Abstract

Identifying Riboregulatory Elements in Long, Single-Stranded, Positive-Sense RNA Viral Genomes

Nicholas Cornell Huston

2022

Single-stranded, positive-sense RNA viruses are a class of pathogens that pose a serious danger to human health. As a group, they have been the focus of research exploring how they infect, evade, and hijack host cellular machinery to propagate. Though these studies have primarily focused on viral proteins, the past two decades have seen a resurgence of interest in the RNA genomes themselves. This is because RNA viral genomes contain functional RNA structures that expand their functional repertoire. While studies of viral RNA structure were originally restricted to 5' and 3' viral termini, recent methodological advancements have facilitated the search for functional structure within extensive viral open-reading frames. Here, these methods are applied to the genomes of two RNA viruses, SARS-CoV-2 and West Nile virus. In pursuit of this work, several methodological advancements were made that will facilitate future studies of functional RNA structure.

In *Chapter 2*, this methodology is applied to the genome of SARS-CoV-2, the etiological agent responsible for the ongoing global pandemic. We develop a novel long-amplicon strategy for the collection of SHAPE-MaP data using a highly processive reverse transcriptase, greatly facilitating structural studies of extremely long viral RNAs. 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 uncover elaborate networks of well-folded RNA secondary structures and 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 that, not only do these architectural features appear to be conserved across the β-coronavirus family, but 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.

In *Chapter 3*, we extend this methodology to the genome of West Nile virus, an arthropod-borne virus that, due to climate change, poses an increasing global health risk. We report for the first time the complete secondary structure of the WNV genome in both arthropod and mammalian cell lines. The resulting genomic secondary structure model recapitulates a conserved motif in the 5'UTR required for viral replication. Along with our SHAPE-MaP data, our structural models provide novel insights into previously studied but poorly understood aspects of flaviviral biology. We describe a global genome architecture that, along with specific regions of well-folded RNA, folds with minimal host dependence. Owing to weak signals of evolutionary conservation, we instead relied on patterns of structural homology to prioritize specific RNA structures for functional validation. Using a highly optimized workflow, we used structure-disrupting LNAs to demonstrate that a subset of novel well-folded RNA structures plays both conserved and host-specific functional roles.

Taken together, the work presented in this dissertation deepens our

understanding of viral biology and functional RNA structure, identifies conserved aspects of the viral life cycle that are readily targetable by a novel class of nucleic acids, and therefore represents an important step forward in our fight against expanding global health threats. Methodological improvements and innovations presented in this dissertation have broad applications beyond the study of viral RNAs and will therefore greatly facilitate the discovery and study of functional RNA structure.

Identifying Riboregulatory Elements in Long, Single-Stranded, Positive-Sense RNA Viral Genomes

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Nicholas Cornell Huston

Dissertation Director: Anna Marie Pyle

December 2022

Table of Contents

Acknowledgements

I would like to first and foremost thank my mentor, Dr. Anna Marie Pyle. From day one, she trusted me with a scientific independence that, while scary at times, allowed me to grow as a scientist in ways I did not think were possible before I matriculated. Anna's steady guidance, brilliant mind, and unwavering support have been both a welcome constant and vital to my success. To that end, I would also like to thank Dr. Douglas Brackney, Dr. Wendy Gilbert, and Dr. Matthew Simon for serving on my committee. Their patience and scientific guidance were instrumental in keeping me on track for graduation. I shudder to think what scientific detours and dead-ends I may have followed were it not for their ability to keep me scientifically focused and motivated.

Members of the Pyle Lab have also been central to my time here. I would like to thank Dr. Rebecca Adams and Dr. Thayne Dickey for being office mates, mentors, and friends when I first joined the lab. Those first couple of shared months in our KBT office still stand out as a highlight of grad school. I'd also like to extend special thanks to Shivali Patel, Rafael Tavares, and Han Wan. Your friendship and scientific insights have helped me weather some of the tougher moments, and I am so grateful. To the rest of the Pyle Lab – I consider myself lucky to have been surrounded by such smart and motivated colleagues, and I will miss you all very much.

Finally, I'd like to thank my family and friends for everything they have done for me. To my parents, Jack and Harriet Huston – I do not think I could have made it through grad school without your love, kindness, and support. I truly cannot thank

you enough. To my siblings Jake, Hattie, Will, and Chris – thank you for helping me enjoy not being in lab when I was not lab. Even if it was simply because you didn't want to hear me talk about science, it meant a lot. Lastly, I'd like to thank my closest friend Sarah Prophet and my boyfriend Frank Wendt – grad school would have been an absolute drag without you both along for the ride.

List of Figures

List of Tables

1. Introduction

As the last three years have acutely demonstrated, single-stranded, positive-sense RNA viruses pose a serious global health risk. While representing an incredibly diverse class of pathogens, the genomes of RNA viruses share a general architecture. Typically, the vast majority of sequence space in viral genome is dedicated to protein-coding information, sequence that is flanked on either side by relatively small, untranslated regions (UTRs). Though much work has been devoted to the study of viral proteins and the functions they effect, it is the UTRs of viruses that led researchers to first appreciate a second type of information encoded in viral genomes: higher-order RNA structure.

One of the best-studied examples of higher-order viral RNA structure is the internal ribosomal entry site (IRES) of Hepatitis C Virus (HCV). Resident in the 5'UTR of the HCV genome, work spanning multiple decades has demonstrated the IRES plays a fundamental role in promoting selective viral translation (Fraser and Doudna, 2007). Elegant structural studies, culminating in a high-resolution Cryo-EM structure of the IRES in complex with the human 40s ribosomal subunit, have revealed that it is the specific tertiary fold of the IRES that allows it to fulfill its function (Quade et al., 2015). In fact, a detailed understanding the HCV IRES's structure-function relationship has facilitated the development of a diverse class of therapeutics that specifically inhibit HCV translation by disrupting the IRES' native structure (Dibrov et al., 2014).

In this way, the HCV IRES represents a powerful case study. Beyond shedding light on a novel mechanism by which viruses co-opt host cellular machinery, it

highlights the central role structured RNAs can play in viral life cycles. Even more, it demonstrates how research directed towards a mechanistic understanding of RNA structure-function relationships can pave the way for development of much-needed antivirals. As such, it serves as both motivation and road map for studies devoted to exploring other viral RNA genomes for functional RNA structure. It is not an accident, however, that some of the best-studied examples of functional viral RNA structure to date are found in viral UTRs. Indeed, the exploration of viral open reading frames (ORFs) for functional RNA structure has only been recently made possible due to methodological advancements in RNA structure-probing methodologies.

1.1 Probing long viral RNA genomes

At the most basic level, the goal of all RNA structure probing methods is to assess the stranded-ness of given RNA nucleotide in a larger RNA chain. In practice, methods developed for RNA structure probing rely on a variety of strategies to make this measurement, and not all of them afford direct read-outs of stranded-ness. Though enzymatic and spectroscopic methods have been successfully deployed to probe RNA structure, the methods that have proved most useful for the study of long viral RNAs instead rely on chemical modification of RNA bases (Hart et al., 2008; Lockard and Kumar, 1981).

SHAPE reagents, named both for both the chemistry they enable (**s**elective 2'-**h**ydroxyl **ac**ylation) and the method by which data collected is collected (**p**rimer **e**xtension), are currently most commonly used for probing the structure of long

viral RNAs. These electrophilic reagents react preferentially with 2'-hydroxyl (2'-OH) moieties of flexible nucleotides, regardless of base identity. Owing to the close proximity of the negatively charged phosphate backbone, the 2'-OH of inflexible nucleotides reacts poorly with SHAPE reagents. Importantly, studies have shown that per-nucleotide flexibility, and by extension reactivity, serves as a reliable proxy for nucleotide stranded-ness (Merino et al., 2005). While a separate class of reagents reports directly on nucleotide stranded-ness by selective modification of unpaired nucleotides, they exhibit base-specificity that necessitates the use of multiple reagents to collect data on all four nucleotides (Wang et al., 2019; Wells et al., 2000).

In all cases, the reaction between the chemical and nucleic acid results in deposition of a bulky adduct on that nucleotide. In the original iterations of chemical probing methods, these bulky adducts were read out using primer extension, as they cause chain termination events when the modified RNA molecule is subsequently reverse transcribed into complimentary DNA (cDNA). By resolving truncated cDNA products on sequencing gels, modified nucleotides can be identified and assessed for modification level. However, due to size constraints imposed by the pore size of polyacrylamide gels, viral genomes cannot be analyzed using this strategy as they often exceed 10kb (Green and Sambrook, 2021).

It was not until these chemical probing methods were adapted for high-throughput sequencing (HTS) that studies of whole viral genomes became tractable. First pass attempts at high-throughput collection of SHAPE data involved addition of a probe ligation step following primer extension and chain termination. These ligated probes, added at the 3' ends of nascent cDNAs, preserved the location of

chain termination events and facilitated down-stream library preparation for HTS (Lucks et al., 2011). However, SHAPE signal decay necessitates the use of RT primers spaced every ~200nt, rendering these so-called "stop-based" methods experimentally intractable for long viral RNAs (Adams et al., 2019; Karabiber et al., 2013).

To bypass the issues of stop-based HTS probing methods, researchers in the Weeks lab discovered that in the presence of manganese, a non-native metal cofactor, the bulky adducts deposited on RNA molecules did not result in chain termination events (Siegfried et al., 2014). Instead, mutations were deposited on nascent cDNAs at the locations of RNA adducts. This method, called mutational profiling (MaP), allows for per-nucleotide reactivity information to be read out by calculating mutation rates in HTS data-sets, and was readily adapted for other chemical probing reagents (Zubradt et al., 2016). Most importantly, because MaP strategies no longer relied on cDNA truncations to read out per-nucleotide reactivity information, studying the structural content of large viral RNA genomes no longer represents the methodological hurdle it once did.

1.2 Secondary structure predictions of long viral RNA genomes

On its own, the data collected using any RNA structure probing experiment is of limited utility for structure discovery. Instead, it is most useful when used to experimentally constrain RNA structure prediction algorithms. And much like chemical probing methodologies, a subset of these RNA structure prediction algorithms have proved particularly useful for the study of long viral RNAs.

The most commonly used RNA structure prediction algorithm relies on an algorithmic framework originally developed in the 1980s. Called nearest-neighbor minimum free energy (MFE) optimization, these algorithms use a dynamic programming strategy to compute an RNA secondary structure with the largest free energy change (ΔG) relative to the linear sequence (Zuker and Stiegler, 1981). Importantly, these algorithms are parameterized with terms that reflect the free energy change of individual nucleotides, whether base-paired or single-stranded, is fundamentally dependent on the identity and stranded-ness of its nearest neighbors (Mathews and Turner, 2002; Mathews et al., 1999; Xia et al., 1998). It is the sum of the free energy changes associated with formation of each individual RNA helix or loop that allows for determination of a minimum free energy structure for the entire RNA sequence. In spite of constant refinement of the energy terms, single MFE predictions have only ~73% accuracy when predicting known structures, and the accuracy decreases for longer RNA sequences (Mathews, 2004).

Two strategies were implemented almost in parallel that drastically improved the accuracy of RNA secondary structure predictions. The first strategy is based on the observation that, for a given nucleic acid sequence, sub-optimal structures may exist that are very close in energy to the computed MFE structure, but that differ drastically in connectivity (Zuker, 1989). As this lead to uncertainty in the overall accuracy of any given MFE structure prediction, sub-optimal or not, researchers implemented partition function calculations (Mathews, 2004). These calculations allow for identification of individual base pairs that are predicted to fold with a high probability and, by extension, afford an empirical way to identify

high-confidence MFE structures. When considering MFE predictions of known RNA structures, base pairs predicted with >99% probability during partition function calculations are correctly predicted 90% of the time. As a result, partition function calculations have become integral aspects of widely used RNA structure prediction algorithms.

The second strategy implemented involved allowing for the inclusion of experimental data, such as chemical probing data, as constraints during both MFE structure prediction and partition function calculation steps. In the context of SHAPE reagents, experimentally determined reactivities are included as pseudo-free energy terms (Mathews, 2004; Mathews et al., 2004). In the simplest terms, the inclusion of SHAPE constraints rewards prediction of highly reactive nucleotides as single-stranded and lowly reactive nucleotides as double-stranded. Regardless of how SHAPE constraints are used to constrain predictions (i.e., 'soft' v. 'hard' constraints), their inclusion improves the accuracy of structure prediction (Swenson et al., 2013).

While significant improvements have been made to individual algorithmic components, the relatively large size of viral RNA genomes still presents a fundamental barrier to RNA structure prediction. This is owed in part to the computational time and power required to compute partition functions and MFE predictions for long nucleic acid sequences. At a more fundamental level, however, this is because prediction accuracy deteriorates for RNAs of increasing size, falling off drastically for RNAs >800nt (Mathews, 2004). To resolve this issue, a prediction pipeline called SuperFold was developed that predicts both SHAPE-constrained

partition function calculations and MFE predictions in sliding windows of 1200 and 3000nt, respectively (Smola et al., 2015). Though this pipeline requires that individual MFE predictions are stitched back together, this is achieved with a fairly elegant strategy. Specifically, base pairs identified by the partition function calculation to fold with >99% probability are forced double-stranded during subsequent MFE prediction steps, essentially nucleating predictions of the remaining sequence around these high confidence base pairs. The result is an experimentally constrained, consensus secondary structure prediction for every single nucleotide in a given sequence, with no upper limit placed on size.

1.3 Identifying RNA secondary structures with functional potential

Considering the difficulty associated with generating secondary structure predictions of whole viral genomes, it is ironic that sorting through these predictions to identify structures of interest represents its own separate challenge. As with structure prediction, two strategies have been developed that allow for identification of RNA structures with functional potential.

The first strategy rests on the assumption the RNA structures with conserved functional roles in a viral life cycle should be 1) highly structured and 2) well-determined. Both of these characteristics are reflected in specific data types generated during the process of constrained secondary structure prediction. As highly structured RNA should contain a large proportion of base-paired nucleotides, the relative structured-ness of an RNA is captured by the SHAPE reactivity data

captured during structure probing steps. By searching for stretches of RNA with low SHAPE reactivity relative to the rest of the RNA, these regions can be identified.

As well-determined RNA should contain base pairs that fold with a high probability, this information is captured during the partition function calculation in a data type called Shannon Entropy. Shannon entropy is calculated for individual nucleotides, and reflects the probabilities of all possible pairing interactions determined for that nucleotide in its conformational ensemble. Nucleotides whose conformational ensembles are dominated by a single, high-probability pairing interaction will have low Shannon entropy (Smola et al., 2015). Therefore, by searching for regions that have low Shannon entropy relative to the rest of the RNA, well-determined regions can be identified. In concert, these so-called low Shannon/SHAPE (lowSS) data signatures have been effective in identifying functional RNA in single-stranded viral genomes (Dethoff et al., 2018; Madden et al., 2020; Mauger et al., 2015; Siegfried et al., 2014).

The second strategy for identifying regions of structured RNA with functional potential relies on identifying evolutionary signals of secondary structure conservation. Comparative sequence analysis is often considered the gold-standard, and relies on identifying base-pairs that co-mutate at a rate that exceeds random chance in a process called covariation (Rivas et al., 2016; Yao et al., 2018). Several software packages have been developed to provide statistical measures of covariation, but they all involve constructing alignments of related RNAs. For RNAs with rich databases of related sequences, like bacterial riboswitches, construction of these alignments is not difficult (Weinberg et al., 2017). However, for RNAs such as

human long, non-coding RNAs (lncRNAs), building these alignments is much more difficult. Indeed, specific adjustments have had to be made to existing pipelines to identify covariation in highly homologous or sparse sequence alignments (Tavares et al., 2018). It is an ongoing debate in the field whether a lack of statistical covariation reflects a lack of evolutionary conservation or simply a low-information alignment (Rivas et al., 2020). A separate method uses preferential accumulation of synonymous mutations at single-stranded regions relative to double-stranded regions as a signal of evolutionary conservation, and it has been successfully applied to studies of viral RNA structure (Assis, 2014; Simmonds and Smith, 1999; Tuplin et al., 2002). However, as this mode of analysis also fundamentally relies on sequence alignments, low-information alignments may render it similarly underpowered. As such, there is an ongoing need for alternate strategies that can identify a subset of regions with functional potential from consensus predictions of whole viral genomes.

1.4 Functional validation of viral RNA secondary structures

Through the use of both lowSS metrics and signals of evolutionary conservation, it is possible to flag a subset of RNA secondary structures with functional potential. Absent any experimental follow-up, however, delineation of these candidate structures is purely descriptive. In this way, functional validation represents the most important aspect of the structure discovery process, and is the avenue of inquiry from which we stand to learn the most about both functional RNA structure and viral biology. However, functional studies of RNA structures in viral

ORFs are complicated by the fact that they contain a code-within-a-code; in addition to carrying protein-coding information, they also contain the information for functional RNA secondary structure. While a fascinating observation in its own right, this feature of viral ORFs severely constrains studies that rely on classical viral genetics strategies to validate the functionality of individual structures.

Classically, RNA structures were functionally assessed by engineering mutations into the viral genome that disrupt an RNA structure of interest, and assay the effect of disruption on various aspects of viral growth. However, in order to unambiguously link defects observed to RNA structure disruption, researchers avoid mutations that alter the viral coding potential, including those that introduce synonymous mutations but also rare codons. This constraint, exemplified in Pirakitikulr et al., 2016, not only restricts the types of mutations that can be introduced, but also limits the severity of structure disruption. These clunky mutational strategies similarly complicate introduction of compensatory mutations that restore RNA structure, a standard aspect of structure validation workflows. Of more recent concern is the possibility of accidentally engineering gain-of-function mutations, though these typically require swapping large viral domains (Menachery et al., 2015).

At a more fundamental level, these mutagenic strategies require that full-length infectious clones are available to mutate. This presents a profound problem for the study of emerging viruses, which often pose serious global health risks. In the case of SARS-CoV-2, the first infectious clone was not constructed until almost 6

months after the virus first emerged, and even then availability of the infectious clone to other researchers was not guaranteed (Xie et al., 2020).

The status of many viruses as biosafety level 3 and 4 pathogens has also necessitated construction of non-infectious viral replicons. In these systems, viral proteins required for completion of full infectious cycles are often replaced with reporter genes, allowing for viral replication in cells but preventing production of infectious virions (Lo et al., 2003; Phan et al., 2009). However, while these subgenomic replicons allow for the study of these viruses at a lower biosafety level, any region of the genome excised cannot be probed or assayed for functional RNA structure. Taken together, these methodological shortcomings highlight the need for strategies to validate candidate RNA structures that do not rely on infectious clones, replicon systems, or clunky mutational strategies.

1.5 Chapter Overview

My dissertation focuses on applying the methods described above to the genomes of SARS-CoV-2 and West Nile virus, two single-stranded, positive sense RNA viruses that represent serious and ongoing global health risks. Experimental difficulties encountered while pursuing this research forced the development of methodological improvements that will facilitate future studies of other viral RNA genomes. More importantly, the application of these methods has provided novel insights into the function of the RNA genome in the viral life cycle of both viruses, deepening our understanding of both viral biology and functional RNA structure.

In *Chapter 2*, this methodology is applied to the genome of SARS-CoV-2, the etiological agent responsible for the ongoing global pandemic. We develop a novel long-amplicon strategy for the collection of SHAPE-MaP data using a highly processive reverse transcriptase, greatly facilitating future structural studies of extremely long viral RNAs (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 also uncover elaborate networks of well-folded RNA secondary structures, and 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 that, not only do its architectural features appear to be conserved across the β-coronavirus family, but 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. Importantly, this method circumvents the need for molecular cloning.

In *Chapter 3*, we extend this methodology to the genome of West Nile virus, an arthropod-borne virus that, due to climate change, poses an increasing global health threat. We report for the first time the complete secondary structure of the WNV genome in both arthropod and mammalian cell lines. The resulting genomic secondary structure model recapitulates a conserved motif in the 5'UTR required for viral replication. Along with the SHAPE-MaP data, it provides novel insights into

previously studied but poorly understood aspects of flaviviral biology including genome cyclization and the structure of the 3'UTR. We describe a global genome architecture that, along with specific regions of well-folded RNA, folds with minimal host dependence. Owing to weak signals of evolutionary conservation, we instead rely on patterns of structural homology to prioritize specific RNA structures for functional validation. Using a highly optimized workflow, we use structure-disrupting LNAs to show that a subset of these well-folded RNA structures plays both pan- and host-specific functional roles.

Taken together, the work presented in this dissertation deepens our understanding of viral biology and functional RNA structure, identifies conserved aspects of the viral life cycle that are readily targetable by a novel class of nucleic acids, and therefore represents an important step forward in our fight against expanding global health threats.

1.6 References

1. Adams, R.L., Huston, N.C., Tavares, R.C.A., and Pyle, A.M. (2019). Sensitive detection of structural features and rearrangements in long, structured RNA molecules (Elsevier Inc.).
2. Assis, R. (2014). Strong Epistatic Selection on the RNA Secondary Structure of HIV. PLoS Pathog. *10*.
3. Dethoff, E.A., Gokhale, N.S., Boerneke, M.A., Muhire, B.M., Martin, D.P., Sacco, M.T., McFadden, M.J., Weinstein, J.A., Messer, W.B., Horner, S.M., et al. (2018). Pervasive Tertiary Structures in the Dengue Virus RNA Genome Modulate Fitness. Submitted *115*, 11513–11518.
4. Dibrov, S.M., Parons, J., Carnevali, M., Zhou, S., Rynearson, K.D., Ding, K., Garcia Sega, E., Brunn, N.D., Boerneke, M.A., Castaldi, M.P., et al. (2014). Hepatitis C Virus Translation Inhibitors Targeting the Internal Ribosomal Entry Site. J. Med. Chem. *57*, 1694–1707.
5. Fraser, C.S., and Doudna, J.A. (2007). Structural and mechanistic insights into hepatitis C viral translation initiation. Nat. Rev. Microbiol. *5*, 29–38.
6. Green, M.R., and Sambrook, J. (2021). Separation of RNA according to size: Electrophoresis of RNA through denaturing urea polyacrylamide gels. Cold

Spring Harb. Protoc. *2021*, 5–14.
7. 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.
8. Hart, J.M., Kennedy, S.D., Mathews, D.H., and Turner, D.H. (2008). NMR-assisted prediction of RNA secondary structure: Identification of a probable pseudoknot in the coding region of an R2 retrotransposon. J. Am. Chem. Soc. *130*, 10233–10239.
9. Karabiber, F., Mcginnis, J.L., Favorov, O. V, and Weeks, K.M. (2013). QuShape : Rapid , accurate , and best-practices quantification of nucleic acid probing information , resolved by capillary electrophoresis. 63–73.
10. Lo, M.K., Tilgner, M., Bernard, K.A., and Shi, P. (2003). Functional Analysis of Mosquito-Borne Flavivirus Conserved Sequence Elements within 3 J Untranslated Region of West Nile Virus by Use of a Reporting Replicon That Differentiates between Viral Translation and RNA Replication. J. Virol. *77*, 10004–10014.
11. Lockard, R.E., and Kumar, A. (1981). Mapping tRNA structure in solution using double-strand-specific ribonuclease V1 from cobra venom. Nucleic Acids Res. *9*, 5125–5140.
12. Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. U. S. A. *108*, 11063–11068.
13. Madden, E.A., Plante, K.S., Morrison, C.R., Kutchko, K.M., Sanders, W., Long, K.M., Taft-Benz, S., Cruz Cisneros, M.C., White, A.M., Sarkar, S., et al. (2020). Using SHAPE-MaP To Model RNA Secondary Structure and Identify 3′UTR Variation in Chikungunya Virus. J. Virol. *94*.
14. 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.
15. Mathews, D.H., and Turner, D.H. (2002). Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. Biochemistry *41*, 869–880.
16. Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. *288*, 911–940.
17. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., and Turner, D.H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A. *101*, 7287–7292.
18. 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. *112*, 201416266.
19. Menachery, V.D., Jr, B.L.Y., Debbink, K., Agnihothram, S., Gralinski, L.E., Plante, J.A., Graham, R.L., Scobey, T., Ge, X., Donaldson, E.F., et al. (2015). A SARS-like

cluster of circulating bat coronaviruses shows potential for human emergence. Nat. Med. *21*.
20. Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. (2005). RNA Structure Analysis at Single Nucleotide Resolution by Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE). 6866–6874.
21. Phan, T., Beran, R.K.F., Peters, C., Lorenz, I.C., and Lindenbach, B.D. (2009). Hepatitis C Virus NS2 Protein Contributes to Virus Particle Assembly via Opposing Epistatic Interactions with the E1-E2 Glycoprotein and NS3-NS4A Enzyme Complexes. J. Virol. *83*, 8379–8395.
22. Pirakitikulr, N., Kohlway, A., Lindenbach, B.D., Pyle, A.M., 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.
23. Quade, N., Boehringer, D., Leibundgut, M., Van Den Heuvel, J., and Ban, N. (2015). Cryo-EM structure of Hepatitis C virus IRES bound to the human ribosome at 3.9-Å resolution. Nat. Commun. *6*, 1–9.
24. Rivas, E., Clements, J., and Eddy, S.R. (2016). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat. Publ. Gr. *14*, 45–48.
25. Rivas, E., Clements, J., and Eddy, S.R. (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. Bioinformatics *36*, 3072–3076.
26. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*.
27. Simmonds, P., and Smith, D.B. (1999). Structural Constraints on RNA Virus Evolution. J. Virol. *73*, 5787–5794.
28. Smola, M.J., Rice, G.M., Busan, S., Siegfried, N.A., and Weeks, K.M. (2015). 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.
29. Swenson, M.S., Kjems, J., and Heitsch, C.E. (2013). Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *41*, 2807–2816.
30. Tavares, R.C.A.A., Pyle, A.M., and Somarowthu, S. (2018). Covariation analysis with improved parameters reveals conservation in lncRNA structures. J. Mol. Biol. #pagerange#.
31. 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.
32. Wang, P.Y., Sexton, A.N., Culligan, W.J., and Simon, M.D. (2019). Carbodiimide reagents for the chemical probing of RNA structure in cells. Rna *25*, 135–146.
33. Weinberg, Z., Christina, E.L., Corbino, K.A., Ames, T.D., Nelson, W., Roth, A., Perkins, K.R., Sherlock, M.E., and Breaker, R. (2017). Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic. *45*, 10811–10823.
34. Wells, S.E., Hughes, J.M.X., Igel, A.H., and Ares, M. (2000). Use of Dimethyl Sulfate

to Probe RNA Structure in Vivo. *318*.
35. Xia, T., SantaLucia, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson - Crick base pairs. Biochemistry *37*, 14719–14735.
36. Xie, X., Muruato, A., Lokugamage, K.G., Leduc, J.W., Menachery, V.D., Xie, X., Muruato, A., Lokugamage, K.G., Narayanan, K., Zhang, X., et al. (2020). An Infectious cDNA Clone of SARS-CoV-2. Cell Host Microbe 1–8.
37. Yao, Z., Weinberg, Z., and Ruzzo, W.L. (2018). Sequence analysis CMfinder — a covariance model based RNA motif finding algorithm. *22*, 445–452.
38. Zubradt, M., Gupta, P., Persad, S., Lambowitz, A.M., Weissman, J.S., and Rouskin, S. (2016). DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Publ. Gr. *14*, 75–82.
39. Zuker, M. (1989). On Finding All Suboptimal Foldings of an RNA Molecule. Science (80-.). *244*.
40. Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. *9*, 133–148.

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

2.1 Preface

This work presented in Chapter 1 represents a collaborative effort between several contributors: Nicholas C. Huston, Han Wan, Madison S. Strine, Rafael de Cesaris Araujo Tavares, Craig Wilen, and Anna Marie Pyle are all authors on the original publication. This work would not have been possible without the additional help of Dr. Li-Tao Guo (Pyle Lab, Yale University) for preparing and sharing MarathonRT enzyme, Dr. Mark Boerneke (Weeks Lab, UNC Chapel Hill) for providing data upon request, and Dr. Ananth Kumar and Gandhar Mahadeshwar (Pyle Lab, Yae University) for thoughtful comments on the manuscript. A version of the work was published in *Molecular Cell,* and is available at online at https://doi.org/10.1016/j.molcel.2020.12.041.

2.2 Abstract

SARS-CoV-2 is the positive-sense RNA virus that causes COVID-19 disease. The genome of SARS-CoV-2 is unique among viral RNAs in its vast potential to form RNA structures and yet, as much as 97% of its 30 kilobases have not been structurally explored. Here, we apply a novel long amplicon strategy to determine for the first time 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 beta 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.

2.3 Introduction

Severe acute respiratory syndrome related coronavirus 2 (SARS-CoV2), which is responsible for the current global pandemic(Zhu et al., 2020), is a positive strand RNA virus in the genus *β-coronavirus*. To date, the outbreak of SARS-CoV2 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 ~30kb genome is comprised of two open reading frames (ORFs) for viral nonstructural proteins (Nsps) and 9 small ORFs that encode

structural proteins and accessory genes (Kim et al., 2020). This extended 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)(Zust 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 has revealed a three-stemmed pseudoknot structure conserved across group II β-coronaviruses(Plant et al., 2005). However, neither the mechanism of frameshifting regulation nor the three-stem pseudoknot PRF conformation has been validated 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) which 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 coronavirus mechanism, improve diagnostics, and identify riboregulatory regions that can be targeted with antivirals drugs.

Recent advances in high-throughput structure probing methods (SHAPE-MaP, 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 and messenger RNA 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-CoV2 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 for the first time the complete secondary structure of SARS-CoV-2 RNA genome using in 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 that, not only do its architectural features appear to be conserved across the β-coronavirus family, but 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.

