the ribosome can unwind it and continue translating. As the long-range base
479 pairs in SARS-CoV-2 do compete with the pseudoknot, they might also have this
480 role, which – for coronaviruses – could not be strictly necessary for
481 frameshifting. One study [74] of translation in SARS-CoV-2 at different time
482 points measured frameshifting around 20% at 4 hours post infection but 60-80%
483 at 12-36 hours. This result is consistent with a previous hypothesis [79] that

484 coronaviruses use frameshifting to time protein synthesis: first translating
485 ORF1a to suppress the immune system, then translating ORF1b containing the
486 RNA polymerase. We surmise the long-range base pairs would form in virions
487 and persist when the virus released its genome into a host cell, where they
488 would initially suppress frameshifting. Once host protein synthesis had been
489 inhibited and the double-membrane vesicles formed, a signal specific to the
490 cytosol would disassemble the long-range base pairs so that frameshifting could
491 occur efficiently and produce the replication machinery from ORF1b. The
492 long-range base pairs would form in viral progeny but not in genomic RNA
493 released into the cytosol for translation, so that more ORF1b could be translated.
494 This possible role of long-range base pairs in the coronaviral life cycle could be
495 tested by probing the RNA structure in subcellular compartments and virions,
496 identifying cytosolic factors that could disassemble the long-range base pairs,
497 and quantifying how they affect frameshifting in the context of a live coronavirus.

498 Future studies could also expand the scope of SEARCH-MaP and SEISMIC-
499 RNA. While all SEARCH-MaP experiments in this study were performed *in vitro*,
500 the method would likely also be feasible *in cellulo*: DMS-MaPseq can detect ASOs
501 binding to RNAs within cells [80]. The main challenges would likely involve op-
502 timizing the ASO probes and transfection protocols to maximize the signal while
503 minimizing unwanted side effects such as immunogenicity. SEARCH-MaP can
504 screen an entire transcript (as in Figure 2), but scaling up to an entire transcrip-
505 tive could prove challenging. One strategy for probing many RNAs simultane-
506 ously could involve adding a pool of ASOs – with no more than one ASO capable
507 of binding each RNA – rather than one ASO at a time. In this manner, a similar
508 number of samples would be needed to search all RNAs as would be needed for
509 the longest RNA. Distinguishing direct from indirect base pairing is another area
510 for development: if segment Q could base-pair with either P or R, then blocking P
511 could perturb R (and vice versa) as a consequence of perturbing Q, even though
512 P and R could not base-pair directly. A solution could be to first block Q with

513 one ASO; then, if blocking P with another ASO caused no change in R (and vice
514 versa), it would suggest that they could only interact indirectly (through Q).

515 We imagine that SEARCH-MaP and SEISMIC-RNA will make it practical to de-
516 termine accurate secondary structure ensembles of entire messenger, long non-
517 coding, and viral RNAs. Collected in a database of long RNA structures, these
518 results would facilitate subsequent efforts to predict RNA structures and bench-
519 mark algorithms, culminating in a real “AlphaFold for RNA” [14] in the hands of
520 every biologist.

521 **Methods**

522 **Development of SEISMIC-RNA**

523 SEISMIC-RNA was written in Python (currently compatible with v3.10 or greater)
524 using PyCharm Community Edition. Its dependencies include Python packages
525 NumPy [81], Numba [82], pandas [83], and SciPy [84]; as well as Samtools [42],
526 Cutadapt [40], Bowtie 2 [41], and RNAstructure [44].

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

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

529 A DNA template of the 2,924 nt segment of SARS-CoV-2,
530 including a T7 promoter, was amplified from a previously
531 constructed plasmid [53] ([Supplementary Data]) in 50 µl using 2X
532 CloneAmp HiFi PCR Premix (Takara Bio) with 250 nM primers
533 TAATACGACTCACTATAGAATAATGAGCTTAGCCTGTTGCACTACG and
534 TAAATTGCGGACATACTTATCGGCAATTTGTTACC (Thermo Fisher
535 Scientific); initial denaturation at 98°C for 60 s; 35 cycles of 98°C for 10 s, 65°C
536 for 10 s, and 72°C for 15 s; and final extension at 72°C for 60 s. The 50 µl PCR
537 product with 10 µl of 6X Purple Loading Dye (New England Biolabs) was
538 electrophoresed through a 50 ml gel – 1% SeaKem Agarose (Lonza), 1X
539 tris-acetate-EDTA (Boston BioProducts), and 1X SYBR Safe (Invitrogen) – at
540 60 V for 60 min. The band at roughly 3 kb was extracted using a Zymoclean Gel
541 DNA Recovery Kit (Zymo Research) according to the manufacturer's protocol,
542 eluted in 10 µl of nuclease-free water (Fisher Bioreagents), and measured with a
543 NanoDrop (Thermo Fisher Scientific). To increase yield, the gel-extracted DNA
544 was fed into a second round of PCR and gel extraction using the same protocol.
545 Due to remaining contaminants, the DNA was further purified using a DNA
546 Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's

547 protocol, eluted in 10 μ l of nuclease-free water (Fisher Bioreagents), and
548 measured with a NanoDrop (Thermo Fisher Scientific).

549 150 ng of DNA template was transcribed using a MEGAscript T7 Transcrip-
550 tion Kit (Invitrogen) according to the manufacturer's protocol, incubating at 37°C
551 for 3 hr. DNA template was then degraded by incubating with 1 μ l of TURBO
552 DNase (Invitrogen) at 37°C for 15 min. RNA was purified using an RNA Clean &
553 Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol,
554 eluted in 20 μ l of nuclease-free water (Fisher Bioreagents), and measured with
555 a NanoDrop (Thermo Fisher Scientific).

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

557 Antisense oligonucleotides (ASOs) were ordered from Integrated DNA Technolo-
558 gies already resuspended to 10 μ M in 1X IDTE buffer (10 mM Tris, 0.1 mM EDTA)
559 in a 96-well PCR plate. Each ASO pool was assembled from 25 pmol of each con-
560 stituent ASO (Supplementary Table 1); volume was adjusted to 12.5 μ l by adding
561 TE Buffer – 10 mM Tris (Invitrogen) with 0.1 mM EDTA (Invitrogen). 450 fmol of
562 2,924 nt SARS-CoV-2 RNA was added to each ASO pool for a total of 13.5 μ l in a
563 PCR tube. The tube was heated to 95°C for 60 s to denature the RNA, placed on
564 ice for several minutes, and transferred to a 1.5 ml tube. To refold the RNA, 35 μ l
565 of 1.4X refolding buffer comprising 400 mM sodium cacodylate pH 7.2 (Electron
566 Microscopy Sciences) and 6 mM magnesium chloride (Invitrogen) was added,
567 then incubated at 37°C for 25 min. For no-ASO control 1, 12.5 μ l of TE Buffer
568 was used instead of an ASO pool. For no-ASO control 2, 12.5 μ l of TE Buffer was
569 added after placing on ice and before refolding to confirm the timing of adding TE
570 Buffer would not alter the RNA structure.

571 RNA was treated with DMS in 50 μ l containing 1.5 μ l (320 nM) of DMS
572 (Sigma-Aldrich) while shaking at 500 rpm in a ThermoMixer C (Eppendorf) at
573 37°C for 5 min. To quench, 30 μ l of beta-mercaptoethanol (Sigma-Aldrich) was
574 added and mixed thoroughly. RNA was purified using an RNA Clean &
575 Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol,

576 eluted in 10 μ l of nuclease-free water (Fisher Bioreagents), and measured with a
577 NanoDrop (Thermo Fisher Scientific).

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

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

586 4 μ l RNA was reverse transcribed in 20 μ l containing 1X First Strand Buffer
587 (Invitrogen), 500 μ M dNTPs (Promega), 5 mM dithiothreitol (Invitrogen), 500 nM
588 FSE primer CTTCGTCCTTTCTTGGAAAGCGACA (Integrated DNA
589 Technologies), 500 nM section-specific reverse primer (Integrated DNA
590 Technologies, Supplementary Table 2), 1 μ l of RNaseOUT (Invitrogen), and 1 μ l
591 of TGIRT-III enzyme (InGex) at 57°C for 90 min, followed by inactivation at 85°C
592 for 15 min. To degrade the RNA, 1 μ l of Hybridase Thermostable RNase H
593 (Lucigen) was added to each tube and incubated at 37°C for 20 min. 1 μ l of
594 unpurified RT product was amplified in 12.5 μ l using the Advantage HF 2 PCR
595 Kit (Takara Bio) with 1X Advantage 2 PCR Buffer, 1X Advantage-HF 2 dNTP
596 Mix, 1X Advantage-HF 2 Polymerase Mix, 250 nM primers (Integrated DNA
597 Technologies) for either the FSE (CCCTGTGGGTTTACACTAAAAAC and
598 CTTCGTCCTTTCTTGGAAAGCGACA) or specific section (Supplementary
599 Table 2); initial denaturation at 94°C for 60 s; 25 cycles of 94°C for 30 s, 60°C for
600 30 s, and 68°C for 60 s; and final extension at 68°C for 60 s. 5 μ l of every
601 amplicon from the same RT product was pooled and then purified using a DNA
602 Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's
603 protocol, eluted in 20 μ l of nuclease-free water (Fisher Bioreagents), and
604 measured with a NanoDrop (Thermo Fisher Scientific).

605 200 ng of pooled PCR product was prepared for sequencing using the NEB-
606 Next Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to
607 the manufacturer's protocol with the following modifications. During size selection
608 after adapter ligation, 27.5 μ l and 12.5 μ l of NEBNext Sample Purification Beads
609 (New England Biolabs) were used in the first and second steps, respectively, to
610 select inserts of 280-300 bp. Indexing PCR was run at half volume (25 μ l) for 3
611 cycles. In lieu of the final bead cleanup, 420 bp inserts were selected using a 2%
612 E-Gel SizeSelect II Agarose Gel (Invitrogen) according to the manufacturer's pro-
613 tocol. DNA concentrations were measured using a Qubit 3.0 Fluorometer (Thermo
614 Fisher Scientific) according to the manufacturer's protocol. Samples were pooled
615 and sequenced using an iSeq 100 Sequencing System (Illumina) with 2 x 150 bp
616 paired-end reads according to the manufacturer's protocol.

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

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

633 **SEARCH-MaP of long-range base pairs in multiple**
634 **coronaviruses**

635 **Computational screen for long-range base pairs in**
636 **coronaviruses**

637 All coronaviruses with reference genomes in the NCBI Reference Sequence
638 Database [56] as of December 2021 were searched for using the following query:

639 refseq[filter] AND ("Alphacoronavirus"[Organism] OR
640 "Betacoronavirus"[Organism] OR
641 "Gammacoronavirus"[Organism] OR
642 "Deltacoronavirus"[Organism])

643 The reference sequences ([https://github.com/rouskinlab/
644 search-map/tree/main/Compute/covs-screen/cov_refseq.fasta](https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/cov_refseq.fasta))
645 and table of features ([https://github.com/rouskinlab/search-map/
646 tree/main/Compute/covs-screen/cov_features.txt](https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/cov_features.txt)) were
647 downloaded and used to locate the slippery site in each genome using a custom
648 Python script ([https://github.com/rouskinlab/search-map/
649 tree/main/Compute/covs-screen/extract_long_fse.py](https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/extract_long_fse.py)). For
650 each genome, up to 100 secondary structure models of the 2,000 nt
651 segment from 100 nt upstream to 1,893 nt downstream of the slippery
652 site (excluding genomes with ambiguous nucleotides in this segment)
653 were generated using Fold v6.3 from RNAstructure [44] via the script
654 [https://github.com/rouskinlab/search-map/tree/main/Compute/
655 covs-screen/fold_long_fse.py](https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/fold_long_fse.py). The fraction of models in which each
656 base paired with any other base between positions 101 and 250 was calculated
657 using the script [https://github.com/rouskinlab/search-map/tree/
658 main/Compute/covs-screen/analyze_interactions.py](https://github.com/rouskinlab/search-map/tree/main/Compute/covs-screen/analyze_interactions.py). Using these
659 fractions, coronaviruses were clustered via the unweighted pair group method
660 with arithmetic mean (UPGMA) and a euclidean distance metric, implemented in
661 Seaborn v0.11 [85] and SciPy v1.7 [84] (Supplementary Figure 5). From each

662 cluster with prominent potential long-range interactions involving the FSE,
663 coronaviruses were manually selected for experimental study.

664 **Synthesis of 239 and 1,799 nt coronaviral RNAs**

665 For each selected coronavirus, the 1,799 nt segment from 290 to 1,502 nt down-
666 stream of the slippery site was ordered from Twist Bioscience as a gene fragment
667 flanked by the standard 5' and 3' adapters CAATCCGCCCTCACTACAACCG and
668 CTACTCTGGCGTCGATGAGGGA, respectively. Gene fragments were resus-
669 pended to 10 ng/ μ l in 10 mM Tris-HCl pH 8 (Invitrogen). Each DNA template
670 for transcription of 1,799 nt RNA segments, including a T7 promoter, was am-
671 plified from 0.5 μ l (5 ng) of a gene fragment in 20 μ l using 2X CloneAmp HiFi
672 PCR Premix (Takara Bio) with 250 nM of each primer TAATACGACTCACTATAG-
673 GCAATCCGCCCTCACTACAACCG and TCCCTCATCGACGCCAGAGTAG; ini-
674 tial denaturation at 98°C for 30 s; 30 cycles of 98°C for 10 s, X°C (see Supple-
675 mentary Table 4) for 10 s, and 72°C for 15 s; and final extension at 72°C for
676 60 s. DNA templates for transcription of 239 nt RNA segments were amplified
677 using the same procedure but with the forward primers with T7 promoters (F+T7)
678 and reverse primers (R) in Supplementary Table 5. For experiments in which
679 the RNAs were transcribed as a pool of all coronaviruses, all PCR products of
680 the same length (i.e. 239 or 1,799 nt) were pooled, then purified using a DNA
681 Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's pro-
682 tocol; concentrations were measured with a NanoDrop (Thermo Fisher Scientific).
683 Otherwise, PCR products were purified individually.

