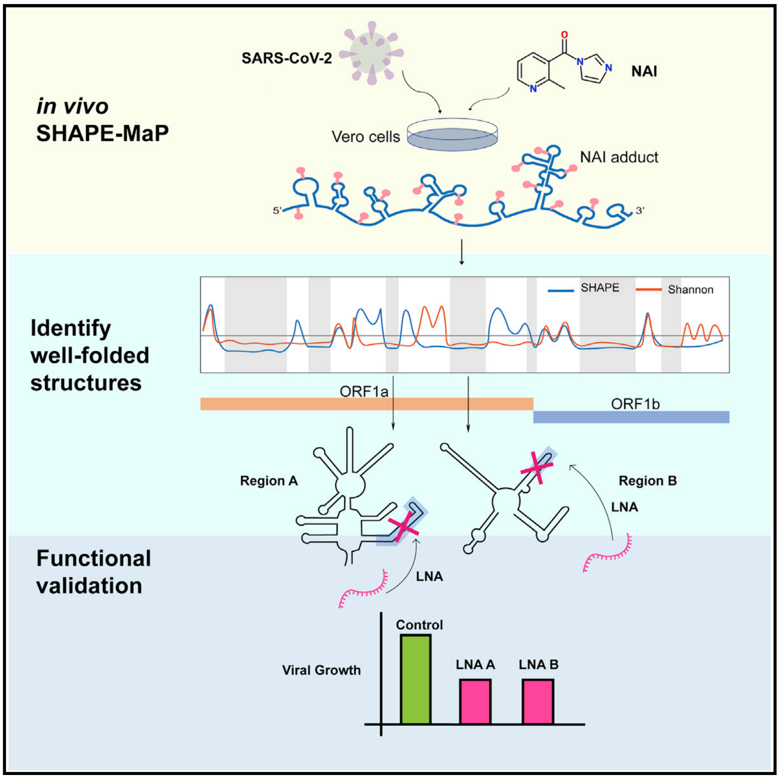
Molecular Cell

Article

Comprehensive *in vivo* secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms

Graphical Abstract



Highlights

- The SARS-CoV-2 genome is probed at single-nucleotide resolution in infected cells
- RNA structure prediction reveals an elaborate SARS-CoV-2 genome architecture
- Networks of well-folded secondary structure are conserved across β-coronaviruses
- Disruption of conserved secondary structures with LNAs inhibits viral growth

Authors

Nicholas C. Huston, Han Wan,
Madison S. Strine,
Rafael de Cesaris Araujo Tavares,
Craig B. Wilen, Anna Marie Pyle

Correspondence

anna.pyle@yale.edu

In Brief

Using cell-based chemical probing, Huston et al. provide an experimentally determined structural map of the SARS-CoV-2 RNA genome in infected cells. The map reveals networks of well-folded RNA structures that are conserved in other coronaviruses. Disruption of these structures inhibits SARS-CoV-2 growth.

Huston et al., 2021, Molecular Cell *81*, 584–598
February 4, 2021 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.molcel.2020.12.041

CellPress

Molecular Cell

Article

Comprehensive *in vivo* secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms

Nicholas C. Huston,[1,7] Han Wan,[2,7] Madison S. Strine,[3,4] Rafael de Cesaris Araujo Tavares,[5] Craig B. Wilen,[3,4] and Anna Marie Pyle[2,5,6,8,*]

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA
[2]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06511, USA
[3]Department of Laboratory Medicine, Yale School of Medicine, New Haven, CT 06510, USA
[4]Department of Immunobiology, Yale School of Medicine, New Haven, CT 06519, USA
[5]Department of Chemistry, Yale University, New Haven, CT 06511, USA
[6]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: anna.pyle@yale.edu

https://doi.org/10.1016/j.molcel.2020.12.041

SUMMARY

Severe-acute-respiratory-syndrome-related coronavirus 2 (SARS-CoV-2) is the positive-sense RNA virus that causes coronavirus disease 2019 (COVID-19). The genome of SARS-CoV-2 is unique among viral RNAs in its vast potential to form RNA structures, yet as much as 97% of its 30 kilobases have not been structurally explored. Here, we apply a novel long amplicon strategy to determine the secondary structure of the SARS-CoV-2 RNA genome at single-nucleotide resolution in infected cells. Our in-depth structural analysis reveals networks of well-folded RNA structures throughout Orf1ab and reveals aspects of SARS-CoV-2 genome architecture that distinguish it from other RNA viruses. Evolutionary analysis shows that several features of the SARS-CoV-2 genomic structure are conserved across β-coronaviruses, and we pinpoint regions of well-folded RNA structure that merit downstream functional analysis. The native, secondary structure of SARS-CoV-2 presented here is a roadmap that will facilitate focused studies on the viral life cycle, facilitate primer design, and guide the identification of RNA drug targets against COVID-19.

INTRODUCTION

Severe-acute-respiratory-syndrome-related coronavirus 2 (SARS-CoV-2), which is responsible for the current pandemic (Zhu et al., 2020), is a positive-strand RNA virus in the genus β-coronavirus. To date, the outbreak of SARS-CoV-2 has infected millions of people globally, causing great economic loss and posing an ongoing public health threat (Dong et al., 2020). Included in the β-coronavirus genus are two related viruses, SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV), that caused global outbreaks in 2003 and 2012, respectively (de Wit et al., 2016). Despite the continued risk posed by β-coronaviruses, mechanistic studies of the family are limited, highlighting the need for research that facilitates the development of therapeutics. With most research efforts focusing on viral proteins (Lan et al., 2020a; Yin et al., 2020; Wan et al., 2020), little is known about the viral RNA genome, especially its structural content.

Like other coronaviruses, the genome of SARS-CoV-2 is incredibly large (Maier et al., 2015; Zhu et al., 2020). The ∼30-kb genome is comprised of two open reading frames (ORFs) for viral nonstructural proteins (Nsps) and nine small ORFs that encode structural proteins and accessory genes (Kim et al., 2020). The entire ORF region is flanked by 5′ and 3′ untranslated regions (UTRs) that have been shown in other coronaviruses to contain conserved RNA structures with important functional roles in the viral life cycle (Madhugiri et al., 2018; Chen and Olsthoorn, 2010; Züst et al., 2008).

One of the best-studied functional RNA elements in β-coronavirus genomes is the programmed ribosomal frameshifting pseudoknot (PRF) that sits at the boundary between Orf1a and Orf1b (Plant and Dinman, 2008). The PRF, found in all coronaviruses, induces a −1 ribosomal frameshift that allows for bypassing of the Orf1a stop codon and production of the Orf1ab polyprotein, which includes the viral replicase. Extensive mutational analysis with reporter constructs and full-length virus has revealed a three-stemmed pseudoknot structure conserved across group II β-coronaviruses (Plant et al., 2005, 2013). However, neither the mechanism of frameshifting regulation nor the SARS-CoV-2 PRF conformation has been validated directly in cells.

While recent computational studies suggest the 5′ UTR, 3′ UTR, and PRF functional elements are conserved in the SARS-CoV-2 genome (Rangan et al., 2020; Andrews et al., 2020), these regions account for a vanishingly small fraction of the total nucleotide content. Studies of other positive-sense viral RNA genomes such as hepatitis C virus (HCV) and human immunodeficiency virus (HIV) have revealed extensive networks of regulatory RNA structures contained within viral ORFs (Siegfried et al., 2014; Pirakitikulr et al., 2016; Friebe and Bartenschlager, 2009; Li et al., 2018; You et al., 2004) that direct critical aspects of viral function. It is therefore vital to characterize structural features of the SARS-CoV-2 ORF, as this knowledge will enhance our understanding of the coronavirus life cycle, improve diagnostics, and identify riboregulatory regions that can be targeted with antiviral drugs.

Recent advances in high-throughput structure probing methods (SHAPE-MaP and DMS-MaP) have greatly facilitated the structural studies of long viral RNAs (Siegfried et al., 2014; Zubradt et al., 2017). Recently, Manfredonia et al. performed full-length SHAPE-MaP analysis on *ex vivo* extracted and refolded SARS-CoV-2 RNA (Manfredonia et al., 2020). However, structural studies on both viral RNA and mRNA have highlighted the importance of probing RNAs in their natural cellular context (Simon et al., 2019; Rouskin et al., 2014). Lan et al. performed full-length *in vivo* DMS-MaPseq on SARS-CoV-2-infected cells (Lan et al., 2020b), but as DMS only reports on A and C nucleotides, the data coverage is necessarily sparse. While both studies reveal important features of the structural content in the SARS-CoV-2 genome and its evolutionary conservation, to date, no work has been published that captures information for every single nucleotide in an *in vivo* context.

Here, we report the complete secondary structure of the SARS-CoV-2 RNA genome using SHAPE-MaP data obtained in living cells. We deploy a novel long-amplicon method, readily adapted to other long viral RNAs, made possible by the highly processive reverse transcriptase MarathonRT (Guo et al., 2020). The resulting genomic secondary structure model reveals functional motifs at the viral termini that are structurally homologous to other coronaviruses, thereby fast tracking our understanding of the SARS-CoV-2 life cycle. We reveal conformational variability in the PRF, highlighting the importance of studying viral structures in their native genomic context and underscoring their dynamic nature. We also uncover elaborate networks of well-folded RNA secondary structures dispersed across Orf1ab, and we reveal features of the SARS-CoV-2 genome architecture that distinguish it from other single-stranded, positive-sense RNA viruses. Evolutionary analysis of the full-length SARS-CoV-2 secondary structure model suggests not only that its architectural features appear to be conserved across the β-coronavirus family but also that individual regions of well-folded RNA may be as well. Using structure-disrupting, antisense locked nucleic acids (LNAs), we demonstrate that RNA motifs within these well-folded regions play functional roles in the SARS-CoV-2 life cycle. Our work reveals the unique genomic architecture of SARS-CoV-2 in infected cells, points to important viral strategies for infection and persistence, and identifies potential drug targets. The full-length structure model we present here thus serves as an invaluable roadmap for future studies on SARS-CoV-2 and other coronaviruses that emerge in the future.

RESULTS

In vivo SHAPE-MaP workflow yields high-quality data suitable for structure prediction

To study the SARS-CoV-2 structure in the context of infected cells, the SARS-CoV-2 isolate USA-WA1/2020, isolated from a symptomatic patient who had returned to the United States from China, was used to infect VeroE6 cells in a BSL3 facility (BEI Resources #NR-52281). At 4 days post-infection, cells were collected and treated with either 2-methylnicotinic acid imizaolide (NAI), which preferentially modifies flexible nucleotides at the 2′ OH, or DMSO as a control. RNA was then extracted and purified. To generate sequencing libraries, 2,000-nt overlapping amplicons were tiled across the entire SARS-CoV-2 genome (Figure 1A). This efficient approach is made possible by the ultra-high-processive reverse transcriptase MarathonRT. Previous work from our lab demonstrated that MarathonRT successfully encodes NAI adducts as cDNA mutations and that structural features of the HCV IRES are perfectly recapitulated when in cell SHAPE-MaP reactivities are used for structure prediction(Guo et al., 2020).

Two independent biological replicates of in cell SHAPE-MaP data were generated and analyzed using the ShapeMapper pipeline (Smola et al., 2015b). Comprehensive datasets were obtained, with median effective read depth >70,000× and effective reactivity data for 99.7% (29,813/29,903) of nucleotides in the SARS-CoV-2 genome in both replicate experiments. To check the SHAPE-MaP data quality, we analyzed the relative mutation rates of NAI-treated and DMSO-treated RNA samples, revealing a significant elevation of mutation rates for NAI-treated samples (Figure 1B; p value < 0.0001). This confirms that the full-length SARS-CoV-2 RNA was successfully modified *in vivo* and that these modifications were encoded as cDNA mutations.

To understand the relative SHAPE reactivity agreement within local regions of the genome, we calculated Pearson correlation coefficients between two biological replicates. The Pearson's correlation across the entire span of Orf1ab is 0.628 (Figure S1A), consistent with those previously reported for reactivities calculated from *in vivo* modified RNAs of this size (Smola et al., 2016). Across the subgenomic RNA ORFs, the Pearson's correlation is poor (Figure S1B). We believe this reflects the fact that amplicons 13–16 will amplify both full-length and subgenomic RNAs, and the difference in context will result in different secondary structures (Tavares et al., 2020). For this reason, despite the fact all data have been obtained globally, subsequent discrete structural analysis will focus on shared features of the viral termini and the Orf1ab region.