2.4 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 four days post-infection, cells were collected and treated with either NAI, which will preferentially modifies flexible nucleotides

at the 2'OH, or DMSO as a control. RNA was then extracted and purified. To generate sequencing libraries, 2000 nucleotide (nt) overlapping amplicons were tiled across the entire SARS-CoV-2 genome (**Fig 2.1A**). 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 > 70,000x and effective reactivity data for 99.7% (29813/29903) 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 (**Fig 2.1B**, 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 (**Fig A2.1A**), consistent with those previously reported for reactivities calculated from *in vivo* modified RNAs of this size(Smola et al., 2016). Across the sub

genomic RNA ORFs, the Pearson 's correlation is poor (**Fig A2.1B**). We believe this reflects the fact that Amplicons 13 - 16 will amplify both full-length *and* sub-genomic 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.

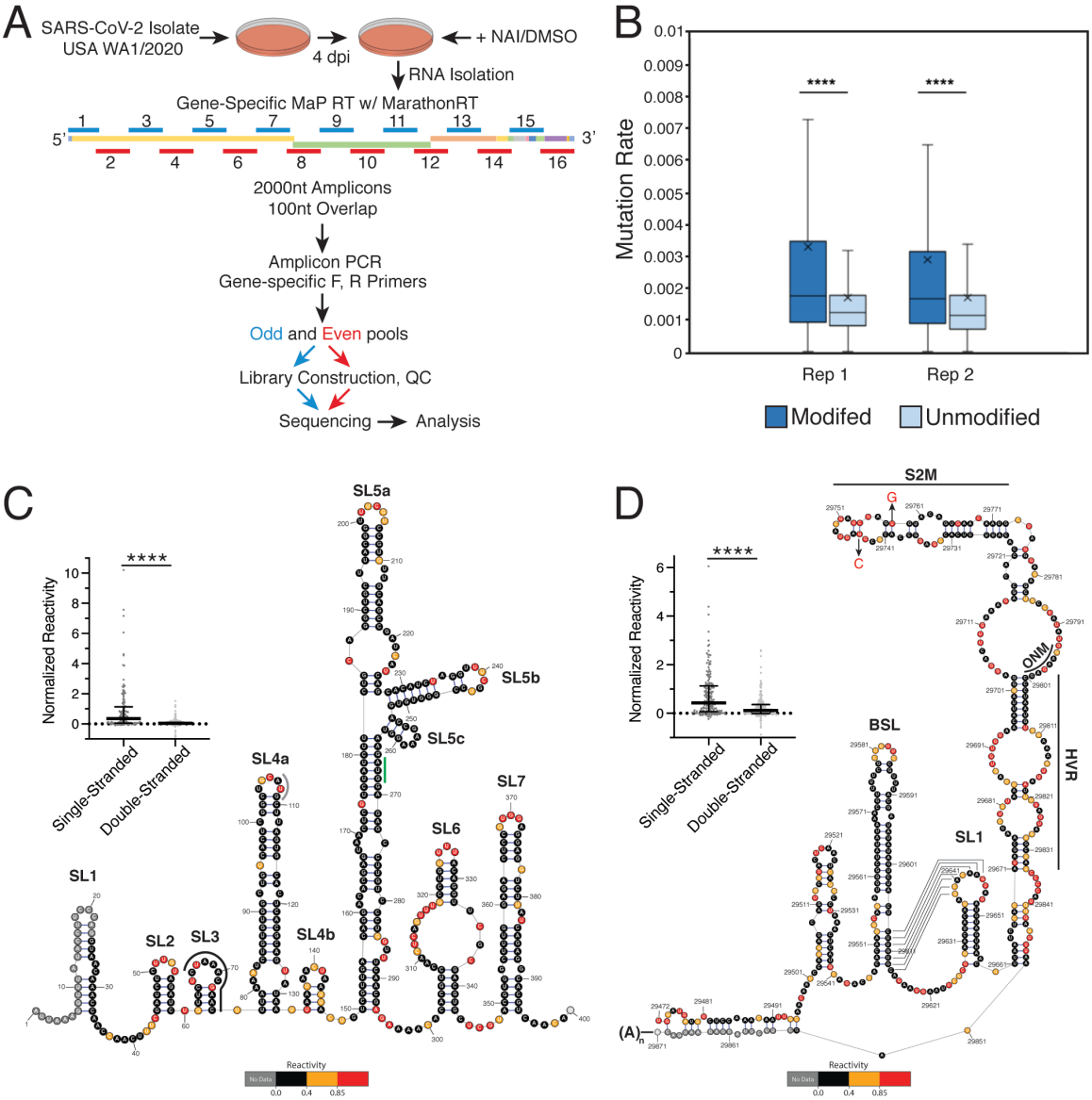

Figure 2.1. **Tiled-amplicon *in vivo* SHAPE-MaP workflow yields high quality data for *de novo* full-length structure prediction**. Structure prediction identifies conserved functional elements at the 5' and 3' viral termini. **A)** Workflow of *in vivo*

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 (TRS sequence = black line; uORF start codon = grey 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 t test.

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 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 (**Fig 2.1C,** 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 ($median_{Nuc(1-400)}$ = $2.7x10^{-5}$; 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 (**Fig 2.1C,** inset). For example, a trifurcated stem is observed at the top of SL5 (**Fig 2.1C**), 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 GroupIIB 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 Fig. 1C (SL1-4, SL6-7), suggesting these structures also play similar functional roles despite having been identified and elucidated other coronaviruses(Yang and Leibowitz, 2015).

The 3' genomic terminus includes three well-studied stems, including the bulged-stem loop (BSL), Stem Loop 1 (SLI), 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 (**Fig 2.1D,** inset) and overall low Shannon entropy ($median_{Nuc(29,472\text{-}29,870)}$ = 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 (**Fig 2.1D)**.

A pseudoknot structure is proposed to exist between the base of the BSL stem loop 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 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 (**Fig 2.1D**). 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 SLI 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 sub-domains. 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 (**Fig 2.1D**). 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 (**Fig 2.1D**). It is possible that the SARS-CoV-2 S2M folds into a unique S2M conformation despite differing by only a two bases, both of which are transversions. (**Fig 2.1D**, 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(Zust et al., 2008). However, structure prediction of the complete stem is not highly accurate, as reactivity information for the downstream stem (nts 29853-29870) is occluded by primer binding and is not constrained by experimental data (**Fig 2.1D**). In addition, the complete stem region (nts 29472-nts 29870) is predicted to have high Shannon entropy ($\text{median}_{\text{Nuc}(29472\text{-}29495,29853\text{-}29870)} = 0.2154$), supporting the notion that this substructure is not well-ordered in the cellular environment.

Structure prediction of the programmed ribosomal frame-shifting element reveals conformational flexibility

One of the most well-studied RNA structures in the coronavirus coding region is the programmed frame-shifting 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 loop (AS), 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. We found that the pseudoknot is successfully predicted in 3 out of 4 windows tested. Moreover, the nucleotides predicted to be involved in the pseudoknotted helix have low SHAPE-reactivity (**Fig 2.2A**, red lines). The frame-shifting 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 **Fig 2.2A**. In our model, the SHAPE reactivity and Shannon entropy calculation support a well-folded AS immediately upstream of the HSS (**Fig 2.2A**). 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).

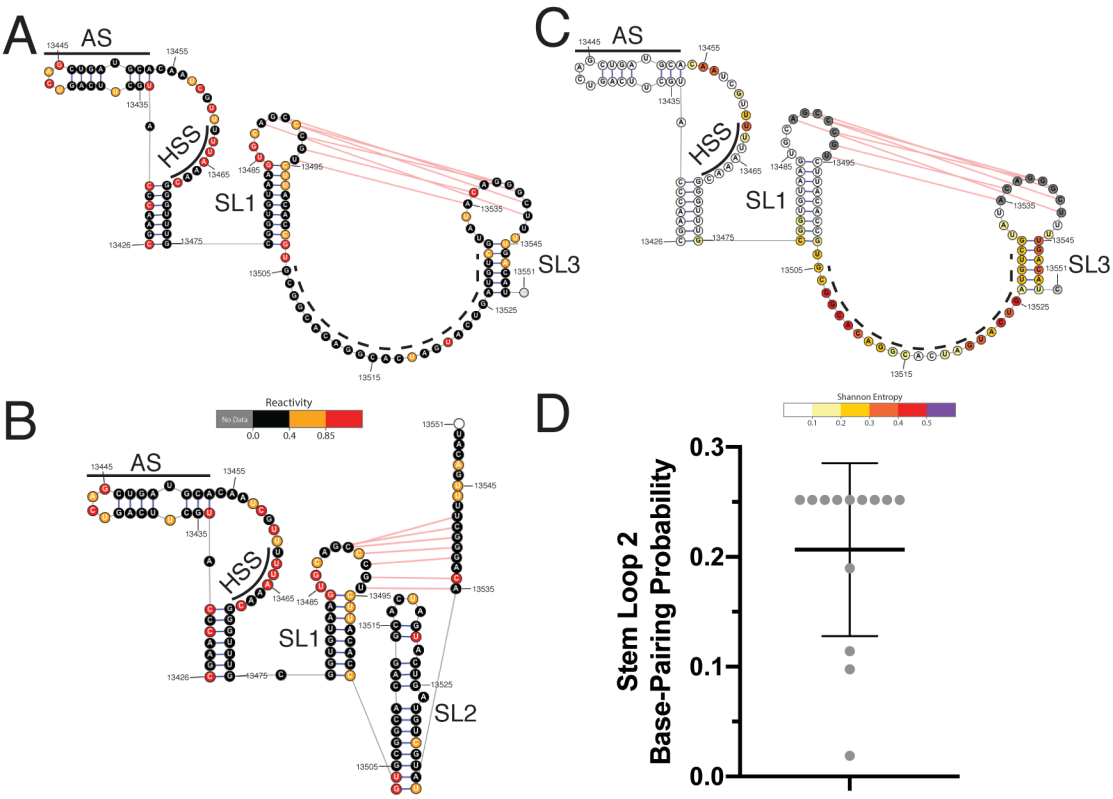

Figure 2.2. Structure prediction of the programmed ribosomal frame-shifting (PRF) element suggests conformational variability of Stem Loop 2. **A)** Dominant PRF structural architecture colored by SHAPE Reactivity. AS = Attenuator Stem; HSS = Heptanucleotide Slippery Sequence; SL1 = Stem Loop 1; dotted line indicates region that forms stem loop 2 (SL2) or long-range interactions outside the PRF; SL3 = Stem Loop 3; 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 Panel A. **D)** Base-pairing probability for alternate SL2 conformation. Each dot represents an individual base pair in SL2.

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 (**Fig 2.2A**) and Shannon entropy (**Fig 2.2C**). 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 (**Fig 2.2A**). 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 (**Fig 2.2C**).

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 (**Fig 2.2A**) forms base-pairing interactions with as many as 6 different regions in the SARS-CoV-2 genome (**Fig A2.2A**), Among these possible interactions is a PRF structure containing the three-stemmed pseudoknot conformation identified across coronaviruses, including a helical SL2 (**Fig 2.2B**) (Plant et al., 2005). The median base-pairing probability calculated for SL2 is 20% (**Fig 2.2D;** individual base-pairs indicated with grey 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 re-calculated the partition function for a 749nt

window in the SARS-CoV-2 genome that surrounds the PRF (**Fig A2.2A**). 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 (**Fig 2.2A**), this conformational cluster contains the AS, a single-stranded HSS, SL1, the pseudoknotted helix, and SL3. However, the SL2 region is base-paired with a region located 470nt upstream (**Fig A2.2B**). The second-best populated cluster contains a nearly identical domain architecture, except that the SL2 region is base-paired with a region 260nt upstream (**Fig A2.2C**). 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 (**Fig A2.2D**).

Taken together, these data suggest that the frame-shifting pseudoknot 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 remaining 97%.

Here we report the first *in vivo*-derived, SHAPE-constrained secondary structural model that includes a description of the base-pairing interactions for all nucleotides within a coronavirus genome (**Fig 2.3A**). 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 (**Fig A2.3**)(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 (**Fig 2.3B**).

To discover additional, well-folded RNA structures within the SARS-CoV-2 genome, we calculated the local median Shannon Entropy and correlated these values with experimentally-determined SHAPE reactivities (**Fig 2.3A**). Only regions with both median Shannon entropy and SHAPE reactivity signals below the global median for stretches longer than 40nt, and which appear in both replicate data sets, were considered well-determined and stable. In total, we identify 40 such regions in

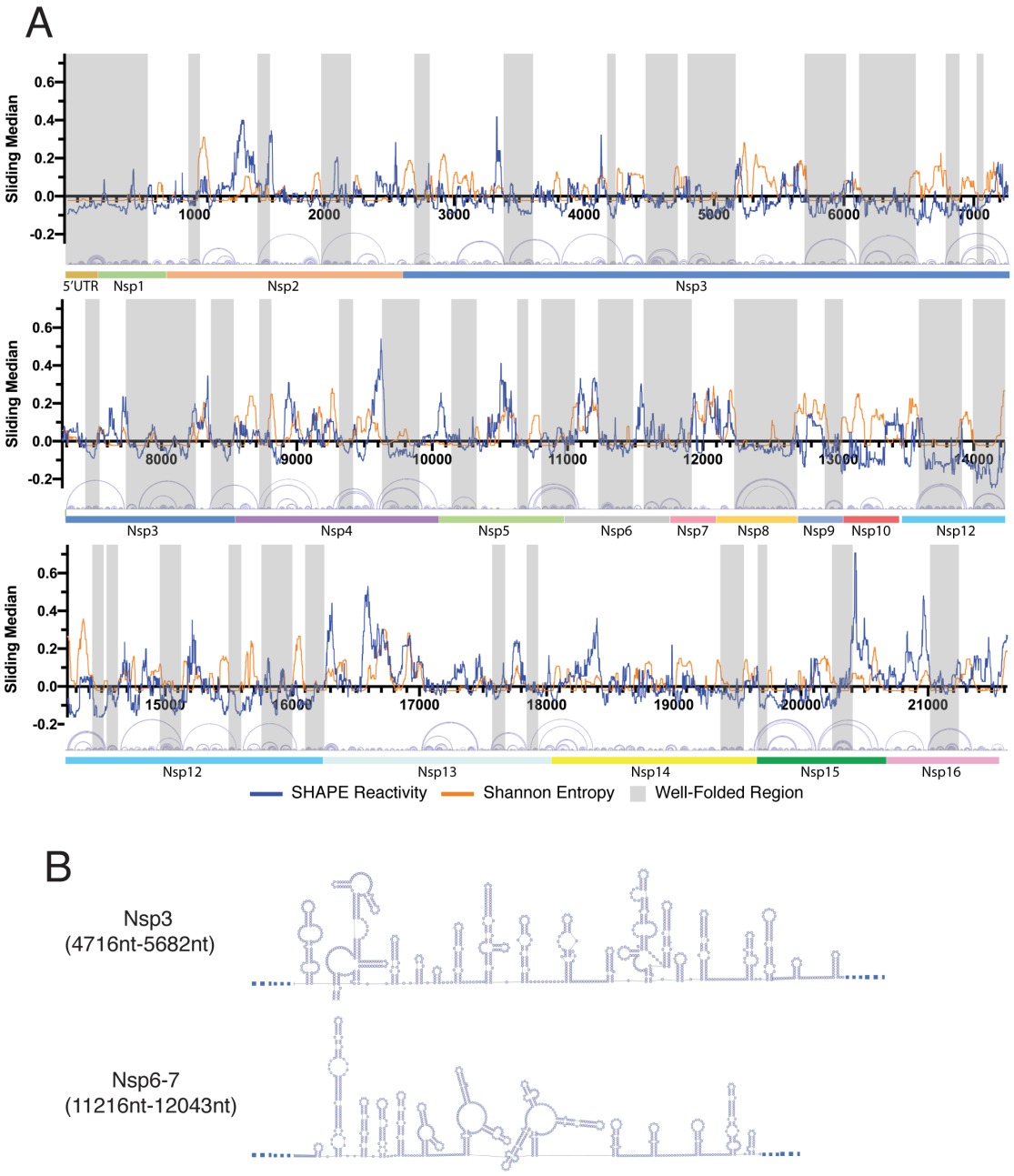

Figure 2.3. Full-length genome structure prediction of SARS-CoV-2 Orf1ab reveals a network of well-folded regions. **A)** Analysis of Shannon Entropy and SHAPE reactivities reveals 40 highly structured, well-determined domains in Orf1ab. Nucleotide coordinates are indicated on the x-axis. Local median SHAPE reactivity and Shannon Entropy are indicated by blue and orange lines, respectively. Well-folded regions are shaded with grey boxes. Arc plots for predicted base-pairing interactions in the structural model are shown below the x-axis. The 5'UTR and non-structural 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.

Orf1ab (**Fig 2.3A**, 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 (**Fig 2.4A**, grey bars). We find 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 (**Fig 2.4A,** 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 lies 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 - **Fig 2.3B**) 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 25nt and is significantly smaller than the median base-pairing distance in the HCV (median=40nt) and Dengue Virus (median=33nt) consensus models (**Fig 2.4B**). This suggests SARS-CoV-2 has fewer long-distance base-paring interactions compared to Dengue and HCV genome.

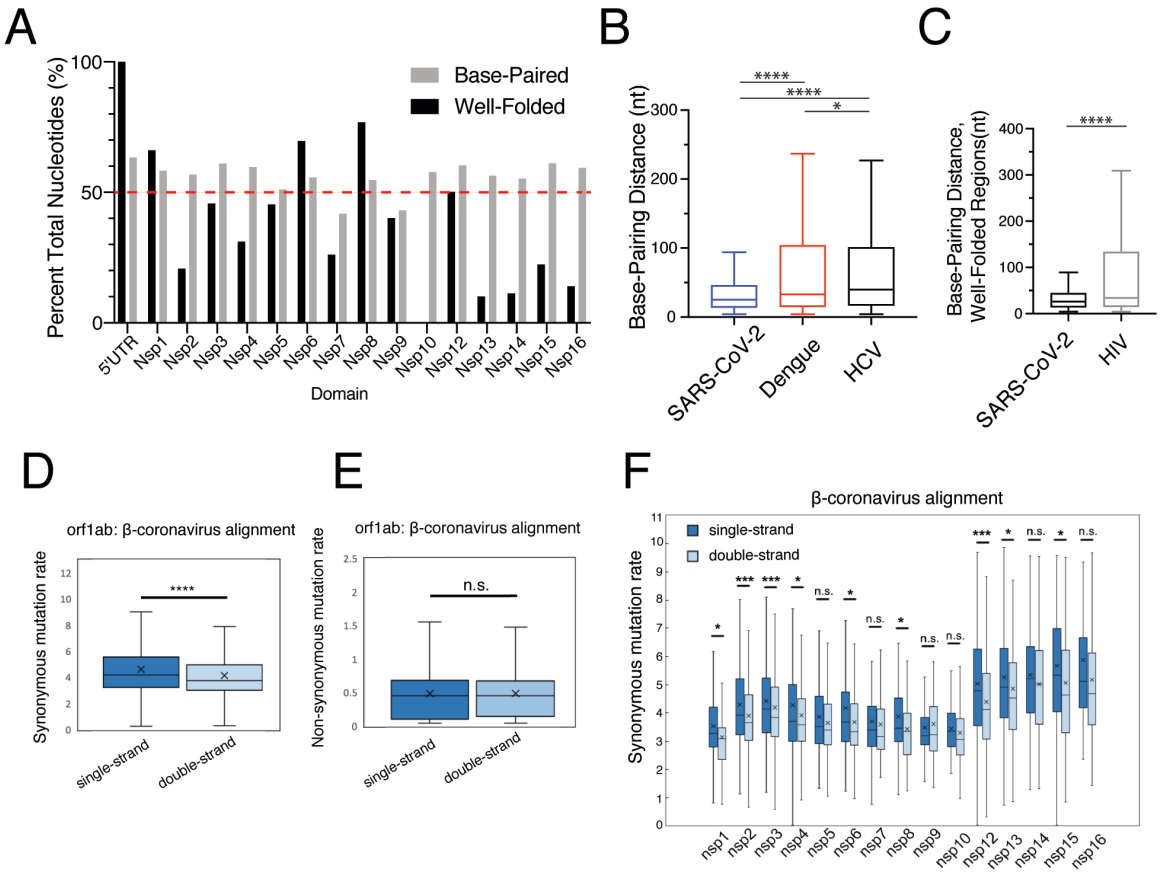

Figure 2.4 Full-length genome structure prediction of SARS-CoV-2 Orf1ab reveals unique and conserved genome architecture. **A)** Base-paired RNA content (grey bars) and well-folded RNA content (black bars) of individual Nsp domains. A dotted line at 50% nucleotide content has been added for clarity. **B)** Median base-pairing distance of the SARS-CoV-2, Hepatitis C virus, and Dengue virus. Dare presented as in Fig 2.1C inset **C)** Median base-pairing distance across well-folded regions identified in SARS-CoV-2 and HIV genomes **D)** Synonymous mutation rates (dS) calculated across β-coronaviruses for single- and double-stranded nucleotides of Orf1ab. **E)** Non-synonymous mutation rates 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. Data are presented as in Fig 2.1B. n.s. not significant,*p<0.05, ***p<0.001 ****p<0.0001 by equal variance unpaired student t test.

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-CoV2 and HIV genomes (data not shown), the median base-pairing distance in the well-folded regions of SARS-CoV-2 (median = 26nt) is significantly lower than the base-pairing distance in well-folded regions of HIV (median = 34nt) (**Fig 2.4C**).

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 (~30kb) relative to those of the positive-sense RNA viruses compared here (~10kb), it is possible that the short base-pairing distance of SARS-CoV2 may carry functional implications for maintaining genomic stability, preserving fidelity of translation, and evading innate immune response.

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

Synonymous mutations 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-CoV2 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 (**Fig 2.4D**). In contrast, there was no significant difference observed for non-synonymous mutation rates (dN) at single- or double-stranded codons (**Fig 2.4E**) 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 (**Fig 2.4F**). Consistent with this, Nsp1, Nsp6, Nsp8, and Nsp12 have >50% of their nucleotides localized within well-folded regions (**Fig 2.4A,** 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 (**Fig 2.4F**). 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 (**Fig 2.4A**). Not only does this analysis support the observation that these regions of RNA are not well-folded in SARS-CoV-2, our data suggest 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 (**Fig 2.3A, Table A2.1**). Four regions showed significantly decreased dS at double-stranded codons across the β-coronavirus alignment (**Fig 2.5A, 2.5B)**. Among those well-folded domains, region 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. Region 23, 34, and 36 (**Fig 2.5C, 2.5E, 2.5F**) contain a series of stem-loops with small bulges. Region 25 contains a long-range duplex that closes a clover-leaf like structure with 8 stem-loops radiating from a central loop (**Fig 2.5D**). 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).

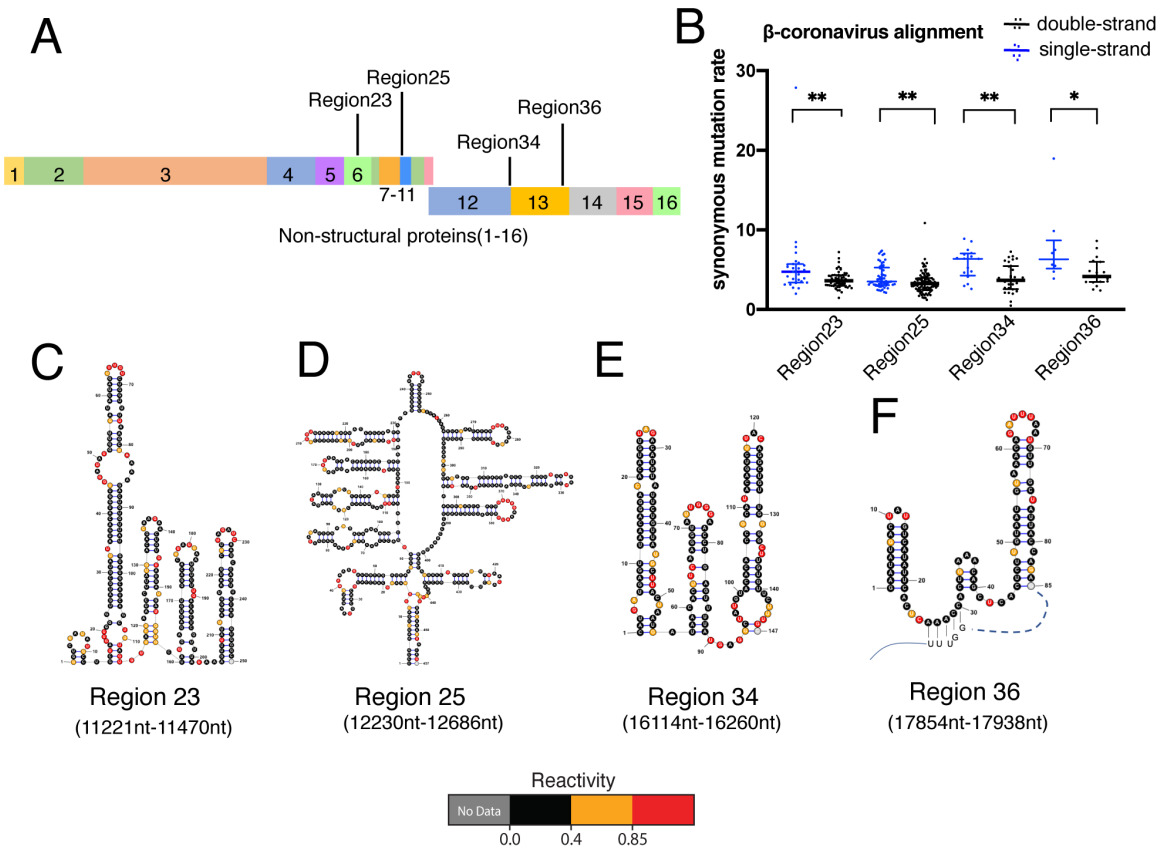

Figure 2.5. Analysis of synonymous mutation rates (dS) within individual well-folded regions of the SARS-CoV-2 genome across β-coronaviruses. A) Schematic of well-folded regions in SARS-COV2 genome supported by dS analysis in β-coronaviruses. **B)** dS separated by stranded-ness in four individual well-folded regions. Data are plotted with as in Fig 2.1C inset. *p<0.05, **p<0.01 by equal variance unpaired student t-test. **C), D), E), 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).

Within the Sarbecovirus subgenus, we were able to identify five well-folded regions with significantly decreased dS in double-stranded codons (**Fig 2.6A, 2.6B**). Among these well-folded domains, Region 24 contains two discrete multi-helix junctions, each with at least three stems radiating from large central loops (**Fig 2.6C**). Region 27 contains a series of six stem-loops (**Fig 2.6D)**. Region 15, like Region 24, contains several well-determined long-range duplexes that segment the region into two discrete multi-helix junctions (**Fig 2.6E**). Region 22 contains a series

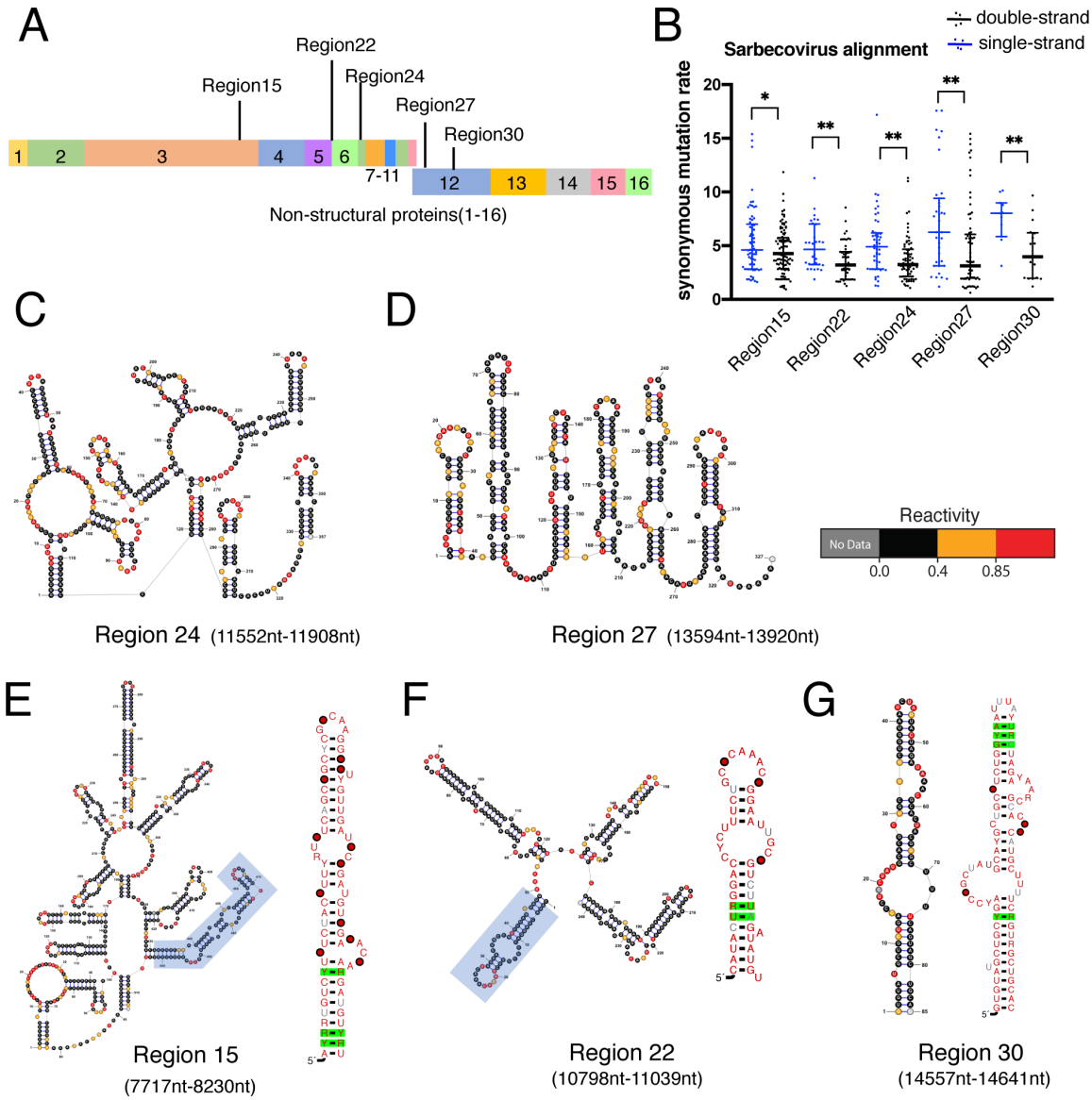

Figure 2.6. Analysis of synonymous mutation rates (dS) and covariation within individual regions of the SARS-CoV-2 genome within the sarbecovirus subgenus. **A)** Schematic of well-folded regions in the SARS-COV2 genome supported by dS analysis. **B)** dS separated by stranded-ness in five individual well-folded regions. Data are plotted Fig 2.1C inset. *p<0.05, **p<0.01 by equal variance unpaired student t test. **C), D)** RNA secondary structures of two well-folded regions colored by SHAPE reactivity **E), F), G)** RNA secondary structure diagrams of three well-folded regions supported by both synonymous mutation rate 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% = grey; 90% = black; 97% = red).

of well-folded loops and it spans the Nsp5/6 boundary (**Fig 2.6F**). Region 30 is a single stem-loop with bulges that divide the stem into distinct duplexes (**Fig 2.6G**).