684 50 ng of DNA template was transcribed using a HiScribe T7 High Yield RNA
685 Synthesis Kit (New England Biolabs) according to the manufacturer's protocol
686 but at one-quarter volume (5 μ l), supplemented with 0.25 μ l RNaseOUT (Invitro-
687 gen), for 16 hr. DNA template was degraded by incubating with 0.5 μ l of TURBO
688 DNase (Invitrogen) at 37°C for 30 min. RNA was purified using an RNA Clean &
689 Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol,

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

692 DMS treatment of 239 and 1,799 nt coronaviral RNAs

693 Antisense oligonucleotides (ASOs) in Supplementary Table 6 were ordered from
694 Integrated DNA Technologies and resuspended to 100 μ M in low-EDTA TE
695 buffer: 10 mM Tris pH 7.4 with 0.1 mM EDTA (Integrated DNA Technologies).
696 For each coronavirus, 5 μ l of each corresponding ASO (Supplementary Table 6)
697 was pooled; the pool of ASOs was diluted with low-EDTA TE buffer to a final
698 volume of 100 μ l, bringing each ASO to 5 μ M. 1X refolding buffer comprising
699 300 mM sodium cacodylate pH 7.2 (Electron Microscopy Sciences) and 6 mM
700 magnesium chloride (Invitrogen) was assembled, then pre-warmed to 37°C.

701 For already-pooled RNA, 300 ng was diluted in 2.5 μ l of nuclease-free water
702 (Fisher Bioreagents) in a PCR tube, heated to 95°C for 1 min to denature, chilled
703 on ice for 3 min, added to 95 μ l of pre-warmed refolding buffer, and incubated at
704 37°C for 20 min to refold. For individually transcribed RNA, 1 pmol was mixed
705 with 10 μ l of either low-EDTA TE buffer (for probing without ASOs) or the ASO
706 pool for the corresponding coronavirus (for probing with ASOs) in a PCR tube,
707 heated to 95°C for 1 min to denature the RNA, chilled on ice for 3 min, added to
708 pre-warmed refolding buffer for a total volume of 100 μ l, and incubated at 37°C for
709 20 min to refold the RNA (possibly with ASOs). Subsequently, equimolar amounts
710 of all refolded RNAs were combined into one 97 μ l pool in a 1.5 ml tube.

711 RNA was treated with DMS (Sigma-Aldrich) – 2.5 μ l (260 mM) for RNAs
712 transcribed as pools or 3 μ l (320 mM) for RNAs pooled after transcription – in
713 100 μ l while shaking at 800 rpm in a ThermoMixer C (Eppendorf) at 37°C for
714 5 min. To quench, 60 μ l of beta-mercaptoethanol (Sigma-Aldrich) was added
715 and mixed thoroughly. DMS-modified RNA was purified using an RNA Clean &
716 Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol,
717 eluted in 16 μ l of nuclease-free water (Fisher Bioreagents), and measured with a
718 NanoDrop (Thermo Fisher Scientific). If added, ASOs were then degraded in

719 50 μ l containing 1X TURBO DNase Buffer (Invitrogen) and 1 μ l of TURBO
720 DNase Enzyme (Invitrogen) at 37°C for 30 min; RNA was purified with an RNA
721 Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's
722 protocol, eluted in 16 μ l of nuclease-free water (Fisher Bioreagents), and
723 measured with a NanoDrop (Thermo Fisher Scientific).

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

726 100 ng of DMS-modified RNA was prepared for sequencing using the xGen Broad-
727 Range RNA Library Preparation Kit (Integrated DNA Technologies) according to
728 the manufacturer's protocol, with the following modifications. During fragmenta-
729 tion, 8 μ l of RNA was combined with 1 μ l of Reagent F1, 4 μ l of Reagent F3, and
730 2 μ l of Reagent F2. For reverse transcription, 1 μ l of Enzyme R1, 2 μ l of TGIRT-III
731 enzyme (InGex), and 1 μ l of 100 mM dithiothreitol (Invitrogen) was used instead
732 of the reaction mix, then incubated at room temperature for 30 minutes before
733 adding 2 μ l of Reagent F2. Reverse transcription was stopped by adding 1 μ l
734 of 4 M sodium hydroxide (Fluka), heating to 95°C for 3 min, chilling at 4°C, then
735 neutralizing with 1 μ l of 4 M hydrochloric acid. Instead of a bead cleanup after the
736 final PCR, unpurified PCR products with 6X DNA loading dye (Invitrogen) were
737 elecrophoresed through an 8% polyacrylamide Tris-borate-EDTA (TBE) gel (Invit-
738 rogen) at 180 V for 55 min. The gel was stained with SYBR Gold (Invitrogen); the
739 section between 250 and 500 bp was excised and placed in a 0.5 ml tube with a
740 hole punctured in the bottom by an 18-gauge needle (BD Biosciences), which was
741 nested inside a 1.5 ml tube and centrifuged at 21,300 x g for 1 min to crush the
742 gel slice into the latter. Crushed gel pieces were suspended in 500 μ l of 300 mM
743 sodium chloride (Boston Bioproducts), shaken at 1,500 rpm in a ThermoMixer C
744 (Eppendorf) at 70°C for 20 min, and centrifuged at 21,300 x g through a 0.22 μ m
745 Costar Spin-X filter column to remove the gel pieces. Filtrate was mixed with
746 600 μ l isopropanol (Sigma-Aldrich) and 3 μ l GlycoBlue Coprecipitant (Invitrogen),
747 vortexed briefly, and stored at -20°C overnight. DNA was then pelleted by cen-

748 trifugation at 4°C at 18,200 x g for 45 min. The supernatant was aspirated, and
749 the pellet was washed with 1 ml of ice-cold 70% ethanol (Sigma-Aldritch), resus-
750 pended in 15 µl nuclease-free water (Fisher Bioreagents), and quantified using
751 the 1X dsDNA High Sensitivity Assay Kit for the Qubit 3.0 Fluorometer (Thermo
752 Fisher Scientific) according to the manufacturer's protocol. Samples were pooled
753 and sequenced using an iSeq 100 Sequencing System (Illumina) with 2 x 150 bp
754 paired-end reads according to the manufacturer's protocol.

755 **Data analysis of 239 and 1,799 nt coronaviral RNAs**

756 Sequencing data were processed with SEISMIC-RNA v0.11 and v0.12 to compute
757 mutation rates, correlations between samples, and secondary structure models
758 using the commands in the shell script [https://github.com/rouskinlab/
759 search-map/tree/main/Compute/covs-1799/run.sh](https://github.com/rouskinlab/search-map/tree/main/Compute/covs-1799/run.sh). For the 239 and
760 1,799 nt RNAs that had been pooled during transcription, the two replicates for
761 each coronavirus for each length were confirmed to give similar results, then
762 merged before comparing the 239 and 1,799 nt RNAs to each other. For the
763 comparison of RNAs with and without ASOs, the no-ASO samples that had been
764 transcribed individually were confirmed to give similar results to those transcribed
765 as a pool; then, all no-ASO samples were pooled before comparing to samples
766 with ASOs. For each coronavirus, the DMS reactivities of the combined no-ASO
767 samples were used to model up to 20 secondary structures of the 1,799 nt seg-
768 ment using Fold from RNAstructure v6.3 [44]. Structure models were checked
769 manually for correspondence with the rolling correlation between the +ASO and
770 no-ASO conditions; the minimum free energy structure was chosen for every coro-
771 navirus except for transmissible gastroenteritis virus, in which the first sub-optimal
772 structure – but not the minimum free energy structure – contained long-range base
773 pairs supported by the rolling correlation. Rolling correlations between +ASO
774 and no-ASO conditions superimposed on secondary structure models (Figure 4)
775 were graphed using the Python script [https://github.com/rouskinlab/
776 search-map/tree/main/Compute/util/pairs_vs_correl.py](https://github.com/rouskinlab/search-map/tree/main/Compute/util/pairs_vs_correl.py).

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

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

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

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

793 1.15X refolding buffer comprising 345 mM sodium cacodylate pH 7.2 (Electron
794 Microscopy Sciences) and 7 mM magnesium chloride (Invitrogen) was assembled
795 and pre-warmed to 37°C. 1 pmol of RNA was mixed with 100 pmol of each ASO
796 (Integrated DNA Technologies, Supplementary Table 3) in 10 µl total, heated to
797 95°C for 60 s to denature, chilled on ice for 5-10 min, and added to 87.1 µl of
798 pre-warmed refolding buffer. If no ASO would be added during refolding, then 1 µl
799 of nuclease-free water (Fisher Bioreagents) was added. RNA was incubated at
800 37°C for 15-20 min to refold. If an ASO would be added during refolding, then
801 100 pmol (1 µl) of ASO was added. RNA was incubated for another 15 min to
802 allow any newly added ASOs to bind.

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

805 To quench, 20 μ l of beta-mercaptoethanol (Sigma-Aldrich) was added and mixed
806 thoroughly. DMS-modified RNA was purified using an RNA Clean & Concentrator-
807 5 kit (Zymo Research) according to the manufacturer's protocol, eluted in 15 μ l of
808 nuclease-free water (Fisher Bioreagents), and measured with a NanoDrop One
809 (Thermo Fisher Scientific).

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

811 1 μ l of DMS-modified RNA was reverse transcribed in 20 μ l using Induro Reverse
812 Transcriptase (New England Biolabs) according to the manufacturer's protocol
813 with 500 nM of primer CTTCGTCCTTTCTTGGAAAGCGACA (Integrated DNA
814 Technologies) at 57°C for 30 min, followed by inactivation at 95°C for 1 min. 1 μ l
815 of unpurified RT product was amplified in 20 μ l using Q5 High-Fidelity 2X Mas-
816 ter Mix (New England Biolabs) with 500 nM of each primer CCCTGTGGGTTT-
817 TACACTAAAAAC and CTTCGTCCTTTCTTGGAAAGCGACA (Integrated DNA
818 Technologies); initial denaturation at 98°C for 30 s; 30 cycles of 98°C for 10 s,
819 65°C for 20 s, and 72°C for 20 s; and final extension at 72°C for 120 s. The PCR
820 product was purified using a DNA Clean & Concentrator-5 kit (Zymo Research)
821 according to the manufacturer's protocol, eluted in 20 μ l of 10 mM Tris-HCl pH 8
822 (Invitrogen), and measured with a NanoDrop One (Thermo Fisher Scientific).

823 50-100 ng of purified PCR product was prepared for sequencing using the
824 NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) accord-
825 ing to the manufacturer's protocol with the following modifications. All steps were
826 performed at half of the volume specified in the protocol, including reactions, bead
827 cleanups, and washes. During size selection after adapter ligation, 14 μ l and 7 μ l
828 of SPRIselect Beads (Beckman Coulter) were used in the first and second steps,
829 respectively, to select inserts of 283 bp. Indexing PCR was run with 400 nM of
830 each primer for 4 cycles. After indexing, PCR products were pooled in pairs; in
831 lieu of the final bead cleanup, 405 bp products were selected using a 2% E-Gel
832 SizeSelect II Agarose Gel (Invitrogen) according to the manufacturer's protocol.
833 DNA concentrations were measured using a Qubit 4 Fluorometer (Thermo Fisher

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

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

838 Sequencing data were processed with SEISMIC-RNA v0.11 and v0.12 to
839 compute mutation rates, clusters, and correlations between samples using the
840 commands in the shell script <https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/run.sh>.

842 Heatmaps of the reproducibility of clustering between replicates
843 (Supplementary Figure 4) were generated using the Python script
844 <https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/compare-clusters.py>. After the two replicates were
846 confirmed to give similar clusters, they were pooled for subsequent analyses.

847 Secondary structures with rolling correlations (Figure 3b) were drawn using the
848 Python script <https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/draw-structure.py>. Alternative

849 structure models (Figure 3c) were selected and created with the help of the
850 Python scripts <https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/choose-model-parts.py>

853 and <https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/make-models.py>. Heatmaps of areas
854 under the curve (Figure 3d) were generated using the Python script
856 <https://github.com/rouskinlab/search-map/tree/main/Compute/sars2-1799/atlas-plot.py>.

858 **DMS-MaPseq of transmissible gastroenteritis virus
859 in ST cells**

860 **Cells and Viruses**

861 Transmissible gastroenteritis virus (TGEV, TC-adapted Miller strain, ATCC VR-
862 1740) and ST cells (ATCC CRL-1746) were ordered from American Type Cul-
863 ture Collection (ATCC). ST cells were maintained in Eagle's Minimum Essential
864 Medium (EMEM, Gibco) supplemented with 10% fetal bovine serum (Gibco), 1%
865 sodium pyruvate (Gibco), and 1% Pen Strep (Gibco) at 37°C with 5% carbon
866 dioxide. For TGEV, the infection medium (IM) comprised EMEM (Gibgo) supple-
867 mented with 10% fetal bovine serum (Gibco), 1% sodium pyruvate, and 1 µg/µl of
868 TPCK trypsin (Thermo Fisher Scientific).

869 **Production and titering of TGEV**

870 A 150 mm dish was seeded with 1×10^7 ST cells, grown overnight, and washed
871 twice with phosphate-buffered saline (PBS, Gibco). Cells were inoculated with
872 8 ml of TGEV in IM at a multiplicity of infection (MOI) of 0.1, which was kept on
873 for 60 min with rocking every 15 min. The inoculum was removed, cells were
874 washed twice with PBS, and 26 ml of IM was added. Cells were checked daily
875 for cytopathic effects (CPE) and were harvested after 5 days upon development
876 of significant (80%) CPE.

877 Harvested TGEV was titrated via tissue culture infectious dose (TCID₅₀).
878 Briefly, ST cells were seeded in a poly-L-lysine coated 96-well plate at 4×10^4
879 cells per well and grown overnight. TGEV was thawed on ice and serially diluted
880 in a 12-well plate from 10₋₁ to 10₋₁₀ in IM. Cells were washed once with PBS, and
881 each well was inoculated with one serial dilution of TGEV (8 replicates per
882 dilution level). The plate was wrapped in parafilm and incubated until CPE
883 appeared. Then, media was aspirated and cells were fixed with 4%
884 paraformaldehyde for 30 min and decanted. 0.5% crystal violet was then added
885 to each well; the plate was rocked for 10 min, submerged in water to remove

886 excess crystal violet, and dried. Wells with CPE were counted and the titer
887 determined using the Spearman-Kärber method.

888 **TGEV infection and DMS treatment**

889 Four 150 mm dishes were each seeded with 1×10^7 ST cells, grown overnight,
890 and washed twice with phosphate-buffered saline. Cells were inoculated with 8 ml
891 of TGEV in IM at a multiplicity of infection (MOI) of 2, which was kept on for 60 min
892 with rocking every 15 min. The inoculum was removed, cells were washed twice
893 with PBS, and 26 ml of IM was added.

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

901 **RNA purification**

902 200 μ l of chloroform was added to each 1 ml technical replicate, vortexed for 20 s,
903 and rested until the phases separated. Samples were then spun at 18,200 x g
904 for 15 min at 4°C; the aqueous phase transferred to a new tube and mixed with
905 an equal volume of 100% ethanol. RNA was purified using a 50 μ g Monarch
906 RNA Cleanup Column (New England Biolabs), eluted in nuclease-free water, and
907 quantified with a NanoDrop.