De novo structure prediction on full-length SARS-CoV-2 RNA identifies conserved functional elements at the 5′ and 3′ genomic termini

We performed secondary structure prediction with the SuperFold pipeline (Smola et al., 2015b) using the *in vivo* SHAPE reactivities to generate an experimentally constrained consensus secondary structure prediction for the entire SARS-CoV-2 genome. As an extensive body of research has elucidated structured RNA elements at the 5′ and 3′ viral termini with conserved functions across β-coronaviruses, we first examined these regions from our consensus prediction to determine whether

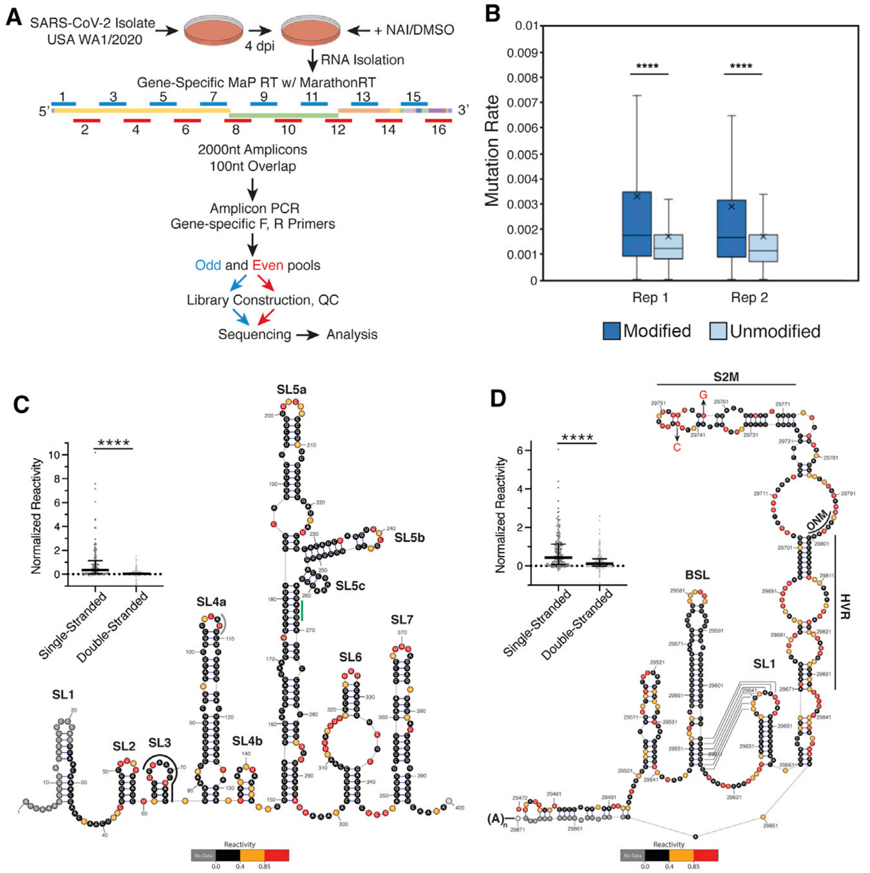


Figure 1. Tiled-amplicon *in vivo* SHAPE-MaP workflow yields high-quality data for *de novo* full-length structure prediction

Structure prediction identifies conserved functional elements at the 5′ and 3′ viral termini.

(A) Workflow of *in vivo* SHAPE-MaP probing of full-length SARS-CoV-2 genomic RNA. The schematic of the SARS-CoV-2 genome is colored by protein-coding domain.

(B) Mutation rates for two biological replicates across the entire SARS-CoV-2 genome (box, interquartile range [IQR]; median indicated by line; average indicated by "x"; whiskers are drawn in the Tukey style, and values outside this range are not shown).

(C) Consensus structure prediction for the 5′ terminus of SARS-CoV-2, colored by SHAPE reactivity. Functional domains are labeled (transcription regulatory sequence (TRS), black line; upstream ORF start codon, gray line; Orf1a start codon, green line). Inset – mapping of SHAPE reactivity data to single- and double-stranded regions. Line indicates median, and whiskers indicate standard deviation.

(D) Structure prediction for the 3′ terminus of SARS-CoV-2, colored by SHAPE reactivity. Functional domains are labeled. The putative pseudoknot is indicated by solid black lines. Inset: mapping of SHAPE reactivity to single- and double-stranded regions. Data are plotted as in (C).

****$p < 0.0001$ by equal variance unpaired Student's t test.

they were stably folded and well determined in the SARS-CoV-2 genome.

The 5′ genomic terminus includes seven regions that have been identified and studied in other coronaviruses (Yang and Leibowitz, 2015). While sequence conservation suggested that these elements might be conserved in SARS-CoV-2, our consensus structure prediction shows this to be the case, and we derived a specific experimentally determined secondary structure for this section of the genome. The *in vivo* SHAPE reactivity data correspond well with the resulting structural model (Figure 1C, inset), and the low overall Shannon entropy values in this region (determined from base-pair probability calculation during the SuperFold prediction pipeline; Smola et al., 2015b) support a well-determined structure for the 5′ genomic terminus ($\text{median}_{\text{Nuc(1-400)}} = 2.7 \times 10^{-5}$; global median = 0.022).

Individual features that typify coronavirus structures are evident in the secondary structure of the SARS-CoV-2 5′ UTR with good SHAPE reactivity agreement (Figure 1C, inset). For example, a trifurcated stem is observed at the top of stem loop 5 (SL5) (Figure 1C), including UUCGU pentaloop motifs in SL5A and SL5B, and a GNRA tetraloop in SLC, as predicted in other coronaviruses. Previous reports suggest that SL5 may represent a packaging signal for group IIB CoVs (Chen and Olsthoorn, 2010). Similarity between SL5 structures reported for other coronaviruses and the experimentally determined structure reported here suggests that SL5 plays a similar role in the SARS-CoV-2 life cycle. The structural homology to other coronaviruses exemplified by the SL5 structure model extends to every other stem loop labeled in Figure 1C (SL1–SL4, SL6, and SL7), suggesting these structures also play similar functional roles despite having been identified and elucidated in other coronaviruses (Yang and Leibowitz, 2015).

The 3′ genomic terminus includes three well-studied stems, including the bulged-stem loop (BSL), stem loop 1 (SL1), and a long bulge stem that includes the hypervariable region (HVR), the S2M domain, the octanucleotide motif (ONM) subdomains, and a pseudoknot (Yang and Leibowitz, 2015). The consensus structure recapitulates the secondary structure of all the three stems with good SHAPE reactivity agreement (Figure 1D, inset) and overall low Shannon entropy ($\text{median}_{\text{Nuc(29,472–29,870)}}$ = 0.016). While the BSL is well determined in our structure model, the low reactivity for bulged nucleotides suggests the possibility of protein-binding partners (Figure 1D).

A pseudoknot structure is proposed to exist between the base of the BSL and the loop of SL1 in coronaviruses (Yang and Leibowitz, 2015). While pseudoknot formation is mutually exclusive with the base of the BSL, studies in Murine Hepatitis virus (MHV) have suggested that both structures contribute to viral replication and may function as molecular switches in different steps of RNA synthesis (Goebel et al., 2004). However, our *in-vivo*-determined secondary structure is inconsistent with formation of the pseudoknot (Figure 1D). The low SHAPE reactivities for the nucleotides at the base of the BSL support formation of the extended BSL stem, while high reactivities of the nucleotides in the loop of SL1 indicate that it is highly accessible. Using the SHAPEKnots program (Hajdin et al., 2013), we found that a pseudoknot is never predicted in three windows that cover the pseudoknotted region. Taken together, our data strongly support the extended BSL conformation, indicating it is probably the dominant conformation *in vivo*.

The third stem in the 3′ UTR includes three subdomains. The HVR, poorly conserved across group II coronaviruses (Goebel et al., 2007), is predicted to be mostly single stranded in our secondary structure, and the high reactivities across the span of this region lends strong experimental support for an unstructured region (Figure 1D). The fact that this region is relatively unstructured may explain why it tolerates deletions, rearrangements, and point mutations in MHV (Goebel et al., 2007).

The S2M region is contained within the apical part of the third stem. We observe that the first three helices of S2M from SARS-CoV-2 exactly match the crystal structure determined for S2M from SARS-CoV (Robertson et al., 2005). However, our *in vivo* secondary structure model deviates significantly at the top of the stem (Figure 1D). It is possible that the SARS-CoV-2 S2M folds into a unique S2M conformation despite differing by only two bases, both of which are transversions (Figure 1D, base changes indicated by arrows; SARS-CoV base identity shown in red). Any base-pairing interaction involving these nucleotides in the SARS-CoV S2M could not be maintained in SARS-CoV-2. Alternatively, this site could interact with factors *in vivo* that are not captured in the crystallographic study.

Finally, we predict a different structure for the terminal stem in the viral 3′ UTR (adjacent to the poly(A) tail) than previously reported for other coronaviruses (Züst et al., 2008). However, structure prediction of the complete stem is not highly accurate, as reactivity information for the downstream stem (nucleotides 29,853–29,870) is occluded by primer binding and is not constrained by experimental data (Figure 1D). In addition, the complete stem region (nucleotides 29,472–29,870) is predicted to have high Shannon entropy ($\text{median}_{\text{Nuc}(29,472\text{–}29,495;29,853\text{–}29,870)}$ = 0.2154), supporting the notion that this substructure is not well ordered in the cellular environment.

Structure prediction of the programmed ribosomal frameshifting element reveals conformational flexibility

One of the most well-studied RNA structures in the coronavirus coding region is the programmed ribosomal frameshifting pseudoknot (PRF). It is located between Orf1a and Orf1b and plays an important role in inducing a −1 frameshift in a translating ribosome, resulting in the synthesis of the polyprotein ab, which includes the SARS-CoV-2 replicase (Plant and Dinman, 2008).

The PRF element previously characterized in SARS-CoV is proposed to contain three parts: an attenuator stem (AS) loop, a conserved heptanucleotide ‘‘slippery’’ sequence (HSS), and a H-type pseudoknot (Plant and Dinman, 2008). We performed SHAPEKnots predictions (Hajdin et al., 2013) over four windows that cover the pseudoknotted region in the SARS-CoV-2 genome. We found that the pseudoknot is successfully predicted in three out of four windows tested. Moreover, the nucleotides predicted to be involved in the pseudoknotted helix have low SHAPE reactivity (Figure 2A, red lines). The frameshifting pseudoknot was thereafter included as a hard constraint during secondary structure prediction.

The most probable, dominant structure of the PRF region, extracted from the full-length *in vivo* secondary structure, is shown in Figure 2A. In our model, the SHAPE reactivity and Shannon entropy calculation support a well-folded AS immediately upstream of the HSS (Figures 2A and 2C). The AS has been demonstrated to be important for attenuating frameshifting in SARS-CoV (Cho et al., 2013), and previous reports suggested that the AS structure is not well conserved between SARS-CoV and SARS-CoV-2 (Kelly et al., 2020). By contrast, our results suggest a SARS-CoV-2-specific fold for the AS. The highly conserved HSS is predicted to be single stranded in our *in vivo* structural model, which is consistent with studies on other coronaviruses (Plant et al., 2005; Plant and Dinman, 2008).

Overall, the dominant structure predicted for the H-type pseudoknot in our structural model differs from the one proposed for SARS-CoV. In SARS-CoV-2, SL1 is well folded, as indicated by SHAPE reactivity mapping (Figure 2A) and Shannon entropy (Figure 2C). However, the region reported to contain the SL2 stem (Rangan et al., 2020, Plant et al., 2005) is predicted as single stranded in our consensus structure. Indeed, the dominant structure predicted for the PRF contains a different stem, which we designate SL3, and this includes the downstream pseudoknot arm (Figure 2A). The single-stranded region expected to contain SL2 is not well determined in our consensus structure, as indicated by Shannon entropy mapping to the region (Figure 2C).

As SuperFold calculates a partition function, low probability base-pairing interactions can be captured during structure prediction steps. We therefore checked the partition function output for alternative, low-probability base-pair interactions captured for the PRF region. We found that the single-stranded region (Figure 2A) forms base-pairing interactions with as many as six different regions in the SARS-CoV-2 genome (Figure S2A). Among these possible interactions is a PRF structure containing the three-stemmed pseudoknot conformation identified across coronaviruses, including a helical SL2 (Figure 2B) (Plant et al., 2005). The median base-pairing probability calculated for SL2 is 20% (Figure 2D; individual base pairs indicated with gray dots). In contrast, the SL3 stem is predicted to form with at least 80% base-pairing probability.

The apparent pairing promiscuity and low SHAPE reactivities within the SL2 region suggests that the PRF region has complex conformational dynamics that are not accurately represented by the single, static structures calculated in SuperFold. We reasoned that explicit modeling of the structural ensemble of the PRF region would reveal more information about the architecture and distribution of actual structural isoforms. To that end, we recalculated the partition function for a 749-nt window in the SARS-CoV-2 genome that surrounds the PRF (Figure S2A). This partition function calculation was then inserted into an ensemble structure modeling framework implemented within RNAstructure (Ding and Lawrence, 2003; Ding et al., 2005; Spasic et al., 2018).