To look for evolutionary evidence that directly supports conservation of specific base-pairing interactions and secondary structures, we performed covariation analysis on the 5 well-folded regions that are supported by Sarbecovirus-specific dS. We identified 3 regions (15, 22, and 30) that have covariation support (**Fig 2.6E-G**; 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 locked nucleic acids (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 analogues that enhance the T_m of a given paired duplex by 2-8°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 RNA structures in both the hepatitis C virus (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 (**Fig 2.6E, 2.6F**). LNAs targeted to

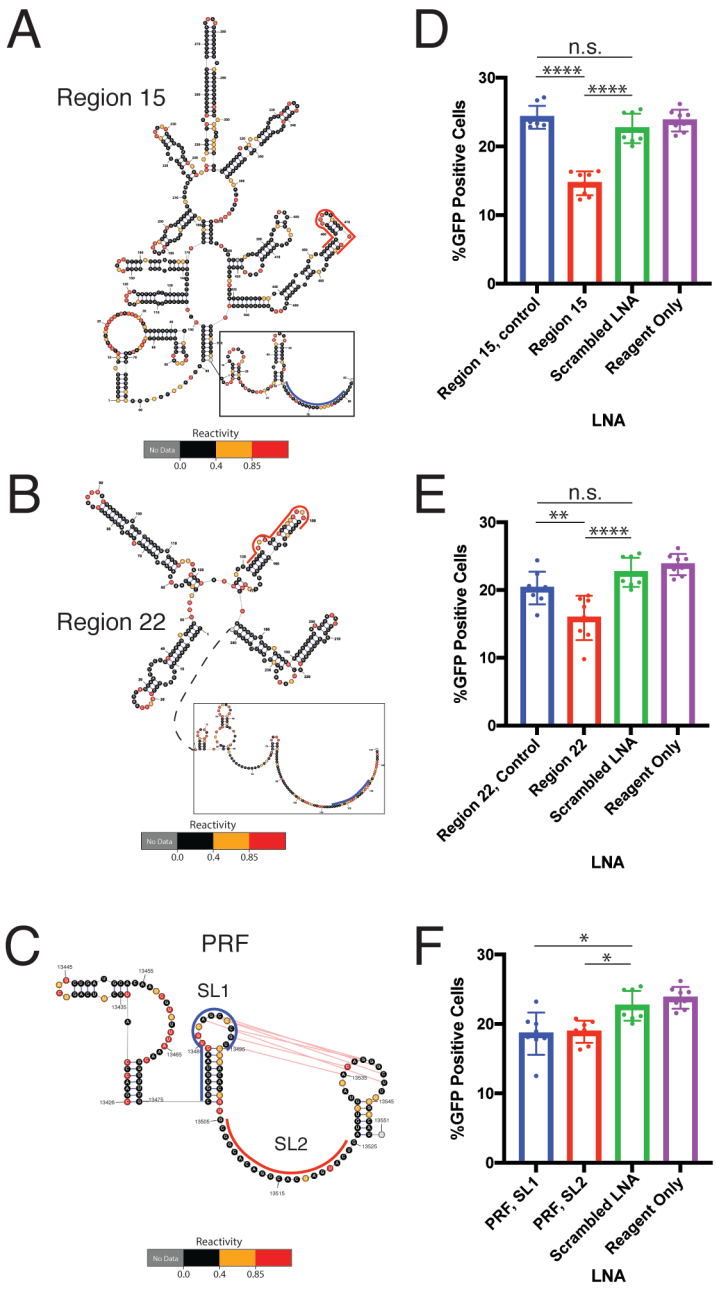

Figure 2.7. RNA structures disrupted by locked nucleotide acids (LNA) exhibit defects in SARS-CoV-2 viral growth. A) Schematic showing Region 15 LNA targeted to the covarying stem (red line) and control LNA (blue line). **B)** Schematic showing region 22 LNA targeted to stem (red line) and the control LNA (blue line). **C)** Schematic showing LNA targeted to the PRF SL1 region and the conformationally flexible SL2 region in SARS-CoV-2 PRF. **D, E, F)** Virus growth as measured and quantified by mNeonGreen expression at 24hpi. 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. Individual data points

represent technical replicates. Asterisk definitions are; *p<0.05, **p<0.01, ****p<0.0001 by ordinary one-way ANOVA with multiple comparisons.

these regions were designed for maximal structure disruption, hybridizing to the top of the stem loop as well as duplex RNA flanking the loop (**Fig 2.7A, 1.7B**; 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 (**Fig 2.7A, 2.7B**; 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 **Fig 2.7D,** 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 (**Fig 2.7E**). 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 (**Fig 2.7C**; red line). In addition, we designed a LNA targeted against the PRF pseudoknot (SL1) (**Fig 2.7C**; blue line) as disruption of the SARS-CoV PRF has

been demonstrated to reduce viral growth (Plant et al., 2013, Plant et al., 2005). This LNA results in an 18% reduction in GFP+ cells relative to the scrambled LNA control (**Fig 2.7F**). 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 region 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, this data lends strong support for a model in which well-folded regions with evolutionary support 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.

2.5 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 which 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 untranslated regions of the genome, as protein-coding sections of the SARS-CoV-2 open reading frame 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”)(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 from 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, and 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).

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

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 for SARS-CoV-2 and other β-coronavirus genomes suggests that the function of these upstream elements is conserved in coronaviruses (reviewed in (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 (reviewed in(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 SLI. 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 programmed ribosomal frameshifting pseudoknot (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 SLI, SL2, and a pseudoknotted helix, with an additional upstream attenuator stem 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 attenuator stem, 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. 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 frame-shifting 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. Secondly, 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 synonymous mutation rates suggests that global architectural features are conserved across β-coronaviruses. Considering the exceptional size of these genomes, the high degree of 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* (up to 65% of total cellular RNA content (Kim et al., 2020)) 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 non-structural proteins, 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 non-structural proteins in the large Orf1a and Orf1ab poly-proteins 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).

To explore structure-function relationships of representative, conserved RNA secondary structures, we used targeted antisense locked nucleic acids (LNAs) to induce structure disruption. Not only is this method faster than reverse genetics, it is 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 icSARS-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-CoV2 secondary structure presented here provides a roadmap for functional studies of the SARS-CoV2 genome and insights into mechanisms of the SARS-CoV-2 life cycle. Evolutionary support for 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.

2.6 Methods

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 450g 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 x 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 $5x10^6$ 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 450g 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 20,000g for 30min at 4°C, washed once with 70% EtOH, and spun again at 20,000g 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 A2.2**. 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 A2.3**. 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 $1/5^{th}$ 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 (accession number: 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 data sets. 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 data sets 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 data sets 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 (MN908947.3), four bat coronaviruses (MG772934.1, JX993987.1, DQ022305.2, DQ648857.1), and five human SARS coronaviruses (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 (MK129253, KP209307, MF598594, MG987420), one HKU-4 sequence (MH002337), three HKU-5 sequence (MH002342, NC009020, MH002341), four HKU1 sequences (KY674942, KF686343, AY597011, DQ415903), three murine hepatitis virus sequences

(AY700211, AF208067, AB551247), three human coronavirus OC43 sequences (AY585229, NC006213, MN026164), two bovine coronavirus sequences (KU558922, KU558923), and one camel coronavirus sequence (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 (accession number: 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 A2.4**).

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 *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.

2.7 Appendix

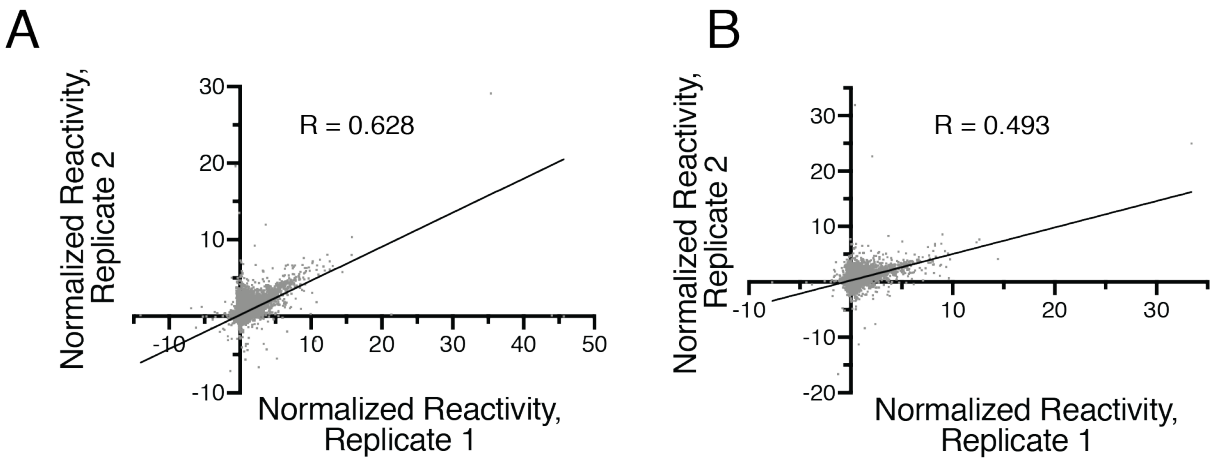

Figure A2.1. Analysis of correlation of normalized SHAPE reactivity reveals good agreement between biological replicates across Orf1ab, but not the subgenomic RNA region. Related to Figure 1. **A), B)** Correlation plot of normalized SHAPE reactivities from two biological replicates determined for the Orf1ab region or subgenomic RNA region, respectively. Lines represent linear regressions fit to the data. Pearson's correlation for each dataset is shown.

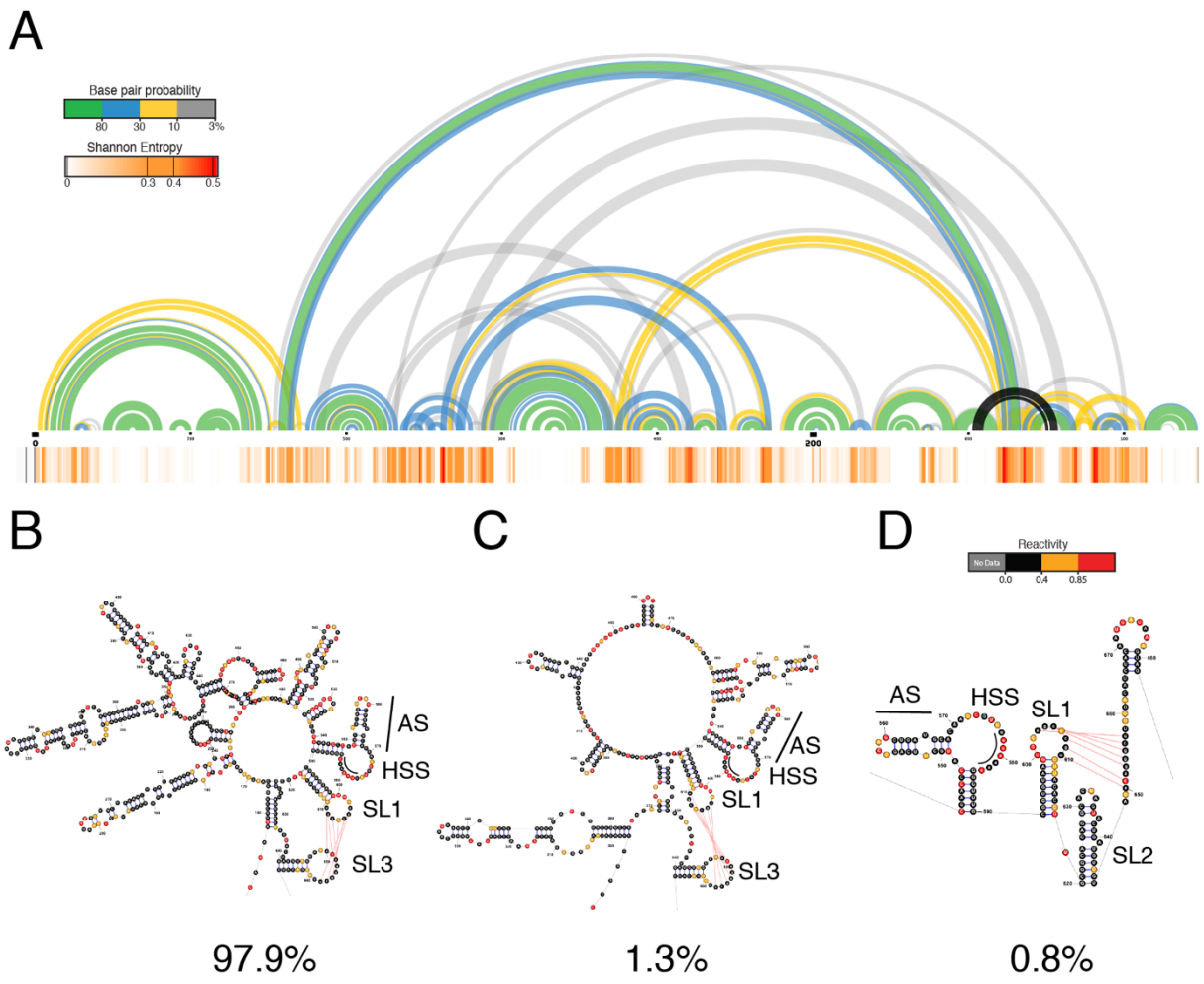

Figure A2.2. Ensemble analysis of the region containing the SARS-CoV-2 PRF confirms that the canonical three-stem pseudoknot structure represents a minority conformation. Related to Figure 3. **A)** Base-pair probabilities calculated with the partition function implemented in the SuperFold pipeline for the region of the SARS-CoV-2 genome containing the PRF. Arcs corresponding to individual base-pairs are colored by base-pairing probability (green = >80%; blue = 80% > base-pair probability > 30%; yellow = 30% > base-pair probability > 10%; grey = 10% > base-pair probability > 3%). A black arc indicates the PRF pseudoknot. Shannon Entropies for individual nucleotides, represented as a heat map, are shown underneath the corresponding nucleotides in the arc plots. **B), C), D)** Structural clusters that comprise the conformational ensemble representing the SARS-CoV-2 PRF. Relative abundances of each cluster are shown as percentages. AS = Attenuator Stem; HSS = Heptanucleotide Slippery Sequence; SL1 = Stem Loop 1; SL2 = Stem Loop 2; SL3 = Stem Loop 3; Red lines indicate pseudoknot interaction.

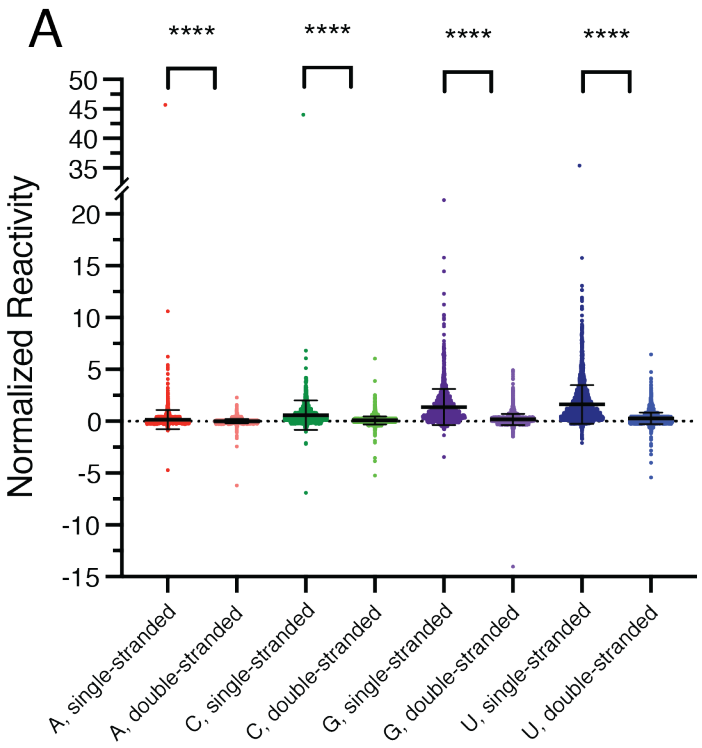

Figure A2.3. Reactivities separated by nucleotide identity and binned by stranded-ness reveals strong agreement between experimentally determined reactivities and the resulting structure prediction for each nucleotide. Related to Figure 4**. A)** Normalized SHAPE reactivity determined by ShapeMapper separated by nucleotide identity and binned by stranded-ness as determined in our consensus structure model. ****p<0.0001 by equal variance unpaired student t test.

Table A2.1. Well-determined region in Orf1ab region, related to **Figure 2.4**.

Region	Window	Start	End	Size	Protein Domain	Median Shannon Entropy
1	1	1	622	622	5'UTR, Nsp1	7.69E-05
2	1	944	1026	83	Nsp2	1.75E-06
3	1	1478	1572	95	Nsp2	2.62E-02
4	1	1968	2188	221	Nsp2	5.07E-04
5	1	2682	2800	119	Nsp2, Nsp3	9.19E-03
6	1	3416	3597	182	Nsp3	5.24E-05
7	1	4169	4232	64	Nsp3	3.01E-06
8	1	4471	4713	243	Nsp3	4.82E-04
9	1	4791	5162	372	Nsp3	4.55E-04
10	1	5693	6013	321	Nsp3	2.07E-03
11	1	6116	6549	434	Nsp3	4.24E-04
12	1	6786	6887	102	Nsp3	4.88E-03
13	1	7025	7072	48	Nsp3	1.06E-03
14	2	7413	7518	106	Nsp3	1.55E-04
15	2	7717	8230	514	Nsp3	1.29E-03
16	2	8350	8512	163	Nsp3	4.90E-04
17	2	8702	8789	88	Nsp4	6.45E-04
18	2	9295	9398	104	Nsp4	2.76E-02
19	2	9612	9886	275	Nsp4	1.70E-02
20	2	10129	10312	184	Nsp5	3.00E-04
21	2	10630	10687	58	Nsp5	3.53E-04
22	2	10798	11039	242	Nsp5, Nsp6	7.67E-04
23	2	11221	11470	250	Nsp6	2.34E-03
24	2	11552	11908	357	Nsp6, Nsp7	2.09E-03
25	2	12230	12686	457	Nsp8, Nsp9	1.86E-03
26	2	12895	13030	136	Nsp9, Nsp10	1.55E-02
27	2	13594	13920	327	Nsp12	7.39E-04
28	2	13993	14230	238	Nsp12	1.19E-02
29	3	14444	14532	89	Nsp12	4.27E-04
30	3	14557	14641	85	Nsp12	1.05E-05
31	3	14973	15136	164	Nsp12	9.07E-04
32	3	15510	15608	99	Nsp12	1.59E-04
33	3	15767	16005	239	Nsp12	4.29E-04
34	3	16114	16260	147	Nsp12, 13	2.69E-03
35	3	17580	17677	98	Nsp13	1.64E-04
36	3	17854	17938	85	Nsp13	6.09E-04
37	3	19373	19550	178	Nsp14	1.52E-02
38	3	19665	19735	71	Nsp15	2.32E-05
39	3	20248	20408	161	Nsp15	7.72E-03
40	3	20668	20792	125	Nsp16	4.91E-06

Table A2.2. Gene-specific RT primers, related to Methods.

Primer Name	Sequence
RT_SC2_Amplicon_1	TTAGTCAAATTCTCAGTGC
RT_SC2_Amplicon_2	TTTGTTGACTATCATCATC
RT_SC2_Amplicon_3	AAACATAAAATGTTTTACC
RT_SC2_Amplicon_4	AATTAGACATTAAAACACC
RT_SC2_Amplicon_5	TACCAACTGCACTAAAAAC
RT_SC2_Amplicon_6	TATCTAAAACGGCAATTCC
RT_SC2_Amplicon_7	AAGCAGTTTGTGTAGTACC
RT_SC2_Amplicon_8	TTAGTAAGTGCAGCTACTG
RT_SC2_Amplicon_9	TAACATTATCGCTACCAAC
RT_SC2_Amplicon_10	TAACTCTGGAAAAATCTGT
RT_SC2_Amplicon_11	AACCACCTAACTGACTATG
RT_SC2_Amplicon_12	TAATACCTATTGGCAAATC
RT_SC2_Amplicon_13	AATCATTTCATCTGTGAGC
RT_SC2_Amplicon_14	TAACATGTTCAACACCAGT
RT_SC2_Amplicon_15	ATGTTGAGTACATGACTGT
RT_SC2_Amplicon_16	TTTTTTTTTGTCATTCTCC

Table A2.3. Gene-specific PCR Primers, related to Methods.

Primer Name	Sequence
F_PCR_SC2_Amplicon_1	ATTAAAGGTTTATACCTTCCCAG
F_PCR_SC2_Amplicon_2	CTCATGAAGTGTGATCATTGTGG
F_PCR_SC2_Amplicon_3	GATTACCAAGGTAAACCTTTGGA
F_PCR_SC2_Amplicon_4	TATGGACAACAGTTTGGTCCAAC
F_PCR_SC2_Amplicon_5	ATAAATATTATAATTTGGTTTTTACTATTA
F_PCR_SC2_Amplicon_6	AAGAGAAGTGGGTTTTGTCG
F_PCR_SC2_Amplicon_7	TGTGGCTATGAAGTACAATTATG
F_PCR_SC2_Amplicon_8	TGTAACAGCTTTAAGGGCCAATT
F_PCR_SC2_Amplicon_9	TAAGGAATTACTTGTGTATGCTG
F_PCR_SC2_Amplicon_10	TTATTGTAAATCACATAAACCAC
F_PCR_SC2_Amplicon_11	ACAGCTAGGTTTTTCTACAGGTG
F_PCR_SC2_Amplicon_12	TTAGAATTAGCTATGGATGAATT
F_PCR_SC2_Amplicon_13	TATATTCTAAGCACACGCCTATT
F_PCR_SC2_Amplicon_14	GTGATTGCCTTGGTGATATT
F_PCR_SC2_Amplicon_15	TCTGGAGTAAAAGACTGTGTTGT
F_PCR_SC2_Amplicon_16	GTCACGCCTAAACGAACATG
R_PCR_SC2_Amplicon_1	AGTCAAATTCTCAGTGCCACAA
R_PCR_SC2_Amplicon_2	TGTTGACTATCATCATCTAACCA
R_PCR_SC2_Amplicon_3	ACATAAAATGTTTTACCTTCATG
R_PCR_SC2_Amplicon_4	TTAGACATTAAAACACCTAAAGC
R_PCR_SC2_Amplicon_5	CCAACTGCACTAAAAACTCTAGG
R_PCR_SC2_Amplicon_6	CTAAAACGGCAATTCCAGTT
R_PCR_SC2_Amplicon_7	CAGTTTGTGTAGTACCGGCA
R_PCR_SC2_Amplicon_8	GTAAGTGCAGCTACTGAAAAGCA
R_PCR_SC2_Amplicon_9	CATTATCGCTACCAACACATGTA
R_PCR_SC2_Amplicon_10	CTCTGGAAAAATCTGTATTATTAGG
R_PCR_SC2_Amplicon_11	CACCTAACTGACTATGACTAAAA
R_PCR_SC2_Amplicon_12	TACCTATTGGCAAATCTACCAAT
R_PCR_SC2_Amplicon_13	CATTTCATCTGTGAGCAAAG
R_PCR_SC2_Amplicon_14	CATGTTCAACACCAGTGTCTGTA
R_PCR_SC2_Amplicon_15	TTGAGTACATGACTGTAAACTACAT
R_PCR_SC2_Amplicon_16	TTTTTTGTCATTCTCCTAAGAAG

Table A2.4. LNAs used in this study (LNA bases are indicated with a "+" on the left), related to Methods.

Region	LNA	LNA content	%GC	RNA Tm
PRF, SL1	+A+C+GG+GC+TGC+ACT+TA+CA+C+C+G	57.9	63.2	90
PRF, SL2	+C+A+GTAC+TAG+TG+CC+TG+TGC+C+G+C	52.4	61.9	89
Region 15	+A+C+AAA+CCC+TTG+CCG+AG+CT+G+C+T	52.4	57.1	91
Region 15, Control	+G+T+TT+TCA+ACT+TTG+TTA+TAG+G+T+G	50.0	31.8	86
Region22	+G+T+CTA+ACA+ACA+TCA+AA+AG+G+T+G	52.4	38.1	86
Region 22, Control	+G+C+TAC+AG+TGG+CAA+GAG+AA+G+G+T	52.4	52.4	86
Scrambled LNA	+G+C+GGC+ACG+TTG+CG+AGT+A+C+T	52.6	63.2	N/A

2.8 References

1. ANDREWS, R. J., PETERSON, J. M., HANIFF, H. S., CHEN, J., WILLIAMS, C., GREFE, M., DISNEY, M. D. & MOSS, W. N. 2020. An in silico map of the SARS-CoV-2 RNA Structurome. *bioRxiv*.
2. ASSIS, R. 2014. Strong epistatic selection on the RNA secondary structure of HIV. *PLoS Pathog,* 10**,** e1004363.
3. BARANOV, P. V., HENDERSON, C. M., ANDERSON, C. B., GESTELAND, R. F., ATKINS, J. F. & HOWARD, M. T. 2005. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology,* 332**,** 498-510.
4. BENSON, D. A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., OSTELL, J., PRUITT, K. D. & SAYERS, E. W. 2018. GenBank. *Nucleic Acids Res,* 46**,** D41-D47.
5. BUSAN, S. & WEEKS, K. M. 2017. Visualization of RNA structure models within the Integrative Genomics Viewer. *RNA,* 23**,** 1012-1018.
6. BUSAN, S. & WEEKS, K. M. 2018. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA,* 24**,** 143-148.
7. CERAOLO, C. & GIORGI, F. M. 2020. Genomic variance of the 2019-nCoV coronavirus. *J Med Virol,* 92**,** 522-528.
8. CHEN, C., ZHANG, H., BROITMAN, S. L., REICHE, M., FARRELL, I., COOPERMAN, B. S. & GOLDMAN, Y. E. 2013. Dynamics of translation by single ribosomes through mRNA secondary structures. *Nat Struct Mol Biol,* 20**,** 582-8.
9. CHEN, S. C. & OLSTHOORN, R. C. 2010. Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology,* 401**,** 29-41.
10. CHO, C. P., LIN, S. C., CHOU, M. Y., HSU, H. T. & CHANG, K. Y. 2013. Regulation of programmed ribosomal frameshifting by co-translational refolding RNA hairpins. *PLoS One,* 8**,** e62283.
11. COLLART, M. A. & WEISS, B. 2020. Ribosome pausing, a dangerous necessity for co-translational events. *Nucleic Acids Res,* 48**,** 1043-1055.

12. DE WIT, E., VAN DOREMALEN, N., FALZARANO, D. & MUNSTER, V. J. 2016. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol,* 14**,** 523-34.
13. 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. & WEEKS, K. M. 2018. Pervasive tertiary structure in the dengue virus RNA genome. *Proc Natl Acad Sci U S A,* 115**,** 11513-11518.
14. DIAS JUNIOR, A. G., SAMPAIO, N. G. & REHWINKEL, J. 2019. A Balancing Act: MDA5 in Antiviral Immunity and Autoinflammation. *Trends Microbiol,* 27**,** 75-85.
15. DING, Y., CHAN, C. Y. & LAWRENCE, C. E. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA,* 11**,** 1157-66.
16. DING, Y. & LAWRENCE, C. E. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res,* 31**,** 7280-301.
17. DONG, E., DU, H. & GARDNER, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis,* 20**,** 533-534.
18. FEDOROVA, O., JAGDMANN, G. E., JR., ADAMS, R. L., YUAN, L., VAN ZANDT, M. C. & PYLE, A. M. 2018. Small molecules that target group II introns are potent antifungal agents. *Nat Chem Biol,* 14**,** 1073-1078.
19. FRIEBE, P. & BARTENSCHLAGER, R. 2009. Role of RNA structures in genome terminal sequences of the hepatitis C virus for replication and assembly. *J Virol,* 83**,** 11989-95.
20. GOEBEL, S. J., HSUE, B., DOMBROWSKI, T. F. & 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**,** 669-82.
21. GOEBEL, S. J., MILLER, T. B., BENNETT, C. J., BERNARD, K. A. & 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**,** 1274-87.
22. GUO, L. T., ADAMS, R. L., WAN, H., HUSTON, N. C., POTAPOVA, O., OLSON, S., GALLARDO, C. M., GRAVELEY, B. R., TORBETT, B. E. & PYLE, A. M. 2020. Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. *J Mol Biol,* 432**,** 3338-3352.
23. HAJDIN, C. E., BELLAOUSOV, S., HUGGINS, W., LEONARD, C. W., MATHEWS, D. H. & WEEKS, K. M. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci U S A,* 110**,** 5498-503.
24. HEWITT, W. M., CALABRESE, D. R. & SCHNEEKLOTH, J. S., JR. 2019. Evidence for ligandable sites in structured RNA throughout the Protein Data Bank. *Bioorg Med Chem,* 27**,** 2253-2260.
25. KELLY, J. A., OLSON, A. N., NEUPANE, K., MUNSHI, S., SAN EMETERIO, J., POLLACK, L., WOODSIDE, M. T. & 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*.
26. KIM, D., LEE, J. Y., YANG, J. S., KIM, J. W., KIM, V. N. & CHANG, H. 2020. The Architecture of SARS-CoV-2 Transcriptome. *Cell,* 181**,** 914-921 e10.
27. KORBIE, D. J. & MATTICK, J. S. 2008. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat Protoc,* 3**,** 1452-6.