908 To remove rRNA, 10 μ g of total RNA was diluted in 6 μ l of nuclease-free water
909 and mixed with 1 μ l of anti-rRNA ASOs (Integrated DNA Technologies) and 3 μ l
910 of HYBE buffer (200 mM sodium chloride, 100 mM Tris-HCl pH 7.5). The mixture
911 was incubated at 95°C for 2 min and cooled by 0.1°C/s until reaching 45°C. A
912 preheated mixture of 10 μ l of RNase H and 2 μ l of RNase H Buffer was added

913 and incubated at 45°C for 30 min. RNA was purified using a 10 µg Monarch
914 RNA Cleanup Column (New England Biolabs) and eluted in 42 µl of nuclease-
915 free water.

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

920 **Library generation for the full TGEV genome**

921 RNA was prepared for sequencing using the xGen Broad-Range RNA Library
922 Preparation Kit (Integrated DNA Technologies) according to the manufacturer's
923 protocol, with the same modifications as described above (5.3.4), notably the
924 substitution of TGIRT-III (InGex) for the kit reverse transcriptase. Samples were
925 pooled and sequenced using a NextSeq 1000 Sequencing System (Illumina) with
926 2 x 150 bp paired-end reads according to the manufacturer's protocol.

927 **Library generation for amplicons**

928 1 µl of rRNA-depleted, DNased RNA was reverse transcribed
929 in 20 µl using Induro Reverse Transcriptase (New England Biolabs)
930 according to the manufacturer's protocol with 500 nM of primer
931 ACAATTCGTCTTAAGGAATTACCAATACACGCAA (Integrated DNA
932 Technologies) at 57°C for 30 min, followed by inactivation at 95°C for
933 1 min. 1 µl of unpurified RT product was amplified in 10 µl using Q5
934 High-Fidelity 2X Master Mix (New England Biolabs) with 1 µM of each
935 primer, either GCCGCTACAAAGGTAAGTCGTGCAAATACCAACT
936 and ACAATTCGTCTTAAGGAATTACCAATACACGCAA or
937 GTGAAAAGTGACATCTATGGTTCTGATTATAAGCAGTA and
938 CTATACCAAGTTGTTGAAATGGTAACCTGCAGTAACA (Integrated DNA
939 Technologies); initial denaturation at 98°C for 30 s; 30 cycles of 98°C for 5 s,

940 69°C for 20 s, and 72°C for 15 s; and final extension at 72°C for 120 s.
941 Amplification was confirmed by electrophoresing 1 μ l of each PCR product. PCR
942 products for both pairs of primers were pooled and then purified using a DNA
943 Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's
944 protocol, eluted in 18 μ l of 10 mM Tris-HCl pH 8 (Invitrogen), and measured with
945 a NanoDrop (Thermo Fisher Scientific).

946 175-225 ng of purified PCR product was prepared for sequencing using the
947 NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) accord-
948 ing to the manufacturer's protocol with the following modifications. All steps were
949 performed at half of the volume specified in the protocol, including reactions, bead
950 cleanups, and washes. During size selection after adapter ligation, 14 μ l and 7 μ l
951 of SPRIselect Beads (Beckman Coulter) were used in the first and second steps,
952 respectively, to select inserts of 295 bp. Indexing PCR was run with 400 nM of
953 each primer for 4 cycles. In lieu of the final bead cleanup, 415 bp products were
954 selected using a 2% E-Gel SizeSelect II Agarose Gel (Invitrogen) according to
955 the manufacturer's protocol. DNA concentrations were measured using a Qubit 4
956 Fluorometer (Thermo Fisher Scientific) according to the manufacturer's protocol.
957 Samples were pooled and sequenced using a NextSeq 1000 Sequencing Sys-
958 tem (Illumina) with 2 x 150 bp paired-end reads according to the manufacturer's
959 protocol.

960 **Data analysis of transmissible gastroenteritis virus in ST cells**

961 The genomic sequence of this TGEV strain was determined using the script
962 <https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/consensus.sh>: reads from the untreated sample were aligned
963 to the TGEV reference genome (NC_038861.1) using Bowtie 2 [41] and the
964 consensus sequence was determined using Samtools [42]. All reads were
965 processed with SEISMIC-RNA v0.15 to compute mutation rates, correlations
966 between samples, and secondary structure models using the commands in the
967 shell script <https://github.com/rouskinlab/search-map/tree/>

969 main/Compute/tgev-virus/run.sh. Positions in the untreated sample with
970 mutation rates greater than 0.01 were masked. Replicates were checked for
971 reproducibility and pooled for clustering and structure modeling. A model of
972 short-range base pairs (maximum distance 300 nt) in the TGEV genome was
973 generated from the DMS reactivities using Fold-smp from RNAstructure [44] in
974 five overlapping 10 kb segments, which were merged using the script
975 <https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/assemble-tgev-ss.py>. Rolling area under the curve
977 superimposed on secondary structure models in Figure 5d was graphed using
978 the script <https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/make-figure-6d.py>, and in Supplementary
979 Figure 9 using the script https://github.com/rouskinlab/search-map/tree/main/Compute/tgev-virus/plot_genome.py.
980

Acknowledgements

982 We thank Miriam L. Rittenberg for assistance with experiments. This research
983 was supported by National Institute of Allergy and Infectious Diseases grant DP2
984 AI175475 (S.R.); National Science Foundation Graduate Research Fellowship
985 Program grants 1745302 (M.F.A.), 2140743 (J.A.), and 2141064 (M.F.A.); and
986 National Institutes of Health grant T32GM145407 (J.A.).
987

Author Contributions

988 S.R. and M.F.A. conceived the project. M.F.A. performed the experiments with
989 SARS-CoV-2 segments. J.A. performed the experiments with other coronavirus
990 segments. J.P. performed the experiments with TGEV-infected ST cells. M.F.A.
991 wrote SEISMIC-RNA with contributions from S.G., Y.J.M.T., A.L., and J.A. M.F.A.
992 and J.A. analyzed the data. M.F.A. drafted the manuscript. All authors reviewed
993 the manuscript and provided comments.
994

Ethics Declarations

995 The authors declare no competing interests.

Data Availability

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

Code Availability

1000 Documentation for SEISMIC-RNA, including instructions for installation, is hosted
1001 on GitHub Pages: <https://rouskinlab.github.io/seismic-rna>.
1002

1003 Source code for SEISMIC-RNA is available from GitHub:
1004 <https://github.com/rouskinlab/seismic-rna>. Shell scripts for
1005 running SEISMIC-RNA, auxiliary scripts for data analysis, supplementary files,
1006 and LaTeX source code for this manuscript are also available from GitHub:
1007 <https://github.com/rouskinlab/search-map>.

1008 References

- 1009 [1] Carla A. Klattenhoff, Johanna C. Scheuermann, Lauren E. Surface, Robert K.
1010 Bradley, Paul A. Fields, Matthew L. Steinhauser, Huiming Ding, Vincent L.
1011 Butty, Lillian Torrey, Simon Haas, Ryan Abo, Mohammadsharif Tabebordbar,
1012 Richard T. Lee, Christopher B. Burge, and Laurie A. Boyer. Braveheart, a
1013 long noncoding RNA required for cardiovascular lineage commitment. *Cell*,
1014 152:570–583, 2013.
- 1015 [2] Blake Wiedenheft, Samuel H. Sternberg, and Jennifer A. Doudna. RNA-
1016 guided genetic silencing systems in bacteria and archaea. *Nature*,
1017 482(7385):331–338, 2012.
- 1018 [3] Harry F Noller. Evolution of protein synthesis from an RNA world. *Cold Spring
1019 Harb Perspect Biol*, 4(4):a003681, Apr 2012.
- 1020 [4] Jens Kortmann and Franz Narberhaus. Bacterial RNA thermometers: molec-
1021 ular zippers and switches. *Nature Reviews Microbiology*, 10(4):255–265,
1022 2012.
- 1023 [5] Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*,
1024 152:17–24, 2013.
- 1025 [6] Arunoday Bhan and Subhrangsu S. Mandal. LncRNA HOTAIR: A master
1026 regulator of chromatin dynamics and cancer. *Biochimica et Biophysica Acta
(BBA) - Reviews on Cancer*, 1856(1):151–164, 2015.
- 1027 [7] Mohammadreza Hajjari and Adrian Salavaty. HOTAIR: an oncogenic long
1028 non-coding RNA in different cancers. *Cancer Biol Med*, 12(1):1–9, Mar 2015.
- 1029 [8] Mark E. J. Woolhouse and Liam Brierley. Epidemiological characteristics of
1030 human-infective RNA viruses. *Scientific Data*, 5(1):180017, 2018.
- 1031 [9] Nicole M. Bouvier and Peter Palese. The biology of influenza viruses. *Vac-
1032 cine*, 26:D49–D53, 2008. Influenza Vaccines: Research, Development and
1033 Public Health Challenges.
- 1034 [10] Dong Yang and Julian L. Leibowitz. The structure and functions of coro-
1035 navirus genomic 3' and 5' ends. *Virus Research*, 206:120–133, 2015.
- 1036 [11] Stefanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. Insights
1037 into RNA structure and function from genome-wide studies. *Nature Reviews
1038 Genetics*, 15(7):469–479, 2014.
- 1039 [12] Kalli Kappel, Kaiming Zhang, Zhaoming Su, Andrew M. Watkins, Wipapat
1040 Kladwang, Shanshan Li, Grigore Pintilie, Ved V. Topkar, Ramya Rangan,
1041 Ivan N. Zheludev, Joseph D. Yesselman, Wah Chiu, and Rhiju Das. Accel-
1042 erated cryo-EM-guided determination of three-dimensional RNA-only struc-
1043 tures. *Nature Methods*, 17:699–707, 2020.
- 1044 [13] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat,
1045 Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data
1046 Bank. *Nucleic Acids Research*, 28:235–242, 1 2000.
- 1047

- 1048 [14] Bohdan Schneider, Blake Alexander Sweeney, Alex Bateman, Jiri Cerny,
1049 Tomasz Zok, and Marta Szachniuk. When will RNA get its AlphaFold mo-
1050 ment? *Nucleic Acids Research*, 51(18):9522–9532, 09 2023.
- 1051 [15] Jamie J Cannone, Sankar Subramanian, Murray N Schnare, James R Col-
1052 lett, Lisa M D’Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V Madabusi,
1053 Kirsten M Müller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin R Gutell.
1054 The comparative RNA web (CRW) site: an online database of comparative
1055 sequence and structure information for ribosomal, intron, and other RNAs.
1056 *BMC Bioinformatics*, 3:2, 2002.
- 1057 [16] Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance
1058 models. *Nucleic Acids Research*, 22(11):2079–2088, 06 1994.
- 1059 [17] Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasins-
1060 ska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-
1061 Nioche, Daniel Gautheret, Zasha Weinberg, Elena Rivas, Sean R Eddy,
1062 Robert D Finn, Alex Bateman, and Anton I Petrov. Rfam 14: expanded cover-
1063 age of metagenomic, viral and microRNA families. *Nucleic Acids Research*,
1064 49(D1):D192–D200, 11 2020.
- 1065 [18] Anthony M. Mustoe, Charles L. Brooks, and Hashim M. Al-Hashimi. Hierar-
1066 chy of RNA functional dynamics. *Annual Review of Biochemistry*, 83(1):441–
1067 466, 2014. PMID: 24606137.
- 1068 [19] Robert C. Spitale and Danny Incarnato. Probing the dynamic RNA structur-
1069 ome and its functions. *Nature Reviews Genetics*, 24(3):178–196, 2023.
- 1070 [20] Jeffrey J. Quinn and Howard Y. Chang. Unique features of long non-coding
1071 RNA biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016.
- 1072 [21] Sita J. Lange, Daniel Maticzka, Mathias Mohl, Joshua N. Gagnon, Chris M.
1073 Brown, and Rolf Backofen. Global or local? predicting secondary structure
1074 and accessibility in mRNAs. *Nucleic Acids Research*, 2012.
- 1075 [22] Beth L Nicholson and K Andrew White. Exploring the architecture of viral RNA
1076 genomes. *Current Opinion in Virology*, 12:66–74, 2015. Antiviral strategies
1077 • Virus structure and expression.
- 1078 [23] Christoph Flamm, Julia Wielach, Michael T. Wolfinger, Stefan Badelt, Ronny
1079 Lorenz, and Ivo L. Hofacker. Caveats to deep learning approaches to RNA
1080 secondary structure prediction. *Frontiers in Bioinformatics*, 2, 2022.
- 1081 [24] Kengo Sato and Michiaki Hamada. Recent trends in RNA informatics: a re-
1082 view of machine learning and deep learning for RNA secondary structure pre-
1083 diction and RNA drug discovery. *Briefings in Bioinformatics*, 24(4):bbad186,
1084 05 2023.
- 1085 [25] David H. Mathews. How to benchmark RNA secondary structure prediction
1086 accuracy. *Methods*, 162-163:60–67, 2019. Experimental and Computational
1087 Techniques for Studying Structural Dynamics and Function of RNA.