Using this mode of analysis, a single conformational cluster overwhelmingly dominates the PRF conformational ensemble. As implied by our previous analysis (Figure 2A), this conformational cluster contains the AS, a single-stranded HSS, SL1, the

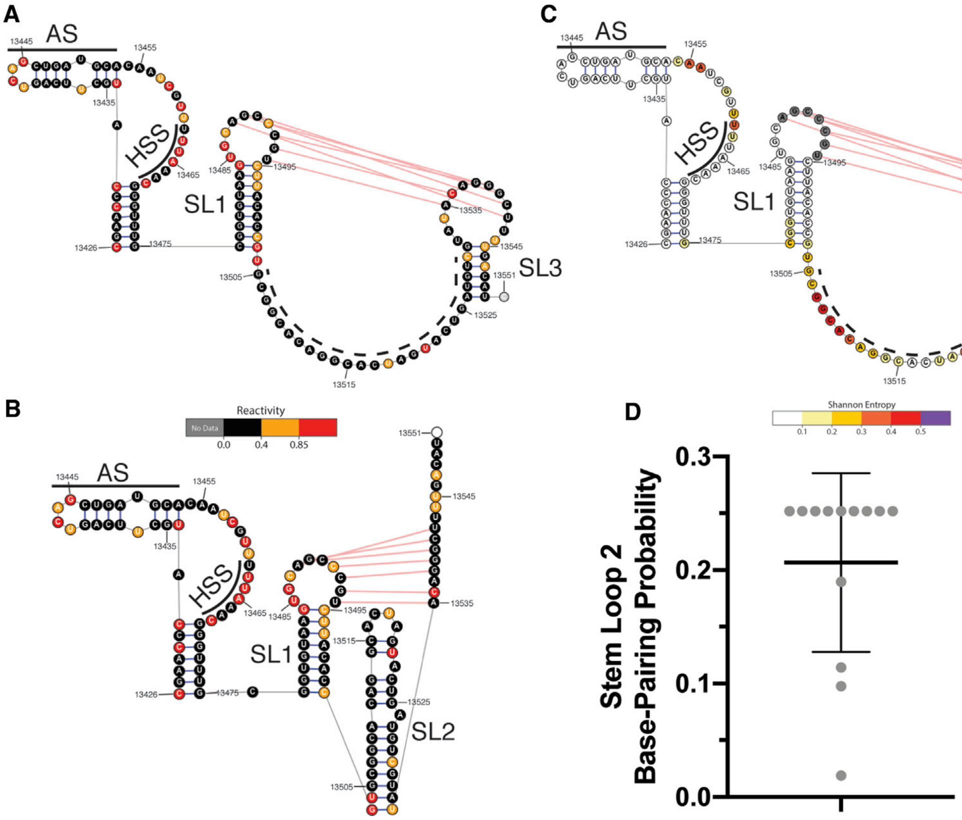


Figure 2. Structure prediction of the programmed ribosomal frameshifting pseudoknot (PRF) suggests conformational variability of stem loop 2

(A) Dominant PRF structural architecture colored by SHAPE reactivity. AS, attenuator stem; HSS, heptanucleotide slippery sequence; SL1, stem loop 1; SL3, stem loop 3. Dotted line indicates region that forms stem loop 2 (SL2) or long-range interactions outside the PRF, and red lines indicate pseudoknot interaction.

(B) Lower probability PRF conformation, with fully formed SL2, colored by SHAPE reactivity.

(C) Dominant PRF structure prediction colored by Shannon entropy, labeled as in (A).

(D) Base-pairing probability for alternate SL2 conformation. Each dot represents an individual base pair in SL2, plotted as in Figure 1C (inset).

pseudoknotted helix, and SL3. However, the SL2 region is base-paired with a region located 470 nt upstream (Figure S2B). The second-best populated cluster contains a nearly identical domain architecture, except that the SL2 region is base-paired with a region 260 nt upstream (Figure S2C). Together, these two clusters represent 99.2% of the PRF conformational ensemble. The least populated cluster is the one that contains the SL2 region imbedded in the canonical three-stemmed pseudoknot conformation, representing 0.8% of the PRF conformational ensemble (Figure S2D).

Taken together, these data suggest that the PRF of SARS-CoV-2 in infected cells includes a well-folded AS, SL1, and the pseudoknot helix but that the region containing the putative SL2 is conformationally variable, with the potential to form a diversity of long-range interactions. Therefore, the three-stem pseudoknot conformation that is conventionally used to characterize β-coronavirus PRFs represents a minority conformation for the SARS-CoV-2 PRF.

The secondary structure of SARS-CoV-2 Orf1ab reveals a network of RNA structural elements

While the successful identification of known, functional RNA structural elements lends strong support for our methodology and for the overall secondary structural model, these known regions account for only 3% of the total nucleotide content of the SARS-CoV-2 genome; little is known about the remaining 97%.

Here, we report the *in-vivo*-derived, SHAPE-constrained secondary structural model that includes a description of the base-pairing interactions for all nucleotides within a coronavirus genome (Figure 3A). To check whether our secondary structure model is in good agreement with experimentally determined *in vivo* SHAPE reactivities, we analyzed the normalized reactivities of each nucleotide separated by strandedness as determined in our model. We observe that, for all four nucleobases, single-stranded nucleotides have significantly higher reactivities than their double-stranded counterparts, which reflects the high quality of the model (Figure S3) (Siegfried et al., 2014, Guo et al., 2020). Representative secondary structural maps of small regions extracted from the consensus prediction exemplify the types of substructures that are observed in the SARS-CoV-2 Orf1ab (Figure 3B).

To discover additional, well-folded RNA structures within the SARS-CoV-2 genome, we calculated the local median Shannon

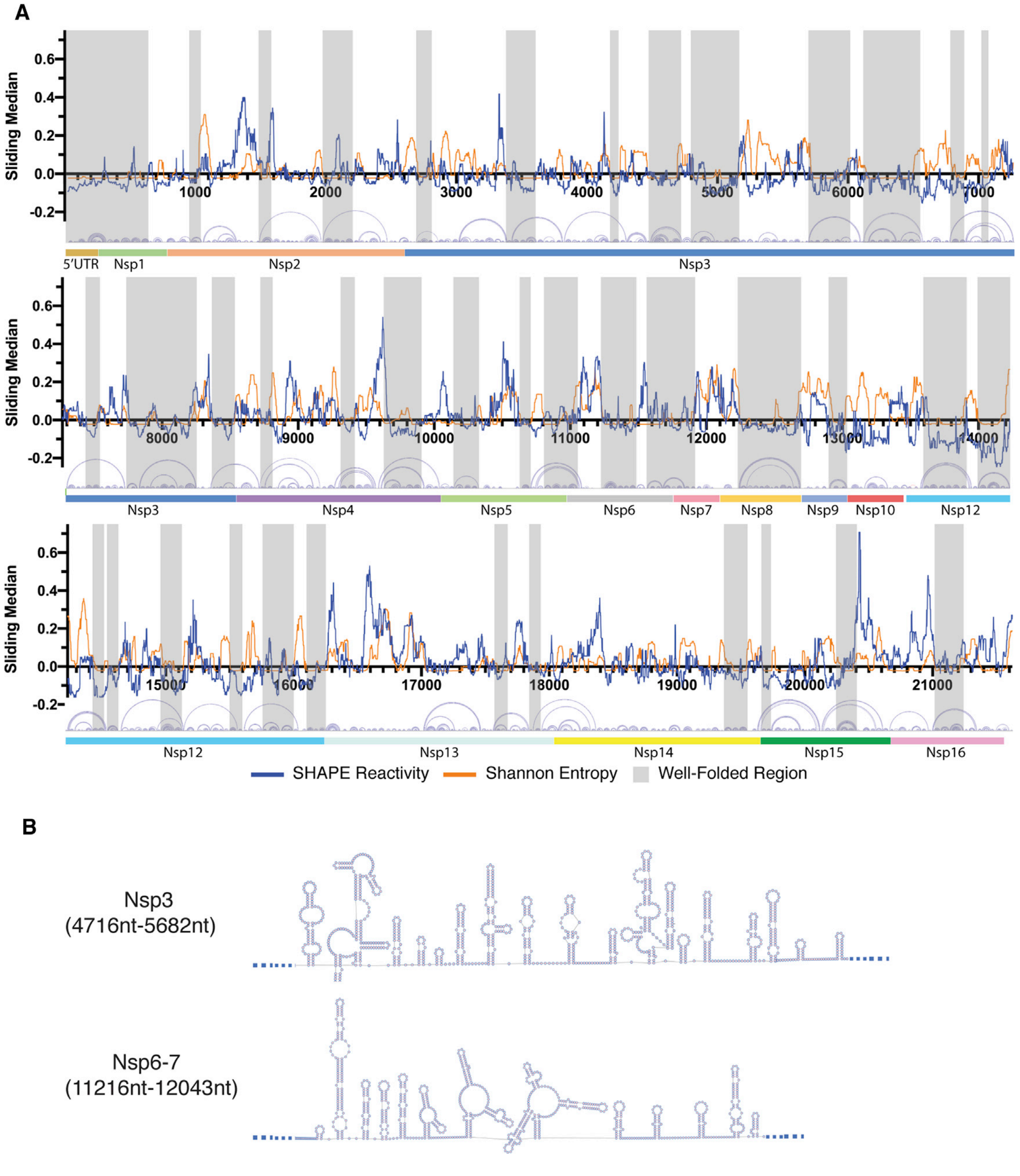


Figure 3. Full-length genome structure prediction of SARS-CoV-2 Orf1ab reveals a network of well-folded regions

(A) Analysis of Shannon entropy and SHAPE reactivities reveals 40 highly structured, well-determined domains in Orf1ab. Nucleotide coordinates are indicated on the x axis. Local median SHAPE reactivity and Shannon entropy are indicated by blue and orange lines, respectively. Well-folded regions are shaded with gray boxes. Arc plots for predicted base-pairing interactions in the structural model are shown below the x axis. The 5′ UTR and nonstructural protein (Nsp) domains are indicated by colored bars underneath arc plot diagrams.

(B) Representative secondary structure predictions of two regions extracted from the full-length consensus structure generated for the SARS-CoV-2 genome.

entropy and correlated these values with experimentally determined SHAPE reactivities (Figure 3A). Only regions with both median Shannon entropy and SHAPE reactivity signals below the global median for stretches longer than 40 nt that appear in both replicate datasets were considered well determined and stable. In total, we identify 40 such regions in Orf1ab (Figure 3A, shaded). Hereafter, any structured region that meets these above criteria will be referred to as ''well folded.''

To understand architectural organization of the overall ''structuredness,'' or base-pair content (BPC) within Orf1ab, we calculated the double-strand content of individual protein domains within this region of the genome (Figure 4A, gray bars). We find

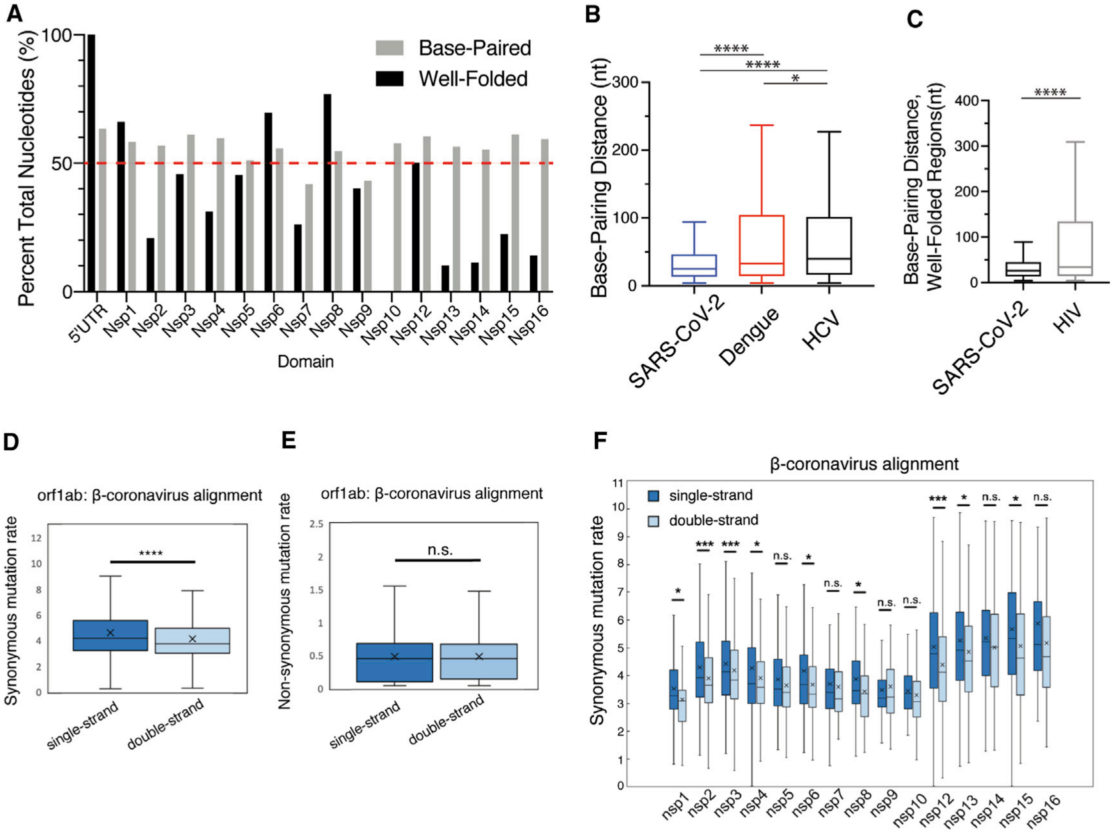


Figure 4. Full-length genome structure prediction of SARS-CoV-2 Orf1ab reveals unique and conserved genome architecture

(A) Base-paired RNA content (gray bars) and well-folded RNA content (black bars) of individual Nsp domains. A dotted line at 50% nucleotide content has been added for clarity.

(B) Median base-pairing distance of the SARS-CoV-2, Hepatitis C virus (HCV), and Dengue virus. Data are plotted as in Figure 1C (inset), though outliers are excluded.

(C) Median base-pairing distance across well-folded regions identified in SARS-CoV-2 and HIV genomes, plotted as in (B).

(D) Synonymous mutation rates (dS) calculated across β-coronaviruses for single- and double-stranded nucleotides of Orf1ab.

(E) Nonsynonymous mutation rates (dNs) calculated across all β-coronaviruses for single- and double-stranded nucleotides of Orf1ab.