28. LAI, D., PROCTOR, J. R., ZHU, J. Y. & MEYER, I. M. 2012. R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res,* 40**,** e95.
29. LAN, J., GE, J., YU, J., SHAN, S., ZHOU, H., FAN, S., ZHANG, Q., SHI, X., WANG, Q., ZHANG, L. & WANG, X. 2020a. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature,* 581**,** 215-220.
30. LAN, T. C. T., ALLAN, M., MALSICK, L., KHANDWALA, S., NYEO, S. Y., BATHE, M., GRIFFITHS, A. & ROUSKIN, S. 2020b. Structure of the full SARS-CoV-2 RNA genome in infected cells. *bioRxiv*.
31. LEAMY, K. A., ASSMANN, S. M., MATHEWS, D. H. & BEVILACQUA, P. C. 2016. Bridging the gap between in vitro and in vivo RNA folding. *Q Rev Biophys,* 49**,** e10.
32. LI, P., WEI, Y., MEI, M., TANG, L., SUN, L., HUANG, W., ZHOU, J., ZOU, C., ZHANG, S., QIN, C. F., JIANG, T., DAI, J., TAN, X. & ZHANG, Q. C. 2018. Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity. *Cell Host Microbe,* 24**,** 875-886 e5.
33. LU, Z. J. & MATHEWS, D. H. 2008. OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. *Nucleic Acids Res,* 36**,** W104-8.
34. LUNDIN, K. E., HOJLAND, T., HANSEN, B. R., PERSSON, R., BRAMSEN, J. B., KJEMS, J., KOCH, T., WENGEL, J. & SMITH, C. I. 2013. Biological activity and biotechnological aspects of locked nucleic acids. *Adv Genet,* 82**,** 47-107.
35. MADHUGIRI, R., KARL, N., PETERSEN, D., LAMKIEWICZ, K., FRICKE, M., WEND, U., SCHEUER, R., MARZ, M. & ZIEBUHR, J. 2018. Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions. *Virology,* 517**,** 44-55.
36. MAIER, H. J., BICKERTON, E. & BRITTON, P. 2015. *Coronaviruses : methods and protocols,* New York, Humana Press ; Springer.
37. 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. & INCARNATO, D. 2020. Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. *bioRxiv*.
38. MATHEWS, D. H. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA,* 10**,** 1178-90.
39. MAUGER, D. M., GOLDEN, M., YAMANE, D., WILLIFORD, S., LEMON, S. M., MARTIN, D. P. & WEEKS, K. M. 2015. Functionally conserved architecture of hepatitis C virus RNA genomes. *Proc Natl Acad Sci U S A,* 112**,** 3692-7.
40. MITCHELL, D., 3RD, ASSMANN, S. M. & BEVILACQUA, P. C. 2019. Probing RNA structure in vivo. *Curr Opin Struct Biol,* 59**,** 151-158.
41. MURRELL, B., MOOLA, S., MABONA, A., WEIGHILL, T., SHEWARD, D., KOSAKOVSKY POND, S. L. & SCHEFFLER, K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol,* 30**,** 1196-205.
42. NISHIKURA, K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem,* 79**,** 321-49.
43. PIRAKITIKULR, N., KOHLWAY, A., LINDENBACH, B. D. & PYLE, A. M. 2016. The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. *Mol Cell,* 62**,** 111-20.

44. PLANT, E. P. & DINMAN, J. D. 2008. The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front Biosci,* 13**,** 4873-81.
45. PLANT, E. P., PEREZ-ALVARADO, G. C., JACOBS, J. L., MUKHOPADHYAY, B., HENNIG, M. & DINMAN, J. D. 2005. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol,* 3**,** e172.
46. PLANT, E. P., SIMS, A. C., BARIC, R. S., DINMAN, J. D. & TAYLOR, D. R. 2013. Altering SARS coronavirus frameshift efficiency affects genomic and subgenomic RNA production. *Viruses,* 5**,** 279-94.
47. RANGAN, R., ZHELUDEV, I. N. & DAS, R. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA*.
48. RANWEZ, V., DOUZERY, E. J. P., CAMBON, C., CHANTRET, N. & DELSUC, F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol Biol Evol,* 35**,** 2582-2584.
49. REGULSKI, E. E. & BREAKER, R. R. 2008. In-line probing analysis of riboswitches. *Methods Mol Biol,* 419**,** 53-67.
50. REUTER, J. S. & MATHEWS, D. H. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics,* 11**,** 129.
51. RIVAS, E., CLEMENTS, J. & EDDY, S. R. 2017. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods,* 14**,** 45-48.
52. ROBERTSON, M. P., IGEL, H., BAERTSCH, R., HAUSSLER, D., ARES, M., JR. & SCOTT, W. G. 2005. The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol,* 3**,** e5.
53. ROUSKIN, S., ZUBRADT, M., WASHIETL, S., KELLIS, M. & WEISSMAN, J. S. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature,* 505**,** 701-5.
54. SIEGFRIED, N. A., BUSAN, S., RICE, G. M., NELSON, J. A. & WEEKS, K. M. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods,* 11**,** 959-65.
55. SIMMONDS, P. & SMITH, D. B. 1999. Structural constraints on RNA virus evolution. *J Virol,* 73**,** 5787-94.
56. SIMON, L. M., MORANDI, E., LUGANINI, A., GRIBAUDO, G., MARTINEZ-SOBRIDO, L., TURNER, D. H., OLIVIERO, S. & INCARNATO, D. 2019. In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res,* 47**,** 7003-7017.
57. SMOLA, M. J., CALABRESE, J. M. & WEEKS, K. M. 2015a. Detection of RNA-Protein Interactions in Living Cells with SHAPE. *Biochemistry,* 54**,** 6867-75.
58. SMOLA, M. J., CHRISTY, T. W., INOUE, K., NICHOLSON, C. O., FRIEDERSDORF, M., KEENE, J. D., LEE, D. M., CALABRESE, J. M. & 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 U S A,* 113**,** 10322-7.
59. SMOLA, M. J., RICE, G. M., BUSAN, S., SIEGFRIED, N. A. & 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**,** 1643-69.

60. SON, J., HUANG, S., ZENG, Q., BRICKER, T. L., CASE, J. B., ZHOU, J., ZANG, R., LIU, Z., CHANG, X., HARASTANI, H. H., CHEN, L., GOMEZ CASTRO, M. F., ZHAO, Y., KOHIO, H. P., HOU, G., FAN, B., NIU, B., GUO, R., ROTHLAUF, P. W., BAILEY, A. L., WANG, X., SHI, P. Y., WHELAN, S. P. J., DIAMOND, M. S., BOON, A. C. M., LI, B. & DING, S. 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*.
61. SPASIC, A., ASSMANN, S. M., BEVILACQUA, P. C. & MATHEWS, D. H. 2018. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res,* 46**,** 314-323.
62. TAVARES, R. C. A., MAHADESHWAR, G., WAN, H., HUSTON, N. C. & PYLE, A. M. 2020. The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. *J Virol*.
63. TAVARES, R. C. A., PYLE, A. M. & SOMAROWTHU, S. 2019. Phylogenetic Analysis with Improved Parameters Reveals Conservation in lncRNA Structures. *J Mol Biol,* 431**,** 1592-1603.
64. TUPLIN, A., STRUTHERS, M., COOK, J., BENTLEY, K. & 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**,** 2914-26.
65. TUPLIN, A., WOOD, J., EVANS, D. J., PATEL, A. H. & SIMMONDS, P. 2002. Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. *RNA,* 8**,** 824-41.
66. VAN TREECK, B., PROTTER, D. S. W., MATHENY, T., KHONG, A., LINK, C. D. & PARKER, R. 2018. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proc Natl Acad Sci U S A,* 115**,** 2734-2739.
67. WAN, Y., KERTESZ, M., SPITALE, R. C., SEGAL, E. & CHANG, H. Y. 2011. Understanding the transcriptome through RNA structure. *Nat Rev Genet,* 12**,** 641-55.
68. WAN, Y., SHANG, J., GRAHAM, R., BARIC, R. S. & 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.
69. WARNER, K. D., HAJDIN, C. E. & WEEKS, K. M. 2018. Principles for targeting RNA with drug-like small molecules. *Nat Rev Drug Discov,* 17**,** 547-558.
70. WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.
71. 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., CHEN, J. S., MANKOWSKI, M. C., FILLER, R. B., RAVINDRA, N. G., GASQUE, V., DE MIGUEL, F. J., PATIL, A., CHEN, H., OGUNTUYO, K. Y., ABRIOLA, L., SUROVTSEVA, Y. V., ORCHARD, R. C., LEE, B., LINDENBACH, B. D., POLITI, K., VAN DIJK, D., KADOCH, C., SIMON, M. D., YAN, Q., DOENCH, J. G. & WILEN, C. B. 2020. Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. *Cell*.
72. XIE, X., MURUATO, A., LOKUGAMAGE, K. G., NARAYANAN, K., ZHANG, X., ZOU, J., LIU, J., SCHINDEWOLF, C., BOPP, N. E., AGUILAR, P. V., PLANTE, K. S., WEAVER, S.

C., MAKINO, S., LEDUC, J. W., MENACHERY, V. D. & SHI, P. Y. 2020. An Infectious cDNA Clone of SARS-CoV-2. *Cell Host Microbe,* 27**,** 841-848 e3.
73. YANG, D. & LEIBOWITZ, J. L. 2015. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res,* 206**,** 120-33.
74. YIN, W., MAO, C., LUAN, X., SHEN, D. D., SHEN, Q., SU, H., WANG, X., ZHOU, F., ZHAO, W., GAO, M., CHANG, S., XIE, Y. C., TIAN, G., JIANG, H. W., TAO, S. C., SHEN, J., JIANG, Y., JIANG, H., XU, Y., ZHANG, S., ZHANG, Y. & XU, H. E. 2020. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science,* 368**,** 1499-1504.
75. YOU, S., STUMP, D. D., BRANCH, A. D. & 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**,** 1352-66.
76. ZHU, N., ZHANG, D., WANG, W., LI, X., YANG, B., SONG, J., ZHAO, X., HUANG, B., SHI, W., LU, R., NIU, P., ZHAN, F., MA, X., WANG, D., XU, W., WU, G., GAO, G. F., TAN, W., CHINA NOVEL CORONAVIRUS, I. & RESEARCH, T. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med,* 382**,** 727-733.
77. ZUBRADT, M., GUPTA, P., PERSAD, S., LAMBOWITZ, A. M., WEISSMAN, J. S. & ROUSKIN, S. 2017. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods,* 14**,** 75-82.
78. ZUST, R., MILLER, T. B., GOEBEL, S. J., THIEL, V. & 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**,** 1214-28.

Chapter 3: West Nile virus genome harbors essential riboregulatory elements with conserved and host-specific functional roles

3.1 Preface

This work presented in Chapter 2 represents a collaborative effort between several contributors: Nicholas C. Huston, Douglas Brackney, and Anna Marie Pyle. This work would not have been possible without the additional help of Han Wan, Rafael Tavares, and Benjamin Gottë for thoughtful comments and advice on experimental design, and Olga Fedorova and Sarah Fergione for synthesizing LNAs.

3.2 Abstract

West Nile virus (WNV) is an arthropod-borne, positive-sense RNA virus that, due to warming climates and lack of effective therapeutics, poses an increasing global threat. Like other enzootic viruses, very little is known about how host context affects the structure of the WNV RNA genome beyond the extreme viral termini. Here, we report a complete secondary structure of the WNV genome in mammalian and arthropod cell lines. Our detailed analysis affords novel structural insights into multiple, conserved aspects of flaviviral biology. By comparing structures obtained in different cellular contexts, we reveal a genome that folds with minimal host-dependence, including regions of well-folded RNA. Using structural homology as a guide, we prioritize well-folded regions for functional validation using structure-disrupting, anti-sense locked nucleic acids. We demonstrate that the WNV genome contains riboregulatory structures with conserved and host-specific functional roles, highlighting the therapeutic potential of a novel class of nucleic acids as both WNV-specific and pan-flaviviral anti-virals.

3.3 Introduction

West Nile virus (WNV) is an arthropod-borne (arbo-) virus with a single-stranded, positive-sense RNA genome. A member of the genus *Flavivirus* (*Flaviviridae*) along with dengue virus and Zike virus, these viruses are maintained in a transmission cycle between a vertebrate reservoir host and invertebrate mosquito vectors. Due to a continued spread across the globe facilitated by a warming climate, *flaviviruses* represent an increasingly serious global health threat (Whitehorn and Yacoub, 2019). WNV poses a particular threat to residents of our institution and state, as it has been detected in Connecticut every year since it was first introduced to the United States in. As no human vaccine or effective therapeutic exists against WNV, there is an urgent need for research that expands our understanding of WNV biology and facilitates the development of both WNV and pan-*flaviviral* anti-virals.

The WNV genome is 11kb, encodes 10 structural and non-structural proteins translated as single poly-protein, and is flanked by 5' and 3' untranslated regions (UTRs). The WNV UTRs are highly structured, and these structures expand the virus' functional repertoire by mediating crucial steps in the viral life cycle. A conserved structure in the 5'UTR, called the Stem Loop A (SLA) promoter, plays an essential role in genome replication by recruiting the viral polymerase (NS5) (Choi, 2021; Dong et al., 2008; Lee et al., 2021). Structures resident in the 3'UTR, one of the best-studied regions of the WNV genome, facilitate innate immune evasion and pathogenesis via liberation of a non-coding RNA species, called the sub-genomic flaviviral RNA (sfRNA), in a process conserved across *flaviviruses* (Göertz et al.,

2016; MacFadden et al., 2018; Pijlman et al., 2008). While the structure of the sfRNA is well understood, it is not known how the region that gives rise to the sfRNA folds in the context of the full-length genome.

Functional RNA structures in the WNV genome can also be dynamic. Complementary regions at the 5' and 3' viral termini form long-range duplexes, allowing the genome to alternate between a linear and cyclized conformation (Basu and Brinton, 2011; Suzuki et al., 2008; Zhang et al., 2008). Genome cyclization is absolutely required for viral replication, as it allows NS5 to pass from the SLA promoter to the 3' viral terminus to initiate negative strand synthesis. Importantly, disruption of WNV genome cyclization does not affect viral translation, suggesting that the linear form is the translation-competent form (Friebe et al., 2011). This gave rise to a model in which genome cyclization functions as a molecular switch, and a subsequent search to identify factors that influence the process. While several groups have separately pointed to both host binding proteins and intrinsic sequence elements as deterministic factors, it has been shown that a balance of linear and cyclized genome conformations is essential for viral fitness (Davis et al., 2013; Iglesias and Gamarnik, 2011; Liu et al., 2016; Villordo et al., 2010). However, the balance of conformational states has never been directly assessed in cells with a full-length *flaviviral* genome.

Though the functional RNA content of the viral termini has been well studied, little else is known about the structural content of the WNV genome. Guided by *in silico* studies, researchers have identified and validated two programmed ribosomal frame-shifting pseudoknots in NS2A and NS4B (Faggioni et al., 2012; Melian et al.,

2010). However, these account for a vanishingly small fraction of WNV open-reading frame (ORF). With the advent of high-throughput RNA structure probing methods, functional RNA elements have been identified throughout the ORFs of several other single-stranded RNA viruses (Dethoff et al., 2018; Huston et al., 2021; Li et al., 2018; Siegfried et al., 2014; Wan et al., 2022). However, these efforts have yet to be extended to the WNV ORF.

Prior work with both viral and messenger RNAs has highlighted the importance of probing RNAs in their natural cellular context (Li et al., 2018; Rouskin et al., 2014; Simon et al., 2019). As WNV is maintained in an enzootic cycle between vertebrate and invertebrate hosts, a full understanding of its genome structure therefore requires studying how it folds in multiple cellular contexts. Indeed, vertebrate and invertebrate model cell systems have evolutionarily distant host proteomes, varying intracellular salt concentrations, and require culturing temperatures that differ by ~10˚C, all of which are features individually known to have important effects on the folding of functional RNAs (Kikovska et al., 2007; Kortmann and Narberhaus, 2012; Pyle, 2002). In fact, careful analysis of dengue and WNV 3'UTRs have identified functional elements that are host-specific and thermally responsive, respectively (de Borba et al., 2019; Meyer et al., 2020; Villordo et al., 2015). To date, no genome-wide study of functional RNA structure in a viral genome has been conducted in multiple hosts.

Here, we report the complete secondary structure of the WNV genome in arthropod and mammalian cell lines using selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) (Siegfried et al., 2014).

The SHAPE-MaP data is of exceptional quality, and the resulting genomic secondary structure model perfectly recapitulates the SLA motif in the 5'UTR. We closely examine genome cyclization *in vivo,* relying on a protein-free *in vitro* system in parallel to elucidate the natural conformational dynamics of the WNV genome in infected cells. We additionally identify a novel tripartite domain architecture at the 3' viral terminus, highlighting the importance of studying viral RNA structures in their native genomic context. We describe a global genome architecture that, along with specific regions of well-folded RNA, folds with minimal host-dependence. Relying on patterns of RNA structural homology between hosts, we prioritize specific RNA structures for functional validation. Using structure disrupting, anti-sense locked nucleic acids (LNAs), we demonstrate that a subset of these well-folded RNA structures play both conserved and host-specific functional roles. Our work deepens our understanding of WNV biology, identifies conserved aspects of the viral life cycle that are readily targetable by a novel class of nucleic acids, and therefore represents an important step forward in our fight against an expanding global health threat.

3.4 Results

In vivo pipeline yields high quality SHAPE-MaP data from two cell types

Like most *flaviviruses,* West Nile virus (WNV) is naturally maintained in an enzootic cycle between mosquitoes and birds, with humans serving as incidental hosts (Brinton, 2014). Therefore, we reasoned it was important to query the WNV genome structure in multiple cellular contexts under physiologically relevant

temperatures. To that end, the WNV genomic RNA was *in vitro* transcribed from an infectious clone of the NY99 strain (Shi et al., 2002), capped with a Type 1 cap, and transfected into either Vero or C6/36 cells . At 4 dpi, cells were collected and modified with NAI, an electrophilic reagent that reacts preferentially with 2'OH moieties in flexible regions of the RNA backbone, or DMSO as a treatment control (**Fig 3.1A**) (Merino et al., 2005).

Following extraction and purification of RNA, sequencing libraries were generated using a tiled-amplicon approach. Specifically, 19 700 nucleotide (nt) amplicons were tiled across the WNV genome with 100nt overlap between adjacent amplicons (**Fig 3.1A**). Reverse transcription (RT) was performed with gene-specific RT primers and manganese, a non-natural co-factor that allows for the encoding of RNA-adducts as cDNA mutations (Siegfried et al., 2014). Subsequent to RT, amplicon PCR with gene-specific primers was performed, with correct sizing confirmed by gel electrophoresis. Two independent biological replicates were prepared for each cell type, and final libraries were sequenced using the Illumina NextSeq 500/550 platform. Sequencing data was analyzed using the ShapeMapper2 pipeline (Busan and Weeks, 2018) .

Resulting data was subjected to stringent quality control metrics to ensure data was of sufficient quality for *de novo* structure prediction. Median effective read depth was >55,000x in both Vero replicates and >35,000x in both C6/36 replicates, far exceeding the read depth threshold required for high confidence reactivity calling (Smola et al., 2015a). As a result, we collected effective reactivity data for 99.8% and 99.7% of the WNV genome in infected Vero or C6/36 cells, respectively.

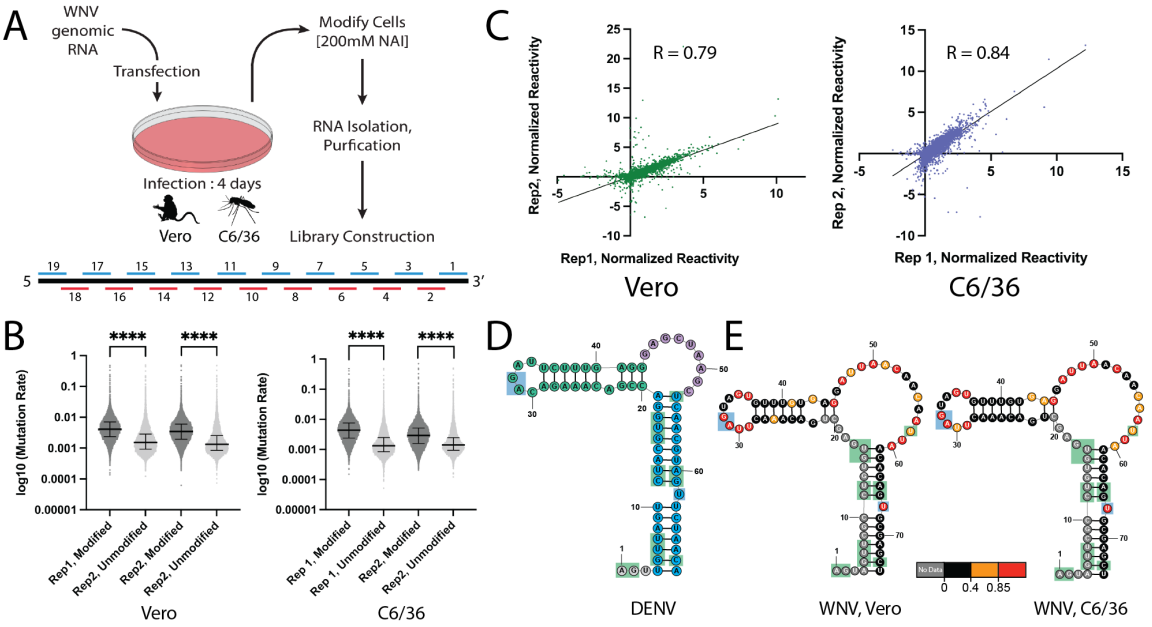

Figure 3.1. Tiled amplicon SHAPE-MaP workflow yields high quality *in vivo* reactivity data from multiple cell types, and *de novo* structure prediction recapitulates a conserved functional motif A) Workflow of *in vivo* SHAPE-MaP probing using *in vitro* transcribed WNV RNA to initiate infection in Vero and C6/36 cells. RNA is modified with NAI on 4 days post-infection (dpi), and a tiled-amplicon approach is used to afford full genome coverage. B) Comparison of mutation rates of NAI-modified or unmodified samples for two independent biological replicates in both Vero and C6/36 cells. Lines indicates mean and whiskers indicate standard deviation. ****p < 0.0001 by equal variance unpaired Student's t test. C) Correlation plot of normalized SHAPE reactivities from two biological replicates collected in either Vero or C6/36 cells. Lines represent linear regressions fit to the data. Pearson's correlation for each dataset is shown. D) Secondary structure of Dengue Virus SLA promoter, adapted from (Lee et al., 2021). Bottom stem = blue; Top stem = teal; Side loop = purple. E) Secondary structure prediction of WNV SLA extracted from full-length prediction in Vero (left) or C6/36 (right) cells, colored by SHAPE reactivity. Green shaded nucleotides = conserved; Blue-shaded nucleotides = conserved, functional.

We next compared the relative mutation rates of modified or unmodified RNA samples. This analysis reveals a significant elevation of mutation rates for modified samples in all four replicates analyzed (**Fig 3.1B;** p-value < 0.0001), thus confirming that WNV RNA was successfully modified with NAI in both *in vivo* contexts, and that these adducts were encoded as cDNA mutations. Finally, we computed genome-wide

Pearson's correlation coefficients to compare replicate normalized reactivity values. Strong correlation was observed across the entire WNV genome in both cell types (**Fig 3.1C;** Vero = 0.79; C6/36 = 0.84), confirming that these biophysical measures of RNA backbone flexibility are highly reproducible and that SHAPE-MaP data is of sufficient quality for genome-wide structure prediction.

We relied on the SuperFold pipeline to generate genome-wide structure predictions using the reactivities generated *in vivo* as experimental constraints (Smola et al., 2015a). To confirm that our structure predictions are of suitable quality for identification of novel riboregulatory elements, we checked if the consensus models generated accurately recapitulate the structure of Stem Loop A (SLA).

Rendered in **Fig 3.1D** is the secondary structure model of the DENV SLA derived from its crystal structure. It includes three domains, a bottom stem (blue), a top stem (teal), and a single-stranded side loop (purple), that are not correctly resolved when performing unconstrained secondary structure predictions (Lee et al., 2021). Both *in vivo* secondary structure models of the WNV SLA, generated either in Vero or C6/36 cells, recapitulate this overall architecture (**Fig 3.1E**). Within SLA, the authors also identify 14 double-stranded nucleotides with no known direct functional role that are nevertheless absolutely conserved (**Fig 3.1D,** shaded green). Our models recapitulate the relative position and strandedness of 12 of these nucleotides. Though an absolutely conserved "U" nucleotide is rendered single-stranded in the side loop of both structure predictions, the medium SHAPE reactivity supports our models, suggesting a WNV-specific SLA fold (**Fig 3.1E**).

Three absolutely conserved nucleotides with known functional roles appear as single-stranded in the DENV SLA crystal structure. Two comprise the unpaired "AG" motif, located within the top stem-loop that is known to function as the NS5 interaction site (**Fig 3.1D,** shaded blue) (Bujalowski et al., 2017). The third nucleotide is a bulged U in the bottom stem that, when mutated or deleted, abolishes viral replication (**Fig 3.1D,** shaded blue) (Filomatori et al., 2011). Both our Vero and C6/36 structure models perfectly recapitulate the relative position and stranded-ness of these nucleotides, with SHAPE reactivity data in strong agreement. Together, these findings suggest our data is of high quality and can confidently be used to identify novel riboregulatory elements.

SHAPE data reveal linear genome conformation dominates *in vivo* while cyclized conformation dominates *in vitro*

Like all *flaviviruses*, the WNV genome alternates between two conformations, linear or cyclized, in a process called genome cyclization (Brinton, 2014). This process is mediated by four pairs of complementary sequence that form non-contiguous, long-range duplexes between the extreme 5' and 3' viral termini (**Fig 3.2A, 3.2B**). Importantly, the linear and cyclized genome conformations are mutually exclusive, and undergo several conformational changes as they interconvert.

Stem loop B (SLB), which houses the start codon and is only folded in the linear genome conformation, contains two complementary regions - the 5' upstream-of-AUG-Region (5'UAR) and the 5'downstream-of-AUG-Region I (5'DAR-

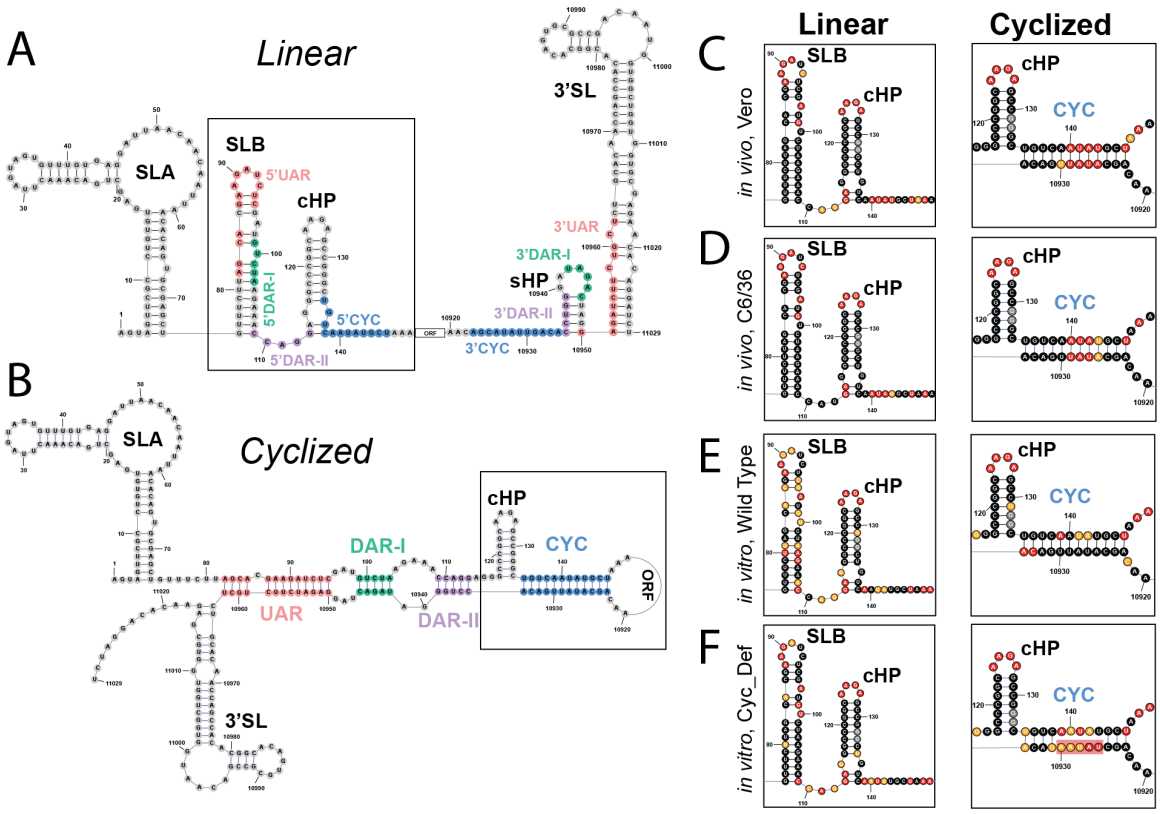

Figure 3.2. The WNV genome favors the linear genome conformation over the cyclized genome conformation in infected mammalian and arthropod cell lines. A) Schematic of secondary structures at the 5' and 3' viral terminus that comprise the linear genome conformation. Sequences that mediate genome cyclization are colored and labeled. Stem Loop A (SLA), Stem Loop B (SLB), Capsid Hairpin (cHP), Open Reading Frame (ORF), Short Hairpin (sHP), 3' Stem Loop (3'SL). B) Schematic of secondary structures and long-range duplexes that comprise the cyclized genome conformation, labeled as in (A). C-F) Normalized SHAPE reactivities mapped to mutually exclusive structural elements of the linear (left) or cyclized (right) genome conformation, labeled as in (A) or (B), respectively. Cyclization Defection (Cyc_Def); Mutated nucleotides are shaded red.