- 1088 [26] Kishore J. Doshi, Jamie J. Cannone, Christian W. Cobaugh, and Robin R.
1089 Gutell. Evaluation of the suitability of free-energy minimization using nearest-
1090 neighbor energy parameters for RNA secondary structure prediction. *BMC
1091 Bioinformatics*, 5(1):105, 2004.
- 1092 [27] Miles Kubota, Catherine Tran, and Robert C Spitale. Progress and chal-
1093 lenges for chemical probing of RNA structure inside living cells. *Nature
1094 Chemical Biology*, 11(12):933–941, 2015.
- 1095 [28] Nathan A. Siegfried, Steven Busan, Greggory M. Rice, Julie A.E. Nelson, and
1096 Kevin M. Weeks. RNA motif discovery by SHAPE and mutational profiling
1097 (SHAPE-MaP). *Nature methods*, 2014.
- 1098 [29] Meghan Zubradt, Paromita Gupta, Sitara Persad, Alan M. Lambowitz,
1099 Jonathan S. Weissman, and Silvi Rouskin. DMS-MaPseq for genome-wide
1100 or targeted RNA structure probing in vivo. *Nature Methods*, 2254:219–238,
1101 2016.
- 1102 [30] Phillip J. Tomezsko, Vincent D.A. Corbin, Paromita Gupta, Harish Swami-
1103 nathan, Margalit Glasgow, Sitara Persad, Matthew D. Edwards, Lachlan
1104 McIntosh, Anthony T. Papenfuss, Ann Emery, Ronald Swanstrom, Trinity
1105 Zang, Tammy C.T. Lan, Paul Bieniasz, Daniel R. Kuritzkes, Athe Tsibris, and
1106 Silvi Rouskin. Determination of RNA structural diversity and its role in HIV-1
1107 RNA splicing. *Nature*, 582:438–442, 2020.
- 1108 [31] Edoardo Morandi, Ilaria Manfredonia, Lisa M. Simon, Francesca Anselmi,
1109 Martijn J. van Hemert, Salvatore Oliviero, and Danny Incarnato. Genome-
1110 scale deconvolution of RNA structure ensembles. *Nature Methods*, 18:249–
1111 252, 2 2021.
- 1112 [32] David H. Mathews, Matthew D. Disney, Jessica L. Childs, Susan J.
1113 Schroeder, Michael Zuker, and Douglas H. Turner. Incorporating chemical
1114 modification constraints into a dynamic programming algorithm for predi-
1115 cition of RNA secondary structure. *Proceedings of the National Academy of
1116 Sciences*, 101:7287–7292, 5 2004.
- 1117 [33] Pablo Cordero, Wipapat Kladwang, Christopher C. Vanlang, and Rhiju Das.
1118 Quantitative dimethyl sulfate mapping for automated RNA secondary struc-
1119 ture inference. *Biochemistry*, 51:7037–7039, 9 2012.
- 1120 [34] Michael F. Sloma and David H. Mathews. Chapter four - improving RNA sec-
1121 ondary structure prediction with structure mapping data. In Shi-Jie Chen and
1122 Donald H. Burke-Aguero, editors, *Computational Methods for Under-
1123 standing Riboswitches*, volume 553 of *Methods in Enzymology*, pages 91–114.
1124 Academic Press, 2015.
- 1125 [35] Clarence Y. Cheng, Wipapat Kladwang, Joseph D. Yesselman, and Rhiju
1126 Das. RNA structure inference through chemical mapping after accidental or
1127 intentional mutations. *Proceedings of the National Academy of Sciences of
1128 the United States of America*, 114:9876–9881, 9 2017.
- 1129 [36] Pablo Cordero and Rhiju Das. Rich RNA structure landscapes revealed by
1130 mutate-and-map analysis. *PLOS Computational Biology*, 11:e1004473, 11
1131 2015.

- 1132 [37] Grzegorz Kudla, Yue Wan, and Aleksandra Helwak. RNA conformation cap-
1133 ture by proximity ligation. *Annual Review of Genomics and Human Genetics*,
1134 21(1):81–100, 2020. PMID: 32320281.
- 1135 [38] Jamie A. Kelly, Alexandra N. Olson, Krishna Neupane, Sneha Munshi, Jo-
1136 sue San Emeterio, Lois Pollack, Michael T. Woodside, and Jonathan D. Din-
1137 man. Structural and functional conservation of the programmed -1 ribosomal
1138 frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *Journal of Biological
1139 Chemistry*, 295:10741–10748, 7 2020.
- 1140 [39] Kaiming Zhang, Ivan N. Zheludev, Rachel J. Hagey, Raphael Haslecker,
1141 Yixuan J. Hou, Rachael Kretsch, Grigore D. Pintilie, Ramya Rangan, Wipa-
1142 pat Kladwang, Shanshan Li, Marie Teng Pei Wu, Edward A. Pham, Claire
1143 Bernardin-Souibgui, Ralph S. Baric, Timothy P. Sheahan, Victoria D’Souza,
1144 Jeffrey S. Glenn, Wah Chiu, and Rhiju Das. Cryo-EM and antisense target-
1145 ing of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA
1146 genome. *Nature Structural & Molecular Biology*, 28:747–754, 8 2021.
- 1147 [40] Marcel Martin. Cutadapt removes adapter sequences from high-throughput
1148 sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.
- 1149 [41] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with
1150 Bowtie 2. *Nature Methods*, 2012.
- 1151 [42] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer,
1152 Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence align-
1153 ment/map format and SAMtools. *Bioinformatics*, 2009.
- 1154 [43] Shalom Hillel Roth, Erez Y. Levanon, and Eli Eisenberg. Genome-wide quan-
1155 tification of adar adenosine-to-inosine rna editing activity. *Nature Methods*,
1156 16(11):1131–1138, 2019.
- 1157 [44] Jessica S. Reuter and David H. Mathews. RNAsstructure: software for RNA
1158 secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129,
1159 2010.
- 1160 [45] Chringma Sherpa, Jason W. Rausch, Stuart F.J. Le Grice, Marie Louise
1161 Hammarskjold, and David Rekosh. The HIV-1 Rev response element (RRE)
1162 adopts alternative conformations that promote different rates of virus replica-
1163 tion. *Nucleic Acids Research*, 43:4676–4686, 3 2015.
- 1164 [46] Beth L. Nicholson and K. Andrew White. Functional long-range RNA-RNA
1165 interactions in positive-strand RNA viruses. *Nature Reviews Microbiology*,
1166 12:493–504, 6 2014.
- 1167 [47] Ewan P. Plant and Jonathan D. Dinman. The role of programmed-1 ribosomal
1168 frameshifting in coronavirus propagation. *Frontiers in Bioscience*, 13:4873–
1169 4881, 2008.
- 1170 [48] Ian Brierley, Paul Digard, and Stephen C. Inglis. Characterization of an ef-
1171 ficient coronavirus ribosomal frameshifting signal: Requirement for an RNA
1172 pseudoknot. *Cell*, 1989.

- 1173 [49] J. Herald and S. G. Siddell. An 'elaborated' pseudoknot is required for high
1174 frequency frameshifting during translation of HCV 229E polymerase mRNA.
1175 *Nucleic Acids Research*, 21:5838–5842, 1993.
- 1176 [50] Ewan P. Plant, Gabriela C. Pérez-Alvarado, Jonathan L. Jacobs, Bani
1177 Mukhopadhyay, Mirko Hennig, and Jonathan D. Dinman. A three-stemmed
1178 mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biology*,
1179 3:e172, 2005.
- 1180 [51] Christina Roman, Anna Lewicka, Deepak Koirala, Nan-Sheng Li, and
1181 Joseph A. Piccirilli. The SARS-CoV-2 programmed -1 ribosomal frameshift-
1182 ing element crystal structure solved to 2.09 Å using chaperone-assisted RNA
1183 crystallography. *ACS Chemical Biology*, 16(8):1469–1481, 08 2021.
- 1184 [52] Christopher P. Jones and Adrian R. Ferré-D'Amaré. Crystal structure of the
1185 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) frameshift-
1186 ing pseudoknot. *RNA*, 28:239–249, 2022.
- 1187 [53] Tammy C.T. Lan, Matty F. Allan, Lauren E. Malsick, Jia Z. Woo, Chi Zhu, Fen-
1188 grui Zhang, Stuti Khandwala, Sherry S.Y. Nyeo, Yu Sun, Junjie U. Guo, Mark
1189 Bathe, Anders Näär, Anthony Griffiths, and Silvi Rouskin. Secondary struc-
1190 tural ensembles of the SARS-CoV-2 RNA genome in infected cells. *Nature
1191 Communications*, 13:1128, 3 2022.
- 1192 [54] Omer Ziv, Jonathan Price, Lyudmila Shalamova, Tsveta Kamenova, Ian
1193 Goodfellow, Friedemann Weber, and Eric A. Miska. The short- and long-
1194 range RNA-RNA interactome of SARS-CoV-2. *Molecular Cell*, 80:1067–
1195 1077.e5, 12 2020.
- 1196 [55] Mei Chi Su, Chung Te Chang, Chiu Hui Chu, Ching Hsiu Tsai, and Kung Yao
1197 Chang. An atypical RNA pseudoknot stimulator and an upstream attenuation
1198 signal for -1 ribosomal frameshifting of SARS coronavirus. *Nucleic Acids
1199 Research*, 33:4265–4275, 2005.
- 1200 [56] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciuffo, Di-
1201 ana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-
1202 White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao,
1203 Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric
1204 Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta,
1205 Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K.
1206 Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey,
1207 Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rang-
1208 wala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Su-
1209 san S. Storz, Hanzen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Ray-
1210 mond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu,
1211 Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul
1212 Kitts, Terence D. Murphy, Kim D. Pruitt, O'Leary NA, Wright MW, Brister JR,
1213 Ciuffo S, Haddad Haft D, McVeigh R, Robbertse Rajput B, Robbertse Rajput
1214 B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova
1215 O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Gold-
1216 farb Gupta T, Goldfarb Gupta T, Haddad Haft D, Hatcher E, Hlavina W, Joar-
1217 dar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR,
1218 O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda

- 1219 A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin
1220 C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts
1221 P, Murphy TD, and Pruitt KD. Reference sequence (RefSeq) database at
1222 NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44:D733–D745, 2016.
1223
- 1224 [57] Gary R. Whittaker, Nicole M. André, and Jean Kaoru Millet. Improving virus
1225 taxonomy by recontextualizing sequence-based classification with biologi-
1226 cally relevant data: the case of the Alphacoronavirus 1 species. *mSphere*,
1227 3(1):10.1128/mspheredirect.00463–17, 2018.
- 1228 [58] Qiang Liu and Huai-Yu Wang. Porcine enteric coronaviruses: an updated
1229 overview of the pathogenesis, prevalence, and diagnosis. *Veterinary Re-
1230 search Communications*, 45(2):75–86, 2021.
- 1231 [59] Michal Legiewicz, Andrei S. Zolotukhin, Guy R. Pilkington, Katarzyna J.
1232 Purzycka, Michelle Mitchell, Hiroaki Uranishi, Jenifer Bear, George N.
1233 Pavlakis, Stuart F. J. Le Grice, and Barbara K. Felber. The RNA transport el-
1234 ement of the murine musD retrotransposon requires long-range intramolecu-
1235 lar interactions for function. *Journal of Biological Chemistry*, 285(53):42097–
1236 42104, 2024/03/11 2010.
- 1237 [60] Eva J. Archer, Mark A. Simpson, Nicholas J. Watts, Rory O’Kane, Bangchen
1238 Wang, Dorothy A. Erie, Alex McPherson, and Kevin M. Weeks. Long-range
1239 architecture in a viral RNA genome. *Biochemistry*, 52(18):3182–3190, 2013.
1240 PMID: 23614526.
- 1241 [61] Yun Bai, Akshay Tambe, Kaihong Zhou, and Jennifer A Doudna. RNA-guided
1242 assembly of Rev-RRE nuclear export complexes. *eLife*, 3:e03656, aug 2014.
- 1243 [62] Jong Ghut Ashley Aw, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim,
1244 Kum Loong Boon, Sidika Tapsin, Yun Shen Chan, Cheng Peow Tan, Ade-
1245 lene Y.L. Sim, Tong Zhang, Teodorus Theo Susanto, Zhiyan Fu, Niranjan
1246 Nagarajan, and Yue Wan. In vivo mapping of eukaryotic RNA interactomes
1247 reveals principles of higher-order organization and regulation. *Molecular Cell*,
1248 62:603–617, 2016.
- 1249 [63] Zhipeng Lu, Qiangfeng Cliff Zhang, Byron Lee, Ryan A. Flynn, Martin A.
1250 Smith, James T. Robinson, Chen Davidovich, Anne R. Gooding, Karen J.
1251 Goodrich, John S. Mattick, Jill P. Mesirov, Thomas R. Cech, and Howard Y.
1252 Chang. RNA duplex map in living cells reveals higher-order transcriptome
1253 structure. *Cell*, 165:1267–1279, 2016.
- 1254 [64] Eesha Sharma, Tim Sterne-Weiler, Dave O’Hanlon, and Benjamin J.
1255 Blencowe. Global mapping of human RNA-RNA interactions. *Molecular Cell*,
1256 62:618–626, 2016.
- 1257 [65] Omer Ziv, Marta M. Gabryelska, Aaron T.L. Lun, Luca F.R. Gebert, Jes-
1258 sica Sheu-Gruttaduria, Luke W. Meredith, Zhong Yu Liu, Chun Kit Kwok,
1259 Cheng Feng Qin, Ian J. MacRae, Ian Goodfellow, John C. Marioni, Grzegorz
1260 Kudla, and Eric A. Miska. COMRADES determines in vivo RNA structures
1261 and interactions. *Nature Methods*, 15:785–788, 9 2018.