(F) Comparison of dS for single- and double-stranded nucleotides within individual protein domains, calculated across all β-coronaviruses. (D-F) Data are plotted as in Figure 1B.

n.s., not significant; *p < 0.05; ***p < 0.001; ****p < 0.0001 by equal variance unpaired Student's t test.

that all protein domains have comparable BPC, with an average of 56% (±6.09%) of nucleotides involved in base-pairing interactions. However, the RNA sequences within each protein domain are not equivalently well folded (Figure 4A, black bars). For example, we observe that >50% of nucleotides within the 5′ UTR, Nsp1, Nsp6, Nsp8, and Nsp12 are concentrated in well-folded regions, suggesting these domains may be hubs for regulatory RNA structures. By contrast, Nsp13, Nsp14, and Nsp16 have <15% of their nucleotide content in discretely well-folded regions. At the most extreme end, Nsp10 contains no nucleotides in well-folded regions.

While analyzing the resulting secondary structural map, we noticed that the SARS-CoV-2 genome contains long stretches of short, locally folded stem loops (for example, see Figure 3B) with few long-distance base-pairing interactions. To determine if this was a quantifiable feature unique to the SARS-CoV-2 genome, we calculated the distance between base-paired nucleotides for every base-pairing interaction in our SARS-CoV-2 structural model. We compared these base-pairing distances to those we calculated from published full-length structural models for HCV (Mauger et al., 2015) and Dengue virus (Dethoff et al., 2018) that used the same structure prediction pipeline and constraints. Interestingly, the median base-pairing distance in our SARS-CoV-2 consensus model is 25 nt and is significantly smaller than the median base-pairing distance in the HCV (median, 40 nt) and Dengue virus (median, 33 nt) consensus models (Figure 4B). This suggests SARS-CoV-2 has fewer long-distance base-paring interactions compared to Dengue and the HCV genome.

We also calculated the median base-pairing distance for the well-folded regions of the SARS-CoV-2 genome and compared the result to well-folded regions previously identified using the same low Shannon/low SHAPE signatures in the HIV genome (Siegfried et al., 2014). We found that although there is no significant difference in the size of well-folded regions in the SARS-CoV-2 and HIV genomes (data not shown), the median base-pairing distance in the well-folded regions of SARS-CoV-2 (median, 26 nt) is significantly lower than the base-pairing distance in well-folded regions of HIV (median, 34 nt) (Figure 4C).

Taken together, these results suggest that the SARS-CoV-2 genome folds into a series of local secondary structures, and it contains fewer long-range base-pairing interactions than observed for positive-sense RNA viruses for which full-length genome structure predictions are available. Given the exceptional size of the coronavirus genome (~30 kb) relative to those of the positive-sense RNA viruses compared here (~10 kb), it is possible that the short base-pairing distance of SARS-CoV-2 may carry functional implications for maintaining genomic stability, preserving fidelity of translation, and evading innate immune response.

The overall structuredness of the SARS-CoV-2 genome is conserved across β-coronaviruses

Synonymous mutation rates (dS) have been used previously to lend evolutionary support for well-folded RNA secondary structures in other positive-sense RNA viruses (Dethoff et al., 2018; Tuplin et al., 2002; Assis, 2014; Simmonds and Smith, 1999). This body of work has suggested lower dS for double-stranded nucleotides when compared to single-stranded nucleotides, likely reflecting an evolutionary pressure to maintain base-pairing interactions of double-stranded nucleotides. We therefore computed relative dS to determine how evolutionary pressure is applied to single- and double-stranded regions of the SARS-CoV-2 genome.

Using an "all β-coronavirus" alignment, we observed a significantly lower dS for double-stranded codons when compared to single-stranded codons in our consensus model (Figure 4D). In contrast, there was no significant difference observed for nonsynonymous mutation rates (dNs) at single- or double-stranded codons (Figure 4E), as dN reflects changes at the amino acid level. This suggests that double-stranded regions of the SARS-CoV-2 genome experience stronger selective pressure against synonymous mutations than single-stranded regions. Because an all-β-coronavirus alignment was used, our results indicate that the structural organization and overall base-pairing content of Orf1ab is a conserved feature of the β-coronavirus family.

When analyzing relative dS within individual protein domains, we observed significantly decreased dS for double-stranded codons in Nsp1, Nsp2, Nsp3, Nsp4, Nsp6, Nsp8, Nsp12, Nsp13, and Nsp15 (Figure 4F). Consistent with this, Nsp1, Nsp6, Nsp8, and Nsp12 have >50% of their nucleotides localized within well-folded regions (Figure 4A, black bars). Taken together, this suggests that certain protein-coding domains contain regions of RNA secondary structure that are conserved across β-coronaviruses. For example, Nsp8, which is the most well-folded domain in SARS-CoV-2, is likely well-folded in other β-coronaviruses.

By contrast, the base-pairing content of Nsp5, Nsp7, Nsp9, Nsp10, Nsp14, and Nsp16 does not appear to be conserved, as there is no significant difference in dS (Figure 4F). Consistent with this, Nsp14 and Nsp16 were shown to have <15% of their nucleotides in well-folded regions, while Nsp10 does not contain any well-folded nucleotides (Figure 4A). This analysis supports the observation that these regions of RNA are not well folded in SARS-CoV-2, and our data suggest that these regions may not be well folded in other β-coronaviruses.

Evolutionary analysis for individual well-folded regions of the SARS-CoV-2 genome identifies several conserved regions

To further prioritize structural elements that may have conserved functional roles in the SARS-CoV-2 life cycle, we next applied our dS analysis to each of the 40 discrete well-folded domains (Figure 3A; Table S1). Four regions showed significantly decreased dS at double-stranded codons across the β-coronavirus alignment (Figures 5A and 5B). Among those well-folded domains, regions 25 and 34 are found at protein domain boundaries. Region 25 ends exactly at the Nsp8/9 domain boundary, while region 34 spans the Nsp12/13 boundary. Regions 23, 34, and 36 (Figure 5C, 5E, and 5F) contain a series of stem loops with small bulges. Region 25 contains a long-range duplex that closes a clover-leaf like structure with eight stem loops radiating from a central loop (Figure 5D). This hub, or multi-helix junction, might represent a promising drug target, as multi-helix junctions often contain binding pockets with high binding affinity and selectivity for small molecules (Warner et al., 2018).

Within the sarbecovirus subgenus, we were able to identify five well-folded regions with significantly decreased dS in double-stranded codons (Figures 6A and 6B). Among these well-folded domains, region 24 contains two discrete multi-helix junctions, each with at least three stems radiating from large central loops (Figure 6C). Region 27 contains a series of six stem loops (Figure 6D). Region 15, like region 24, contains several well-determined long-range duplexes that segment the region into two discrete multi-helix junctions (Figure 6E). Region 22 contains a series of well-folded loops, and it spans the Nsp5/6 boundary (Figure 6F). Region 30 is a single stem loop with bulges that divide the stem into distinct duplexes (Figure 6G)

To look for evolutionary evidence that directly supports conservation of specific base-pairing interactions and secondary structures, we performed covariation analysis on the five well-folded regions that are supported by sarbecovirus-specific dS. We identified three regions (15, 22, and 30) that have covariation support (Figures 6E–6G; covarying pairs shaded green). Taken together, these results suggest the existence of stable, evolutionarily conserved structural elements that merit subsequent functional analysis.

Functional validation of candidate structures by targeted LNA disruption

To provide a rapid method for evaluating the functional significance of predicted RNA structures, we developed an antisense-based reporter method that relies on the use of LNAs to disrupt putative structures within the genome. An infectious clone of SARS-CoV-2 with mNeonGreen inserted into Orf7 was used to monitor viral growth (Xie et al., 2020). LNAs are non-natural base analogs that enhance the T_m of a given paired duplex by 2°C–8°C for each LNA nucleotide(Lundin et al., 2013), enabling them to dominate over competing RNA-RNA duplexes. This strategy has been successfully deployed to study functional

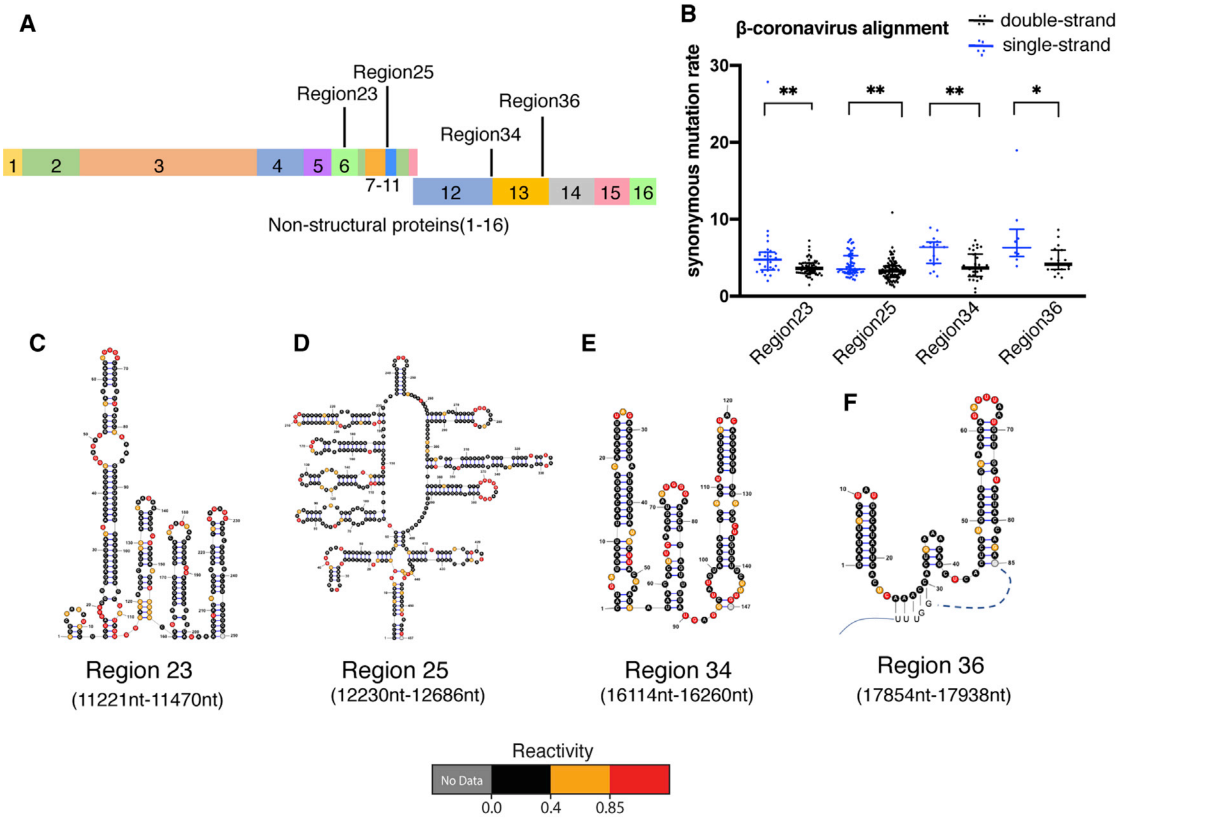


Figure 5. Analysis of dS within individual well-folded regions of the SARS-CoV-2 genome across β-coronaviruses

(A) Schematic of well-folded regions in SARS-CoV-2 genome supported by dS analysis in β-coronaviruses.

(B) dS separated by strandedness in four individual well-folded regions. Data are plotted as in Figure 1C (inset). *p < 0.05, **p < 0.01 by equal variance unpaired Student's t test.

(C–F) RNA secondary structure diagrams of four well-folded regions with dS support, colored by SHAPE reactivities, with genomic coordinates indicated below and in (A).

RNA structures in both the HCV and Dengue virus genomes (Dethoff et al., 2018, Tuplin et al., 2015).

For functional targets, we focused on two well-folded ORF regions, 15 and 22, each of which has strong evolutionary support (Figures 6E and 6F). LNAs targeted to these regions were designed for maximal structure disruption, hybridizing to the top of the stem loop as well as duplex RNA flanking the loop (Figures 7A and 7B, red lines). Importantly, we also designed a negative control that targets high Shannon entropy regions immediately downstream of each well-folded region but still within the ORF (Figures 7A and 7B, blue lines). We do not expect hybridization of negative control LNAs to have an effect on viral growth unless overall translation is disrupted. We included a scrambled LNA that should not bind to the SARS-CoV-2 genome as a global negative control.

As shown in Figure 7D, the LNA targeting the covarying stem in region 15 results in a 40% decrease in GFP$^+$ cells when compared to the region 15 control and a 35% decrease when compared to the scrambled LNA control. The region 15 control LNA has no effect on viral growth relative to the scrambled LNA control. A similar trend is observed for region 22 (Figure 7E). The LNA targeting the stem within region 22 results in a 22% decrease in GFP$^+$ cells when compared to the region 22 LNA control and a 30% reduction when compared to the scrambled LNA control. As before, there is no significant difference observed between the region 22 control and the scrambled LNA control.

Our structural modeling of the PRF suggests it contains a conformationally flexible SL2. In order to evaluate the functional importance of SL2, we tested whether an LNA targeted against the SL2 region resulted in a measurable defect in viral growth (Figure 7C; red line). In addition, we designed an LNA targeted against the PRF pseudoknot (SL1) (Figure 7C, blue line), as disruption of the SARS-CoV PRF has been demonstrated to reduce viral growth (Plant et al., 2005, 2013). This LNA results in an 18% reduction in GFP$^+$ cells relative to the scrambled LNA control (Figure 7F). Interestingly, the LNA targeted against the SL2 region results in a 17% decrease in GFP$^+$ cells when compared to the scrambled LNA control.