1) (**Fig 3.2A;** boxed, pink and green, respectively). Immediately 3' of SLB is the capsid hairpin (cHP), a structure known to direct correct start codon usage by a scanning ribosome (Clyde and Harris, 2006). The base of cHP, which melts upon genome cyclization, overlaps with the two remaining complementary regions - 5' downstream-of-AUG-Region-II (5'DAR-II) and the 5' cyclization sequence (5'CYC)

(**Fig 3.2A**; boxed, purple and blue, respectively). Importantly, the top portion of cHP remains folded upon cyclization (**Fig 3.2B**, boxed).

The formation of the CYC duplex (**Fig 3.2B,** blue) is thought to be an essential first step for genome cyclization, providing the binding energy required for unwinding of stems folded in the linear conformation (Friebe et al., 2011). Indeed, mutations introduced into the 3'CYC region completely abrogate viral growth (Basu and Brinton, 2011). Importantly, the 5' and 3' arms of the CYC duplex are primarily single-stranded in the linear genome conformation (**Fig 3.2A**, blue). While the secondary structures that comprise the linear conformation can be extracted from structure predictions prepared above, the cyclized genome conformation cannot be predicted due to distant constraints imposed during structure prediction (Smola et al., 2015a). We therefore reasoned that evaluating the mapping quality of our *in vivo* SHAPE reactivity to the mutually exclusive elements in the linear and cyclized conformations may reveal which conformation dominates *in vivo* (Spasic et al., 2018).

The *in vivo* reactivities strongly support formation of the linear conformation. Indeed, reactivities from both cell types, when mapped to the predicted SLB structures, show that highly reactive nucleotides are almost entirely restricted to single-stranded regions, while lowly reactive nucleotides are restricted to double-stranded regions (**Fig 3.2C, 3.2D,** Linear). Interestingly, G99, found in the start codon, is highly reactive even though it is predicted as double-stranded. It is possible this flexibility may afford access to a scanning ribosome. Similarly, with the exception of two nucleotides at the base of cHP, the *in vivo* reactivities from both cell

types agree with a folded cHP (**Fig 3.2C, 3.2D,** Linear). Importantly, a highly flexible G113 does not lend support to either genome conformation, as it is double-stranded in both.

Low quality mapping of reactivities to the CYC duplex reveal that the cyclized genome conformation is disfavored *in vivo*. Specifically, nucleotides 140 - 144, contained in the 5'CYC duplex arm and single-stranded in the linear conformation, are highly reactive in both cell types (**Fig 3.2C, 3.2D**, Cyclized). Similarly, nucleotides 10926 – 10930, contained in the 3'CYC duplex arm and single-stranded in the linear conformation, are also highly reactive in both cell types (**Fig 3.2C, 3.2D,** Cyclized). Taken together these data suggest that the CYC duplex is not formed *in vivo*.

Because multiple host proteins have been implicated in genome cyclization, we performed *in vitro* probing experiments on natively purified, full-length genomic RNA to better understand the cyclization dynamics of the genomic RNA in a protein-free system (Davis et al., 2013; Dong et al., 2008; Polacek et al., 2009). Reactivities derived from the wild-type WNV construct show that the cyclized genome conformation dominates *in vitro*. This is evidenced by strong disagreement between our *in vitro* reactivity data and a folded SLB, including highly reactive nucleotides at the base of SLB that become single-stranded upon genome cyclization (**Fig 3.2E,** Linear). Even more, low reactivity values support the formation of the long-range CYC duplex, including nucleotides that were shown to have high reactivity *in vivo* (**Fig 3.2E,** Cyclized).

Importantly, the use of an *in vitro* system allows us to probe a cyclization defective mutant (Cyc_Def) that cannot grow, and therefore cannot be probed *in vivo*. The Cyc_Def mutant is generated by inverting five nucleotides in the 3'CYC duplex, thus disrupting formation of the CYC duplex (**Fig 3.2F;** mutated nucleotides shaded red). This mutation renders the virus replication-incompetent and unable to revert, suggesting a profound replication defect (Basu and Brinton, 2011). Our *in vitro* probing data of the Cyc_Def mutant reveal striking observations. First, reactivity mapping to both the SLB and CYC duplex confirm that the mutations introduced have the intended structural consequence. Specifically, reactivity mapping to SLB supports the presence of the linear conformation, while reactivity mapping to the CYC duplex shows highly reactivity nucleotides present in both the 5' and 3'CYC duplex arms (**Fig 3.2F**). Put another way, this pattern of reactivity mapping represents the data signature of the linear genome conformation. In this context, this data further bolsters the observation that the linear genome conformation dominates *in vivo*, as the *in vivo* reactivity mapping of the actively replicating, wild type construct almost perfectly recapitulates that of the Cyc_Def construct *in vitro*.

Taken together, these data suggest that, in the absence of any host factors, the WNV genome naturally favors the cyclized genome conformation. In a cellular context, host factors may disrupt this natural equilibrium, causing the genome to favor its linear form.

Genome-wide structure prediction of the WNV genome reveals a novel tripartite domain architecture of the 3' viral terminus

A large body of work, spanning multiple decades, has provided strong support for the existence of four pseudoknots in the 3'UTR of WNV. These pseudoknots provide mechanistic stability that ultimately results in liberation of a non-coding RNA, called the subgenomic flaviviral RNA (sfRNA), that plays a crucial role in innate immune evasion and viral pathogenesis (de Borba et al., 2019; Funk et al., 2010; MacFadden et al., 2018; Pijlman et al., 2008). Our *in vivo* data support the presence of all four pseudoknots, as the majority of all pseudoknotted nucleotides are lowly reactive in both replicate data sets in both cell types (**Fig A3.1A, A3.1B**). For these reasons, these pseudoknots were forced in all genome-wide structure predictions performed. The structure of the 3' viral terminus, extracted from these genome-wide predictions, reveals a novel, tripartite domain architecture with striking structural homology between cell types (**Fig 3.3A, 3.3B**). Each of the three structural domains is identified on the basis of a unique and reproducible combination of SHAPE reactivity, Shannon Entropy, and stranded-ness.

Domain I (**Fig 3.3A, 3.3B**, pink) is situated upstream of the stop codon in the coding region of NS5. The majority of the nucleotides contained in Domain I exhibit low reactivity, low Shannon entropy, and are double-stranded; ~60% of nucleotides have reactivities <0.4 (**Fig 3.3C;** Vero = 58.6%, C6/36 = 63.4%), ~85% of nucleotides have Shannon Entropy <0.1 (**Fig 3.3D**; Vero = 84.8%, C6/36 = 83.4%), and ~55% of nucleotides are double-stranded (**Fig 3.3E**; Vero = 55.9%, C6/36 = 58.6%). In fact, a sub-section of this domain qualifies as well-folded as defined in the

methods section (**Table 3.1**; Vero - Region18, C6/36 - Region19). Most interestingly, the 5' end of Domain I is engaged in a longer distance interaction with the 5' end of Domain III. This reveals that the 3'UTR does not fold independently of the viral ORF, confirming a similar observation made regarding the dengue virus 3' viral terminus (Dethoff et al., 2018).

Domain II includes both coding and non-coding nucleic acid sequence, and therefore contains the stop codon of the viral ORF. Further, it is unique among the three domains in that it is the most flexible and least folded (**Fig 3.3A**, 3.**3B**, green). The majority of the nucleotides contained in Domain II exhibit high reactivity, high Shannon entropy, and are single-stranded; <45% of nucleotides have reactivities <0.4 (**Fig 3.3C;** Vero = 43.4%, C6/36 = 35.2%), ~45% of nucleotides have Shannon Entropy <0.1 (**Fig 3.3D**; Vero = 42.6%, C6/36 = 46.7%), and ~35% of nucleotides are double-stranded (**Fig 3.3E**; Vero = 36.1%, C6/36 = 23%). This region was previously reported to include Stem Loop I (SLI), though it is absent from all structural models generated (Pijlman et al., 2008). Instead, this region includes a long, highly reactive single-stranded region that sits immediately upstream of Domain III (**Fig A3.2A, A3.2B**).

Domain III perfectly corresponds to the sfRNA and exhibits data signatures that suggest it is the most structured domain at the 3' viral terminus (**Fig 3.3A, 3.3B,** purple). The majority of the nucleotides contained in Domain III exhibit low reactivity, low Shannon entropy, and are double-stranded; ~85% of nucleotides have reactivities <0.4 (**Fig 3.3C;** Vero = 85.5%, C6/36 = 86.1%), ~80% of nucleotides have Shannon Entropy <0.1 (**Fig 3.3D**; Vero = 78.9%, C6/36 = 79%),

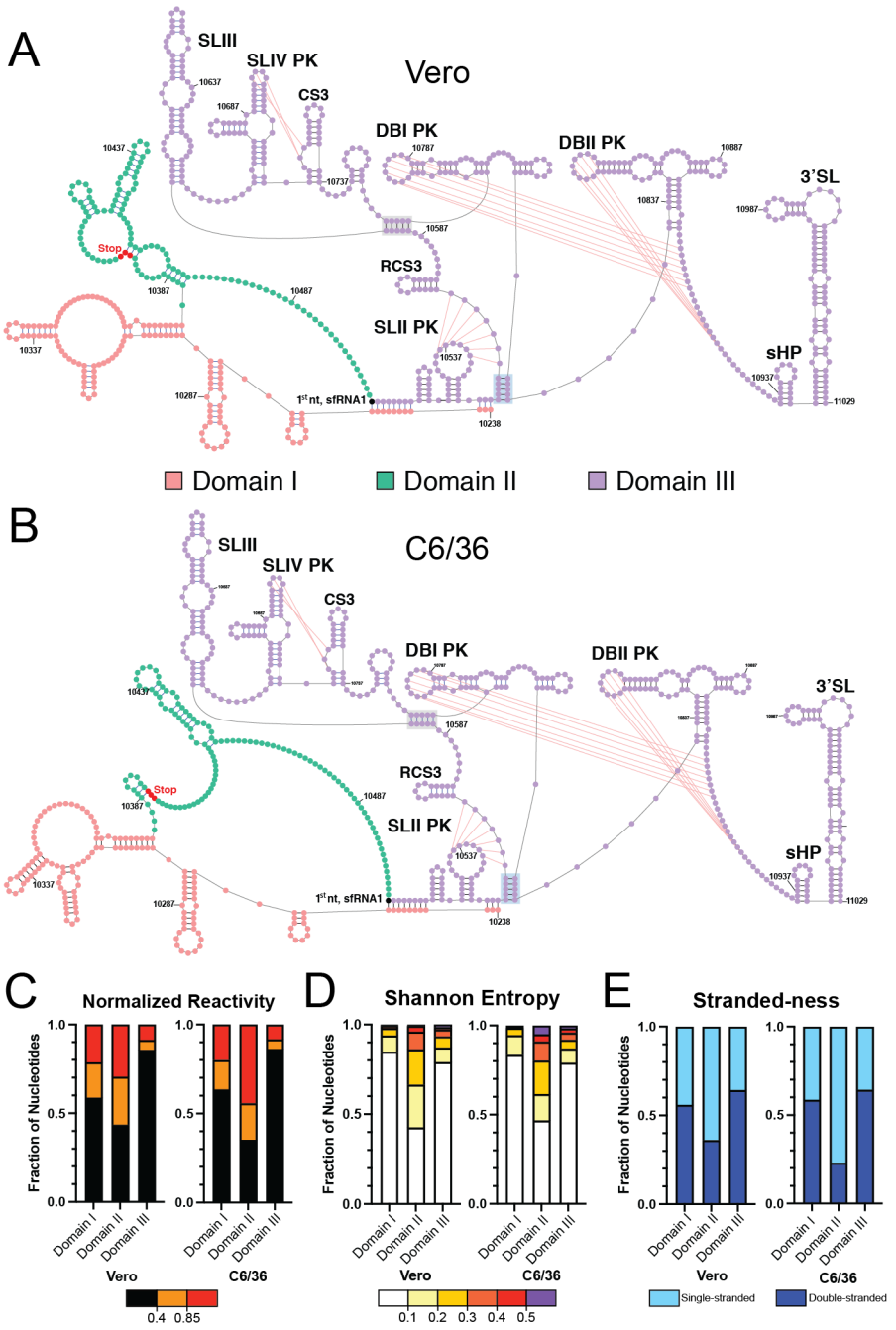

Figure 3.3. The 3' viral terminus of WNV is comprised of three distinct RNA structural domains. Structure of the 3' viral terminus determined in infected (A) Vero or (B) C6/36 cells, color-coded by domain. Stop codon = red; 1st nucleotide of the largest sfRNA = black; Pseudoknotted bases indicated by pink lines; Previously identified structural motifs are labeled. C) Nucleotides in each domain binned by normalized reactivity, with bin size expressed as a fraction of total nucleotides in that domain. D) Shannon entropy signatures of each domain, plotted as in (C). E) Stranded-ness of each domain, plotted as in (C).

and 64.2% of nucleotides are double-stranded (**Fig 3.3E**; Vero = 64.2%, C6/36 = 64.2%). (**Fig 3.3A, 3.3B**, purple). Strikingly, the structure of Domain III is identical in both cell types (**Table 3.1**, Vero – Region 19, C6/36 – Region 20; SENS = 100%, PPV = 100%).

However, our model deviates slightly from existing models of the WNV sfRNA. As mentioned above, a long-range duplex forms between Domains I and III. Formation of this duplex liberates nucleotides that would otherwise form the base of the Stem Loop II (SLII), allowing them to engage in previously unreported, longer range interactions (**Fig 3.3A, 3.3B,** shaded blue). As a direct consequence, the duplex reported to form the base of Dumbbell I (DBI) also does not fold, instead base-pairing with sequence ~100nt upstream (**Fig 3.3A, 3.3B,** shaded grey). Our models otherwise recapitulate all of the structural elements previously reported for the WNV 3'UTR, including RCS3, SLIII, SLIV, CS3, DBII, sHP, and the 3'SL (**Fig 3.3A, 3.3B,** labeled). Overall, our model suggests Domain III is extensively base-paired and highly compact, and suggests a context-specific fold for both the 3'UTR and specifically the sfRNA.

West Nile Virus genome folds into networks of well-folded regions with little apparent host dependency

The high structural homology observed for functional elements in the 5' and 3'UTR of WNV suggested that, despite being cultured in different cell types, structural homology might extend into the viral ORF. To test this, we first analyzed the global correlation of Vero and C6/36 reactivities. Remarkably, we observed

correlations between cell types that were as strong as those observed for reactivities collected in the same cell type (**Fig 3.4A,** $0.76 \leq R \leq 0.89$; **Fig 3.1C,** $0.79 \leq R \leq 0.84$). This suggests that the per-nucleotide backbone flexibility across the entire WNV genome is largely cell-type independent.

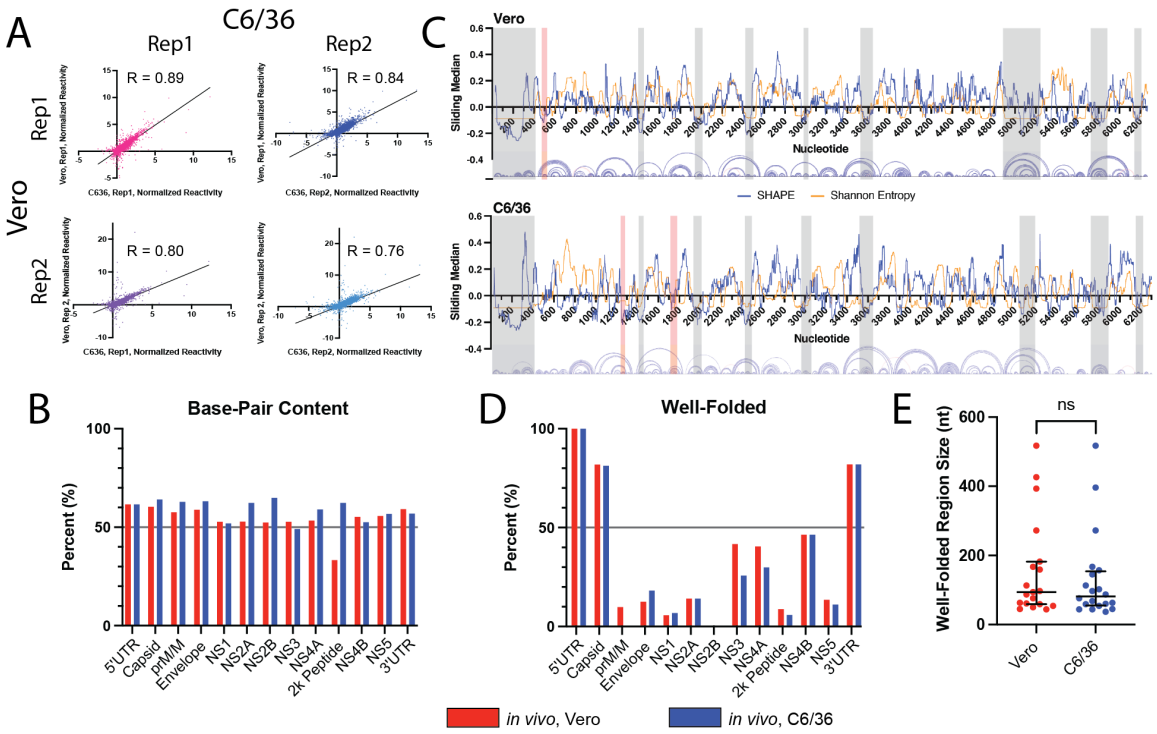

Figure 3.4. West Nile Virus genome folds into networks of well-folded regions with little apparent host dependency A) Comparison of normalized SHAPE reactivities made between biological replicates collected in either Vero or C6/36 cells. Lines represent linear regressions fit to the data. Pearson's correlation for each dataset is shown. B) Base-pairing content in the UTRs and individual protein domains determined in Vero (red) or C6/36 (blue) cell types. C) Analysis of SHAPE reactivities and Shannon entropy in reveals the presence of highly structured, well-determined domains in WNV. Nucleotide coordinates are indicated on the x-axis – only the first half of the WNV genome is shown. Local median SHAPE reactivity and Shannon entropy are indicated by blue and orange lines, respectively. Well-folded regions that appear in both or only a single cell type are shaded with gray or red boxes, respectively. Arc plots for predicted base-pairing interactions in the structural model are shown below the x-axis. D) Well-folded RNA content in the UTRs and individual protein domains determined in Vero (red) or C6/36 (blue) cell types. E) The size of well-folded regions determined in Vero (red) or C6/36 (blue) cell types.

We sought to understand whether this strong agreement in backbone flexibility is reflected in discrete secondary structure predictions. At a global level, we did not observe differences in the base-pair content of the WNV genome, with an average double-stranded content of 54.3% (±7.07%) or 59.0% (±5.18%) across protein domains in Vero and C6/36 cells, respectively (**Fig 3.4B**). The only domain in which a difference was observed was the 2K peptide, a 69nt region nested between NS4A and NS4B; 23nt of this region are double-stranded in Vero cells, while 43nt are double-stranded in C6/36 cells (**Fig 3.4B**). Therefore, the apparent difference in %BPC likely reflects the small size of this region.

As our goal is ultimately to identify novel riboregulatory structures, we next focused on identifying regions of the genome that are highly structured and well determined. These criteria have been successfully deployed to identify riboregulatory regions in other single-stranded viruses (Dethoff et al., 2018; Huston et al., 2021; Siegfried et al., 2014; Wan et al., 2022). Briefly, we identified regions with local SHAPE reactivity and Shannon Entropy below the global median for $\geq$40nt that appear in both replicate data sets from either Vero or C6/36 cells. Any region that meets these "lowSS" criteria is hereafter referred to as "well-folded."

When comparing the well-folded RNA content between cell types, we observe minimal host-dependence. For example, 10 well-folded regions appear in the first half of the genome when actively replicating in Vero cells. When considering this same span in C6/36 cells, we identify 11 well-folded regions. Of these, 9 well-folded regions appear in both cell types (**Fig 3.4C,** shaded grey boxes). In total, we identify 19 well-folded regions across the WNV genome in infected Vero

cells, and 20 well-folded regions in infected C6/36 cells. Eighteen appear in both cell types, ten of which have identical nucleotide boundaries, with the remaining 8 overlapping (**Table 3.1**, shaded dark or light green, respectively). Of all the well-folded regions identified, only 3 three well-folded regions appear in a single cell type (**Fig 3.4C**, shaded red boxes). These three regions are small (<55nt) and cluster exclusively at the 5' end of the genome. It follows, therefore, that we do not observe large differences in the percent of each protein domain that are well-folded (**Fig 3.4D**). Furthermore, we did not find significant differences in the size of well-folded domains identified in each cell-type (**Fig 3.4E**). Taken together, these data demonstrate that WNV folds independent of host context and suggest that, much like proteins, nucleic acid structure is hard-coded in primary sequence.

Table 3.1. Database of well-folded regions identified in WNV

in Vivo, Vero Well-folded*					in Vivo, C636 Well-folded*					Sens, PPV Calculations**	
Region	**Start**	**End**	**Size**	**Domain**	**Region**	**Start**	**End**	**Size**	**Domain**	**SENS**	**PPV**
n/a	1	149	149	5' Terminus	n/a	1	149	149	5' Terminus		
1	1	393	393	5'UTR/Capsid	1	1	396	396	5'UTR/Capsid		
UV_1	461	514	54	Capsid/M							
					UC_1	1227	1270	44	Envelope		
2	1385	1445	61	Envelope	2	1402	1445	44	Envelope		
					UC_2	1709	1762	54	Envelope		
3	1926	2001	76	Envelope	3	1926	2001	76	Envelope	100% (27/27)	100% (27/27)
4	2414	2463	50	Envelope	4	2415	2482	68	Envelope/NS1		
5	2969	3027	59	NS1	5	2969	3027	59	NS1	100% (17/17)	100% (17/17)
6	3530	3626	97	NS2A	6	3530	3626	97	NS2A	71.79% (28/39)	71.79% (28/39)
7	4893	5318	426	NS3	7	5039	5183	145	NS3		
8	5721	5887	167	NS3	8	5721	5887	167	NS3	98.11% (52/53)	100% (52/52)
9	6153	6215	63	NS3	9	6153	6215	63	NS3	100% (17/17)	100% (17/17)
10	6352	6510	159	NS3/NS4A	10	6352	6453	102	NS3		
11	6736	6848	113	NS4A/2k peptide	11	6736	6848	111	NS4A	85.71% (30/35)	85.71% (30/35)
12	6912	7183	272	2K peptide/NS4B	12	6912	7183	272	2K Peptide/NS4B	98.70% (76/77)	100.0% (76/76)
13	7455	7541	87	NS4B	13	7455	7541	87	NS4B	100% (26/26)	100% (26/26)
14	7757	7938	182	NS5	14	7772	7928	157	NS5		
15	8821	8864	44	NS5	15	8829	8865	37	NS5		
16	9065	9158	94	NS5	16	9083	9142	60	NS5		
17	10249	10293	45	NS5	17	10249	10293	45	NS5	100% (13/13)	100% (13/13)
18	10513	11029	517	3'UTR	18	10513	11029	517	3'UTR	100% (163/163)	100% (163/163)
n/a	10238	11029	792	3' Terminus	n/a	10238	11029	792	3' Terminus	96.44 (217/225)	93.94% (217/231)

*****Shaded dark green = identical nucleotides boundaries; shaded light green = overlapping nucleotide boundaries; not shaded = cell-type specific
**For SENS and PPV calculations, Vero structure = "Predicted", C6/36 structure = "Accepted"

Structural homology of well-folded regions serves as sorting criteria for functional validation

In order to provide evolutionary support for well-folded RNA secondary structures, researchers studying positive-sense RNA viruses often analyze synonymous rates (dS) (Assis, 2014; Dethoff et al., 2018; Huston et al., 2021; Simmonds and Smith, 1999; Tuplin et al., 2002). This mode of analysis relies on the assumption that, if a secondary structure is evolutionarily conserved, synonymous mutations should accumulate more quickly at single-stranded regions relative to double-stranded. Put another way, higher relative dS at single-stranded regions would reflect an evolutionary pressure to maintain double-stranded base-pairing interactions. We therefore constructed an alignment of 47 mosquito-borne *flaviviral* ORFs and computed relative dS for well-folded regions in the WNV genome.

We observed significantly elevated dS for single-stranded codons relative to double-stranded codons for a single well-folded region (**Fig 3.5A,** inset). This region, designated as Region 12, is the largest well-folded region identified in either cell type. It contains two small stem-loop structures that flank a large stem capped with a multi-helix junction, RNA motifs that have shown promise as drug targets (Warner et al., 2018). Importantly, the reactivity data from both cell types agrees strongly with the predicted structure (**Fig 3.5A**). As such, Region 12 represents an ideal target for functional validation.

However, as dS analysis flagged no other well-folded regions, we recognized the need to generate criteria to prioritize other regions for functional validation. Because we expect conserved riboregulatory elements to fold in both cell types, we reasoned the overall structural homology of Region 12 might represent such a criteria. First, both regions have identical domain boundaries (Genome coords. =

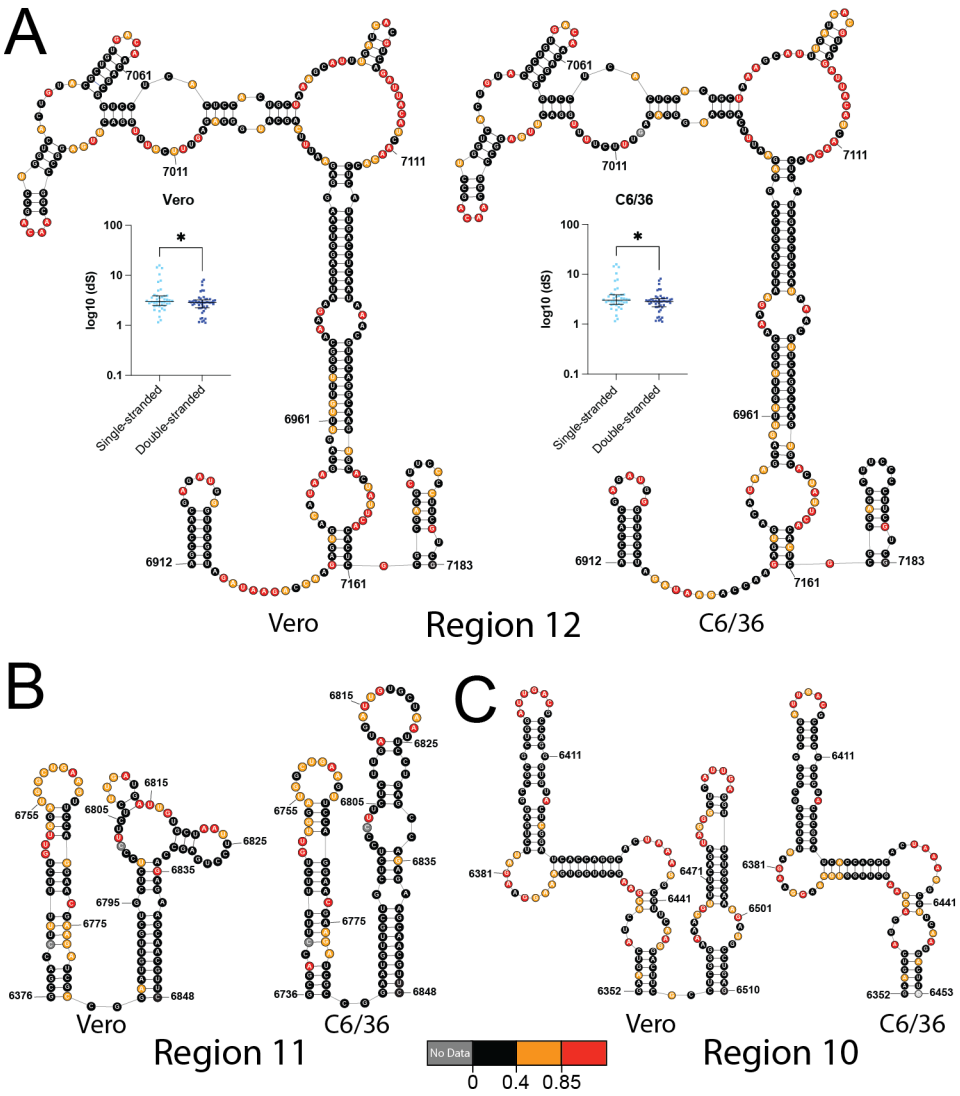

Figure 3.5. Patterns of structural homology of well-folded regions between cell types allows for prioritization of putative riboregulatory elements A) RNA secondary structure of well-folded Region 12, which has identical nucleotide boundaries and near identical connectivity in Vero (left) and C6/36 (right) cells. Inset- dS separated by stranded-ness in Region 12 using the appropriate secondary structure. Lines indicate median and whiskers indicate interquartile range. *$p < 0.05$ by equal variance unpaired Student's t test. B) RNA secondary structure of well-folded Region 11, which has identical nucleotide boundaries and non-identical connectivity in Vero (left) and C6/36 (right) cells. C) RNA secondary structure of well-folded Region 10, which has non-identical nucleotide boundaries, but identical connectivity in the region shared in Vero (left) and C6/36 (right) cells.