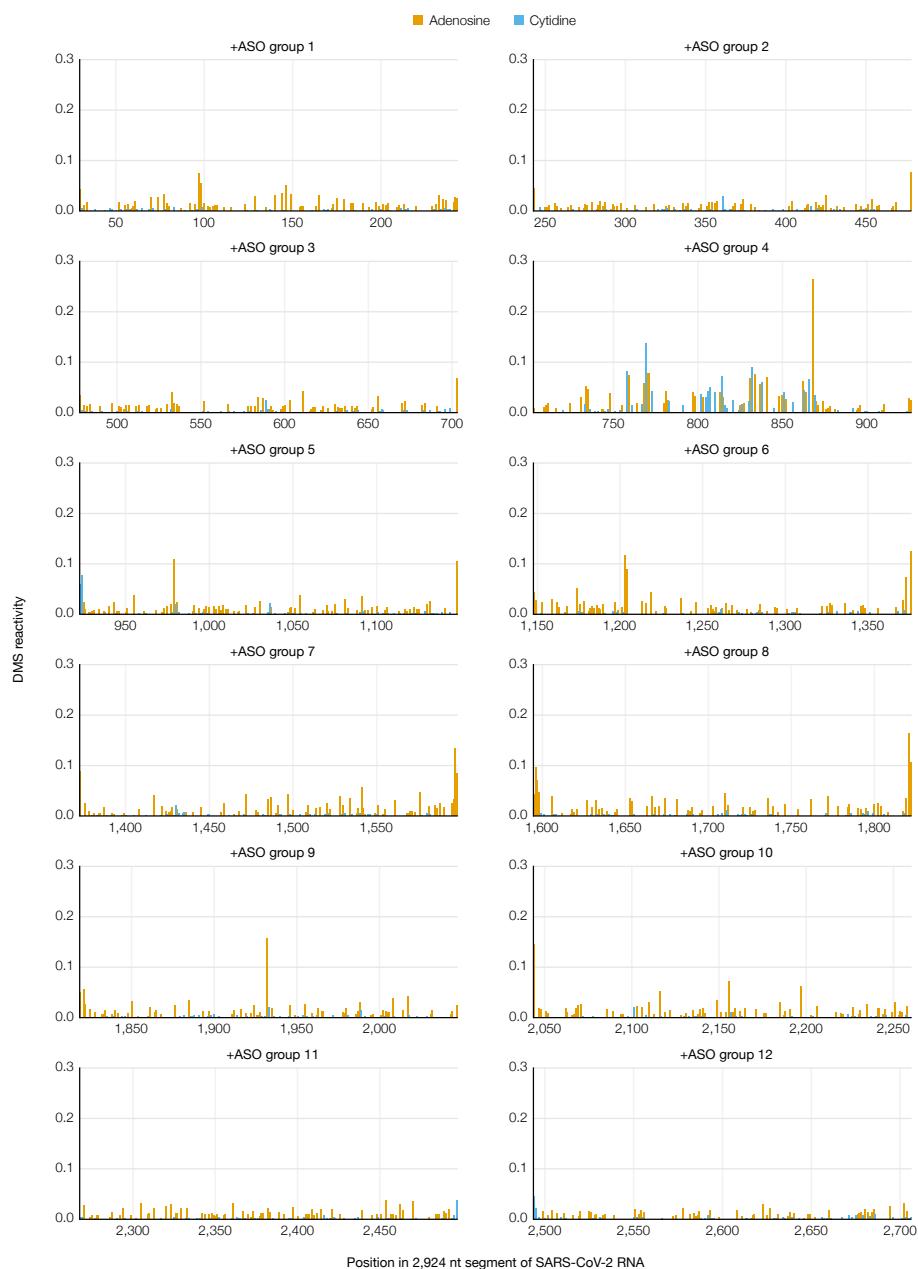
- 1262 [66] Ryan Van Damme, Kongpan Li, Minjie Zhang, Jianhui Bai, Wilson H. Lee,
1263 Joseph D. Yesselman, Zhipeng Lu, and Willem A. Velema. Chemical re-
1264 versible crosslinking enables measurement of RNA 3D distances and alter-
1265 native conformations in cells. *Nature Communications*, 13(1):911, 2022.
- 1266 [67] Jennifer K. Barry and W. Allen Miller. A -1 ribosomal frameshift element
1267 that requires base pairing across four kilobases suggests a mechanism of
1268 regulating ribosome and replicase traffic on a viral RNA. *Proceedings of the*
1269 *National Academy of Sciences of the United States of America*, 99:11133–
1270 11138, 8 2002.
- 1271 [68] Yuri Tajima, Hiro oki Iwakawa, Masanori Kaido, Kazuyuki Mise, and Tetsuro
1272 Okuno. A long-distance RNA-RNA interaction plays an important role in pro-
1273 grammed - 1 ribosomal frameshifting in the translation of p88 replicase pro-
1274 tein of Red clover necrotic mosaic virus. *Virology*, 417:169–178, 8 2011.
- 1275 [69] Anna A Mikkelsen, Feng Gao, Elizabeth Carino, Sayanta Bera, and Anne E
1276 Simon. -1 programmed ribosomal frameshifting in Class 2 umbravirus-like
1277 RNAs uses multiple long-distance interactions to shift between active and
1278 inactive structures and destabilize the frameshift stimulating element. *Nucleic*
1279 *Acids Research*, 51(19):10700–10718, 09 2023.
- 1280 [70] Hafeez S. Haniff, Yuquan Tong, Xiaohui Liu, Jonathan L. Chen, Blessy M.
1281 Suresh, Ryan J. Andrews, Jake M. Peterson, Collin A. O’Leary, Raphael I.
1282 Benhamou, Walter N. Moss, and Matthew D. Disney. Targeting the SARS-
1283 COV-2 RNA genome with small molecule binders and ribonuclease targeting
1284 chimera (RiboTAC) degraders. *ACS Central Science*, 6:1713–1721, 2020.
- 1285 [71] Pramod R. Bhatt, Alain Scaiola, Gary Loughran, Marc Leibundgut, Annika
1286 Kratzel, Romane Meurs, René Dreos, Kate M. O’Connor, Angus McMillan,
1287 Jeffrey W. Bode, Volker Thiel, David Gatfield, John F. Atkins, and Nenad Ban.
1288 Structural basis of ribosomal frameshifting during translation of the SARS-
1289 CoV-2 RNA genome. *Science*, 372:1306–1313, 5 2021.
- 1290 [72] Yu Sun, Laura Abriola, Rachel O. Niederer, Savannah F. Pedersen, Mia M.
1291 Alfajaro, Valter Silva Monteiro, Craig B. Wilen, Ya-Chi Ho, Wendy V. Gilbert,
1292 Yulia V. Surovtseva, Brett D. Lindenbach, and Junjie U. Guo. Restriction of
1293 SARS-CoV-2 replication by targeting programmed -1 ribosomal frameshif-
1294 ting. *Proceedings of the National Academy of Sciences of the United States*
1295 *of America*, 118:e2023051118, 6 2021.
- 1296 [73] Yaara Finkel, Orel Mizrahi, Aharon Nachshon, Shira Weingarten-Gabbay,
1297 David Morgenstern, Yfat Yahalom-Ronen, Hadas Tamir, Hagit Achdout,
1298 Dana Stein, Ofir Israeli, Adi Beth-Din, Sharon Melamed, Shay Weiss, Tomer
1299 Israeli, Nir Paran, Michal Schwartz, and Noam Stern-Ginossar. The coding
1300 capacity of SARS-CoV-2. *Nature*, 589:125–130, 1 2021.
- 1301 [74] Doyeon Kim, Sukjun Kim, Joori Park, Hee Ryung Chang, Jeeyoon Chang,
1302 Junhak Ahn, Heedo Park, Junehee Park, Narae Son, Gihyeon Kang,
1303 Jeonghun Kim, Kisoon Kim, Man Seong Park, Yoon Ki Kim, and Daehyun
1304 Baek. A high-resolution temporal atlas of the SARS-CoV-2 translatome and
1305 transcriptome. *Nature Communications*, 12:5120, 8 2021.

- [75] 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.
- [76] Ricarda J Rieger and Neva Caliskan. Thinking outside the frame: Impacting genomes capacity by programmed ribosomal frameshifting. *Frontiers in Molecular Biosciences*, 9:842261, 2022.
- [77] Matthew F. Allan, Amir Brivanlou, and Silvi Rouskin. RNA levers and switches controlling viral gene expression. *Trends in Biochemical Sciences*, 48, 2023.
- [78] 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.
- [79] Jamie A. Kelly, Michael T. Woodside, and Jonathan D. Dinman. Programmed -1 ribosomal frameshifting in coronaviruses: A therapeutic target. *Virology*, 554:75–82, 2021.
- [80] 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.
- [81] 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.
- [82] 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.
- [83] Wes McKinney. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010.
- [84] 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,

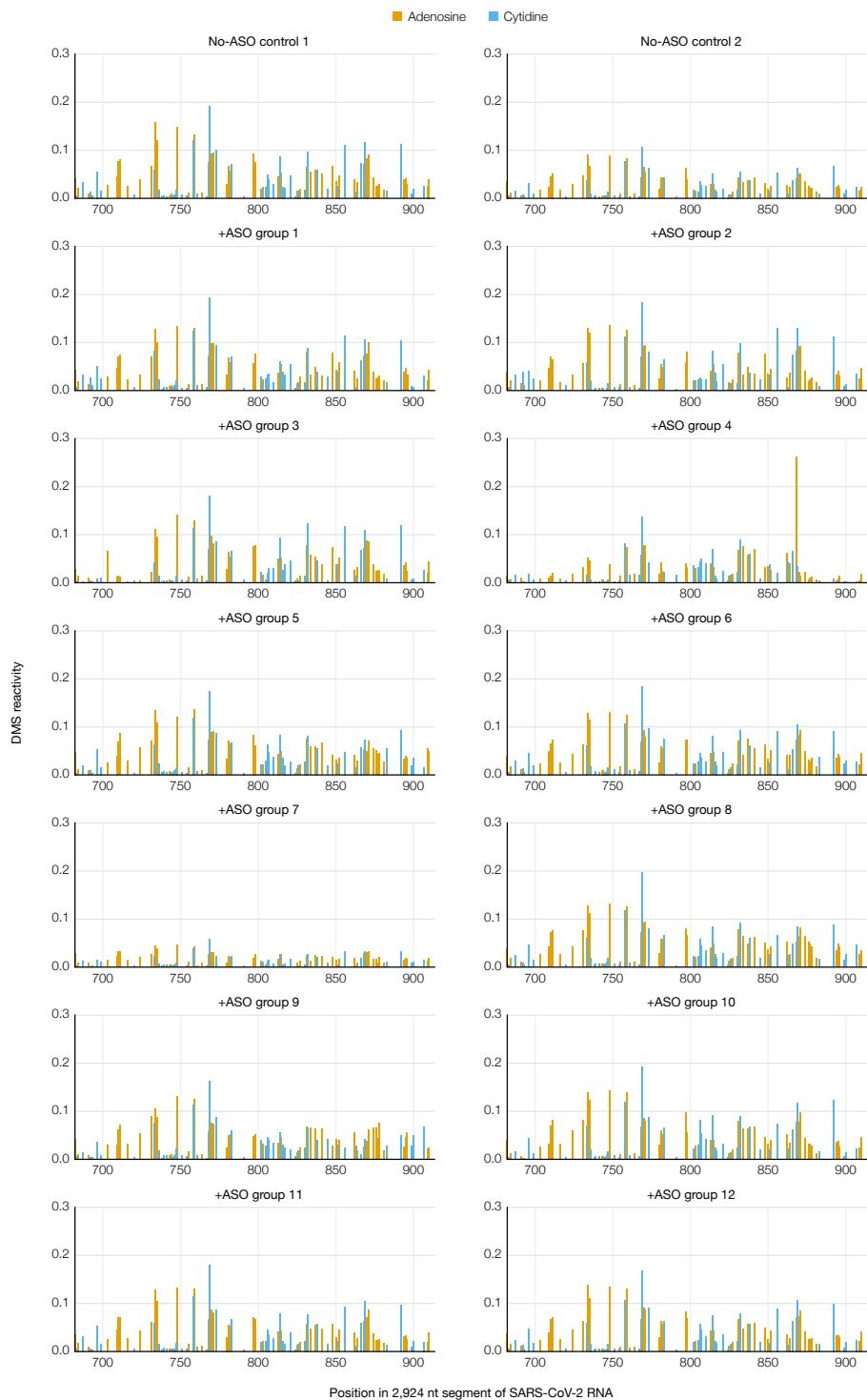
- 1350 Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, An-
1351 tônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Con-
1352 tributors. SciPy 1.0: Fundamental algorithms for scientific computing in
1353 Python. *Nature Methods*, 17:261–272, 2020.
- 1354 [85] Michael Waskom. seaborn: statistical data visualization. *Journal of Open*
1355 *Source Software*, 6, 2021.
- 1356 [86] Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and
1357 editing of the RNA secondary structure. *Bioinformatics*, 25:1974–1975, 2009.
- 1358 [87] Sara Alonso, Ander Izeta, Isabel Sola, and Luis Enjuanes. Transcription
1359 regulatory sequences and mRNA expression levels in the coronavirus trans-
1360 missible gastroenteritis virus. *Journal of Virology*, 76(3):1293–1308, 2002.
- 1361 [88] K. Nakagawa, K.G. Lokugamage, and S. Makino. Viral and cellular mRNA
1362 translation in coronavirus-infected cells. *Advances in Virus Research*,
1363 96:165, 12 2016.
- 1364 [89] D.A. Knoll and D.E. Keyes. Jacobian-free Newton-Krylov methods: a sur-
1365 vey of approaches and applications. *Journal of Computational Physics*,
1366 193(2):357–397, 2004.