Taken together, our data suggest that RNA stem loops in regions 15 and 22 play functional roles in the SARS-CoV-2 viral life cycle, as their disruption results in a significant decrease in GFP$^+$ cells. Even more, these data lend strong support for a model in which well-folded regions with evolutionary support

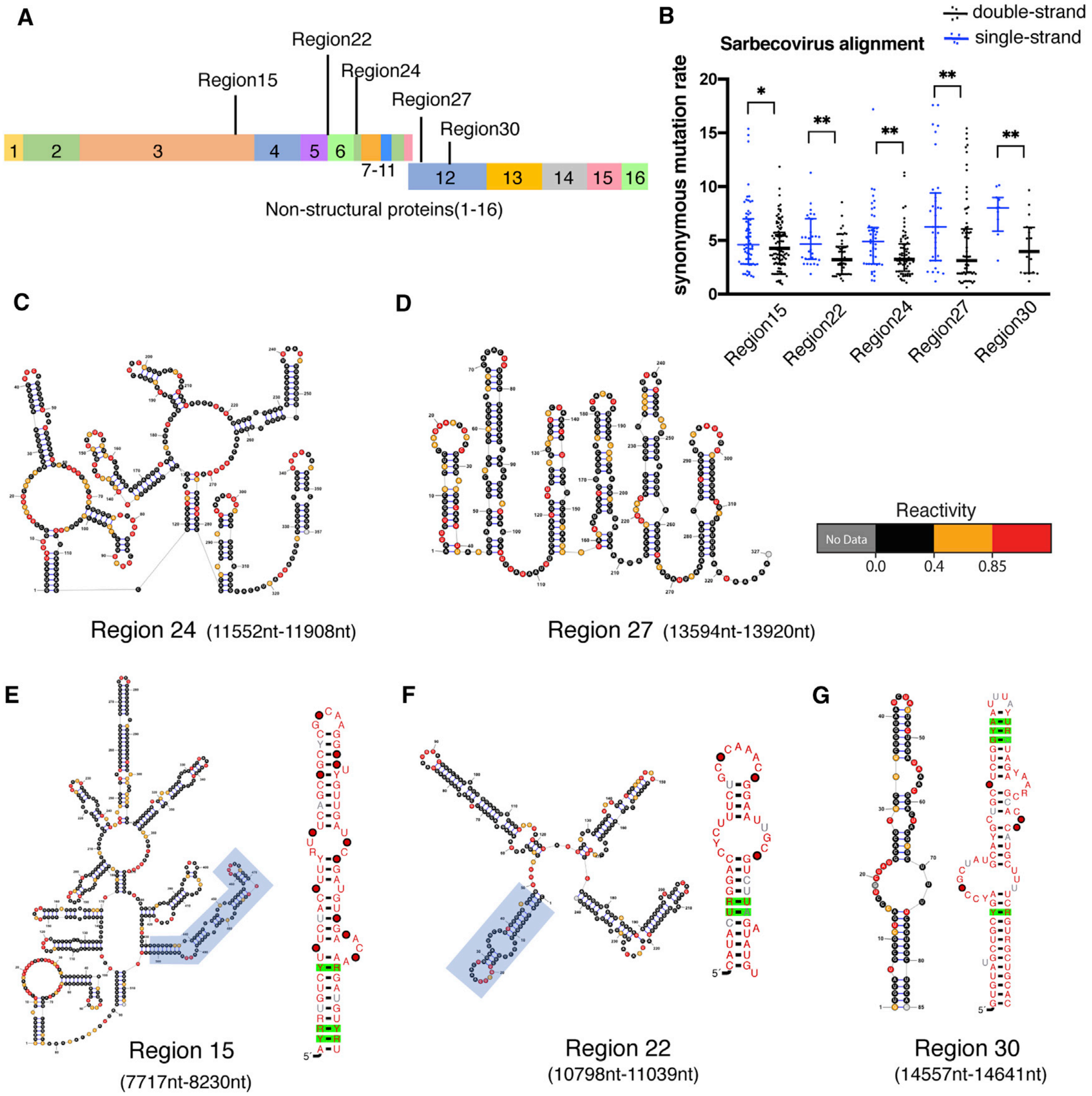


Figure 6. Analysis of dS and covariation within individual regions of the SARS-CoV-2 genome within the sarbecovirus subgenus

(A) Schematic of well-folded regions in the SARS-CoV-2 genome supported by dS analysis.

(B) dS separated by strandedness in five individual well-folded regions. Data are plotted as in Figure 1C (inset). *p < 0.05, **p < 0.01 by equal variance unpaired Student's t test.

(C and D) RNA secondary structures of two well-folded regions colored by SHAPE reactivity.

(E–G) RNA secondary structure diagrams of three well-folded regions supported by both dS analysis and covariation in sarbecoviruses, colored by SHAPE reactivities. Green boxes indicate significantly covarying base pairs tested by Rscape-RAFSp (e-value < 0.05).

Consensus nucleotides are colored by degree of sequence conservation (75% = gray; 90% = black; 97% = red). Circles indicate positional conservation and percentage occupancy thresholds (50% = white; 75% = gray; 90% = black; 97% = red).

represent hubs of regulatory RNA secondary structures. Finally, our data confirm that both the PRF pseudoknot and base-pairing interactions involving the SL2 region are crucial for viral growth.

DISCUSSION

Here, we establish that the SARS-CoV-2 genomic RNA has a complex molecular architecture, filled with elaborate secondary and tertiary structural features that persist *in vivo* and are conserved through time, suggesting that this network of RNA secondary structural elements plays a functional role in the virus life cycle. This RNA secondary structural complexity is not just confined to UTRs of the genome, as protein-coding sections of the SARS-CoV-2 ORF are among the most well-structured regions. Thus, as observed for HCV, coronavirus reading frames experience evolutionary pressure that simultaneously shapes both protein sequence and the surrounding RNA structures in which the proteins are encoded (a "code within the code")

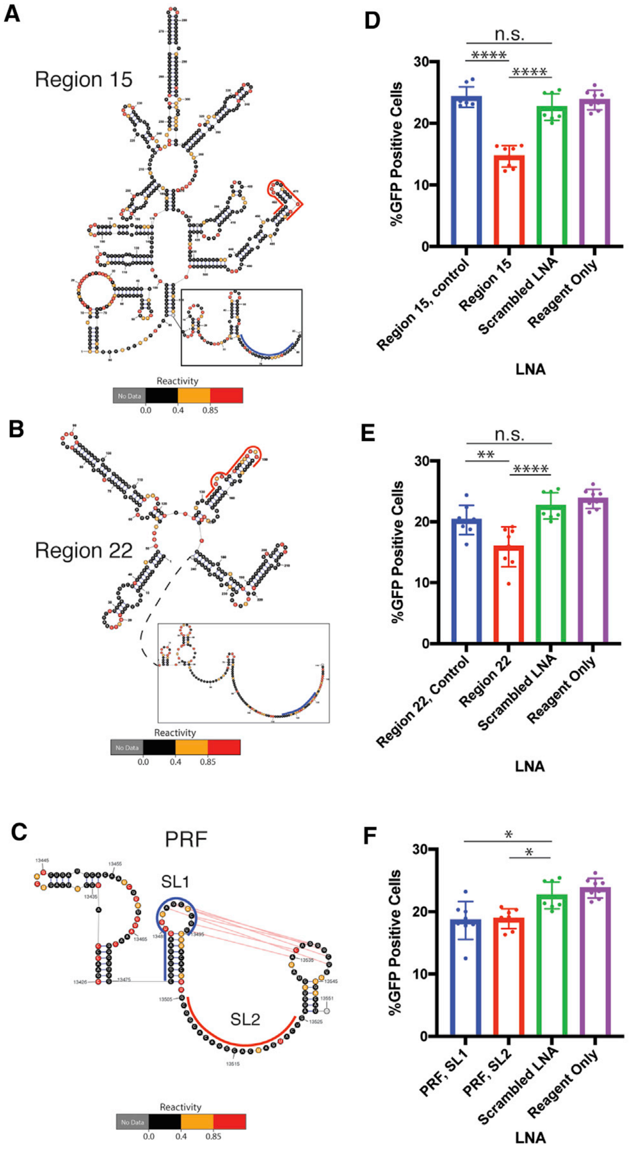


Figure 7. RNA structures disrupted by locked nucleic acids (LNAs) exhibit defects in SARS-CoV-2 viral growth

(A) Schematic showing region 15 LNA targeted to the covarying stem (red line) and control LNA (blue line).

(B) Schematic showing region 22 LNA targeted to stem (red line) and the control LNA (blue line).

(C) Schematic showing LNA targeted to the PRF SL1 region and the conformationally flexible SL2 region in the SARS-CoV-2 PRF.

(D–F) Virus growth as measured and quantified by mNeonGreen expression at 24 hours post-infection. All LNAs were tested concurrently and are split into subpanels for clarity. The same negative controls (scrambled LNA, reagent only) are shown in all subpanels for comparison. Data are plotted as in Figure 1C (inset). Individual data points represent technical replicates. n.s., not significant; *p < 0.05, **p < 0.01, ****p < 0.0001 by ordinary one-way ANOVA with multiple comparisons.

(Pirakitikulr et al., 2016). The secondary structure that we report is well determined based on available metrics in the field (Siegfried et al., 2014). It is both a roadmap for navigating the vast RNA landscape in coronaviruses and a resource for orthogonal studies by others. As such, the data reported here are all publicly available for analysis and comparison by others: https://github.com/pylelab/SARS-CoV-2_SHAPE_MaP_structure.

Well-determined secondary structures of long RNA molecules are typically difficult to obtain *in vivo* (Mitchell et al., 2019; Leamy et al., 2016). Experimental secondary structures are usually derived from transcripts that have been refolded and probed *in vitro* or isolated cellular transcripts that have been stripped of cellular components (Smola et al., 2015a; Siegfried et al., 2014). What is particularly surprising about this SARS-CoV-2 study, as well as the high quality of the resulting secondary structure, is the fact that it was entirely determined *in vivo*, using infected cells that were treated directly with chemical probes. The success of this effort is likely attributable to the fact that SARS-CoV-2 genomic RNA is so abundant in the infected cell, ultimately becoming ~65% of the total cellular RNA (Kim et al., 2020). The abundance of SARS-CoV-2 RNA may overwhelm the cell's ability to coat transcripts with nonspecific RNA-binding proteins, which can otherwise limit accessibility of chemical probes. That said, it will be interesting to compare the consensus structure reported here with that obtained "*ex vivo*" (stripped of protein), as the ΔSHAPE approach provides a useful way to flag possible protein-binding sites (Smola et al., 2015a).

The resulting experimental secondary structure provides new insights into known coronaviral RNA motifs and leads to the prediction of new ones that are likely to regulate viral function. The near-perfect structural homology of motifs at the 5′ terminus of SARS-CoV-2 with other β-coronavirus genomes suggests that the function of these upstream elements is conserved in coronaviruses (Yang and Leibowitz, 2015). Furthermore, because our SARS-CoV-2 secondary structure was determined *in vivo*, our findings validate previous coronavirus structural models of 5′ elements, as our data were obtained in a biologically relevant context.

Our SARS-CoV-2 secondary structure at the 3′ viral terminus largely agrees with previous studies on other β-coronavirus genomes (Yang and Leibowitz, 2015). However, our model of the 3′ viral terminus deviates in one important way. Neither the raw SHAPE reactivity data nor the subsequent secondary structure prediction supports formation of a pseudoknot proposed between the base of the BSL and SL1. Similarly, data collected in a recent study of SARS-CoV-2 RNA interactions in infected cells does not support pseudoknot formation (Ziv et al., 2020). Indeed, the putative pseudoknot conformation is mutually exclusive with the well-structured stem that we report at the base of the BSL. However, both conformations are proposed to be essential in MHV (Goebel et al., 2004), so it is possible that the pseudoknot exists as a minority conformation or is transiently folded in SARS-CoV-2.

Arguably the best-studied structural element in coronaviruses is the PRF. Required for proper replicase translation in all coronavirus family members, the PRF adopts different conformations in the various coronaviruses, including three-stemmed, two-stemmed, and kissing-loop pseudoknots (Baranov et al., 2005; Plant and Dinman, 2008). The core of the SARS-CoV PRF, which shares an almost identical sequence with SARS-CoV-2, is predicted to form a three-stem pseudoknot comprised of SL1, SL2, and a pseudoknotted helix, with an additional upstream AS that is poorly conserved in SARS-CoV-2 (Kelly et al., 2020). Our SHAPE reactivity and structure prediction are consistent with the existence of an AS, SL1, and the pseudoknot. However, our consensus model suggests that the region containing SL2 is conformationally flexible. When the PRF is modeled explicitly as a conformational ensemble, the three-stemmed pseudoknot of the SARS-CoV-2 PRF appears as a minority conformation. Consistent with our reported distribution of structural isoforms, Kelly et al. use a reporter assay to confirm that frameshifting mediated by the SARS-CoV-2 PRF occurs in a minority of read-through events by the ribosome (Kelly et al., 2020), indicating that the observed conformational variability of SL2 may be functional. Indeed, SL2 might function like a switch; when SL2 is formed (a minority of the time), frameshifting occurs, but when unfolded or forming base pairs with structures outside the PRF region, frameshifting would not occur. LNA hybridization results in this region are consistent with this model. However, further studies are required to fully explore the relationship between SL2 formation and SARS-CoV-2 frameshifting efficiency.