6912 – 7183nt), reflecting a strong agreement of SHAPE and Shannon Entropy data

signatures between cell types. More importantly, these two regions are predicted to

have near identical connectivity. Indeed, using the Vero secondary structure as the "predicted" structure and the C6/36 as the "accepted" structure, we calculate SENS and PPV to be 98.7% (76/77) and 100% (76/76), respectively; they differ only by a single base-pair predicted between U7092 and G7102 in C6/36 cells (**Fig 3.5A**). One other well-folded region, Region 8, has identical nucleotide boundaries (Genome coords. = 5721 – 5887nt) and near identical connectivity (SENS = 98.11% (52/53), PPV = 100% (52/52)); this region will also be prioritized for functional validation.

We relied on two other patterns of structural homology to flag regions for functional validation. The first pattern is exemplified by Region 11 (**Fig 3.5B**), which has identical boundaries in both cell types (Genome coords. = 6736 – 6848nt), but non-identical connectivity. It features two sequential stem-loop structures, with the more 3' stem displaying cell type dependent connectivity (SENS = 85.71% (30/35); PPV = 85.71% (30/35)). Importantly, this difference in base-pairing arises from true differences in reactivity; nucleotides 6808 – 6810 are medium reactivity in Vero cells, but lowly reactive in C6/36 cells (**Fig 3.5B**). One other well-folded region, Region 6, has identical nucleotide boundaries (Genome coords. = 3530 – 3626) and non-identical connectivity (SENS = 71.79% (28/39), PPV = 71.79% (28/39). Both of these regions will be prioritized for functional validation.

The final pattern of structural homology used to identify priority targets is exemplified by Region 10, a region with non-identical boundaries (Vero, Genome coords. = 6532 – 6510; C6/36, genome coords. = 6532 – 6453) (**Fig 3.5C**). In Vero cells, this region consists of two sequential stems, with the second stem absent from the well-folded region identified in C6/36 cells. However the region that appears as

well-folded in both has identical connectivity (SENS = 100% (31/31); PPV = 100% (31/31), with strikingly similar patterns of SHAPE reactivities (**Fig 3.5C**). One other well-folded region, Region 16, has non-identical nucleotide boundaries (Vero, Genome coords. = 9065 – 9158; C6/36, Genome coords = 9083 - 9142) but near identical connectivity for the region shared between cell types (SENS = 93.3% (14/15), PPV = 100% (14/14)). Both of these regions will be prioritized for functional validation. Taken together, these results show that evolutionary support exists for one well-folded region, but also that patterns of structural homology can be used to identify regions that merit subsequent functional interrogation.

Functional validation of candidate structures by targeted LNA disruption

Targeted disruption of RNA secondary structures using anti-sense locked nucleic acids (LNAs) has emerged as a powerful tool to identify functional RNA structures in viral genomes (Dethoff et al., 2018; Huston et al., 2021; Tuplin et al., 2015). This method relies on the ability of LNAs, non-natural RNA base analogues that increase the T_m of a given duplex by 2-8°C, to out-compete naturally occurring RNA-RNA duplexes (Lundin et al., 2013). While previous studies used reporter constructs to measure viral growth, we instead generated a completely novel workflow. Specifically, we rely on co-transfection of *in vitro* transcribed, capped WNV genomic RNA (WNVic) and LNAs, with an established qRT-PCR assay used to afford direct and highly accurate quantitation of WNV growth (Brackney et al., 2009; Lanciotti et al., 2000).

To establish the efficacy and dynamic range of our co-transfection system, we designed an LNA that anneals across the 3'CYC, 3'DAR-II, and 3'DAR-I cyclization elements (**Fig 3.2A**). This LNA should result in disruption of genome cyclization and produce a profound replication defect in both Vero and C6/36 cells. Indeed, we observe a significant, >3 $\log_{10}$ fold reduction in viral growth in both cell types when 3'CYC LNA is co-transfected with WNVic compared to a non-targeting LNA (**Fig 3.6A, 3.6B;** 3'CYC v non-targeting). Importantly, there is no significant difference in viral growth observed in either cell type when WNVic is transfected alone or with the same non-targeting LNA (**Fig 3.6A, 3.6B**). This confirms that targeted disruption of an RNA structure with conserved, pan-host function mediates strong defects in WNV growth using our co-transfection system in both cell types.

We next turned to the six candidate ORF structures described above, prioritized for functional validation based on patterns of structural homology in different cellular contexts. All LNAs targeting these regions were designed with similar lengths (20nt), % LNA content (52%), and predicted RNA:LNA T_m (89°C). LNAs were designed for maximal structure disruption, and target both single- and double-stranded regions of the well-folded RNA structures (**Fig 3.6C – 3.6H**; colored lines). A list of all LNAs used in this study is available in Table A3.5.

Of the 6 LNAs targeted to well-folded regions in the WNV ORF, 4 mediate significant defects in WNV growth in both cell types tested (**Fig 3.6A, 3.6B**). These include the LNAs targeted against Region 12 and Region 8 (**Fig 3.6C, 3.6D**)**,** well-folded regions that share identical nucleotide boundaries and near identical connectivity (>98% SENS, 100% PPV). Significant defects were also observed when

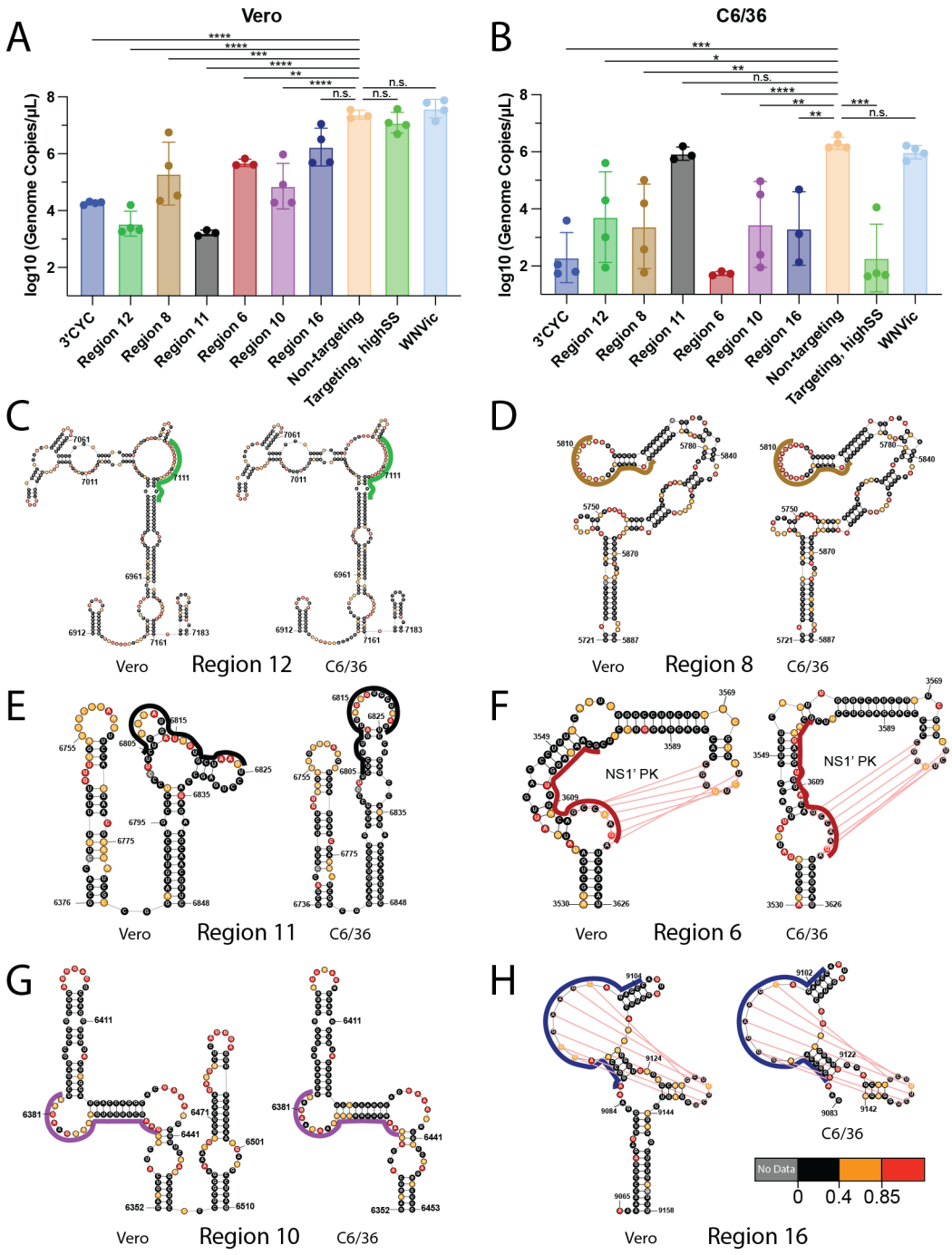

Figure 3.6. Targeted disruption of RNA structures with anti-sense locked nucleic acids (LNAs) results in potent viral growth defects A-B) Virus growth as measured by quantifying viral genomes in cell supernatant with qRT-PCR in (A) Vero cells at 3dpi or (B) C6/36 cells at 6dpi. Data points represent independent technical replicates. Bar height is the mean, and whiskers represent standard deviation. n.s., not significant; *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001 by ordinary one-way ANOVA with multiple comparisons. C-F) Schematic showing LNAs (colored lines) targeted to well-folded regions determined in Vero (left) or C6/36 (right) cells. Region-specific LNAs are colored as in A/B.

targeting a sub-structure in Region 10 that appears with identical connectivity (100% SENS, 100% PPV) in both cell types, though the well-folded region is larger in Vero cells (**Fig 3.6G**). Most interestingly, significant defects were observed when targeting Region 6, though the defect is ~3 $\log_{10}$-fold stronger in C6/36 cells. This region contains a previously reported pseudoknot in NS2A (NS1' PK), disrupted upon LNA annealing, that mediates programmed ribosomal frameshifting (PRF) and generates a functional NS1 variant (**Fig 3.6F**; pink lines; **Fig A3.2C, A3.2D**) (Melian et al., 2010, 2014; Winkelmann et al., 2011).

Of the 6 LNAs targeted to well-folded regions in the WNV ORF, 2 mediate significant defects in WNV growth in only a single cell type tested (**Fig 3.6A, 3.6B**). The first such LNA targets Region 16, a region with non-identical boundaries that contains a novel pseudoknot predicted in both cell types using a workflow described in the methods (**Fig 3.6H**, **A3.2C, A3.2D)**. The LNA that targets region 16, however, results in a significant growth defect only in C6/36 cells (**Fig 3.6A, 3.6B**). The second LNA with host-specific effects targets Region 11, a region with identical boundaries but non-identical connectivity (**Fig 3.6E**). The region 11 LNA, which targets the stem with a cell-type specific fold, mediates a significant, ~4 $\log_{10}$-fold reduction in WNV growth in Vero cells, but has no effect on viral growth in C6/36 cells.

As all of the RNA structures described above were identified on the basis of their "lowSS" signature, we were interested to target a region that appears with highSS signatures in both cell types. To that end, we designed an LNA against a region with Shannon Entropy and SHAPE reactivity *above* the global median in both

cell types (Genome coords. = 8920 – 8939). Interestingly, this LNA mediated no effect in Vero cells, but mediates a significant ~4 $\log_{10}$-fold defect in C6/36 cells (**Fig 3.6A, 3.6B**). While prior work with SARS-CoV-2 showed that LNAs targeted against highSS regions mediate no effect, these results suggests that highSS regions might harbor regulatory RNA sequence, though not necessarily RNA structure, that can play host-specific functional roles (Huston et al., 2021).

Taken together, we identify 6 novel RNA structures in the WNV ORF that play functional roles in the viral life cycle. Four structures mediate conserved regulatory roles, while two function with apparent host specificity. Interestingly, we also identify a highSS region that may harbor RNA sequence with a host-specific functional role. Our data confirm that the novel LNA workflow presented here is a potent, highly sensitive method for validating viral riboregulatory elements. Most importantly, our data validate a model in which putative riboregulatory elements can be identified on the basis of structural homology between cell types.

3.5 Discussion

We demonstrate that the WNV genome adopts a global fold with little apparent host-dependence, and is replete with highly structured, stable RNA secondary structures, a subset of which mediate functional roles in the viral life cycle. Though we observe only subtle differences in the structures of riboregulatory regions between cell types, we show that those differences can have profound impacts on function. This adds a fascinating new layer to a growing body of work showing viral ORFs contain conserved riboregulatory elements, and are therefore

placed under remarkable evolutionary pressure (Dethoff et al., 2018; Huston et al., 2021; Wan et al., 2022). As WNV, and other *flaviviruses*, represent an expanding threat to global health due to rapidly warming climates (Whitehorn and Yacoub, 2019), this work will serve as a valuable resource to other researchers.

While collecting *in vivo* reactivities for RNAs often presents a technical hurdle, the tiled-amplicon approach deployed here allowed for collection of exceptionally high quality SHAPE-MaP data in two evolutionarily distant cell lines (Leamy et al., 2016; Mitchell et al., 2019). This suggests that this method, readily adapted for any virus of interest, is also generalizable to other cell lines. As such, it should serve as a useful scaffold for researchers aiming to study viruses without restrictions placed on cellular context.

Both the SHAPE-Map data and the resulting experimental secondary structures provide novel insights into previously studied but poorly understood aspects of *flaviviral* biology. Our work demonstrates for the first time that the linear conformation dominates *in vivo* but that, in a protein-free system, the cyclized genome dominates its conformational ensemble. While several studies have identified protein-binding partners as promoters of genome cyclization, our data suggest that host factors, such as processing ribosomes, may actually push the genome away from its preferred cyclized conformation (Blackwell and Brinton, 1997; Davis et al., 2013). In this context, sequestration of the genome away from host factors inside replication complexes would allow the genome to naturally cyclize. In this model, the "molecular switch" that mediates genome cyclization is formation of replication complexes, a process itself known to be dependent on

translating and accumulating sufficient levels of WNV non-structural proteins (Brinton, 2014). This accords with studies that show altering the balance of linear and cyclized genomes negatively alters viral replication kinetics, and represents an elegant strategy for sequential timing of the viral life cycle in any host context (Liu et al., 2016; Villordo et al., 2010).

Our structures of the WNV 3' viral terminus reveal a novel tripartite domain architecture, with each domain displaying a unique structural profile. Overall, it includes a flexible, single-stranded region (Domain II) sandwiched between two highly structured, well-determined domains (Domain I&III). As Domain III perfectly corresponds to the sfRNA, it is possible that its positioning immediately downstream of Domain II is important for stalling of Xrn1 at the appropriate position. Of additional note is a long-range base-pairing interaction that forms between Domain I and the 5' end of Domain III and results in the formation of previously unreported long-range duplexes within Domain III. While a deviation from canonical depictions of sfRNA structure, our structure may simply represent a genome-specific fold of the 3'UTR (Funk et al., 2010; MacFadden et al., 2018; Pijlman et al., 2008). Indeed, our lab has previously highlighted the importance of upstream sequence context in rendering accurate structure predictions of viral RNAs (Tavares et al., 2021). As these long-range duplexes have high Shannon entropy, it is likely that liberation of the smaller sfRNA species may be owed to local refolding of the DBI pseudoknot that occurs after Xrn1 has chewed through upstream duplexes.

We report a structure prediction for every nucleotide in the WNV genome in two different, biologically relevant cell types, allowing us to interrogate features of

its genome architecture that promote viral fitness. Overall, we report a minimal host-dependence of the WNV genome structure despite vastly different cellular contexts. This is reflected in both strong correlations of SHAPE-MaP data as well as almost identical levels of base-pairing retained across the WNV genome between cell types. Even more, there is remarkable agreement in the location, size, and connectivity of well-folded regions that appear across the genome (Table 3.1). The data suggest that, much like proteins, RNA structure is hard-coded into primary nucleic acid sequence. It is worth noting that this general feature of RNA structure may confer durable fitness advantages to enzootic viruses as they alternate between hosts.

Though methods used to provide evolutionary support for RNA secondary structure identified only a single well-folded region, we do not believe this points to an absence of conserved riboregulatory elements in the WNV genome. Rather, we believe this may reflect the slow evolutionary rate of vector borne viruses, thought to arise due to host-specific evolutionary pressures experienced by the virus as it replicates within a single host as well as genetic bottlenecks encountered as the virus switches hosts (Grubaugh and Ebel, 2016; Grubaugh et al., 2015; Woelk and Holmes, 1998). These slow evolutionary rates result in highly homologous sequence alignments, a type of low-information alignment known to impede efforts to identify evolutionary patterns of RNA secondary structure conservation (Rivas et al., 2020; Tavares et al., 2018). The presence of functional secondary structure in the ORF of a vector borne virus would additionally constrain its evolutionary rate, further hindering attempts at flagging patterns of evolutionary conservation.

Instead, we relied on our lab's "structure first" approach, using patterns of structural homology between cell types to identify structures that may mediate conserved functional roles. With this novel strategy, we delineate 3 unique patterns of structural homology that allow us to prioritize 6 well-folded regions for functional validation. To that end, we develop a novel method that relies on co-transfection of structure-disrupting anti-sense LNAs along with *in vitro* transcribed WNVic. Not only is this method faster and more scalable than traditional viral genetics systems, it allows for more potent structure disruption as mutational strategies in viral ORFs are necessarily limited to synonymous base changes to avoid changes in coding potential. Even more, the use of qRT-PCR to monitor viral growth allows for sensitive, highly quantitative measurement of viral growth defects.

In total, we identify 4 well-folded RNA structures in the WNV ORF that, upon disruption, result in severe growth defects in both cell types tested. Of these, 3 exhibit growth defects of comparable magnitude in both cell types; because many of the proteins shared in these cellular contexts are viral, it stands to reason that the RNA structures in Regions 2, 8, and 10 mediate their function via recruitment of viral proteins. . As such, these regions would be ideal targets for follow-up studies using RNA anti-sense purification coupled with mass spectrometry (RAP-MS) methodologies (McHugh et al., 2015). Of special interest, however, is Region 6 which contains the NS1' PK. Though careful molecular virology has confirmed production of an extended NS1 variant, our *in vivo* reactivities provide direct confirmation that WNV NS1' PK folds inside infected cells (**Fig A3.2C, A3.2D**) (Melian et al., 2010). As

NS1' PK disruption was shown to attenuate WNV replication in live mosquitos and birds, our data suggest that the anti-NS1' PK LNA may have applications as an anti-viral therapeutic (Melian et al., 2014). By the same token, the LNA that blocks genome cyclization (3'CYC) and resulted in profound replication defects could prove equally efficacious as a pan-*flaviviral* therapeutic agent.

We additionally identify 2 well-folded RNA structures in the WNV ORF that, upon disruption, mediate cell-type specific growth defects. To date, functional RNA structures that mediate host-specific effects in *flaviviruses* have only been identified in UTRs (de Borba et al., 2019; Villordo and Gamarnik, 2013). As functional RNA often acts through protein binding, the observed host-specific function of these two regions may simply reflect the recruitment of host-specific proteins. The data for Region 11, however, suggests a more interesting mechanistic explanation. If the profound (~4 $\log_{10}$-fold) defect observed in Vero cells was attributable to protein binding, one would expect lower reactivities in this region in Vero cells because protein binding is known to occlude access of SHAPE reagents (Smola et al., 2015b). However, the opposite is true; the region targeted by the LNA contains three contiguous nucleotides that are *more* highly reactive in Vero cells. As Vero cells are cultured at a higher temperature than C6/36 cells (37°C v 28°C, respectively), one possible explanation for the enhanced reactivity of these nucleotides may be due to RNA unfolding. The presence of a functionally important, thermally responsive RNA structure may suggest the WNV genome harbors an RNA thermometer, host-sensing elements first identified in pathogenic bacteria (Kortmann and Narberhaus, 2012). While prior work has pointed to the base of 3'SL in the WNV 3'UTR as a thermally-

responsive element that influences genome cyclization, Region 11 would represent the first thermally responsive element discovered in a viral ORF (Meyer et al., 2020). However, more work is required to establish a conclusive link between the observed structural and functional differences and culturing temperature.

The work presented here deepens our understanding of the West Nile Virus life cycle, revealing a global genome fold with minimal host dependence We identify six riboregulatory elements that fold with only subtle differences between cell types that function with either conserved or host-specific functional roles. Our study also demonstrates that patterns of structural homology can serve as powerful indicators of functional RNA structure, a method readily extended to other enzootic viruses. Finally, the identification of LNAs that mediate potent defects in WNV growth, with targeting strategies generalizable to other *flaviviruses*, represents an exciting development in the field of anti-viral nucleic acid therapeutics.

3.6 Methods

Cell Culture

Vero cells (ATCC, CCL-81) were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% heat-inactivated fetal-bovine serum (HI FBS) supplemented with NEAA, L-glutamine, and sodium bicarbonate and incubated at 37°C/5% CO_2. C6/36 cells (ATCC, CRL-1660) were cultured in DMEM supplemented with 10% HI FBS DMEM, NEAA, L-glutamine, and sodium bicarbonate and incubated at 28°C/5% CO_2. Unless otherwise stated, cell lines were cultured as

described above. All experiments using live WNV were performed in the BSL3 facility at the Connecticut Agriculture Experimental Station (CAES).

Construct Design and Preparation

Previous work has shown that disruption of the CYC duplex by inverting 5 consecutive nucleotides in the 3'CYC arm results in a replication incompetent virus (Basu and Brinton, 2011). To generate this mutant, we introduced the mutation into a plasmid containing the full-length infectious cDNA of WNV. Mutations were introduced using site-directed mutagenesis, with mutations encoded as partly overlapping handles in the forward and reverse primers (Table A3.1). Successful mutagenesis was verified at the Yale Keck Facility.

***In vivo* SHAPE-MaP, reverse transcription, and library preparation**

A plasmid containing the full-length infectious cDNA clone of the WNV NY99 (AF404756) was used to generate full-length genomic RNA. Briefly, the plasmid was linearized using Xba1 (NEB, Cat. No. R015S), followed by ethanol precipitation and resuspension in TE pH 7.5. The linearized plasmid serves as a template in a run-off, *in vitro* transcription using a T7 RNA polymerase variant P266L (Tang et al., 2014) with reaction conditions previously described (Adams et al., 2019). Following transcription, plasmid template was digested with RQ1 DNase (Promega, Cat. No. M6101) and RNA purified using the RNeasy kit (Qiagen, Cat. No. 74104) according to the manufacturer's protocols. A type-1 cap, including an inverted 7-methylguanosine and a 2'-OMe on the first nucleotide, are added to the RNA using the Vaccinia Capping Enzyme (NEB, Cat. No. M2080S) and 2'-OMe transferase (NEB,

Cat. No. M0366S) in a one-pot reaction according to manufacturer's protocols. Capped RNA was purified using an RNeasy column, eluted in 1xME buffer (8mM MOPS, 0.1mM EDTA, pH 6.5), diluted to 1μg/μL, and frozen down at -80°C until needed. Final capped RNA products were visualized on a denaturing agarose gel to ensure production of full-length products.

Prior to transfection, Vero or C6/63 cells were seeded into eight 10cm tissue culture treated plates and grown to ~90% confluency. Transfection of *in vitro* transcribed, capped viral RNA was performed using the Mirus *TransIT®-mRNA Transfection Kit* (Cat. No. MIR2225). Specifically, 6 μg of viral RNA was transfected per 10cm plate according to manufacturer's protocol. Four hours post-transfection (hpt), cells were gently washed once with cold 1xPBS and complete media replaced.

Four days post-infection (dpi), media was aspirated from tissue culture plates and cells washed once with cold 1xPBS. The contents of 4 x 10cm plates were collected in 2 mL total of 1xPBS with cell scrapers, spun down at 1000g for 5 min at 4°C, resuspended in 2 mL 1xPBS, and transferred to 15 mL falcon tubes. Subsequently, 200 μL of 2M NAI or an equivalent volume of DMSO was added to the suspension. Samples were pipetted vigorously to ensure sufficient mixing and incubated in the dark for 10 min.

Following the 10 min. incubation, 6 mL of Trizol reagent was added. Samples were incubated for 10 min to ensure complete viral inactivation. RNA was extracted with addition of 1.2 mL chloroform:isoamyl alcohol (24:1), spun at full speed for 15 min, and the aqueous phase was transferred to a fresh tube. To this tube was added 16 mL of 100% EtOH (70% final), and samples were incubated overnight at -20*C.

RNA was precipitated by spinning tubes at 20,000g for 30 min at 4°C, washing once with 70% EtOH, and spun again at 20,000g for 15 min at 4°C. The EtOH was aspirated off, and pellets were air-dried for 5 min. The pellet was resuspended in 1xME buffer, purified using the Qiagen RNeasy kit according to manufacturer's protocol, and stored at -80°C until needed.

To prepare sequencing libraries, we relied on a tiled-amplicon approach. Specifically, we designed 19 700nt amplicons tiled across the WNV genome to achieve full sequencing coverage. Adjacent amplicons overlapped by 100nt, with additional overlap at the 5' and 3' viral termini to ensure sufficient sequencing coverage. A list of all gene-specific RT and PCR primers used is available in Table A3.2.

Reverse transcription (RT) reactions were prepared using 1.5 µg of total cellular RNA, SuperScript II (SSII) (Invitrogen, Cat. No. 18064014), SSII-MaP reaction buffer (50mM 1M Tris-HCl pH 8.0, 75mM KCl, 10mM DTT, 6mM $MnCl_2$, 0.5mM dNTP), and 1µM gene-specific RT primer. RT reactions were incubated at 42°C for 3 hrs. Following RT, viral genomic RNA was degraded enzymatically at 37°C using an equal mix of RNaseA (NEB, Cat. No. T3018L), RNaseT1 (NEB, Cat. No. EN0541), and RNaseH (NEB, Cat. No. M0297S). Single-stranded cDNA was purified using AmpureXP beads (Agencourt, Cat. No. A63881) with a bead to sample ration of 1.8:1.

Tiled amplicons were generated using 5 µL of cDNA, gene-specific forward and reverse primers, and NEBNext Ultra II Q5 Mastermix (Cat. No. M0544L). 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. Even and odd tiled amplicons were subsequently pooled for downstream library preparation.

In vitro SHAPE-MaP, reverse transcription, and library preparation

In vitro transcriptions of genomic WNV RNA was performed as described above, though reaction volume was increased 4x (1mL total) to ensure sufficient yields for downstream purification steps. After transcription, plasmid template was digested using RQ1 DNase, followed by addition of 30mg/mL Proteinase K (ThermoFisher, Cat. No. 17916) to inactivate all enzymes. To this reaction was added 25mM final EDTA at pH 8.0 to chelate Mg^{2+}. Samples were divided in half and applied to a 100kDa Amicon Ultra filtration column (Amicon, Cat. No. UFC510096) and spun to half volume. Filtration buffer (50mM K-HEPES pH 7.5, 150mM KCl, 100μM EDTA pH 8.0) was added to the sample, and spun to half volume. This step was repeated a total of 8 times to ensure removal of unincorporated nTPs and all products of enzymatic digestion.

Subsequently, RNA was subjected to size-exclusion chromatography, performed at room temperature, using a self-packed Sephacryl-1000 column with a 24 mL bed volume pre-equilibrated with filtration buffer. RNA from the peak fraction was diluted to 100ng/μL and folded in the presence of 10mM Mg^{2+} at 37°C for 30 min. Following folding, RNA was modified with either a final concentration of 10mM 1M7 for 3 min at 37°C (synthesized in-house) or 100mM NAI (EMD

Millipore) for 10 min at 37°C. In both cases, probing reactions were quenched by EtOH precipitation and incubated overnight at -20°C. RNA was then spun at 20,000g for 30 min at 4°C, washed once with 70%, spun again at 20,000g for 15 min at 4°C, and resuspended in 1x ME.

Reverse transcription (RT) reactions were prepared using 1 µg of *in vitro* purified RNA, SuperScript II (SSII), SSII-MaP reaction buffer (50mM 1M Tris-HCl pH 8.0, 75mM KCl, 10mM DTT, 6mM $MnCl_2$, 0.5mM dNTP), and random nonamers (NEB, Cat. No. S1254S). RT reactions were incubated at 42°C for 3hrs. Second strand synthesis was performed using the NEBNext Ultra II Non-Directional Second Strand synthesis module (NEB) according to manufacturer's protocol. Double-stranded cDNA was purified using Monarch DNA cleanup kits and a 5:1 binding buffer : sample ratio.

Library quantification, sequencing, and data analysis

Following generation of double-stranded cDNA *in vitro* libraries or dsDNA *in vivo* odd and even amplicon pools, samples were diluted to 0.2 ng/µL. Libraries were fragmented and tagged with Illumina sequencing adaptors using the NexteraXT DNA library preparation kit (Illumina, Cat. No. FC-131-1024) according to manufacturer's protocols, but at $1/5^{th}$ the recommended volume.

Libraries were quantified using a Qubit dsDNA HS Assay Kit (ThermoFisher, Cat. No. Q32851) to determine library concentration and a BioAnalyzer High Sensitivity DNA Analysis kit (Agilent, Cat. No. 5067-4636) to determine the average library member size. Library concentration and average size were used to dilute

libraries to 4nM which are subsequently denatured. Final library dilutions were prepared according to manufacturer's protocols, and libraries were sequenced on the NextSeq 500/550 platform.