Supplementary Information

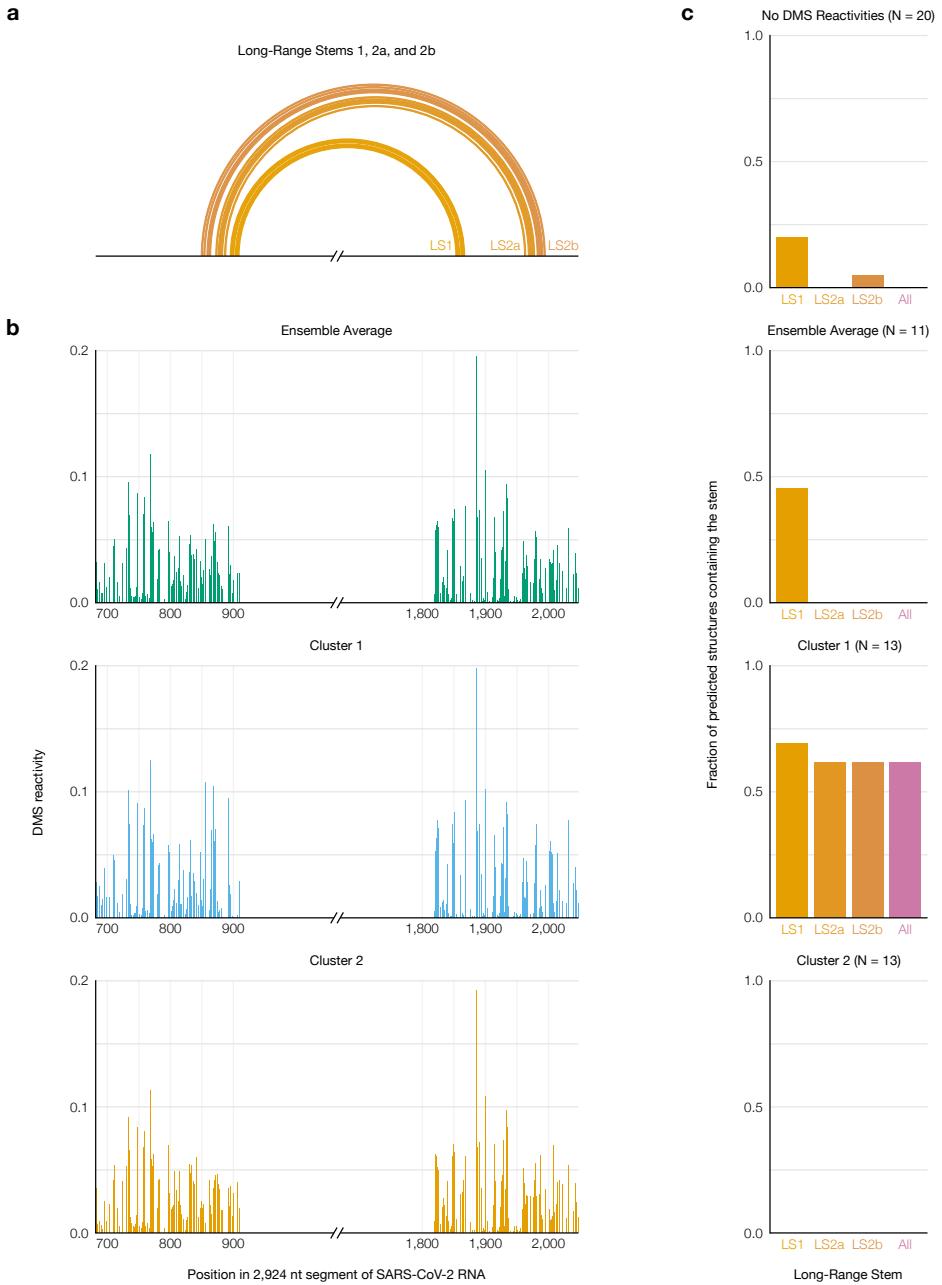
Supplementary Figures



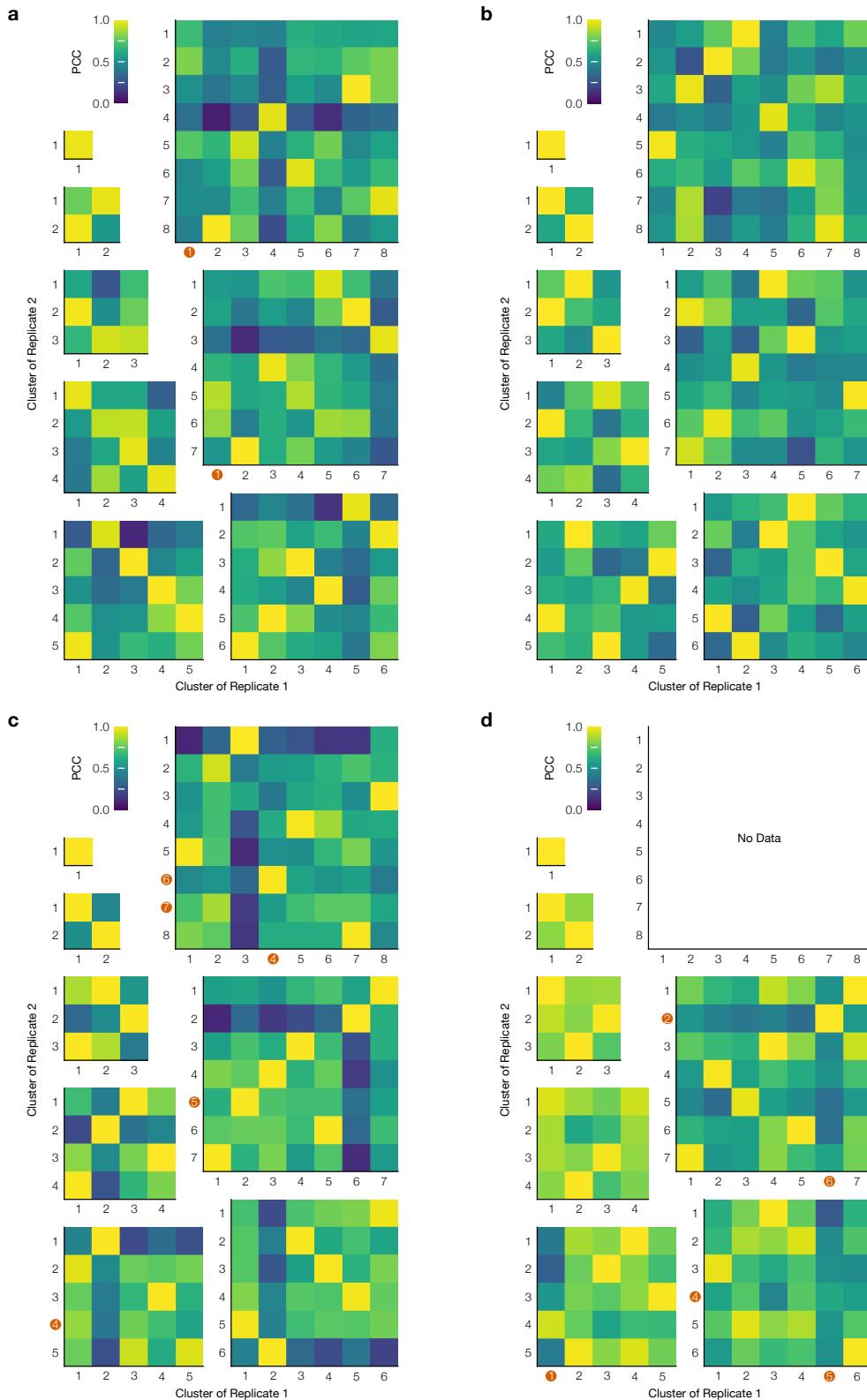
Supplementary Figure 1: Mutational profile of each ASO target section upon adding the corresponding group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA. Positions are colored based on the RNA sequence.



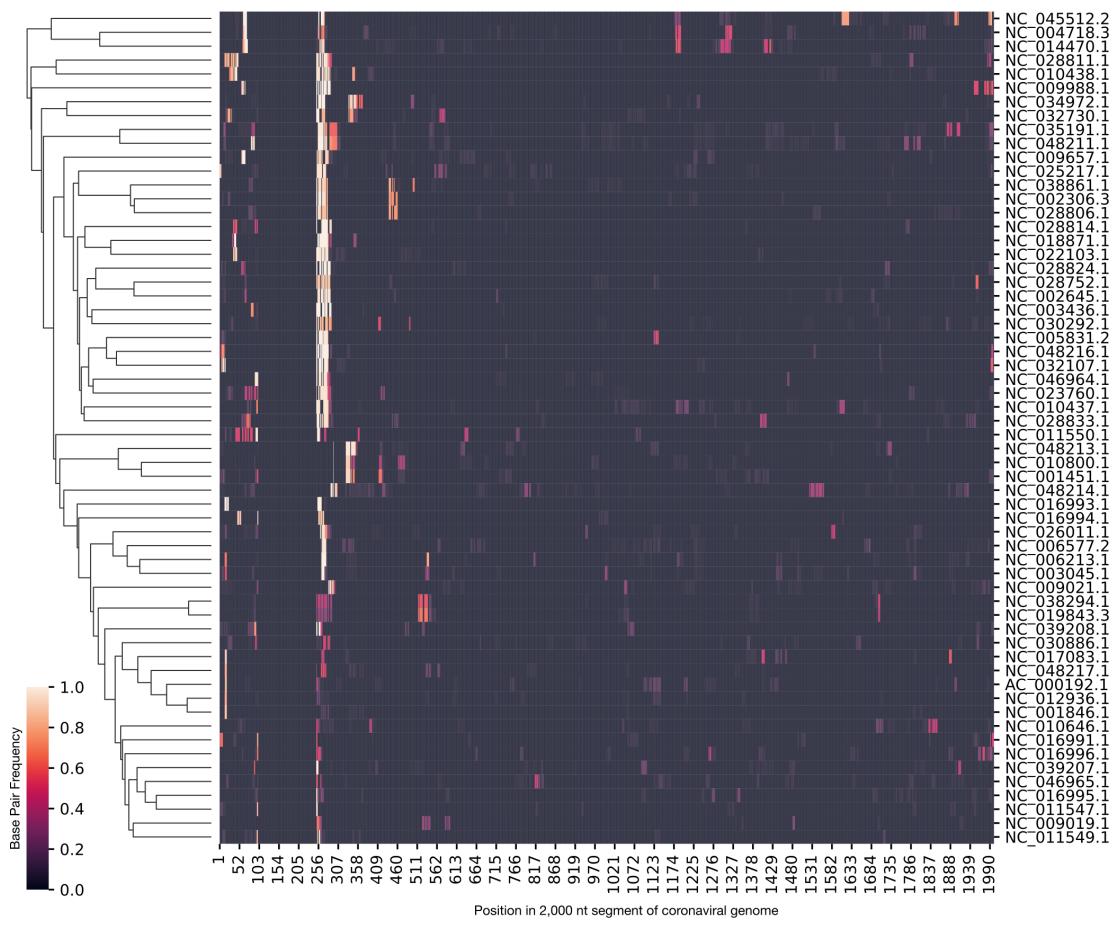
Supplementary Figure 2: Mutational profiles of the FSE section upon adding each group of ASOs to the 2,924 nt segment of SARS-CoV-2 genomic RNA.
Positions are colored based on the RNA sequence.



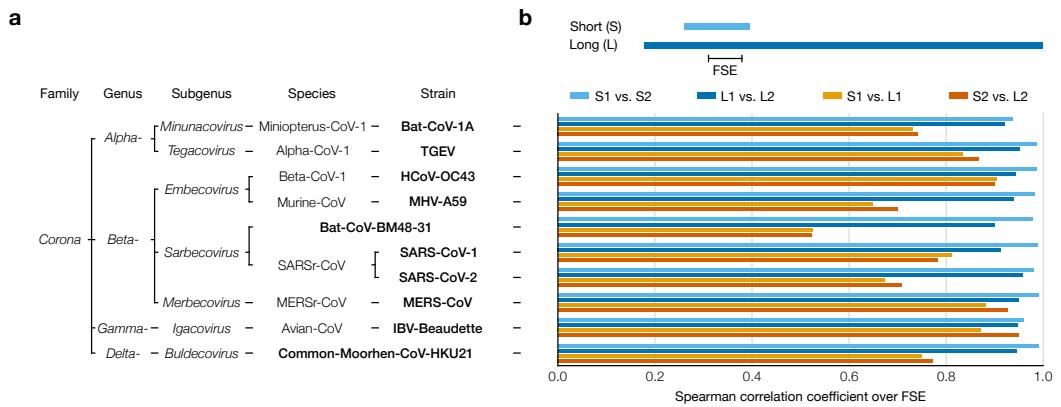
Supplementary Figure 3: Improved prediction of long-range stems in SARS-CoV-2 using clustered DMS reactivities. **(a)** Model of the two inner stems of the FSE-arch [54], denoted long stems (LS) 1 and 2a/b. **(b)** Mutational profiles of the ensemble average and of clusters 1 and 2 on both sides of the 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.



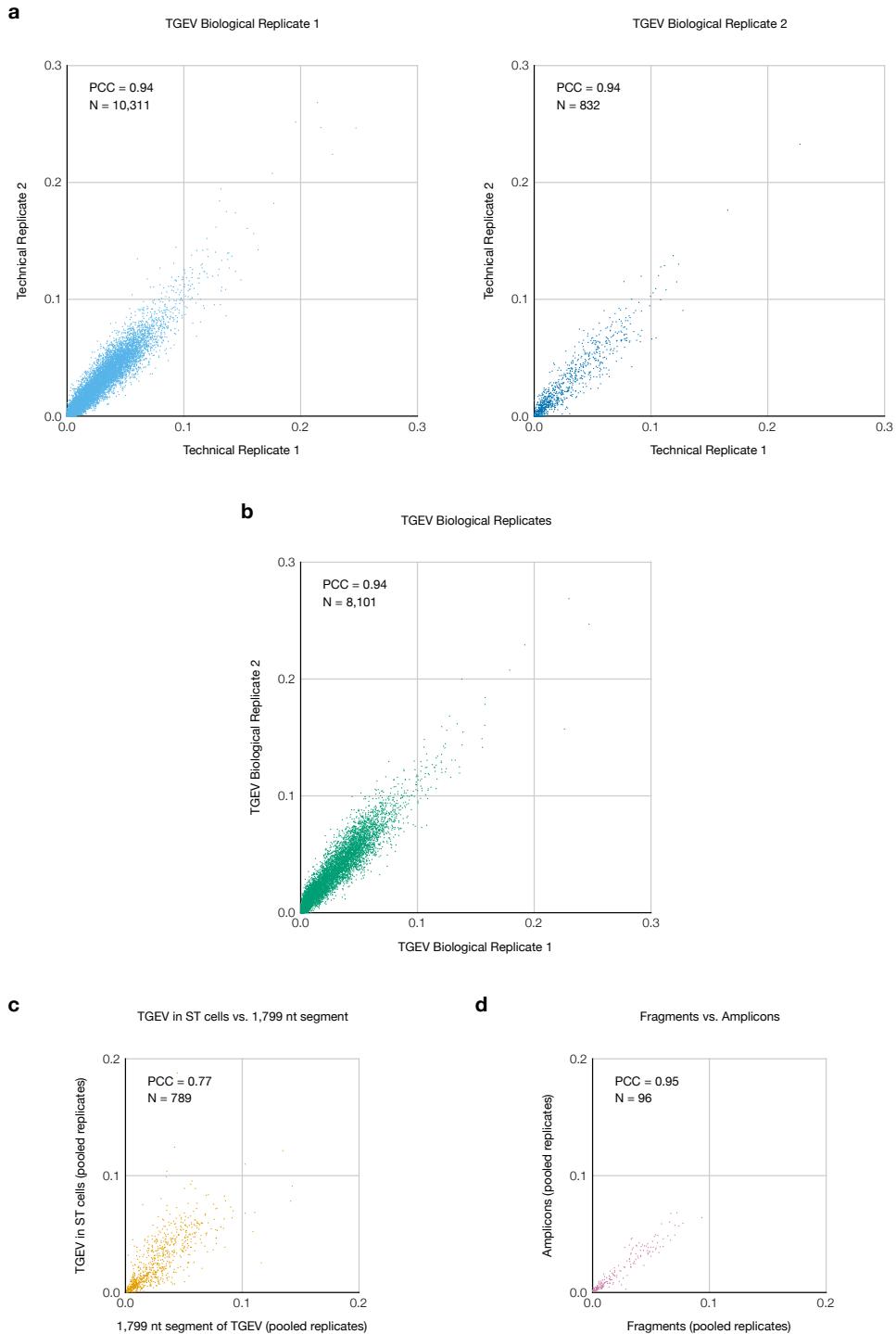
Supplementary Figure 4: Reproducibility of clustering the SARS-CoV-2 FSE after adding ASOs. (a) Heatmaps of the Pearson correlation coefficient (PCC) between each pair of clusters from two replicates of the 1,799 nt segment of SARS-CoV-2. Each heatmap corresponds to one order (i.e. number of clusters).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