The study reported here provides a structure prediction for every nucleotide in the SARS-CoV-2 genome, enabling us to simultaneously interrogate both global and local features of genome architecture. One can make two major observations about the global architecture the SARS-CoV-2 genome. First, this *in-vivo*-derived, SHAPE-constrained model strongly agrees with the high double-strand RNA content predicted from the entirely *in silico* model recently reported by our lab (Tavares et al., 2020). Because the data herein were obtained *in vivo*, this work confirms that the unusually high double-strand content is maintained in a cellular context. Second, analysis of the experimental secondary structure reveals that the SARS-CoV-2 genome has a shorter median base-pairing distance when compared with other positive-sense RNA viruses for which full-length genome structure predictions are available, suggesting a role for extreme compaction in the function of coronaviral genomes. Downstream analysis of dS suggests that global architectural features are conserved across β-coronaviruses. Considering the exceptional size of these genomes, the high degree of double-stranded RNA (dsRNA) content may represent an evolutionary strategy to enhance genome stability, as duplex RNA undergoes self-hydrolysis at a much slower rate than single-stranded RNA and it is more resistant to cellular nucleases (Regulski and Breaker, 2008; Wan et al., 2011). Interestingly, single-stranded regions in mRNA have been shown to mediate phase separation at high cellular RNA concentrations (Van Treeck et al., 2018). Because SARS-CoV-2 RNA is very abundant *in vivo*, it is possible the high dsRNA content may provide a strategy to avoid phase separation during infection. The preference for abundant locally folded, short stem-loop structures in β-coronavirus genomes may also provide a conserved strategy for innate immune evasion. Pattern recognition receptors such as MDA5 (Dias Junior et al., 2019) and ADAR modification (Nishikura, 2010) proteins recognize long RNA duplexes as part of host defense processes, which could obviously be avoided by keeping duplex lengths short.

Analysis of local features within the genome pinpoints 40 well-folded regions within the SARS-CoV-2 Orf1ab region. Of these 40 regions, four are conserved across all β-coronaviruses, and five are sarbecovirus specific. Four of the nine regions span boundaries between Nsps, which may have relevance for polyprotein translation. Previous studies have shown that RNA secondary structures can slow the rate of ribosome translocation (Chen et al., 2013), and ribosome stalling is known to be important for proper folding of nascent polypeptides (Collart and Weiss, 2020). Conserved, well-folded regions at protein domain boundaries may therefore slow or stall translocating ribosomes, thus allowing individual Nsps in the large Orf1a and Orf1ab polyproteins to fold into their native conformations.

Intriguingly, three of the nine well-folded regions contain complex, multi-helix junctions, or structural hubs. This is significant, because multi-helix junctions often comprise the core of RNA tertiary structures, like group II self-splicing introns, riboswitches, and other regulatory elements. Because these elements are likely to contain well-defined pockets, they often bind specifically to small molecules and therefore serve as possible drug targets (Warner et al., 2018; Hewitt et al., 2019; Fedorova et al., 2018; Parsons et al., 2009; Haniff et al., 2020).

To explore structure-function relationships of representative, conserved RNA secondary structures, we used targeted antisense LNAs to induce structure disruption. This method is not only faster than reverse genetics but also more scalable and can be used in cases for which genetic systems have not yet been optimized. Using this strategy, we showed that disruption of RNA stems in regions 15 and 22 result in significant inhibition of viral growth, indicating they likely play novel regulatory roles in the SARS-CoV-2 life cycle. Importantly, the magnitude of reduction we observe in these cases is the same as that reported for cases of pharmacological inhibition (∼30%) of the same ic-SARS-CoV-2-mNG construct (Son et al., 2020; Wei et al., 2020). This indicates that the LNAs developed in this study may themselves have potential as antiviral therapeutics.

The *in-vivo*-determined SARS-CoV-2 secondary structure presented here provides a roadmap for functional studies of the SARS-CoV-2 genome and insights into mechanisms of the SARS-CoV-2 life cycle. Evolutionary support for our consensus model across β-coronaviruses hints at conserved strategies for genome stability, translation fidelity, and innate immune evasion. Finally, the identification of individual well-folded regions conserved across β-coronaviruses, and within the sarbecovirus subgenus, provide potential targets for the study of regulatory elements and the search for much-needed therapeutically active small molecules.

Limitations

One important cautionary observation from our work is the poor correlation of SHAPE reactivities between two *in vivo* biological replicates for regions encoding the subgenomic RNAs. Previous *in silico* work from our lab has shown that individual subgenomic RNAs (sgRNAs), such as the N sgRNA, fold differently than the corresponding regions in the genomic RNA due to differences in upstream sequence context (Tavares et al., 2020). Though our tiled-amplicon design affords sequencing coverage for the entire SARS-CoV-2 genome, it precludes deconvolution of reactivity signals for regions shared between genomic and subgenomic RNAs. This underscores the need for methodological innovations that accurately assess the structural content specific to individual subgenomic RNA molecules. Absent such methodological advances, we caution others when interpreting reactivities from the subgenomic region from bulk sequencing data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Cell culture and SARS-CoV-2 infection
  - RNA probing and purification
  - Tiled-amplicon design
  - Reverse transcription with MarathonRT
  - SHAPE-MaP library construction
  - Structure prediction
  - Ensemble structure modeling for the PRF region
  - Identification of well-folded regions
  - Multiple sequence alignment
  - Synonymous mutation rate analysis
  - Covariation analysis
  - Design of antisense locked nucleic acids
  - LNA transfection and icSARS-CoV-2-mNG infection
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2020.12.041.

ACKNOWLEDGMENTS

We would like to thank Dr. Li-Tao Guo for preparing and sharing MarathonRT enzyme and Dr. Ananth Kumar and Gandhar Mahadeshwar (Pyle lab, Yale University) for thoughtful comments on the manuscript. We would also like to thank Dr. Mark Boerneke (Weeks Lab, UNC Chapel Hill) for providing data upon request. This work was supported by the Howard Hughes Medical Institute, the National Institutes of Health (grant R01 HG009622 to A.M.P. and grant T32AI055403 to N.C.H.), the China Scholarship Council (CSC)-Yale World Scholars Program in Biomedical Sciences (H.W.), and the National Science Foundation (fellowship DGE1752134 to M.S.S). Funding for open access charge was provided by the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

N.C.H., H.W., M.S.S., and C.B.W. conducted experiments. N.C.H., H.W., M.S.S., C.B.W., and A.M.P. designed experiments. N.C.H., H.W., R.d.C.A.T., and A.M.P. wrote the paper.

DECLARATION OF INTERESTS

A patent application on MarathonRT has been filed by Yale University.

Received: July 13, 2020
Revised: November 6, 2020
Accepted: December 21, 2020
Published: January 1, 2021

REFERENCES

Andrews, R.J., Peterson, J.M., Haniff, H.S., Chen, J., Williams, C., Grefe, M., and Disney, M.D. (2020). An in silico map of the SARS-CoV-2 RNA structurome. bioRxiv. https://doi.org/10.1101/2020.04.17.045161.

Assis, R. (2014). Strong epistatic selection on the RNA secondary structure of HIV. PLoS Pathog. *10*, e1004363.

Baranov, P.V., Henderson, C.M., Anderson, C.B., Gesteland, R.F., Atkins, J.F., and Howard, M.T. (2005). Programmed ribosomal frameshifting in decoding the SARS-CoV genome. Virology *332*, 498–510.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D., and Sayers, E.W. (2018). GenBank. Nucleic Acids Res. *46* (D1), D41–D47.

Busan, S., and Weeks, K.M. (2017). Visualization of RNA structure models within the Integrative Genomics Viewer. RNA *23*, 1012–1018.

Busan, S., and Weeks, K.M. (2018). Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA *24*, 143–148.

Ceraolo, C., and Giorgi, F.M. (2020). Genomic variance of the 2019-nCoV coronavirus. J. Med. Virol. *92*, 522–528.

Chen, S.C., and Olsthoorn, R.C. (2010). Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. Virology *401*, 29–41.

Chen, C., Zhang, H., Broitman, S.L., Reiche, M., Farrell, I., Cooperman, B.S., and Goldman, Y.E. (2013). Dynamics of translation by single ribosomes through mRNA secondary structures. Nat. Struct. Mol. Biol. *20*, 582–588.

Cho, C.P., Lin, S.C., Chou, M.Y., Hsu, H.T., and Chang, K.Y. (2013). Regulation of programmed ribosomal frameshifting by co-translational refolding RNA hairpins. PLoS ONE *8*, e62283.

Collart, M.A., and Weiss, B. (2020). Ribosome pausing, a dangerous necessity for co-translational events. Nucleic Acids Res. *48*, 1043–1055.

de Wit, E., van Doremalen, N., Falzarano, D., and Munster, V.J. (2016). SARS and MERS: recent insights into emerging coronaviruses. Nat. Rev. Microbiol. *14*, 523–534.

Dethoff, E.A., Boerneke, M.A., Gokhale, N.S., Muhire, B.M., Martin, D.P., Sacco, M.T., McFadden, M.J., Weinstein, J.B., Messer, W.B., Horner, S.M., and Weeks, K.M. (2018). Pervasive tertiary structure in the dengue virus RNA genome. Proc. Natl. Acad. Sci. USA *115*, 11513–11518.

Dias Junior, A.G., Sampaio, N.G., and Rehwinkel, J. (2019). A balancing act: MDA5 in antiviral immunity and autoinflammation. Trends Microbiol. *27*, 75–85.

Ding, Y., and Lawrence, C.E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. *31*, 7280–7301.

Ding, Y., Chan, C.Y., and Lawrence, C.E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA *11*, 1157–1166.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. *20*, 533–534.

Fedorova, O., Jagdmann, G.E., Jr., Adams, R.L., Yuan, L., Van Zandt, M.C., and Pyle, A.M. (2018). Small molecules that target group II introns are potent antifungal agents. Nat. Chem. Biol. *14*, 1073–1078.

Friebe, P., and Bartenschlager, R. (2009). Role of RNA structures in genome terminal sequences of the hepatitis C virus for replication and assembly. J. Virol. *83*, 11989–11995.

Goebel, S.J., Hsue, B., Dombrowski, T.F., and Masters, P.S. (2004). Characterization of the RNA components of a putative molecular switch in the 3′ untranslated region of the murine coronavirus genome. J. Virol. *78*, 669–682.

Goebel, S.J., Miller, T.B., Bennett, C.J., Bernard, K.A., and Masters, P.S. (2007). A hypervariable region within the 3′ cis-acting element of the murine coronavirus genome is nonessential for RNA synthesis but affects pathogenesis. J. Virol. *81*, 1274–1287.

Guo, L.T., Adams, R.L., Wan, H., Huston, N.C., Potapova, O., Olson, S., Gallardo, C.M., Graveley, B.R., Torbett, B.E., and Pyle, A.M. (2020). Sequencing and structure probing of long RNAs using MarathonRT: a next-generation reverse transcriptase. J. Mol. Biol. *432*, 3338–3352.

Hajdin, C.E., Bellaousov, S., Huggins, W., Leonard, C.W., Mathews, D.H., and Weeks, K.M. (2013). Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc. Natl. Acad. Sci. USA *110*, 5498–5503.

Haniff, H.S., Tong, Y., Liu, X., Chen, J.L., Suresh, B.M., Andrews, R.J., Peterson, J.M., O'Leary, C.A., Benhamou, R.I., Moss, W.N., and Disney, M.D. (2020). Targeting the SARS-CoV-2 RNA genome with small molecule binders and ribonuclease targeting chimera (RIBOTAC) degraders. ACS Cent. Sci. *6*, 1713–1721.

Hewitt, W.M., Calabrese, D.R., and Schneekloth, J.S., Jr. (2019). Evidence for ligandable sites in structured RNA throughout the Protein Data Bank. Bioorg. Med. Chem. *27*, 2253–2260.

Kelly, J.A., Olson, A.N., Neupane, K., Munshi, S., San Emeterio, J., Pollack, L., Woodside, M.T., and Dinman, J.D. (2020). Structural and functional conservation of the programmed −1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). J. Biol. Chem. Published online June 22, 2020. https://doi.org/10.1074/jbc.AC120.013449.

Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. Cell *181*, 914–921.e10.

Korbie, D.J., and Mattick, J.S. (2008). Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat. Protoc. *3*, 1452–1456.

Lai, D., Proctor, J.R., Zhu, J.Y., and Meyer, I.M. (2012). R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic Acids Res. *40*, e95.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., and Wang, X. (2020a). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature *581*, 215–220.

Lan, T.C.T., Allan, M., Malsick, L., Khandwala, S., Nyeo, S.Y., Bathe, M., Griffiths, A., and Rouskin, S. (2020b). Structure of the full SARS-CoV-2 RNA genome in infected cells. bioRxiv.

Leamy, K.A., Assmann, S.M., Mathews, D.H., and Bevilacqua, P.C. (2016). Bridging the gap between in vitro and in vivo RNA folding. Q. Rev. Biophys. *49*, e10.

Li, P., Wei, Y., Mei, M., Tang, L., Sun, L., Huang, W., Zhou, J., Zou, C., Zhang, S., Qin, C.F., et al. (2018). Integrative analysis of zika virus genome RNA structure reveals critical determinants of viral infectivity. Cell Host Microbe *24*, 875–886.e5.

Lu, Z.J., and Mathews, D.H. (2008). OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. Nucleic Acids Res. *36*, W104-8.