All sequencing data were analyzed using ShapeMapper 2 (Busan and Weeks, 2018), aligning reads to the WNV genome (AF404756). Default quality control benchmarks implemented in ShapeMapper2 were used to ensure data is of high quality. Mutation rates of 1M7- or NAI-modified were compared to unmodified samples and tested for significance using the equal variance t-test.

Structure Prediction

ShapeKnots was used to examine the West Nile Virus genome for pseudoknots (Hajdin et al., 2013). Specifically, *in silico* predictions were performed across the entire genome in 500nt windows separated by a 100nt slide, with the 20 lowest minimum-free energy (MFE) structures output for each window. For each window, the coordinates of pseudoknots that appeared in the five most stable pseudoknotted structures were extracted. Pseudoknots were considered plausible if a given pseudoknot appeared in the majority of extracted pseudoknotted structures in the all of the windows that covered it. In addition to identifying five novel pseudoknots, these filtering criteria successfully capture two previously reported pseudoknots, including a programmed ribosomal-frameshifting pseudoknot contained in NS4b (Faggioni et al., 2012) and an exoribonuclease-resistant pseudoknot found in the 3'UTR (Funk et al., 2010; Pijlman et al., 2008).

Superfold (Smola et al., 2015a) was used to generate a unique consensus structure prediction for each replicate of *in vitro* and *in vivo* SHAPE data collected. Normalized reactivities were included as experimental constraints using default slope and intercept values and a maximum pairing distance of 500nt. Novel pseudoknots flagged with ShapeKnots were only included as hard constraints in individual predictions if the majority of pseudoknotted nucleotides had low reactivity (<0.4) in the corresponding datasets. Four additional pseudoknots for which an abundance of functional data exists were included as hard constraints after evaluation using the same criteria (**Fig A3.2**) (de Borba et al., 2019; Funk et al., 2010; MacFadden et al., 2018; Melian et al., 2014; Pijlman et al., 2008). A list of pseudoknots included in all SuperFold predictions is detailed in Table A3.3. Structures output from the SuperFold prediction pipeline were visualized using StructureEditor, a tool in the RNAstructure software suite (Reuter and Mathews, 2010). Full-length structures (.ct file) and SHAPE reactivities from *in vitro*, Vero, and C6/36 models are available at the PyleLab GitHub repository.

Identification of Well-Folded Regions

We relied on two on SHAPE reactivity and Shannon Entropy data signatures to identify regions that are highly structured and stably folded, respectively. SHAPE reacitivty is calculated using the ShapeMapper analysis tool described above. Shannon entropy is derived from base-pairing probabilities calculated using the SuperFold partition function (Smola et al., 2015a). Each of the replicate SHAPE

datasets collected in a given cell type were used to generate separate SuperFold predictions.

For a given cell-type, local median SHAPE reactivity and Shannon entropy were calculated in 55nt sliding windows. The global median SHAPE reactivity and Shannon entropy were subtracted to facilitate subsequent analysis steps. Regions were considered "well-folded" if they met two criteria; 1) local SHAPE and Shannon Entropy signals were below the global median for stretches $\geq$40 nucleotides and 2) these regions appeared in both replicate data sets. A region was not disqualified if the local SHAPE reactivity or Shannon Entropy rose above the global median for <40 nucleotides. Replicate consensus structure predictions are compared for regions that meet the above criteria to ensure agreement between structure models. Well-folded regions identified in each cell type are reported in Table 3.1**.**

Multiple Sequence Alignment

To analyze evolutionary support for consensus structure predictions generated in either Vero or C6/36 cell types, we compiled a codon-based multiple sequence alignment (MSA) for genomes of mosquito-borne *flaviviruses* (MBFV). All sequences were chosen based on a phylogenetic study of *flaviviruses* (Moureau et al., 2015). Sequences included in the MSA are detailed in Table A3.4, and were downloaded from NCBI.

The open-reading frames of each virus were extracted from the full-length sequences based on the GenBank annotations. Codon alignments were generated

using MACSE v2.0.3 (Ranwez et al., 2018) and default parameters (-prog alignSequences).

Synonymous Mutation Rate Analysis

Codon alignments generated with MACSE were visualized with Jalview v2.11.0 (Waterhouse et al., 2009). Importantly, sequences that corresponded to gaps in the parental WNV NY99 sequence were deleted. Synonymous mutation rates for each codon in the WNV genome were estimated using the phylogenetic-based parametric maximum likelihood (FUBAR) method (Murrell et al., 2013). Using a representative consensus structure prediction derived from either Vero or C6/36 cells, each codon was categorized as single- or double-stranded as determined by the stranded-ness of the nucleotide in the third position. Statistically significant differences between synonymous mutation rates separated into single- and double-stranded bins were determined using two-tailed, equal variance t-test.

LNA Design and Transfection

Locked nucleic acids were designed to anneal to target sequences within the WNV genome. All LNAs were designed with three consecutive LNA bases at the 5' and 3' ends, with internal stretches of unlocked bases limited to three consecutive nucleotides. All LNAs were designed with similar thermodynamic properties, such as length, %GC content, %LNA content, and LNA:RNA duplex T_m. A list of LNAs deployed in this study is included in Table A3.5**.** All LNAs were synthesized in-house.

Prior to LNA transfection, Vero and C6/36 cells were plated and grown to ~90% confluency in 12-well tissue culture plates. LNAs were co-transfected in quadruplicate at a final concentration of 400nM/well along with 0.5 ug of *in vitro* transcribed, Type-1 capped WNV genomic RNA, prepared as described above, using the Transit-mRNA reagent. Four hours post-transfection, cells were washed with cold 1xPBS, and complete media was replaced.

Supernatant samples were collected on 3 dpi from Vero cells, or 6dpi from C6/36 cell. Samples were spun down at 1000g for 5 min at 4°C to remove any cellular debris. To remove any RNAs not encapsulated in virions, supernatant samples were subjected to a 30 min RNase A degradation at 37°C. RNase A was deactivated with addition of 20mg/mL Proteinase K and incubation at 37°C for 1 hour. Viral RNA was extracted using the Mag-Bind Viral DNA/RNA 96 Kit (Omega Bio-Tek, M6246) and a Kingfisher Flex liquid-handling robot and frozen at -80°C prior to use.

Quantification of viral genomes

To monitor viral growth, we relied on a quantitative-RT-PCR assay adapted from (Lanciotti et al., 2000) that allows for absolute quantitation of viral genomes. We first prepared a 2.4kb standard from a portion of the WNV genome that includes the E gene (Genome Coords: 1031-3431). RNA standards were generated as described above, purified using an RNeasy column per the manufacturer's instructions, quantified and aliquoted in serial 10-fold dilutions.

Viral RNA copy numbers in supernatant or total cellular RNA samples were determined using a FAM-labeled TaqMan probe targeted against the E gene, primers WNV1160F and WNV1229R, and the Luna Universal Probe One-Step RT-qPCR kit (NEB, Cat. No. E3006L). Using a linear regression derived from the standard curve, we calculated viral RNA genome copy number per microliter of supernatant, and statistical outliers were removed using the ROUT outlier test available in GraphPad Prism. Primer and probe sequences used are available in Table A3.1.

3.7 Appendix

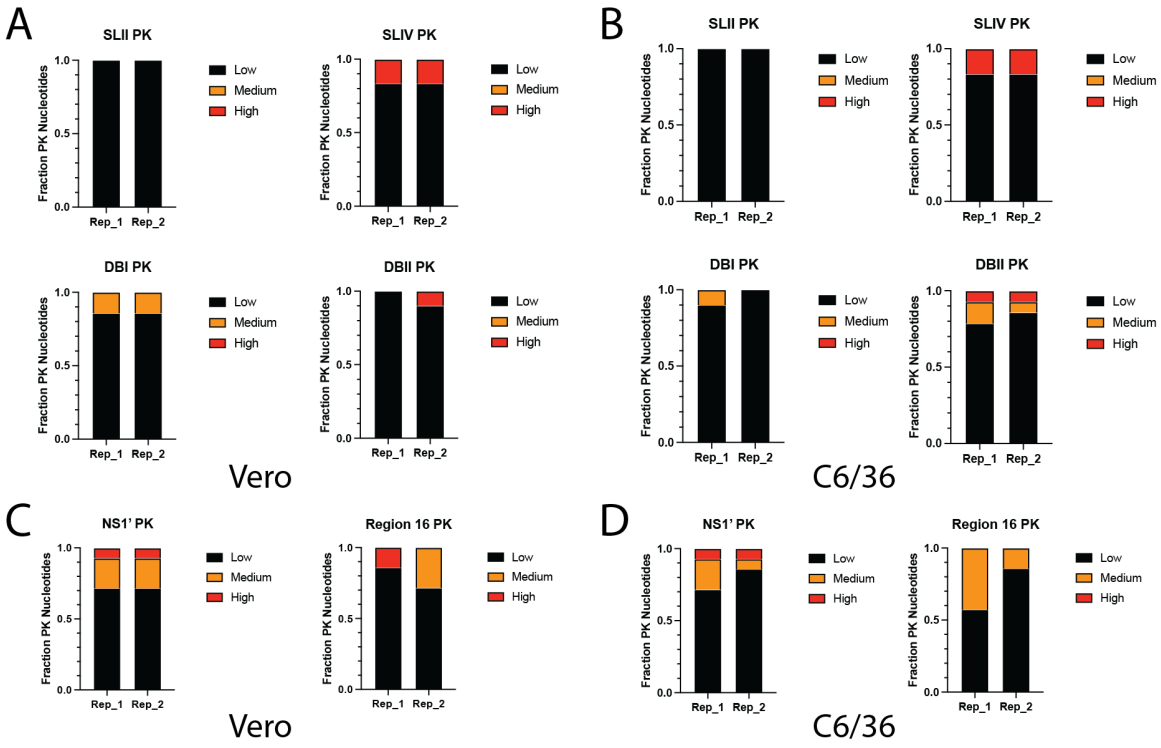

Figure A3.1. Analysis of SHAPE-MaP reactivities of pseudoknotted nucleotides in the WNV genome confirms the formation of these pseudoknots *in vivo*. A) Nucleotides of each pseudoknot in the WNV 3'UTR, binned by normalized reactivity collected in Vero cells, with bin size expressed as a fraction of total nucleotides in that pseudoknot. B) Nucleotides of each pseudoknot in the WNV 3'UTR, binned by normalized reactivity collected in C6/36 cells, plotted as in (A). C) Nucleotides of the NS1' pseudoknot (left) or a novel pseudoknot predicted to fold in Region 16 (right), binned by normalized reactivity collected in Vero cells, plotted as in (A). D) Nucleotides of the NS1' pseudoknot (left) or a novel pseudoknot predicted to fold in Region 16 (right), binned by normalized reactivity collected in C6/36 cells, plotted as in (A).

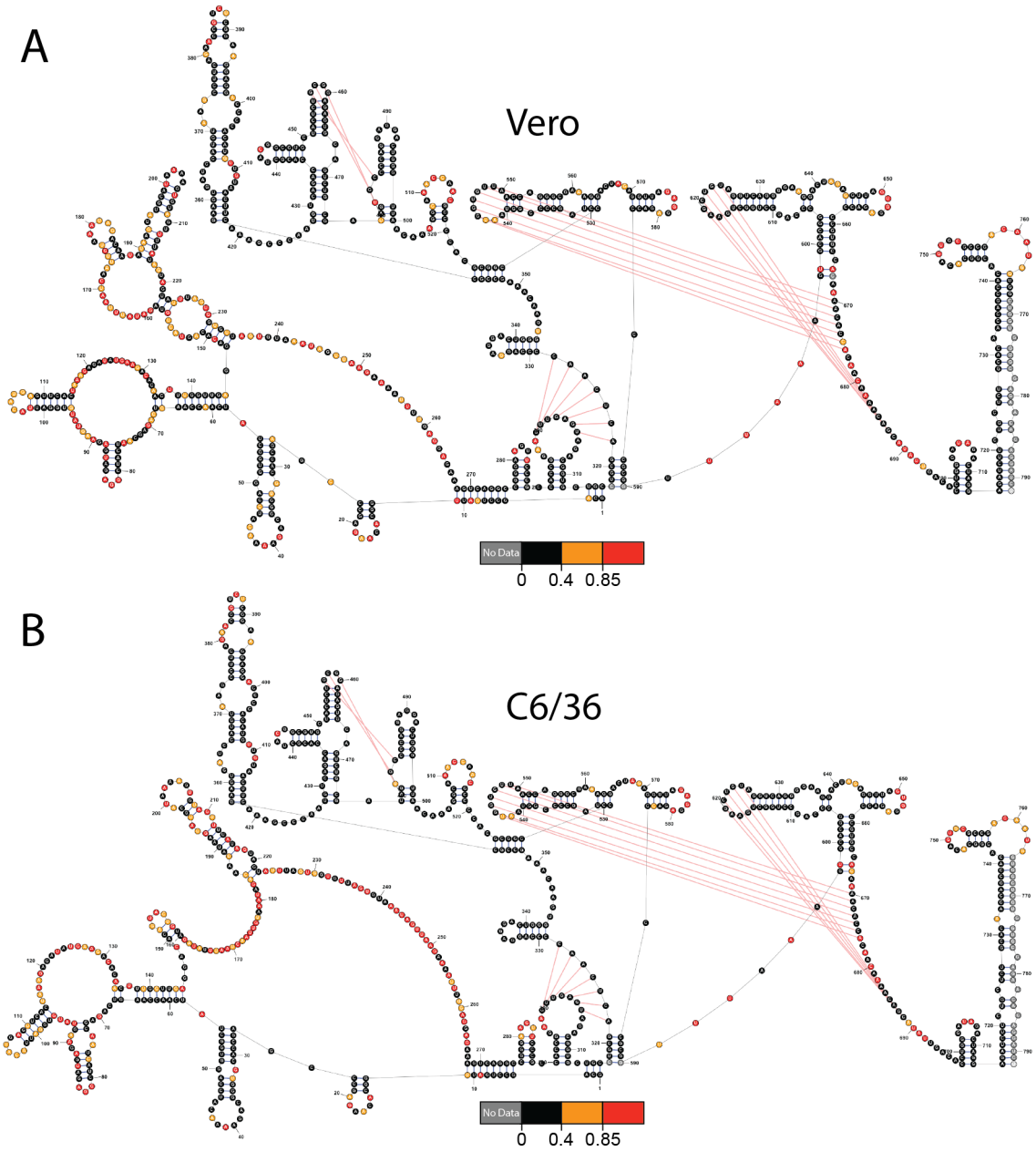

Figure A3.2. Normalized SHAPE reactivity mapped to the structure prediction of the 3' viral terminus reveals domain-specific patterns of RNA backbone flexibility Structure of the 3' viral terminus determined in infected (A) Vero cells or (B) C6/36, color-coded by normalized SHAPE reactivity.

Table A3.1. Mutagenic, qRT-PCR primers

Primer Name	Sequence	Purpose
Cyc_Def_F	aaGACACCTGGGATAGACTAGG	Cyc_Defect_F
Cyc_Def_R	ataGCTGTTTTGTTGTGGTGTTTTG	Cyc_Defect_R
WNV_1160_F	TCAGCGATCTCTCCACCAAAG	for TaqMan Assay
WNV_1229_R	GGGTCAGCACGTTTGTCATTG	for TaqMan Assay
WNV_1031_F	TAATACGACTCACTATAGATTTGGTTCTCGAAGGCGACAG	Amplify E gene + T7 promoter
WNV_3430_R	GTGGTGGTAAGGTGCAGCTC	Amplify E Gene, R

Table A3.2. Gene-specific primers for SHAPE-MaP

Primer Name	Sequence	Primer Name	Sequence	Primer Name	Sequence
RT_WNV_Amplicon_1	AGATCCTGTGTTCTCGCAC	R_PCR_WNV_Amplicon_1	TCCTGTGTTCTCGCACCAC	F_PCR_WNV_Amplicon_1	GAAGTATGTGGATTACATGAGTTCA
RT_WNV_Amplicon_2	AAGATCTCCTAGTCTATCC	R_PCR_WNV_Amplicon_2	AGATCTCCTAGTCTATCCCAG	F_PCR_WNV_Amplicon_2	AGGAAAACGAGAGGACATC
RT_WNV_Amplicon_3	TAAAACTACACTTTTATGC	R_PCR_WNV_Amplicon_3	CTATAAAACTACACTTTTATGCATA	F_PCR_WNV_Amplicon_3	CTTCCTCAATGCTATGTCA
RT_WNV_Amplicon_4	TTCATGATCAATTCAGTGA	R_PCR_WNV_Amplicon_4	TCATGATCAATTCAGTGAAAT	F_PCR_WNV_Amplicon_4	AGAGAGAGAAAAAACCCGG
RT_WNV_Amplicon_5	TGAGTTCTTTCTTCCAAGC	R_PCR_WNV_Amplicon_5	GAGTTCTTTCTTCCAAGCCAG	F_PCR_WNV_Amplicon_5	TCAGTGAATATGACCAGCC
RT_WNV_Amplicon_6	TACATCTTCCTCGTATTGG	R_PCR_WNV_Amplicon_6	ACATCTTCCTCGTATTGGGGT	F_PCR_WNV_Amplicon_6	ATAACATGGACACTCATAAAGAACA
RT_WNV_Amplicon_7	TTTCTTTCCAAACCTCTCC	R_PCR_WNV_Amplicon_7	TTCTTTCCAAACCTCTCCCAA	F_PCR_WNV_Amplicon_7	GGGAGAGTTTCTTTTGGAC
RT_WNV_Amplicon_8	TTTATTGAGGTCAATGAGG	R_PCR_WNV_Amplicon_8	TTATTGAGGTCAATGAGGTGT	F_PCR_WNV_Amplicon_8	CGAAGCTTGGTGAAAGGAA
RT_WNV_Amplicon_9	TATCTGAGAACGTTTTCCC	R_PCR_WNV_Amplicon_9	ATCTGAGAACGTTTTCCCGAG	F_PCR_WNV_Amplicon_9	AAGTAGTCCAATTGAACAGAAAGTC
RT_WNV_Amplicon_10	TTTCAGATATGTCTGTTGT	R_PCR_WNV_Amplicon_10	TTCAGATATGTCTGTTGTGAT	F_PCR_WNV_Amplicon_10	TAGTGCAGGGTGAAAGGAT
RT_WNV_Amplicon_11	TCAGTCTTCTGTTTATGGC	R_PCR_WNV_Amplicon_11	CAGTCTTCTGTTTATGGCCTC	F_PCR_WNV_Amplicon_11	GGAGCACCTTGGAAGATAT
RT_WNV_Amplicon_12	TTATCCAAAATCCAACTAC	R_PCR_WNV_Amplicon_12	TATCCAAAATCCAACTACTGA	F_PCR_WNV_Amplicon_12	AATGGCTTATCACGATGCC
RT_WNV_Amplicon_13	TATGGCTCTCAGTATCATC	R_PCR_WNV_Amplicon_13	ATGGCTCTCAGTATCATCCAA	F_PCR_WNV_Amplicon_13	CTGGGTACAAGACACAAAA
RT_WNV_Amplicon_14	TTATCAACTTTCCGCTCTC	R_PCR_WNV_Amplicon_14	TATCAACTTTCCGCTCTCTGT	F_PCR_WNV_Amplicon_14	GGAAGCAGTGAAGGACGAG
RT_WNV_Amplicon_15	TTTAGGTGCTGACTTGTAC	R_PCR_WNV_Amplicon_15	TTAGGTGCTGACTTGTACATT	F_PCR_WNV_Amplicon_15	TTGGTCACTGTCAACCCTT
RT_WNV_Amplicon_16	TTGATCTGTTGTTCTCCTC	R_PCR_WNV_Amplicon_16	TGATCTGTTGTTCTCCTCTGC	F_PCR_WNV_Amplicon_16	AACTACTCCACACAGGTTG
RT_WNV_Amplicon_17	TATTCTCCAAGCTTTAGTG	R_PCR_WNV_Amplicon_17	ATTCTCCAAGCTTTAGTGTGT	F_PCR_WNV_Amplicon_17	CACTGACAGTGCAGACACA
RT_WNV_Amplicon_18	TCTGTTTTTACCAAATACC	R_PCR_WNV_Amplicon_18	CTGTTTTTACCAAATACCTTG	F_PCR_WNV_Amplicon_18	CTCGATGTCTAAGAAACCA
RT_WNV_Amplicon_19	TTGGTGCATCTTCCATACC	R_PCR_WNV_Amplicon_19	GTGCATCTTCCATACCTGA	F_PCR_WNV_Amplicon_19	AGTAGTTCGCCTGTGTGAGCT

Table A3.3. Pseudoknot coordinates/constraints for structure prediction

PK_1		Region 16 PK	
5' Arm, nt Coord.	3' Arm, nt Coord.	5' Arm, nt Coord.	3' Arm, nt Coord.
753	932	9093	9137
754	931	9094	9136
755	930	9095	9135
756	929	9096	9134
757	928	9097	9133
758	928	9098	9132
759	926	9099	9131
760	925	DBI PK	
761	924	5' Arm, nt Coord.	3' Arm, nt Coord.
PK_2		10780	10913
5' Arm, nt Coord.	3' Arm, nt Coord.	10781	10912
2198	2365	10782	10911
2199	2364	10783	10910
2200	2363	10784	10909
2201	2362	10785	10908
2202	2361	10786	10907
2203	2360	NS1' PK	
PK_3		5' Arm, nt Coord.	3' Arm, nt Coord.
5' Arm, nt Coord.	3' Arm, nt Coord.	3575	3619
3711	4001	3576	3618
3712	4000	3577	3617
3713	3999	3578	3616
3714	3998	3579	3615
3715	3997	3580	3614
3716	3996	3581	3613
PK_4		SLII PK	
5' Arm, nt Coord.	3' Arm, nt Coord.	5' Arm, nt Coord.	3' Arm, nt Coord.
5998	6142	10535	10565
5999	6141	10536	10564
6000	6140	10537	10563
6001	6139	10538	10562
6002	6138	10539	10561
6003	6137	10540	10560
6004	6136	10541	10559
NS4B'		SLIV PK	
5' Arm, nt Coord.	3' Arm, nt Coord.	5' Arm, nt Coord.	3' Arm, nt Coord.
7336	7364	10694	10719
7337	7363	10695	10718
7338	7362	10696	10717
7339	7361	DBII PK	
7340	7360	5' Arm, nt Coord.	3' Arm, nt Coord.
7341	7359	10857	10925
		10858	10924
		10859	10923
		10860	10922
		10861	10921

Table A3.4. Viral genome sequences used for MSA construction

Virus	Abbreviation	Accession Number
Banzi Virus	BANV	DQ859056
Uganda S Virus	UGSV	DQ859065
Jugra Virus	JUGV	DQ859066
Potiskum Virus	POTV	DQ859067
Saboya Virus	SABV	DQ859062
Bouboui Virus	BOUV	DQ859057
Sepik Virus	SEPV	DQ859063
Wesselsbron Virus	WESSV	DQ859058
Yellow Fever Virus	YFV	JX898878
Ilomantsi Virus	ILOV	KC692067
Donggang Virus	DNGV	NC_016997
Chaoyang Virus	CHOV	FJ883471
Lammi Virus	LAMV	KC692068
Koutango Virus	KOUV	EU082200
Kunjin Virus	KUNV	AY274504
West Nile Virus	WNV	DQ318020.1
Yaounde Virus	YAOV	EU082199
West Nile Virus	WNV	AF404756
Alfuy Virus	ALFV	AY898809
Murray Valley Encephalitis Virus	MVEV	NC_000943
Usutu Virus	USUV	NC_006551
Japenese Encephalitis Virus	JEV	NC_001437
Cacipacore Virus	CPCV	KF917536.1
St. Louis Encephalitis Virus	SLEV	NC_007580
Ilheus Virus	ILHV	AY632539
Rocio Virus	ROCV	AY632542
Bagaza Virus	BAGV	AY632545
Israel Turkey Meningoencephalomyelitis Virus	ITV	KC734550.1
Ntaya Virus	NTAV	JX236040
Baiyangdian Virus	BYD	JF312912
Sitiawan Virus	STWV	JC477686
Tembusu Virus	TMUV	JX577685
Naranjal Virus	NJLV	KF917538
Bussuquara Virus	BSQV	NC_009026
Iguape Virus	IGUV	AY632538
Aroa Virus	AROAV	KF917535
Spondweni Virus	SPOV	DQ859064
Zika Virus	ZKV	EU545988
Kedougou Virus	KEDV	DQ859061
Kokobera Virus	KOKV	NC_009029
Stratford Virus	STRV	KM225263
Dengue Virus	DENV_1	EU081265
Dengue Virus	DENV_3	AY648961
Dengue Virus	DENV_2	EU687249
Dengue Virus	DENV_4	AY618991
Nounane Virus	NOUV	EU159426
New Mapoon Virus	NMV	KC788512

Table A3.5. LNAs used in this study (LNA bases are indicated with a "+")

Region	LNA	Length	LNA Content	%GC	RNA Tm
3'CYC	+T+C+TA+TCC+CAGG+TGT+CAA+T+A+T	20	50.0	40.0	90.0
Region 12	+A+A+TGA+GG+TGT+TGA+TG+TA+A+T+C	20	55	35	87
Region 8	+C+G+TTC+TTA+CAT+TT+TG+GG+T+A+C	20	55	40	88
Region 11	+A+T+TAG+CAC+AA+TCA+TC+AA+G+A+G	20	55	35	83
Region 6	+C+T+TGGC+TG+TC+CAC+CTCT+T+G+C	20	50	60	90
Region 10	+A+T+TCC+CCA+TCCA+TGG+TA+T+A+T	20	50	40	89
Region 16	+G+G+CTC+TGC+TTC+CCT+TGG+C+C+T	20	50	65	94
Non-targeting	+G+T+TTA+TC+TAG+TAA+TAG+ATT+A+C+C	22	50.0	27.3	90
Targeting, highSS	+T+G+ATT+CTG+CTCT+TCA+AA+C+A+T	20	50.0	35.0	88

3.8 References

1. Adams, R.L., Huston, N.C., Tavares, R.C.A., and Pyle, A.M. (2019). Sensitive detection of structural features and rearrangements in long, structured RNA molecules (Elsevier Inc.).
2. Assis, R. (2014). Strong Epistatic Selection on the RNA Secondary Structure of HIV. PLoS Pathog. *10*.
3. Basu, M., and Brinton, M.A. (2011). West Nile virus (WNV) genome RNAs with up to three adjacent mutations that disrupt long distance 5′-3′ cyclization sequence basepairs are viable. Virology *412*, 220–232.
4. Blackwell, J.L., and Brinton, M.A. (1997). Translation Elongation Factor-1 Alpha Interacts with the 3 J Stem-Loop Region of West Nile Virus Genomic RNA. J. Virol. *71*, 6433–6444.
5. de Borba, L., Lequime, S., Sánchez Vargas, I., Gebhard, L.G., Lambrechts, L., Marsico, F.L., Carballeda, J.M., Blair, C.D., Filomatori, C. V., Villordo, S.M., et al. (2019). RNA Structure Duplication in the Dengue Virus 3' UTR: Redundancy or Host Specificity? MBio *10*, 1–18.
6. Brackney, D.E., Beane, J.E., and Ebel, G.D. (2009). RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. PLoS Pathog. *5*, 1–9.
7. Brinton, M.A. (2014). Replication Cycle and Molecular Biology of the West Nile Virus. 13–53.
8. Bujalowski, P.J., Bujalowski, W., and Choi, K.H. (2017). Interactions between the Dengue Virus Polymerase NS5 and Stem-Loop A. J. Virol. *91*, 1–11.
9. 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.
10. Choi, K.H. (2021). The role of the stem-loop a rna promoter in flavivirus replication. Viruses *13*.
11. Clyde, K., and Harris, E. (2006). RNA secondary structure in the coding region of Dengue Virus type 2 directs translation start codon selection and is required for viral replication. J. Virol. *80*, 2170–2182.
12. Davis, W.G., Basu, M., Elrod, E.J., Germann, M.W., and Brinton, M.A. (2013). Identification of cis-Acting Nucleotides and a Structural Feature in West Nile Virus 3-Terminus RNA That Facilitate Viral Minus Strand. *87*, 7622–7636.
13. Dethoff, E.A., Gokhale, N.S., Boerneke, M.A., Muhire, B.M., Martin, D.P., Sacco, M.T.,