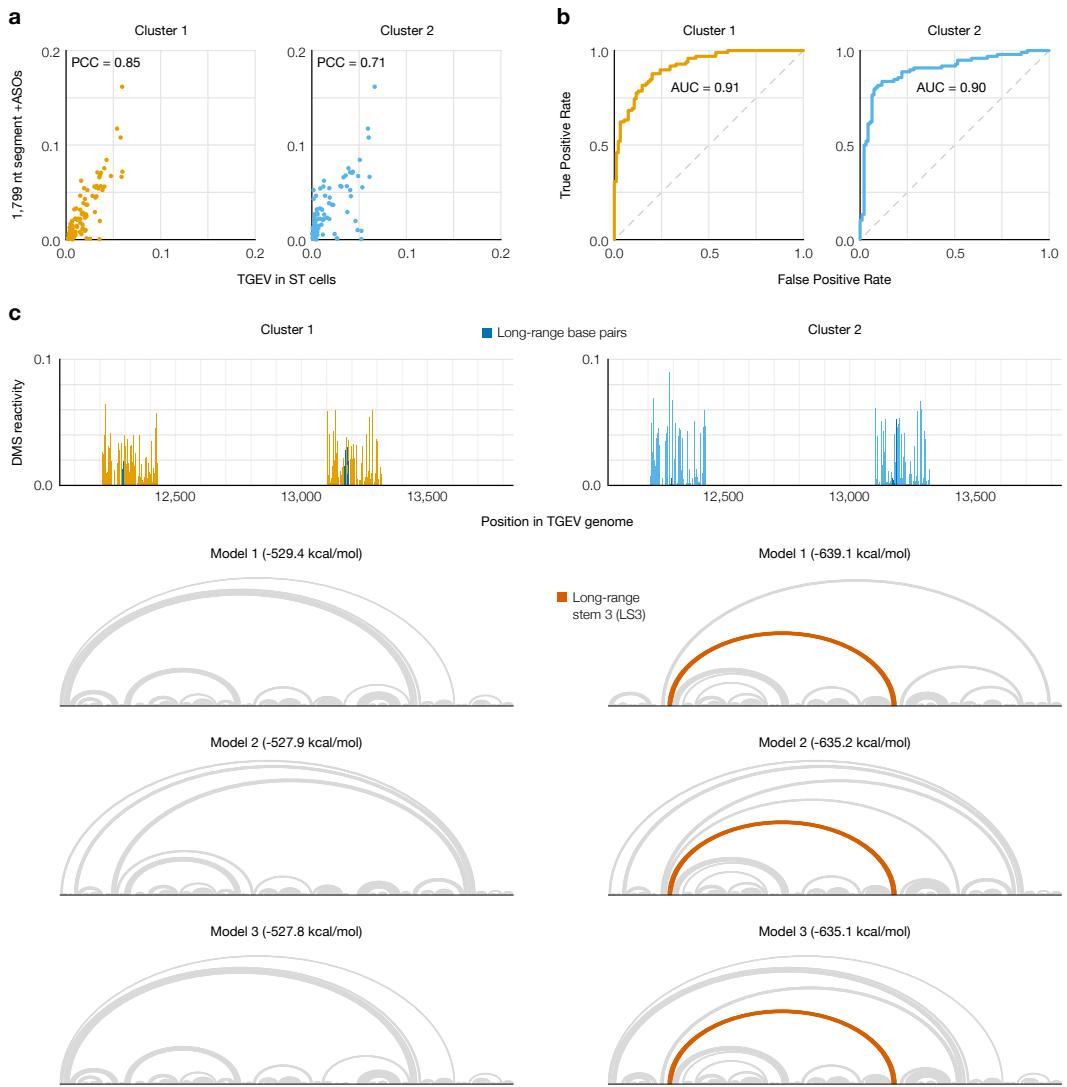
Supplementary Figure 5: Computational screen of long-range base pairing near the FSE in 60 coronaviruses. For each 2,000 nt segment of each coronaviral genome, the fraction of predicted structures in which each position outside the range 101-250 base-paired with any position in the range 101-250 is indicated. Genomes are clustered by their base-pairing frequencies. For each genome, the accession number for NCBI [56] is indicated.



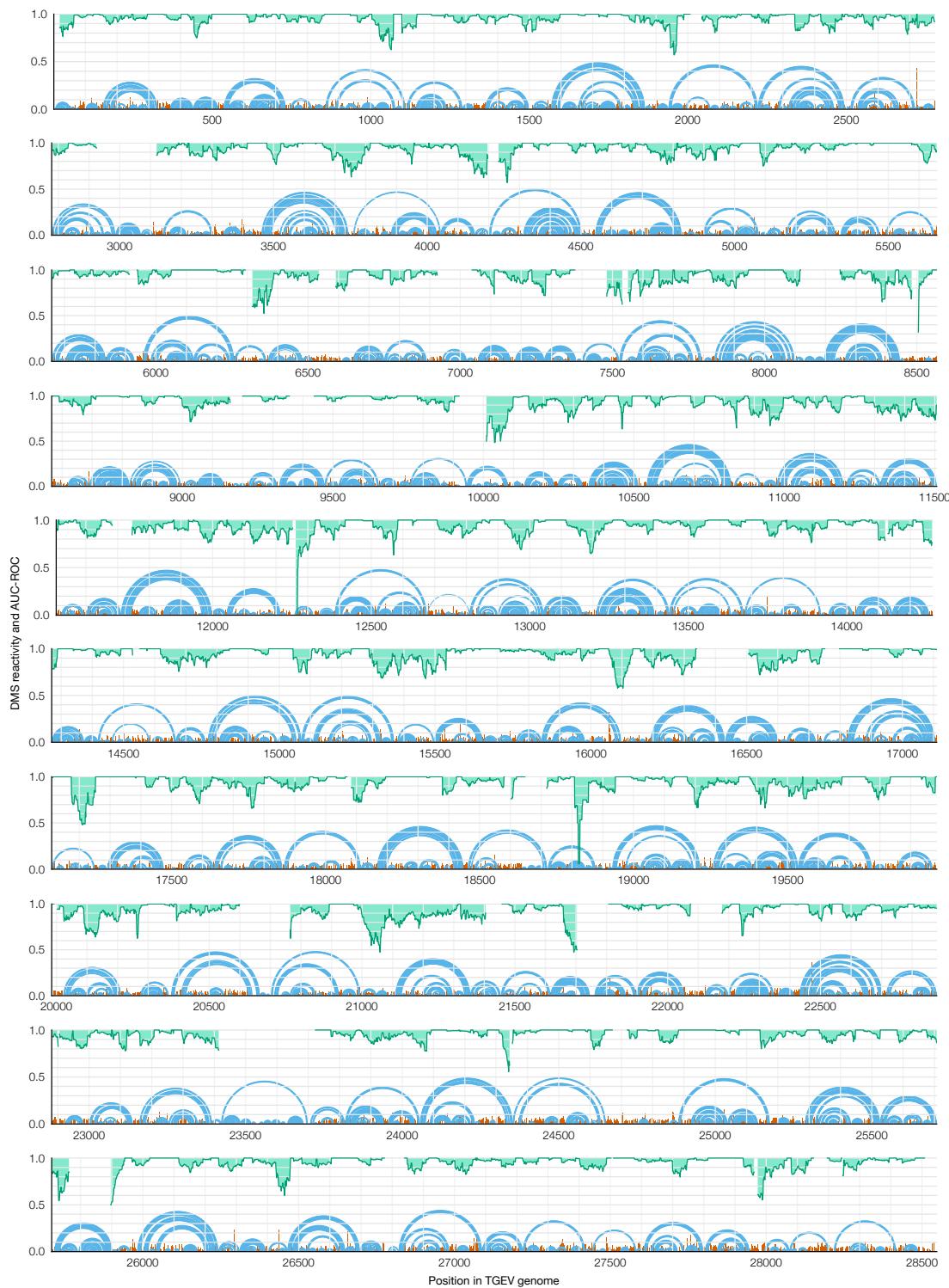
Supplementary Figure 6: Experimental screen of long-range base pairing near the FSE in 10 coronaviruses. (a) Taxonomy of the ten coronavirus species/strains in this screen; the lowest-level group for each virus is bolded. Bat-CoV-1A: bat coronavirus 1A (NC_010437.1), TGEV: transmissible gastroenteritis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



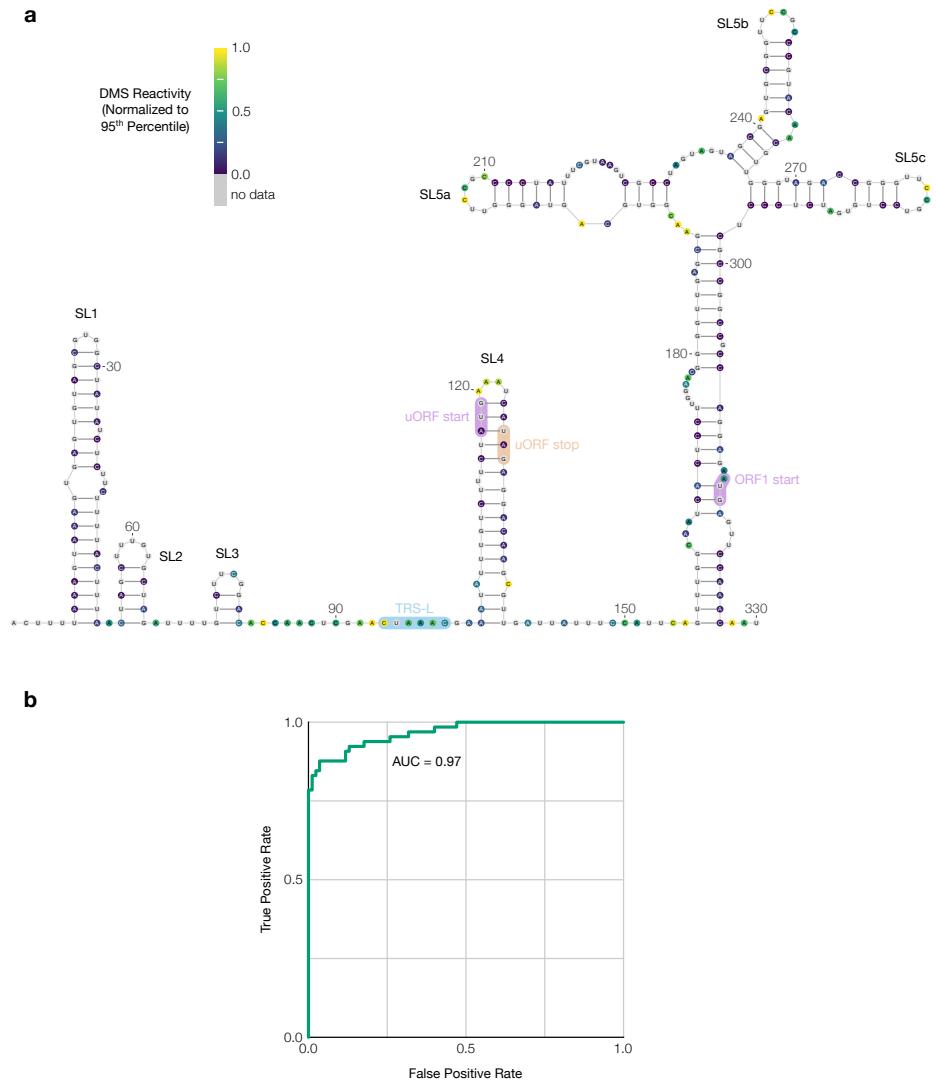
Supplementary Figure 7: Replicates of TGEV in ST cells and comparison to the 1,799 nt segment. (a) Comparison of DMS reactivities of the two technical replicates for each biological replicate of TGEV in ST cells. Each point represents one base in the sequence. The number of points (N) and Pearson correlation 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 [86].



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) [87], upstream open reading frame (uORF) [88], and start codon of ORF1 are also labeled. The model was drawn using VARNA [86]. **(b)** Receiver operating characteristic curve showing agreement between the DMS reactivities and the secondary structure model; the area under the curve (AUC) is indicated.

¹³⁶⁹ Supplementary Methods

¹³⁷⁰ Correcting observer bias due to drop-out of reads

¹³⁷¹ Let N reads from K clusters align to a reference sequence of length L . Let the
¹³⁷² proportion of reads whose 5' and 3' ends align, respectively, to coordinates a and
¹³⁷³ b ($1 \leq a \leq b \leq L$) be η_{ab} (assuming these proportions are equal for all clusters).
¹³⁷⁴ Let the mutation rate of base j ($1 \leq j \leq L$) in cluster k ($1 \leq k \leq K$) be μ_{jk} . Let
¹³⁷⁵ the proportion of cluster k in the ensemble be π_k . To express these quantities as
¹³⁷⁶ probabilities, let C_k be the event that a read comes from cluster k ; let E_{ab} be the
¹³⁷⁷ event that a read aligns with 5' and 3' coordinates a and b , respectively; let S_j be
¹³⁷⁸ the event that a read contains position j (i.e. its alignment coordinates a and b
¹³⁷⁹ satisfy $1 \leq a \leq j \leq b \leq L$); let M_j be the event that a read has a mutation at
¹³⁸⁰ position j ; and let G_g be the event that a read has no two mutations separated by
¹³⁸¹ fewer than g non-mutated bases.

¹³⁸² Deriving mutation rates of reads with no two mutations too ¹³⁸³ 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 , respectively ($1 \leq a \leq b \leq L$), 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, $P(G_g|S_j C_k)$ is the average of d_{abk} over every read that contains position j , weighted by the proportions η_{ab} :

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

The term $P(G_g|M_j E_{ab} C_k)$ 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 $P(G_g|M_j E_{ab} C_k)$ (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, $P(G_g|S_j M_j C_k)$ is the average of f_{abjk} over every read that contains position j , weighted by the proportions η_{ab} .

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

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

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

1384 **Deriving end coordinate proportions of reads with no two 1385 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 $P(E_{ab}|G_g C_k)$. 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 $P(G_g C_k)$ depends on k . Using these definitions and Bayes' theorem yields a probabilistic formula for e_{abk} :

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

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}):

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

¹³⁸⁶ 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 :

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

1389 **Solving total mutation rates and cluster and coordinate
1390 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:

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

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

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

1391 where x_{ij} is 1 if read i has a mutation at position j , otherwise 0; y_{abi} is 1 if read
1392 i aligns to coordinates a and b , otherwise 0; and z_{ik} is the probability that read i
1393 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:

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

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

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

1394 Solving this entire system at once has proven computationally impractical for all
1395 but extremely short sequences. A more feasible approach is to first solve for μ_{jk}
1396 given an initial guess for η_{ab} , next solve for η_{ab} given the updated μ_{jk} , then solve
1397 for π_k given the updated μ_{jk} and η_{ab} , and iterate until all three sets of parameters
1398 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

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

¹³⁹⁹ and applying the Netwon-Krylov method [89] implemented in SciPy [84].

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 :

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

1400 Clustering reads with the expectation-maximization 1401 algorithm

- 1402 Let N reads from K clusters align to a reference sequence of length L . Let the
- 1403 proportion of reads whose 5' and 3' ends align, respectively, to coordinates a and
- 1404 b ($1 \leq a \leq b \leq L$) be η_{ab} (assuming these proportions are equal for all clusters).
- 1405 Let the mutation rate of base j ($1 \leq j \leq L$) in cluster k ($1 \leq k \leq K$) be μ_{jk} . Let
- 1406 the proportion of cluster k in the ensemble be π_k .