Lundin, K.E., Højland, T., Hansen, B.R., Persson, R., Bramsen, J.B., Kjems, J., Koch, T., Wengel, J., and Smith, C.I. (2013). Biological activity and biotechnological aspects of locked nucleic acids. Adv. Genet. *82*, 47–107.

Madhugiri, R., Karl, N., Petersen, D., Lamkiewicz, K., Fricke, M., Wend, U., Scheuer, R., Marz, M., and Ziebuhr, J. (2018). Structural and functional conservation of cis-acting RNA elements in coronavirus 5′-terminal genome regions. Virology *517*, 44–55.

Maier, H.J., Bickerton, E., and Britton, P. (2015). Coronaviruses: Methods and Protocols (Humana Press).

Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T., Marinus, T., Ogando, N.S., Snider, E.J., Van Hemert, M.J., Bujnicki, J.M., and Incarnato, D. (2020). Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Research *48*, 12436–12452.

Mathews, D.H. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA *10*, 1178–1190.

Mauger, D.M., Golden, M., Yamane, D., Williford, S., Lemon, S.M., Martin, D.P., and Weeks, K.M. (2015). Functionally conserved architecture of hepatitis C virus RNA genomes. Proc. Natl. Acad. Sci. USA *112*, 3692–3697.

Mitchell, D., 3rd, Assmann, S.M., and Bevilacqua, P.C. (2019). Probing RNA structure in vivo. Curr. Opin. Struct. Biol. *59*, 151–158.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., and Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. *30*, 1196–1205.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu. Rev. Biochem. *79*, 321–349.

Parsons, J., Castaldi, M.P., Dutta, S., Dibrov, S.M., Wyles, D.L., and Hermann, T. (2009). Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA. Nat. Chem. Biol. *5*, 823–825.

Pirakitikulr, N., Kohlway, A., Lindenbach, B.D., and Pyle, A.M. (2016). The coding region of the HCV genome contains a network of regulatory RNA structures. Mol. Cell *62*, 111–120.

Plant, E.P., and Dinman, J.D. (2008). The role of programmed-1 ribosomal frameshifting in coronavirus propagation. Front. Biosci. *13*, 4873–4881.

Plant, E.P., Pérez-Alvarado, G.C., Jacobs, J.L., Mukhopadhyay, B., Hennig, M., and Dinman, J.D. (2005). A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. PLoS Biol. *3*, e172.

Plant, E.P., Sims, A.C., Baric, R.S., Dinman, J.D., and Taylor, D.R. (2013). Altering SARS coronavirus frameshift efficiency affects genomic and subgenomic RNA production. Viruses *5*, 279–294.

Rangan, R., Zheludev, I.N., Hagey, R.J., Pham, E.A., Wayment-Steele, H.K., Glenn, J.S., and Das, R. (2020). RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. RNA *26*, 937–959.

Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Mol. Biol. Evol. *35*, 2582–2584.

Regulski, E.E., and Breaker, R.R. (2008). In-line probing analysis of riboswitches. Methods Mol. Biol. *419*, 53–67.

Reuter, J.S., and Mathews, D.H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics *11*, 129.

Rivas, E., Clements, J., and Eddy, S.R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat. Methods *14*, 45–48.

Robertson, M.P., Igel, H., Baertsch, R., Haussler, D., Ares, M., Jr., and Scott, W.G. (2005). The structure of a rigorously conserved RNA element within the SARS virus genome. PLoS Biol. *3*, e5.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*, 959–965.

Simmonds, P., and Smith, D.B. (1999). Structural constraints on RNA virus evolution. J. Virol. *73*, 5787–5794.

Simon, L.M., Morandi, E., Luganini, A., Gribaudo, G., Martinez-Sobrido, L., Turner, D.H., Oliviero, S., and Incarnato, D. (2019). In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. Nucleic Acids Res. *47*, 7003–7017.

Smola, M.J., Calabrese, J.M., and Weeks, K.M. (2015a). Detection of RNA-protein interactions in living cells with SHAPE. Biochemistry *54*, 6867–6875.

Smola, M.J., Rice, G.M., Busan, S., Siegfried, N.A., and Weeks, K.M. (2015b). Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat. Protoc. *10*, 1643–1669.

Smola, M.J., Christy, T.W., Inoue, K., Nicholson, C.O., Friedersdorf, M., Keene, J.D., Lee, D.M., Calabrese, J.M., and Weeks, K.M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proc. Natl. Acad. Sci. USA *113*, 10322–10327.

Son, J., Huang, S., Zeng, Q., Bricker, T.L., Case, J.B., Zhou, J., Zang, R., Liu, Z., Chang, X., Harastani, H.H., et al. (2020). Nitazoxanide and JIB-04 have broad-spectrum antiviral activity and inhibit SARS-CoV-2 replication in cell culture and coronavirus pathogenesis in a pig model. bioRxiv.

Spasic, A., Assmann, S.M., Bevilacqua, P.C., and Mathews, D.H. (2018). Modeling RNA secondary structure folding ensembles using SHAPE mapping data. Nucleic Acids Res. *46*, 314–323.

Tavares, R.C.A., Pyle, A.M., and Somarowthu, S. (2019). Phylogenetic analysis with improved parameters reveals conservation in lncRNA structures. J. Mol. Biol. *431*, 1592–1603.

Tavares, R.C.A., Mahadeshwar, G., Wan, H., Huston, N.C., and Pyle, A.M. (2020). The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. J. Virol., JVI.02190-20.

Tuplin, A., Wood, J., Evans, D.J., Patel, A.H., and Simmonds, P. (2002). Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. RNA *8*, 824–841.

Tuplin, A., Struthers, M., Cook, J., Bentley, K., and Evans, D.J. (2015). Inhibition of HCV translation by disrupting the structure and interactions of the viral CRE and 3′ X-tail. Nucleic Acids Res. *43*, 2914–2926.

Van Treeck, B., Protter, D.S.W., Matheny, T., Khong, A., Link, C.D., and Parker, R. (2018). RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. Proc. Natl. Acad. Sci. USA *115*, 2734–2739.

Wan, Y., Kertesz, M., Spitale, R.C., Segal, E., and Chang, H.Y. (2011). Understanding the transcriptome through RNA structure. Nat. Rev. Genet. *12*, 641–655.

Wan, Y., Shang, J., Graham, R., Baric, R.S., and Li, F. (2020). Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. *94*, 94.

Warner, K.D., Hajdin, C.E., and Weeks, K.M. (2018). Principles for targeting RNA with drug-like small molecules. Nat. Rev. Drug Discov. *17*, 547–558.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview version 2: a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Wei, J., Alfajaro, M.M., Deweirdt, P.C., Hanna, R.E., Lu-Culligan, W.J., Cai, W.L., Strine, M.S., Zhang, S.M., Graziano, V.R., Schmitz, C.O., et al. (2020). Genome-wide CRISPR screens reveal host factors critical for SARS-CoV-2 infection. Cell. Published online October 20, 2020. https://doi.org/10.1016/j.cell.2020.10.028.

Xie, X., Muruato, A., Lokugamage, K.G., Narayanan, K., Zhang, X., Zou, J., Liu, J., Schindewolf, C., Bopp, N.E., Aguilar, P.V., et al. (2020). An infectious cDNA clone of SARS-CoV-2. Cell Host Microbe *27*, 841–848.e3.

Yang, D., and Leibowitz, J.L. (2015). The structure and functions of coronavirus genomic 3′ and 5′ ends. Virus Res. *206*, 120–133.

Yin, W., Mao, C., Luan, X., Shen, D.D., Shen, Q., Su, H., Wang, X., Zhou, F., Zhao, W., Gao, M., et al. (2020). Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science *368*, 1499–1504.

You, S., Stump, D.D., Branch, A.D., and Rice, C.M. (2004). A cis-acting replication element in the sequence encoding the NS5B RNA-dependent RNA polymerase is required for hepatitis C virus RNA replication. J. Virol. *78*, 1352–1366.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al.; China Novel Coronavirus Investigating and Research Team (2020). A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. *382*, 727–733.

Ziv, O., Price, J., Kamenova, T., Goodfellow, I., Weber, F., and Miska, E.A. (2020). The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2. Molecular Cell *80*, 1067–1077, https://doi.org/10.1016/j.molcel.2020.11.004.

Zubradt, M., Gupta, P., Persad, S., Lambowitz, A.M., Weissman, J.S., and Rouskin, S. (2017). DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Methods *14*, 75–82.

Züst, R., Miller, T.B., Goebel, S.J., Thiel, V., and Masters, P.S. (2008). Genetic interactions between an essential 3′ cis-acting RNA pseudoknot, replicase gene products, and the extreme 3′ end of the mouse coronavirus genome. J. Virol. *82*, 1214–1228.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
SARS-CoV-2 isolate USA-WA1/2020	BEI Resources	#NR-52281
icSARS-CoV-2mNG	Xie et al., 2020	N/A
Chemicals, peptides, and recombinant proteins		
DMEM	ThermoFisher	11965118
DPBS	ThermoFisher	14190144
NAI	MilliporeSigma	03-310
DMSO	Sigma-Aldrich	276855-100ml
Trizol	ThermoFisher	15596018
Chloroform:Isoamyl alcohol 24:1	MilliporeSigma	C0459-1PT
MarathonRT	Kerafast	EYU007
AmpureXP beads	Beckman coulter	A63880
NEB next UltraIIQ5 Master Mix	NEB	M0544S
Monarch PCR & DNA Cleanup Kit	NEB	T1030S
Nextera XT DNA Library Preparation Kit	Illumina	FC-131-1096
Qubit^{TM} dsDNA HS Assay Kit	ThermoFisher	Q32851
RNeasy Mini kit	QIAGEN	74104
Bioanalyzer High Sensitivity DNA Kit	Agilent	5067-4626
Nextseq 500/550 Mid Output kit v2.5 (150 cycles)	Illumina	20024904
TransIT-Oligo Transfection Reagent	Mirus	MIR 2164
Deposited data		
Raw and analyzed data	This paper	GEO:GSE154171
Experimental models: cell lines		
VeroE6 cells	ATCC	CRL1586
Oligonucleotides		
Gene-specific RT primers	This paper	Table S2
Gene-specific PCR primers	This paper	Table S3
Locked nucleic acids	This paper	Table S4
Software and algorithms		
OligoWalk	Lu and Mathews, 2008	http://rna.urmc.rochester.edu/cgi-bin/server_exe/oligowalk/oligowalk_form.cgi
ShapeMapper2	Busan and Weeks, 2018	https://github.com/Weeks-UNC/shapemapper2
ShapeKnots	Hajdin et al., 2013	https://rna.urmc.rochester.edu/RNAstructure.html
SuperFold	Smola et al., 2015b	https://weekslab.com/software/
MACSE v2.0.3	Ranwez et al., 2018	https://bioweb.supagro.inra.fr/macse/
R-scape v0.2.1	Rivas et al., 2017	http://www.eddylab.org/R-scape/
R-CHIE	Lai et al., 2012	https://bioconductor.org/packages/release/bioc/html/R4RNA.html
Jalview v2.11.0	Waterhouse et al., 2009	https://www.jalview.org/
FUBAR	Murrell et al., 2013	http://www.datamonkey.org/fubar
GraphPad Prism	GraphPad	https://www.graphpad.com/scientific-software/prism/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
StructureEditor	Reuter and Mathews, 2010	https://rna.urmc.rochester.edu/RNAstructure.html
Other		
Sequence data, analyses, and resources related to SHAPEmap data collection and SARS-CoV-2 structure prediction	This paper	https://github.com/pylelab/SARS-CoV-2_SHAPE_MaP_structure

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Anna Marie Pyle (anna.pyle@yale.edu).

Materials availability

No unique reagents were generated in this study.

Data and code availability

All ShapeMapper outputs, secondary structure files, and multiple sequence alignments used in this work are available at the GitHub repository: https://github.com/pylelab/SARS-CoV-2_SHAPE_MaP_structure. SHAPE-MaP data has been deposited on the Gene Expression Omnibus (GEO) database under the accession GSE154171.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

To generate SARS-CoV-2 viral stocks, Huh7.5 cells were inoculated with SARS-CoV-2 isolate USA-WA1/2020 (BEI Resources #NR-52281) at an MOI of 0.01 for three days to generate a PI stock. The P1 stock was used to inoculate Vero-E6 (ATCC) cells for three days. Supernatant was harvest and clarified by centrifuging at 450 g for 5min. Clarified supernatant was filtered through a 0.45-micron filter, aliquoted, and stored at −80°C.

Virus titer was determined by plaque assay. VeroE6 cells were seeded at 7.5×10^5 cells/well in 6-well plates.The following day, media were removed and replaced with 100µL of 10-fold serially diluted viral stock. Plates were incubated at 37°C for 1 hour with gentle rocking. Following the incubation, each well was overlaid with overlay media (DMEM, 2%FBS, 0.6% Avicel RC-581). Two days post-infection, plates were fixed with 10% formaldehyde for 30min followed by staining with crystal violet solution (0.5% crystal violet in 20% EtOH) for 30min. After staining, wells were rinsed with deionized water to visualize plaques.

METHOD DETAILS

Cell culture and SARS-CoV-2 infection

VeroE6 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) with 10% heat-inactivated fetal bovine serum (FBS). Approximately 5×10^6 cells were plated in each of four T150 tissue culture treated flasks. The following day media was removed and 10^5 PFU in 4mL of media of SARS-CoV-2 isolate USA-WA1/2020 (BEI Resources #NR-52281) was added to each flask. Virus was adsorbed for 1 hour at 37°C and then 16mL of fresh media was added to each flask.