McFadden, M.J., Weinstein, J.A., Messer, W.B., Horner, S.M., et al. (2018). Pervasive Tertiary Structures in the Dengue Virus RNA Genome Modulate Fitness. Submitted *115*, 11513–11518.
14. Dong, H., Zhang, B., and Shi, P.Y. (2008). Terminal structures of West Nile virus genomic RNA and their interactions with viral NS5 protein. Virology *381*, 123–135.
15. Faggioni, G., Pomponi, A., De Santis, R., Masuelli, L., Ciammaruconi, A., Monaco, F., Di Gennaro, A., Marzocchella, L., Sambri, V., Lelli, R., et al. (2012). West Nile alternative open reading frame (N-NS4B/WARF4) is produced in infected West Nile Virus (WNV) cells and induces humoral response in WNV infected individuals. Virol. J. *9*, 1–14.
16. Filomatori, C. V., Iglesias, N.G., Villordo, S.M., Alvarez, D.E., and Gamarnik, A. V. (2011). RNA sequences and structures required for the recruitment and activity of the dengue virus polymerase. J. Biol. Chem. *286*, 6929–6939.
17. Friebe, P., Shi, P., and Harris, E. (2011). The 5' and 3' Downstream AUG Region Elements Are Required for Mosquito-Borne Flavivirus RNA Replication. *85*, 1900–1905.
18. Funk, A., Dong, H., Shi, P.-Y., Edmonds, J., Khromykh, A.A., Floden, N., Torres, S., Truong, K., Nagasaki, T., and Balmori Melian, E. (2010). RNA Structures Required for Production of Subgenomic Flavivirus RNA. J. Virol. *84*, 11407–11417.
19. Göertz, G.P., Fros, J.J., Miesen, P., Vogels, C.B.F., van der Bent, M.L., Geertsema, C., Koenraadt, C.J.M., van Rij, R.P., van Oers, M.M., and Pijlman, G.P. (2016). Noncoding Subgenomic Flavivirus RNA Is Processed by the Mosquito RNA Interference Machinery and Determines West Nile Virus Transmission by Culex pipiens Mosquitoes. J. Virol. *90*, 10145–10159.
20. 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. *110*, 5498–5503.
21. Huston, N.C., Wan, H., Strine, M.S., de Cesaris Araujo Tavares, R., Wilen, C., and Pyle, A.M. (2021). Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Mol. Cell *81*, 1–15.
22. Iglesias, N.G., and Gamarnik, A. V (2011). Dynamic RNA structures in the dengue virus genome. RNA Biol. *8*, 249–257.
23. Korbie, D.J., and Mattick, J.S. (2008). Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat. Protoc. *3*, 1452–1456.
24. Kortmann, J., and Narberhaus, F. (2012). Bacterial RNA thermometers : molecular zippers and switches. Nat. Publ. Gr. *10*, 255–265.
25. Lanciotti, R.S., Kerst, A.J., Nasci, R.S., Godsey, M.S., Mitchell, C.J., Savage, H.M., Komar, N., Panella, N.A., Allen, B.C., Volpe, K.E., et al. (2000). Rapid Detection of West Nile Virus from Human Clinical Specimens. J. Clin. Pathol. *38*, 4066–4071.
26. 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*, 1–26.
27. Lee, E., Bujalowski, P.J., Teramoto, T., Gottipati, K., Scott, S.D., Padmanabhan, R., and Choi, K.H. (2021). Structures of flavivirus RNA promoters suggest two binding modes with NS5 polymerase. Nat. Commun. *12*, 1–12.
28. 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.
29. Liu, Z., Li, X., Jiang, T., Deng, Y., and Ye, Q. (2016). Viral RNA switch mediates the dynamic control of flavivirus replicase recruitment by genome cyclization. Elife 1–27.
30. Lundin, K.E., Højland, T., Hansen, B.R., Persson, R., Bramsen, J.B., Kjems, J., Koch, T., Wengel, J., and Smith, C.I.E. (2013). Biological Activity and Biotechnological Aspects of Locked Nucleic Acids. In Advances in Genetics, pp. 47–107.
31. MacFadden, A., Òdonoghue, Z., Silva, P.A.G.C., Chapman, E.G., Olsthoorn, R.C., Sterken, M.G., Pijlman, G.P., Bredenbeek, P.J., and Kieft, J.S. (2018). Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. Nat. Commun. *9*, 1–11.
32. Melian, E.B., Hinzman, E., Nagasaki, T., Firth, A.E., Wills, N.M., Nouwens, A.S., Blitvich, B.J., Leung, J., Funk, A., Atkins, J.F., et al. (2010). NS1' of Flaviviruses in the Japanese Encephalitis Virus Serogroup Is a Product of Ribosomal Frameshifting and Plays a Role in Viral Neuroinvasiveness. *84*, 1641–1647.
33. Melian, E.B., Hall-mendelin, S., Du, F., Owens, N., Bosco-lauth, A.M., Nagasaki, T., Rudd, S., Brault, A.C., Bowen, R.A., Hall, R.A., et al. (2014). Programmed Ribosomal Frameshift Alters Expression of West Nile Virus Genes and Facilitates Virus Replication in Birds and Mosquitoes. *10*.
34. Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. (2005). RNA Structure Analysis at Single Nucleotide Resolution by Selective 2 ′ -Hydroxyl Acylation and Primer Extension (SHAPE). 6866–6874.
35. Meyer, A., Freier, M., Schmidt, T., Rostowski, K., Zwoch, J., Lilie, H., Behrens, S.E., and Friedrich, S. (2020). An RNA thermometer activity of the West Nile virus genomic 30-terminal stem-loop element modulates viral replication efficiency during host switching. Viruses *12*, 1–22.
36. Mitchell, D., Assmann, S.M., and Bevilacqua, P.C. (2019). Probing RNA structure in vivo. Curr. Opin. Struct. Biol. *59*, 151–158.
37. Moureau, G., Cook, S., Lemey, P., Nougairede, A., Forrester, L., Khasnatinov, M., Charrel, R.N., Firth, A.E., and Gould, E.A. (2015). New Insights into Flavivirus Evolution , Taxonomy and Biogeographic History , Extended by Analysis of Canonical and Alternative Coding Sequences. PLoS One *10*.
38. 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.
39. Pijlman, G.P., Funk, A., Kondratieva, N., Leung, J., Torres, S., van der Aa, L., Liu, W.J., Palmenberg, A.C., Shi, P.Y., Hall, R.A., et al. (2008). A Highly Structured, Nuclease-Resistant, Noncoding RNA Produced by Flaviviruses Is Required for Pathogenicity. Cell Host Microbe *4*, 579–591.
40. Polacek, C., Friebe, P., and Harris, E. (2009). Poly(A)-binding protein binds to the non-polyadenylated 3' untranslated region of dengue virus and modulates translation efficiency. J. Gen. Virol. *90*, 687–692.
41. 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.
42. Reuter, J.S., and Mathews, D.H. (2010). RNAstructure: Web servers for RNA secondary structure prediction and analysis. BMC Bioinformatics *11*.
43. 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.
44. Shi, P., Tilgner, M., Lo, M.K., Kent, K.A., and Bernard, K.A. (2002). Infectious cDNA Clone of the Epidemic West Nile Virus from New York City. *76*, 5847–5856.
45. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*.
46. Simmonds, P., and Smith, D.B. (1999). Structural Constraints on RNA Virus Evolution. J. Virol. *73*, 5787–5794.
47. 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.
48. Smola, M.J., Rice, G.M., Busan, S., Siegfried, N.A., and Weeks, K.M. (2015a). 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.
49. Smola, M.J., Calabrese, J.M., and Weeks, K.M. (2015b). Detection of RNA-Protein Interactions in Living Cells with SHAPE. Biochemistry *54*, 6867–6875.
50. 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.
51. Suzuki, R., Fayzulin, R., Frolov, I., and Mason, P.W. (2008). Identification of mutated cyclization sequences that permit efficient replication of West Nile virus genomes: use in safer propagation of a novel vaccine candidate. J. Virol. *82*, 6942–6951.
52. Tang, G.Q., Nandakumar, D., Bandwar, R.P., Lee, K.S., Roy, R., Ha, T., and Patel, S.S. (2014). Relaxed rotational and scrunching changes in P266L mutant of T7 RNA polymerase reduce short abortive RNAs while delaying transition into elongation. PLoS One *9*, 1–12.
53. Tavares, R. de C.A., Mahadeshwar, G., Wan, H., Huston, N.C., and Pyle, A.M. (2021). The Global and Local Distribution of RNA Structure throughout the SARS-CoV-2 Genome. J. Virol. *95*, 1–17.
54. 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.
55. 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.
56. Villordo, S.M., and Gamarnik, A. V. (2013). Differential RNA Sequence Requirement for Dengue Virus Replication in Mosquito and Mammalian Cells. J. Virol. *87*, 9365–9372.

57. Villordo, S.M., Alvarez, D.E., and Gamarnik, A. V. (2010). A balance between circular and linear forms of the dengue virus genome is crucial for viral replication. RNA *16*, 2325–2335.
58. Villordo, S.M., Filomatori, C. V, Sánchez-vargas, I., and Blair, C.D. (2015). Dengue Virus RNA Structure Specialization Facilitates Host Adaptation. PLoS Pathog. *11*, 1–22.
59. Wan, H., Adams, R.L., Lindenbach, B.D., and Pyle, A.M. (2022). The In Vivo and In Vitro Architecture of the Hepatitis C Virus RNA Genome Uncovers Functional RNA Secondary and Tertiary Structures . J. Virol. *96*, 1–23.
60. 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.
61. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.
62. Whitehorn, J., and Yacoub, S. (2019). Global warming and arboviral infections. Clin. Med. J. R. Coll. Physicians London *19*, 149–152.
63. Winkelmann, E.R., Widman, D.G., Suzuki, R., and Mason, P.W. (2011). Analyses of mutations selected by passaging a chimeric flavivirus identify mutations that alter infectivity and reveal an interaction between the structural proteins and the nonstructural glycoprotein NS1. Virology *421*, 96–104.
64. Zhang, B., Dong, H., Stein, D.A., Iversen, P.L., and Shi, P. (2008). West Nile virus genome cyclization and RNA replication require two pairs of long-distance RNA interactions. *373*, 1–13.

4. Conclusions

4.1 Summary of Findings

The work presented here adds to a growing body of research demonstrating that the genomes of single-stranded, positive-sense RNA viruses do far more than simply code for proteins. Instead, they are replete with functional RNA secondary structures that mediate important aspects of the viral life cycle (Dethoff et al., 2018; Huber et al., 2019; Li et al., 2018; Madden et al., 2020; Pirakitikulr et al., 2016; Siegfried et al., 2014; Wan et al., 2022). My dissertation work expands this observation to two additional viruses, SARS-CoV-2 and West Nile virus, through application and modification of the SHAPE-MaP workflow.

The experimentally constrained secondary structure models of both SARS-CoV-2 and West Nile virus genomes capture conserved structures known to mediate important roles in *β-coronaviruses* and *flaviviruses*, respectively. Searching through these genomes, we showed that regions of well-folded RNA with functional potential are scattered throughout the viral ORFs. Even more, we demonstrated that a subset of these structures mediate functional roles in the viral life cycle. Beyond allowing us to discover novel aspects of viral biology, these structural models will serve as invaluable roadmaps to the virology community writ large. More importantly, the work under-taken here resulted in, and often times required, methodological innovations discussed below that will greatly facilitate future studies of viral RNA structure.

4.1.1 SHAPE-MaP data collection

The study of SARS-CoV-2 genome structure necessitated several methodological innovations, in large part due to the relatively large size of its RNA genome. The original protocol for *in vivo* SHAPE-MaP data collection called for the use of SuperScript II (SSII), a reverse transcriptase (RT) enzyme derived from the Moloney murine leukemia virus (MMLV). However, SSII possess an intrinsically low processivity, and can only reverse transcribe ~700 nucleotides along a modified RNA in the standard SHAPE-MaP protocol. The use of SSII would therefore require >50 amplicons to afford full genome coverage of SARS-CoV-2, and >100 amplicons for SHAPE-MaP experiments. As each amplicon requires a unique RT and PCR reaction, SHAPE-MaP data collection would become tedious and reagent intensive to the point of intractability.

The Pyle lab has previously discovered MarathonRT, a Group II-encoded maturase with exceptional processivity (Zhao et al., 2018). Not only does MarathonRT out-perform SSII on unmodified RNAs, preliminary work from our lab demonstrated its applicability for SHAPE-MaP data collection with long amplicons as well (Guo et al., 2020). Using this enzyme, we were able to quickly adapt MaP RT conditions for the use of total cellular RNA from SARS-CoV-2 infected cells. As a result of MarathonRT's processivity, 16 amplicons were sufficient to afford full genome coverage. As such, a single researcher can perform MaP RT and subsequently library construction with fewer reactions and reduced input cellular RNA. The method deployed in pursuit of the SARS-CoV-2 work therefore makes the

study of extremely long viral RNAs feasible, and our hope is that will serve as a valuable tool for other researchers.

However, this method has important implications for structural studies of long viral RNAs beyond simple issues of feasibility. Researchers have long appreciated the functional importance of RNA elements with complex 3D folds, such as the HCV IRES. However, methods that allow for discovery of these elements, especially in the context of viral RNA genomes, are still being developed. One such method, adapted from existing chemical probing protocols, allows for identification of higher-order structures in viral RNAs. Briefly, it relies on identification of nucleotide pairs with correlated chemical probing signals (Dethoff et al., 2018; Homan et al., 2014). As this method requires that chemical modifications appear on the same molecule, it is impossible to detect long-range interactions that exceed the processivity of the RT enzyme used. As such, using MarathonRT for correlated chemical probing would allow for detection of even longer-distance through-space interactions. It is worth noting that the read-format requirements of Illumina sequencing platforms also restrict the upper limit of detection. To that end, adapting sequencing-based structure probing methodologies for Nanopore platforms would similarly improve the study of long-distance through-space interactions.

4.1.2 Prioritizing well-folded RNAs for functional validation

While recent methodological advancements have facilitated the study of entire viral RNA genomes, the sheer volume of secondary structures contained in genome-wide structure predictions requires the use of sorting criteria to flag

regions that merit follow-up analysis. Typically, researchers have relied on signals of evolutionary conservation to identify candidates with functional potential (Dethoff et al., 2018; Kutchko et al., 2018; Pirakitikulr et al., 2016). Indeed, these methods proved successful in flagging regions in the SARS-CoV-2 genome with functional potential (**Figure 2.5, 2.6**). However, it is well understood that the use of low information alignments can hamstring these analysis pipelines (Rivas et al., 2020; Tavares et al., 2019). Assessing conservation of RNA structures in viral coding regions is made harder because the evolutionary pressures placed on sequence with coding potential further constrains the mutation rate of these sequences.

It was no surprise, then, to see these methods prove underpowered when applied to well-folded regions identified in the West Nile virus genome; analysis of synonymous mutation rates flagged only a single region. However, comparison of well-folded regions identified in both mammalian and arthropod cell lines suggested a path forward. Specifically, we reasoned that patterns of structural homology might be useful indicators of functional potential. Remarkably, all six regions identified on the basis of structural homology were shown to play a functional role in at least one cell type tested as WNV growth was reduced upon structure disruption (**Figure 3.6**). These results therefore highlight that, at least for enzootic viruses, patterns of structural homology can be used as a powerful sorting criteria, especially in instances where evolutionary data is weak or non-existent.

Though the application of this strategy to other viruses that alternate between hosts makes intuitive biological sense, it is possible that it may be applicable to viruses that infect a single host. For example, in the case of SARS-CoV-

2, it would be interesting to probe and predict the entire genome in multiple cells types it can infect (Ravindra et al., 2021). A comparison of the structures adopted in each cell type would allow for identification of ones that appear in all cellular contexts, suggesting functionally important folds. A comparison of validated structures identified either by evolutionary conservation of structural homology might yield an understanding of which signal is better correlated with RNA function.

4.1.3 Functional validation of viral RNA secondary structures

The study of SARS-CoV-2 and West Nile virus represent two excellent working examples highlighting the shortcomings of classical viral genetics strategies for validating candidate RNAs. In the case of SARS-CoV-2, the lack of an infectious clone early in the pandemic meant that, despite having candidate structures to validate, it was impossible to perform viral mutagenesis. Though an infectious clone of West Nile virus has existed since 2002, *Flaviviruses* are incredibly difficult to work with. Indeed, these plasmids are toxic to bacteria and prone to recombination, necessitating the use of very low copy plasmids and specialized bacterial strains. In my own experience, it took over 4 months of cloning optimization to successfully introduce a single, contiguous 5-nucleotide substitution in the WNV genome.

It was this difficulty with WNV cloning that inspired me to explore alternate strategies for validating candidate structures early on in my dissertation. Thankfully, I stumbled upon two examples of anti-sense, locked nucleic acids (LNAs) being deployed to study functional RNA secondary structure (Dethoff et al., 2018; Tuplin et al., 2015). With these studies serving as a roadmap, we were able to

successfully adapt the structure-disrupting LNA strategy to both viruses. In the context of both SARS-CoV-2 and WNV, we demonstrated that LNAs targeted against highly conserved functional elements mediated potent defects in viral growth *in vivo*. Along with the use of careful controls, were able to extend this strategy to our candidate RNA structures to identify novel functional elements in both viral ORFs. As this strategy is faster and more scalable than classic mutagenic strategies, our hope is that the method is widely adopted for structural inquiries of other RNA viruses.

More exciting, however, is the possible application of structure-disrupting LNAs as anti-viral therapeutics. Indeed, the field of nucleic acid therapeutics has exploded in the past several years, with the most notable example being the SARS-CoV-2 mRNA vaccine (Corbett et al., 2020). The Pyle lab has also contributed to this field, identifying a small synthetic RNA that has shown promise as an anti-tumor drug (Jiang et al., 2019). It is not hard to imagine, then, that LNAs that mediate profound replication defects in viral growth could be adapted to clinical settings. Even more, as functional RNA structures are often conserved across entire viral families, individual strategies of LNA targeting could prove efficacious against a wide variety of human pathogens. For example, LNAs that disrupt genome cyclization could be deployed against any *flavivirus*, albeit with virus-specific sequence. Either way, LNAs represent an exciting class of nucleic acids therapeutics.

4.2 Future Directions & Perspectives

The experimentally constrained, genome-wide secondary structure models of both SARS-CoV-2 and West Nile virus represent a wealth of data from which focused studies of viral biology can be launched. As an example, RNA elements in the WNV genome that mediate pan-host functions could be further studied using RNA antisense purification coupled with mass spectrometry (RAP-MS) to assess if they mediate their function via recruitment of viral proteins. However, it has become evident in the process of conducting this work that several methods used for structure discovery in long viral RNAs require further improvement.

4.2.1 Pseudoknot prediction

Pseudoknots are a class of RNA tertiary structure and play a diverse set of conserved roles in the life cycles of RNA viruses. The genomes of both SARS-CoV-2 and West Nile virus contain well-studied examples of programmed ribosomal frameshifting pseudoknots, RNA structures that cause ribosomes to slip into different reading frames (Faggioni et al., 2012; Kelly et al., 2020; Melian et al., 2014). The WNV 3'UTR, like all *flaviviruses*, contains a unique class of pseudoknots that rely on their mechanstic stability stall an exoribonuclease (Göertz et al., 2016; MacFadden et al., 2018; Pijlman et al., 2008).

Owing to assumptions hard-coded into RNA structure prediction algorithms that rely on dynamic programming, however, pseudoknots require prediction pipelines distinct from the ones implemented in SuperFold (Mathews, 2006; Smola et al., 2015). While pseudoknot prediction algorithms that accept SHAPE-constraints have been developed, there is no accepted field standard for evaluating pseudoknot

predictions, if they're even performed (Hajdin et al., 2013). Studies that do predict pseudoknots rely on a similar strategy. Generally, pseudoknot predictions are made in overlapping windows tiled across a given viral genome. By evaluating how many times a given pseudoknot appears in all of the prediction windows that cover it, plausible pseudoknots are identified. However, criteria with widely variable stringencies have been implemented in this sorting step, and the sorting has to be performed manually (Dethoff et al., 2018; Siegfried et al., 2014; Wan et al., 2022). Not only does this prevent comparison of pseudoknots predicted in different viruses, it also prevents comparison of pseudoknots predicted in different viruses. As a result, it makes assessing the validity of any one set of sorting criteria very difficult.

As such, the field of viral RNA structure prediction is in dire need of an automated, standardized pipeline to predict and evaluate pseudoknot predictions. The ideal framework would rely on sorting criteria calibrated on a diverse set of pseudoknots with known function. Whether it is developed as a standalone pipeline or implemented directly in the SuperFold framework, a standardized pseudoknot prediction pipeline would be of great utility. It would therefore be well-worth the requisite time and effort required to build.

4.2.2 Long-range structure probing and prediction

Long-range interactions are known to be functionally important in viral RNA genomes. *Flavivirus* genome cyclization, discussed at length in Chapter 2, is a well-studied example. However, owing to a combination of windowed structure

prediction and constraints imposed on max-pairing distance, long-range base-pairing interactions are hard to model. Because structure predictions of both SARS-CoV-2 and West Nile virus are prepared using this pipeline, we report no base-pairing interactions in either genome that exceeds 500 nucleotides.

Several methods have been developed that allow for identification of long-range RNA-RNA interactions in viral genomes. As mentioned above, correlated chemical probing was used to identify functional long-range interactions in the Dengue virus genome (Dethoff et al., 2018). Though this method requires slightly modified sequencing library construction, it otherwise requires no changes to existing SHAPE- and DMS-MaP workflows. Several other methods have been developed that directly detect long-range RNA-RNA interactions. Generally, these methods rely on probes that, upon UV activation, cross-link RNA duplexes. Cross-linked duplexes are then fragmented, purified, and sequenced to identify both duplex arms. (Lu et al., 2018; Ziv et al., 2018). More recently, a bi-functional probe was developed that, much like SHAPE probes, reacts selectively with flexible 2'-OH moieties. Owing to the presence of two electrophilic groups, this reagent is able to cross-link RNA duplexes that, like other strategies, are identified by generating and aligning chimeric reads in HTS datasets (Christy et al., 2021).

Datasets generated with these long-range probing methods have demonstrated utility in constraining molecular dynamics simulations of RNAs. Unfortunately, these simulations are not suitable for structure discovery in long viral genomes. In the context of secondary structure prediction, these datasets are also of limited utility; no prediction pipeline currently exists that accept this data as

constraints during prediction steps. Instead, long-range probing data is primarily used to validate secondary structure models. A recent study found that SHAPE-constrained structure predictions that did not conflict with long-range probing data were more accurate (Huber et al., 2019). As this sorting was done manually, however, extending this analysis to long viral genomes would be tedious to the point of intractability. As such, there is an unmet need for secondary structure prediction pipelines that accept long-range probing data as constraints. Generation of these pipelines would not only yield more accurate structure predictions, but also reveal fascinating new insights into higher-order genome organization in viruses.

4.2.3 Expanding the search for functional RNA structures

Currently, discovery of candidate RNA structures in whole viral genome structure predictions relies heavily on identifying regions with low SHAPE/Shannon Entropy (lowSS). This is in part owed to the assumption that functional RNAs should be both highly structured with well-determined predictions. However, an LNA targeted against a highSS region resulted in a profound WNV growth defect in C6/36 cells, suggesting that functional RNA sequence and structure exists outside the bounds of lowSS regions (**Fig 36**). As the work presented in this dissertation is focused entirely on regions with lowSS signatures, there remain large portions of both the SARS-CoV-2 and WNV genomes that need to be searched for functional RNA structures.

At a more basic reason, however, that we focus our search for functional RNA structures on lowSS regions. Specifically, lowSS regions are the only regions for

which discrete secondary structure predictions are appropriate. Put another way, static secondary structure predictions do not serve as accurate models of regions with high Shannon Entropy. Indeed, because a high Shannon Entropy indicates that multiple structures co-exist for the same sequence, these regions would be better modeled by conformational ensembles. As SuperFold only explicitly considers conformational ensembles during the partition function step, separate prediction pipelines are required to depict RNA conformational ensembles.

Luckily, secondary structure prediction algorithms have been made that use SHAPE reactivity information to model RNA folding ensembles (Spasic et al., 2018). Though these are not suitable for entire viral RNA genomes, they represent an attractive strategy for targeted inquiries of smaller regions. In this way, defining well-folded regions may have served the unintentional purpose of delineating natural bounds for structural studies of the intervening sequence. Conformational ensembles generated for these regions will provide a more robust depiction of their structural content, and could open up new avenues of inquiry. For example, it is possible that the highSS LNA described above targets a structure that dominates the region's conformational ensemble in C6/36 cells, but that is weakly populated in Vero cells. Not only would this provide an explanation for the cell type-specific effect observed, it would serve as yet another example of how evolutionary pressures have carefully tuned the conformational dynamics of viral RNA genomes. To that end, more detailed structural modeling of viral RNA genomes represents an exciting path forward for the discovery and study of viral RNA structure.

4.3 References

1. Christy, T.W., Giannetti, C.A., Houlihan, G., Smola, M.J., Rice, G.M., Wang, J., Dokholyan, N. V., Laederach, A., Holliger, P., and Weeks, K.M. (2021). Direct Mapping of Higher-Order RNA Interactions by SHAPE-JuMP. Biochemistry.
2. Corbett, K.S., Edwards, D.K., Leist, S.R., Abiona, O.M., Boyoglu-Barnum, S., Gillespie, R.A., Himansu, S., Schäfer, A., Ziwawo, C.T., DiPiazza, A.T., et al. (2020). SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. Nature *586*, 567–571.
3. Dethoff, E.A., Gokhale, N.S., Boerneke, M.A., Muhire, B.M., Martin, D.P., Sacco, M.T., McFadden, M.J., Weinstein, J.A., Messer, W.B., Horner, S.M., et al. (2018). Pervasive Tertiary Structures in the Dengue Virus RNA Genome Modulate Fitness. Submitted *115*, 11513–11518.
4. Faggioni, G., Pomponi, A., De Santis, R., Masuelli, L., Ciammaruconi, A., Monaco, F., Di Gennaro, A., Marzocchella, L., Sambri, V., Lelli, R., et al. (2012). West Nile alternative open reading frame (N-NS4B/WARF4) is produced in infected West Nile virus (WNV) cells and induces humoral response in WNV infected individuals. Virol. J. *9*, 1–14.
5. Göertz, G.P., Fros, J.J., Miesen, P., Vogels, C.B.F., van der Bent, M.L., Geertsema, C., Koenraadt, C.J.M., van Rij, R.P., van Oers, M.M., and Pijlman, G.P. (2016). Noncoding Subgenomic Flavivirus RNA Is Processed by the Mosquito RNA Interference Machinery and Determines West Nile virus Transmission by Culex pipiens Mosquitoes. J. Virol. *90*, 10145–10159.
6. 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.
7. 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. *110*, 5498–5503.
8. Homan, P.J., Favorov, O. V., Lavender, C.A., Kursun, O., Ge, X., Busan, S., Dokholyan, N. V., and Weeks, K.M. (2014). Single-molecule correlated chemical probing of RNA. Proc. Natl. Acad. Sci. U. S. A. *111*, 13858–13863.
9. Huber, R.G., Lim, X.N., Ng, W.C., Sim, A.Y.L., Poh, H.X., Shen, Y., Lim, S.Y., Sundstrom, K.B., Sun, X., Aw, J.G., et al. (2019). Structure mapping of dengue and Zika viruses reveals functional long-range interactions. Nat. Commun. *10*.
10. Jiang, X., Muthusamy, V., Fedorova, O., Kong, Y., Kim, D.J., Bosenberg, M., Pyle, A.M., and Iwasaki, A. (2019). Intratumoral delivery of RIG-I agonist SLR14 induces robust antitumor responses. J. Exp. Med. *216*, 2854–2868.
11. Kelly, J.A., Olson, A.N., Neupane, K., Munshi, S., Emeterio, J.S., 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. *295*, 10741–10748.
12. Kutchko, K.M., Madden, E.A., Morrison, C., Plante, K.S., Sanders, W., Vincent, H.A., Cruz Cisneros, M.C., Long, K.M., Moorman, N.J., Heise, M.T., et al. (2018). Structural divergence creates new functional features in alphavirus genomes. Nucleic Acids Res. *46*, 3657–3670.

13. 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.
14. Lu, Z., Gong, J., and Zhang, Q.C. (2018). PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution. *1649*, 59–84.
15. MacFadden, A., Òdonoghue, Z., Silva, P.A.G.C., Chapman, E.G., Olsthoorn, R.C., Sterken, M.G., Pijlman, G.P., Bredenbeek, P.J., and Kieft, J.S. (2018). Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. Nat. Commun. *9*, 1–11.
16. Madden, E.A., Plante, K.S., Morrison, C.R., Kutchko, K.M., Sanders, W., Long, K.M., Taft-Benz, S., Cruz Cisneros, M.C., White, A.M., Sarkar, S., et al. (2020). Using SHAPE-MaP To Model RNA Secondary Structure and Identify 3′UTR Variation in Chikungunya Virus. J. Virol. *94*.
17. Mathews, D.H. (2006). Revolutions in RNA Secondary Structure Prediction. J. Mol. Biol. *359*, 526–532.
18. Melian, E.B., Hall-mendelin, S., Du, F., Owens, N., Bosco-lauth, A.M., Nagasaki, T., Rudd, S., Brault, A.C., Bowen, R.A., Hall, R.A., et al. (2014). Programmed Ribosomal Frameshift Alters Expression of West Nile virus Genes and Facilitates Virus Replication in Birds and Mosquitoes. *10*.
19. Pijlman, G.P., Funk, A., Kondratieva, N., Leung, J., Torres, S., van der Aa, L., Liu, W.J., Palmenberg, A.C., Shi, P.Y., Hall, R.A., et al. (2008). A Highly Structured, Nuclease-Resistant, Noncoding RNA Produced by Flaviviruses Is Required for Pathogenicity. Cell Host Microbe *4*, 579–591.
20. Pirakitikulr, N., Kohlway, A., Lindenbach, B.D., Pyle, A.M., 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.
21. Ravindra, N.G., Alfajaro, M.M., Gasque, V., Huston, N.C., Wan, H., Szigeti-Buck, K., Yasumoto, Y., Greaney, A.M., Habet, V., Chow, R.D., et al. (2021). Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells , alterations in gene expression , and cell state changes. PLoS Biol. *19*, 1–24.
22. Rivas, E., Clements, J., and Eddy, S.R. (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. Bioinformatics *36*, 3072–3076.
23. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*.
24. Smola, M.J., Rice, G.M., Busan, S., Siegfried, N.A., and Weeks, K.M. (2015). 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.
25. 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.

26. 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.
27. 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.
28. Wan, H., Adams, R.L., Lindenbach, B.D., and Pyle, A.M. (2022). The In Vivo and In Vitro Architecture of the Hepatitis C Virus RNA Genome Uncovers Functional RNA Secondary and Tertiary Structures . J. Virol. *96*, 1–23.
29. Zhao, C., Liu, F., and Pyle, A.M. (2018). An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. RNA *24*, 183–185.
30. Ziv, O., Gabryelska, M.M., Lun, A.T.L., Gebert, L.F.R., Sheu-Gruttadauria, J., Meredith, L.W., Liu, Z.-Y., Kwok, C.K., Qin, C.-F., MacRae, I.J., et al. (2018). COMRADES determines in vivo RNA structures and interactions. Nat. Methods 1.