1407 Maximization step

- 1408 The maximization step updates the parameters (μ_{jk} , η_{ab} , and π_k) using the current
- 1409 cluster memberships (z_{ik}). The observed estimates of the parameters m_{jk} , e_{ab} ,
- 1410 and p_k are first computed; then, the underlying parameters μ_{jk} , η_{ab} , and π_k are
- 1411 solved for as described in 10.1.4.

1412 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:

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

¹⁴¹³ **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	GGCAGCACAAGACATCTGTCGTAGTGCAACAGGACTAA-GCTCATTATT
	2	TGTAGTAAGCTAACGCATTGTCATCAGTGCAAGCAGTTT-GTGTAGTACC
	3	TGTAAATCGGATAACAGTGCAAGTACAAACCTACCTCCC-TTTGTTGTGT
	4	GATAGTACCAGTTCCATCACTCTTAGGGAATCTAGCCCCA-TTCAAATCC
	5	CTTAGGTGTCTGTAACAAACCTACAAGGTGGTTCCA-GTTCTGTATA
2	1	ATACCTCTATTTAGGTTTTAACCTTTAATAAAGTATAA-ATACTTCACTTAGGAC
	2	CACTCTGTTGCATTACCAGCTTGTAGACGTACTGTGGC-AGCTAAACTACCAAGTACC
	3	AAGCTTAGCAGCATCTACAGCAAAAGCACAGAAAGATA-ATACAGTTGAATTGGCAGG
	4	CACAACATCTAACACAATTAGTGATTGGTTGTCCCCCA-CTAGCTAGATAATCTTG
3	1	GATCCATATTGGCTTCCGGTGTAACTGTTATTGCCTGAC-CAGTACCAAGTGTGTGA
	2	ATGATCTATGTGGCAACGGCAGTACAGACAACACGATG-CACCACCAAAGGATTCTT

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

3	GTTGTAGGTATTTGTACATACTTACCTTTAAGTCACAAA- ATCCTTTAGGATTGG
4	CCGCAGACGGTACAGACTGTGTTTAAGTGTAAAACCC- ACAGGGTCATTAGCACAA
4	1 CTGAAGCATGGGTTCGCGGAGTTGATCACAACACTACAGC- CATAACCTTCCACATA
	2 GGAAGCGACAACAATTAGTTTAGGAATTAGCAAAAC- CAGCTACTTATCATTG
5	1 TGTCTCTTAACTACAAAGTAAGAATCAATTAAATTGTCAT- CTTCGTCCTTTCTT
	2 GACAATCCTTAAGTAAATTATAAATTGTTCTTCATGTTG- GTAGTTAGAGAAAGTG
	3 GGTACCATGTCACCGTCTATTCTAAACTAAAGAAGTCA- TGTTTAGCAACAGCTG
	4 AAGCATAGACGAGGTCTGCCATTGTTGATTAGTAAGAC- GTTGACGTGATATATGT
6	1 TGTATGTACAAGTATTCTTTAATGTCACAATTACC- TTCATCAAAATGCCTTA
	2 GGTTTCTACAAATCATACCAGTCCTTTATTGAAATA- ATCATCATCACAACAAT
	3 TTAACAAAGCTGGCGTACACGTTCACCTAACGTTGGCGT- ATACGCGTAATATATCTG
	4 ATGTCAGTACACCAACAATACCAGCATTGCGATGGCAT- CACAGAATTGTAATGTTT

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

7	1 GTTTGTATGAAATCACCGAAATCATACCAGTTACCATTG- AGATCTTGATTATCTA 2 TAGGCATTAACAATGAATAATAAGAACATCTACAAACAGGAA- CTCCACTACCTGGCGTG 3 GTTAAGTCAGTGTCAACATGTGACTCTGCAGTTAAAGCC- CTGGTCAAGGTTAATA 4 TTAACCTCTCTCCGTGAAGTCATATTTAACAAATCCC- CTTAATGTAAGGCTT
8	1 AACACAATTGGGTGGTATGTCTGATCCAATATTTAAAAA- TAACGGTCAAAGAGTT 2 GAGAATAAAACATTAAAGTTGCACAATGCAGAACATGCAT- CTGTCATCCAAACAGTT 3 CATCAACAAATATTTCTCACTAGTGGTCCAAAACTTGT- AGGTGGGAACACTGTA 4 ATGTACAACACCTAGCTCTGAAGTGGTATCCAGTTGA- AACTACAAATGGAACAC
9	1 TACACAAGTAATTCTAAACTAAGTCTAGAGCTATGTA- AGTTTACATCCTGATT 2 TGCCTTATCTAGTAATAGATTACCAGAACAGCAGCGTGCA- TAGCAGGGTCAGCAGCA 3 TTTGACAGTTGAAAAGCAACATTGTTAGTAAGTGCAGC- TACTGAAAAGCACGTAG 4 CTTAAAGAAACCCTAGACACAGCAAAGTCATAGAAGTC- TTTGTAAAATTACCGGG

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

10	1 CAGCATTACCATCCTGAGCAAAGAAGAAGTGTAAATT- CAACAGAACCTTCCTTC 2 CTGATATCACACATTGTTGGTAGATTATAACGATAGTAGT- CATAATCGCTGATAG 3 ACCATCGTAACAATCAAAGTACTTATCAACAACTTCAACT- ACAAATAGTAGTTGT 4 AACCAGCTGATTGTCTAGGTTGTTGACGATGACTTGGT- TAGCATTAAATACAGCC
11	1 CCTCATAACTCATTGAATCATAATAAAAGTCTAGCCTTACC- CCATTATTAAATGGAA 2 ATTTGAGTTAGTAGGGATGACATTACGTTGTATATG- CGAAAAGTGCATCTTGAT 3 GAGACACCAGCTACGGTGCGAGCTCTATTCTTGCACTA- ATGGCATACTTAAGATT 4 GGCTATTGATTCAATAATTTGATGAAACTGTCTATTG- GTCATAGTACTACAGATA
12	1 CAACCACCATAAGAATTGCTTGTCCAATTACTACAGTA- GCTCCTCTAGTGGC 2 CCATAAGGTGAGGGTTTCTACATCACTATAAACAGTTT- TAACATGTTGTGC 3 CATAATTCTAACATGTTAGGCATGGCTCTATCACATTAA- GGATAATCCCAAC 4 ACGGTGTGACAAGCTACAACACGTTGTATGTTGCGAG- CAAGAACAAAGTGAGGC

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

13 1 ACACATGACCATTCACTCAATACTTGAGCACACTCATTAGCTAATCTATAGAA
 2 AGTTGTGGCATCTCCTGATGAGGTTCCACCTGGTTAAC-ATATAGTGAACCGCC
 3 ATTAACATTGGCCGTGACAGCTTGACAAATGTTAAAAAC-ACTATTAGCATAAGC
 4 TAAATTGCGGACATACTTATCGGCAATTTGTTACCATCAGTAGATAAAAGTGC

1418

1419

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	AATAATGAGCTTAGCCTGTTGCACTACG
	R	AGGTTGTTAACCTTAATAAAGTATAAACTTCACT-TAGG
2	F	ACCTTGTAGGTTGTTACAGACACACCTAA
	R	TTGCCTGACCAGTACCACTAGTGTGTG
3	F	GGACAACCAATCACTAATTGTTAAGATGTTG
	R	TCACAACATACAGCCATAACCTTCCACA
4	F	CTTAAAAACACAGTCTGTACCGTCTGC
	R	GTAAGAACATTAAATTGTCATCTCGTCCTTTC
5	F	TGCTAAATTCTAAAAACTAATTGTTGTCGCTT
	R	ATGTGTCACAATTACCTTCATCAAAATGCCT
6	F	CAATGGCAGACCTCGTCTATGC
	R	GAAATCATACCAGTTACCATTGAGATCTTGATTATC
7	F	CGAAATGCTGGTATTGTTGGTGTACTGAC
	R	GTCTGATCCAATATTAAAATAACGGTCAAAGAG
8	F	TGTTAAAATATGACTTCACGGAAGAGAGGTT
	R	AAGTCTAGAGCTATGTAAGTTACATCCTGA
9	F	CCACTTCAGAGAGCTAGGTGTTGTAC
	R	CAAAGAAGAAGTGTAAATTCAACAGAACCTCCT
10	F	TGACTTGCTGTCTAAGGGTTCTTA
	R	CATAATAAAGTCTAGCCTACCCCATTATTAAATGG
11	F	CGTCAACAAACCTAGACAAATCAGCTGG

1420

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	GACCAATAGACAGTTCATCAAAAATTATTGAAATCAA-TAG
	R	ATACTTGAGCACACTCATTAGCTAATCTATAG
13	F	ACAACGTGTTGTAGCTTGTACACACC
	R	TAAATTGCGGACATACTTATCGGCAATTTG

1421

1422

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

1423

1424

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 Hepatitis Virus	60
SARS Coronavirus 1	60
SARS Coronavirus 2	55
Transmissible Gastroenteritis Virus	55

1425

1426

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	GGACCTATACGGTTTGT- CTTGAAAA
	F+T7	TAATACGACTCACTATAGGA- CCCTATACGGTTTGTCTTG- AAAA
	R	TTTTACAATAAAAGAAAGCAT- CATGCTT
Bat Coronavirus BM48-31	F	GGGTTTTATTCTTAGAAACA- CAGTCTG
	F+T7	TAATACGACTCACTATAGG- GTTTTATTCTTAGAAACACA- GTCTG
	R	GGAGTCTAATAAGTTGCC- TCTTCATC
Common Moorhen Coronavirus	F	GGATAAAGATAAGGAACCT- GTTTCTTT
	F+T7	TAATACGACTCACTATAGGA- TAAAGATAAGGAACCTGTTT- CTTT
	R	ACTATTAGGTATTGGCAAAT- TAATGCG
Human Coronavirus OC43	F	GGCTGTGTCTTATGTTTGA- CACATGA

1427

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	TAATACGACTCACTATAAGG- CTGTGTCTTATGTTTGACA- CATGA
	R	ATCTAATTATCACCGTTCT- CATCAAC
Infectious Bronchitis Virus	F	GGTTTGCAGTGTGCCAG- TGTTGGAT
	F+T7	TAATACGACTCACTATAAGGT- TTGCACTGTTGCCAGTGTT- GGAT
	R	CTCAAGATTCCATCTTCAG- TATCGCG
MERS Coronavirus	F	GGGATTTGTTGTCAAATA- CCCCCTG
	F+T7	TAATACGACTCACTATAAGG- GATTTGTTGTCAAATACC- CCCTG
	R	ATGATGCCCTGGTCATCT- AATTCTAC
Murine Hepatitis Virus	F	GGCTGTGTCATATGTGTTG- ACGCATGA
	F+T7	TAATACGACTCACTATAAGG- CTGTGTCATATGTGTTGAC- GCATGA

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	ATCCAACCTGTTGCCGTCC- TCATCTAC
SARS Coronavirus 1	F	GGGTTTTACACTTAGAAACA- CAGTCTG
	F+T7	TAATACGACTCACTATAGG- GTTTTACACTTAGAAACACA- GTCTG
	R	AGAGTCTAATAAATTGCCTT- CCTCATC
SARS Coronavirus 2	F	GGGTTTTACACTAAAAACA- CAGTCTG
	F+T7	TAATACGACTCACTATAGG- GTTTTACACTAAAAACACA- GTCTG
	R	AGAATCAATTAAATTGTCAT- CTTCGTC
Transmissible Gastroenteritis Virus	F	GGCAATTCGGTTCTGTATT- GAAAATGA
	F+T7	TAATACGACTCACTATAGG- CAATTGGTTCTGTATTGAA- AATGA
	R	TTTGACAATGTAGTAGGCAT- CATGTTT

Table 6: Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs.

Coronavirus	ASO	Sequence
Bat Coronavirus 1A	1	CAGGGCTCTAGTCGAGCTGC- ACTAGAGCCCCTTGCTCGTT- AAATAAGCCTGATCAACAG
	2	GCAACTTCTTATTGTAAATAT- CAAAGGCGCGTACAACATGC- TCCGGTTCAGTACCATTA
Bat Coronavirus BM48-31	1	GACATCAGTGCTTGTGCCCTGT- GCCGCACGGTGTAAGACGGG- CCGCACTTACACCGCAAAC
	2	TTTAGGAACCTTGCAAAACC- AGCAACTTCTCATTATAAATA- TCAAAAGCCCTGTAAAC
	3	AAAATAGGAGTCTAATAAGTT- GCCCTCTTCATCAACTTCCTG- GAAACGGCAACAATTGT
Common Moorhen Coronavirus	1	TGGGGTTCTAGACGGGCATC- ACTAGAACCCCTTACTCGTT- AAATAAGCTGTATTTGCA
	2	GTTATATTATTATGTACATGAA- ACGCCCTTTTACAATATCCG- GCTGAGTGCCAGACTGT
Infectious Bronchitis Virus	1	ACATCAAAGGCTCGCTTACA- ACATCAGGATCACATCCACTA- GCAAGGGGTATCAGCCGA

Table 6: Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs. (Continued)

Murine Hepatitis Virus	1 AAGCCACTGGCACAGGGTAC- AAGACGGGCATTTACACTTGT- ACCCCGAATCCGTTAAAA 2 CCAATGCCAGCTCGATTAGCA- TTACAAATGTCAAATGCCCTT- AATTGAACATCAGTGTCC 3 AACTTGTTGCCGTCTCATCT- ACACGCTGGAAGCGGCAGCA- ATTCACTTATAATACAAA
SARS Coronavirus 1	1 TTTGCAAAACCAGCAACTTTT- TCGTTGTAATATCAAAAGCC- CTGTAGACGACATCAGTA 2 TCTAATAAATTGCCCTCCTCAT- CCTTCTCCTGGAAGCGACAG- CAATTAGTTTTAGGAAC
SARS Coronavirus 2	1 GACATCAGTACTAGTGCCTGT- GCCGCACGGTGTAAGACGGG- CTGCACTTACACCGCAAAC 2 TTTTAGGAATTAGCAAAACC- AGCTACTTTATCATTGTAGAT- GTCAAAAGCCCTGTATAC
Transmissible Gastroenteritis Virus	1 TAAATAACTTGATCAACAGT- AAAACCTCTGCATAGAAGTACG- ATCGCACATGCAACCATT

Continued on next page

Table 6: Sequences of the antisense oligonucleotides (ASOs) targeting the 1,799 nt segments of coronaviral RNAs. (Continued)

2 GGTCTGGATCAGTACCATTGC-
AGGGTTCTAGTCGAGCTGCA-
CTAGAACCCCGCACTCGTT

1433