RNA probing and purification

Four days post-infection (dpi), the supernatant was aspirated from each flask, cells were washed with 10mL of cold PBS−/− and then dislodged in 10ml PBS−/− with a cell scraper. The contents were collected and centrifuged at 450 g x 5 min at 4°C. The supernatant was removed and the cell pellet was resuspended into 2ml of PBS−/− with 200µl DMSO or 2ml PBS with 200µl of 2M NAI (final concentration = 200mM). Cells were incubated for 10 minutes at room temperature followed by addition of 6mL of Trizol. RNA was extracted with the addition of 1.2mL of chloroform:isoamyl alcohol (24:1). The aqueous phase was transferred to a new tube, followed by the addition of 12mL of 100% EtOH (70% final) and incubated overnight at −20°C. RNA was pelleted at 20,000 g for 30min at 4°C, washed once with 70% EtOH, and spun again at 20,000 g for 15min at 4°C. RNA was resuspended in 1xME buffer and purified using the QIAGEN RNeasy kit according to the manufacturer's protocol. RNA was eluted in 1xME buffer (8mM MOPS, 0.1mM EDTA, pH 6.5).

Tiled-amplicon design

Leveraging the extreme processivity of MarathonRT, a highly processive group II intron-encoded RT(Guo et al., 2020), we designed fifteen 2000nt amplicon and a single 1300nt amplicons tiled across the SARS-CoV-2 genome for full sequencing coverage. Adjacent amplicons were designed with a 100nt overlap to ensure data is collected for regions otherwise masked by primer binding. Primers for reverse transcription (RT) were designed using the OligoWalk tool (Lu and Mathews, 2008) to avoid highly-structured primers and highly-structured regions of the SARS-CoV-2 genome. Forward and reverse primer sets were designed for an optimal T_m of 58°C. Reverse primers were inset 3nt from the 5′end of the RT primer to enhance specificity of the PCR reaction.

Reverse transcription with MarathonRT

MarathonRT purification was performed as described in (Guo et al., 2020). For each amplicon, 500ng of total cellular RNA was mixed with 1μL of the corresponding 1μM RT primer. Gene-specific primers used for RT are listed in Table S2. Primers were annealed at 65°C for 5min then cooled to room temperature, followed by addition of 8μL of 2.5x MarathonRT SHAPE-Map Buffer (125mM 1M Tris-HCl pH 7.5, 500mM KCl, 12.5mM DTT, 1.25mM dNTPs, 2.5mM Mn^{2+}), 4μL of 100% glycerol, and 0.5μL of MarathonRT. RT reactions were incubated at 42°C for 3 hours. 1μL 3M NaOH was added to each reaction and incubated at 95°C for 5min to degrade the RNA, followed by the addition of 1μL 3M HCl to neutralize the reaction. cDNA was purified using AmpureXP beads (Cat. No. A63880) according to manufacturer's protocol and a 1.8x bead-to-sample ratio. Purified cDNA was eluted in 10μL nuclease-free water.

SHAPE-MaP library construction

Amplicons tiling the SARS-CoV-2 genome were generated using NEBNext UltraII Q5 MasterMix (Cat. No. M0544L), gene-specific forward and reverse PCR primers, and 5μL of purified cDNA. Gene-specific primers used for PCR are listed in Table S3. Touchdown cycling PCR conditions were used to enhance PCR specificity (68-58°C annealing temperature gradient) (Korbie and Mattick, 2008). PCR reaction products were purified with Monarch PCR&DNA Clean-up Kits (NEB, Cat. No. T1030S) with a binding buffer:sample ratio of 2:1 to remove products smaller than 2kb. PCR products were visualized on 0.8% agarose gels to confirm production of correctly sized amplicons. Amplicons were diluted to 0.2ng/uL and then pooled into two odd and two even amplicon pools for downstream library preparation. Sequencing libraries were generated using a NexteraXT DNA Library Preparation Kit (Illumina) according to manufacturer's protocol, but with 1/5th the recommended volume. Libraries were quantified using a Qubit dsDNA HS Assay Kit (ThermoFisher, Cat. No. Q32851) to determine the concentration and a BioAnalyzer High Sensitivity DNA Analysis (Agilent, Cat. No. 5067-4626) to determine average library member size. Using these two values, libraries were diluted to 4nM, denatured, and final library dilutions prepared according to manufacturer's protocols. Amplicon pools were recombined and sequenced on a NextSeq 500/550 platform using a 150 cycle mid-output kit.

Structure prediction

All libraries were analyzed using ShapeMapper 2 (Busan and Weeks, 2018), aligning reads to SARS-CoV-2 genome (GenBank: MN908947). The default read-depth threshold setting of 5000x was used as a quality control benchmark. Mutation rates between NAI-modified and unmodified samples were tested for significance using the equal variance t test. Using reactivities output from ShapeMapper, ShapeKnots (Hajdin et al., 2013) was used to determine whether two previously reported pseudoknots contained in the SARS-CoV-2 genome were predicted with experimental SHAPE constraints. The two pseudoknots tested were the programmed ribosomal frameshifting element that exists at the Orf1a/b boundary, and a pseudoknot in the 3′UTR that was identified in the MHV and B-CoV genomes (Goebel et al., 2004). We analyzed all 500nt windows separated by a 100nt slide that contained each of the putative pseudoknots to determine if the pseudoknot was successfully predicted.

SuperFold (Smola et al., 2015b) was used to generate a consensus structure prediction for the entire SARS-CoV-2 genome with both replicate datasets. We imposed a maximum pairing distance of 500nt. As our data only supported formation of the pseudoknot contained in the programmed ribosomal frameshifting element, only this pseudoknot was forced in this prediction. All structures output from the SuperFold prediction were visualized and drawn using StructureEditor, a tool in the RNAStructure software suite (Reuter and Mathews, 2010).

Base-pairing distances were calculated from .ct structure files output from SuperFold full-length SARS-CoV-2 consensus predictions, and compared to previously published, publically available full-length genome structures for Dengue and Hepatitis C Virus generated with SHAPE constraints, a max-pairing distance of 500nt, and the SuperFold pipeline (Mauger et al., 2015; Dethoff et al., 2018).

Ensemble structure modeling for the PRF region

A region surrounding the SARS-CoV2 PRF (Genomic coordinate: 12886-13635) was used to model the structural ensemble of the PRF. The region boundaries were determined based on base-paring probabilities output from partition function calculation performed in the SuperFold pipeline. Specifically, we ensured all nucleotides involved in base-pairing interactions with the PRF were included for the ensemble modeling.

To perform ensemble structure modeling, we followed step 6, 7 and 8 from the Rsample program (Spasic et al., 2018). To elaborate, first we used the Partition program (implemented in RNA structure v6.1, Mathews (Mathews, 2004)) to generate the partition saved file (PFS) for the region described. Replicate 1 SHAPE reactivity was used as a soft constraint (using the same slope and intercept as we used in the Superfold prediction) and the pseudoknotted base pairs were forced single strand. The PFS file was used to sample 1000 probable structures in proportion to their Boltzmann weights using the stochastic program (implemented in RNA structure v6.1) (Ding and Lawrence, 2003). This sample was then clustered using the hierarchical divisive method (Ding et al., 2005) and was asked to output 10 clusters with a representative conformation. A cluster is defined as a subset of structures with similar base pairs. The PFS file was visualized using IGV v2.8.2(Busan and Weeks, 2017).

Identification of well-folded regions

Two data signatures were used to identify well-folded regions: The first is the SHAPE reactivity data generated with the SHAPE-MaP workflow and the ShapeMapper analysis tool (Busan and Weeks, 2018). The second is the Shannon entropy calculated from base-pairing probabilities determined during the SuperFold partition function calculation (Smola et al., 2015b). Two replicate datasets were used, including separate SuperFold predictions.

Local median SHAPE reactivity and Shannon Entropy were calculated in 55nt sliding windows. The global median SHAPE reactivity or Shannon Entropy were subtracted from calculated values to aid in data visualization. Regions with local SHAPE and Shannon Entropy signals 1) below the global median 2) for stretches longer than 40 nucleotides 3) that appear in both replicate datasets were considered well-folded. Disruptions, or regions where local SHAPE or Shannon Entropy rose above the global median, are not considered to disqualify well-folded regions if they extended for less than 40 nucleotides. Arc plots generated from each replicate consensus structure predication were compared for regions that meet sorting criteria described above in order to ensure agreement between secondary structure models generated from each replicate SHAPE-MaP dataset.

Base-pairing distances of well-folded regions were calculated from .ct structure files output from SuperFold consensus predictions, and compared to previously published, publicly available structures for well-folded regions of the HIV genome generated with SHAPE constraints, a max-pairing distance of 500nt, and the SuperFold pipeline (Siegfried et al., 2014).

Multiple sequence alignment

To analyze evolutionary support for our consensus secondary structure prediction of the SARS-CoV-2 genome, we generated two codon-based multiple sequence alignments (MSA) for Orf1a and Orf1b constructed from genomes of closely related viral species (Ranwez et al., 2018). All sequences were chosen based on a phylogenetic study of SARS-CoV-2 (Ceraolo and Giorgi, 2020). All sequences referenced below were downloaded from the NCBI Taxonomy browser (Benson et al., 2018).

A sarbecovirus MSA was generated using SARS-CoV-2 isolate Wuhan-Hu-1 (GenBank: MN908947.3), four bat coronaviruses (GenBank: MG772934.1, JX993987.1, DQ022305.2, DQ648857.1), and five human SARS coronaviruses (GenBank: AY515512.1, AY274119.3, NC_004718.3, GU553363.1, DQ182595.1).

We also generated an “All β-coronavirus Alignment” using the sarbecovirus sequences described above in addition to four MERS-CoV sequences (GenBank: MK129253, KP209307, MF598594, MG987420), one HKU-4 sequence (MH002337), three HKU-5 sequence (GenBank: MH002342, NC009020, MH002341), four HKU1 sequences (GenBank: KY674942, KF686343, AY597011, DQ415903), three murine hepatitis virus sequences (GenBank: AY700211, AF208067, AB551247), three human coronavirus OC43 sequences (GenBank: AY585229, NC006213, MN026164), two bovine coronavirus sequences (GenBank: KU558922, KU558923), and one camel coronavirus sequence (GenBank: MN514966).

The orf1a and orf1b region were extracted from the full-length sequences based on the GenBank annotation. Separate codon alignments for both Orf1a and orf1b were generated using MACSE v2.0.3 (Ranwez et al., 2018) and default parameters (*-prog alignSequences*).

Synonymous mutation rate analysis

All codon alignments were visualized and edited using Jalview v 2.11.0 (Waterhouse et al., 2009). Synonymous mutation rates for each codon were estimated using the phylogenetic-based parametric maximum likelihood (FUBAR) method (Murrell et al., 2013). Each codon was categorized as base-paired or unpaired depending on strandedness of the nucleotide at the third position of each codon in our SARS-CoV-2 consensus structure model (Dethoff et al., 2018). The significance of synonymous mutation rates between single- and double-stranded regions was determined using two-tailed, equal variance t test.

Covariation analysis

Covariation calculation and visualization was performed using R-chie (Lai et al., 2012). The Sarbecovirus codon alignment described above was used for covariation analysis. Identification of base-pairs with statistically significant evidence of covariation was performed on individual structures using R-Scape (version 0.2.1) (Rivas et al., 2017) with the RAFSp statistics by using the “– RAFSp” flag(default E-value:0.05) (Tavares et al., 2019).

Design of antisense locked nucleic acids

Antisense locked nucleic acids (LNAs, Integrated DNA Technologies) were designed to anneal to target sequences within the SARS-CoV-2 genome (GenBank: MN908947). All LNAs were designed with three consecutive LNA bases at the 5′ and 3′ ends of each oligonucleotide, with stretches of unlocked bases within the oligonucleotide limited to three consecutive nucleotides. All LNAs were designed with similar thermodynamic properties, including length, %GC content, %LNA content, and LNA:RNA duplex T_m (Table S4).

LNA transfection and icSARS-CoV-2-mNG infection

Vero-E6 were grown in DMEM+10% FBS+1% PBS and incubated at 37°C/5% CO_2. Approximately 7.5x10^5 Vero-E6 cells were plated per well in a 6-well plate prior to transfection. LNAs were transfected at a final concentration of 400nM per well using the TransIT-Oligo reagent, including a reagent only transfection control (Mirus, MIR 2164). One day post-transfection, transfected or control Vero-E6 cells were plated at 2.5 × 10^3 cells per well in a 384-well plate in phenol free media and were then infected with icSARS-CoV-2mNG at a MOI of 1.0 (Xie et al., 2020)

Infected cell frequencies, as quantified by mNeonGreen expression, were assessed at 24 hours post-infection by high content imaging (Cytation 5, BioTek) configured with bright field and GFP cubes. Total cell numbers were determined from bright field images using Gen5 software. Object analysis measured the number of mNeonGreen positive cells. Percent infection was calculated as the ratio between the total number of mNeonGreen+ cells and total cells.

QUANTIFICATION AND STATISTICAL ANALYSIS

Graphs and statistical analysis were made using GraphPad Prism 8 and Microsoft Excel v16. The results are expressed as Tukey plots with median and interquartile range indicated with bars. Specific values are reported in the Results. All statistically significant differences were calculated using the unpaired t test assuming both populations have equal variance unless otherwise stated. Significance of comparisons is indicated in figures and supplemental data as *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.