USING SHAPE-MaP TO IDENTIFY FUNCTIONAL RNA SECONDARY STRUCTURES IN RNA VIRUSES

Emily A. Madden

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Microbiology & Immunology in the School of Medicine.

Chapel Hill
2021

Approved by:

Mark Heise

Aravinda Desilva

Stanley Lemon

Alain Laederach

Nathaniel Moorman

ABSTRACT

Emily A. Madden: Using SHAPE-MaP to Identify Functional RNA Secondary Structures in RNA Viruses
(Under the direction of Mark Heise)

Alphaviruses are a genus of arboviruses often transmitted by mosquitos. Alphaviruses are responsible for multiple outbreaks over the last two decades and continue to pose a serious threat to human health. The 2013 outbreak of chikungunya virus (CHIKV), the most notable alphavirus, caused over one million infections. Despite the frequency with which these viruses re-emerge, there are no effective therapies or vaccines against alphaviruses. Like other RNA viruses, alphavirus genomes contain functionally important RNA secondary structures that contribute to immune evasion, RNA transcription, RNA translation, and virion assembly. However, very little of alphavirus genomes have been characterized due to a previous inability to accurately and quickly model long RNAs. This work used the RNA structure probing technique SHAPE-MaP to produce experimentally informed RNA secondary structure models of multiple RNA virus genomes. We probed genomes of closely related alphaviruses to identify conserved structured and unstructured regions. We found that alphaviruses are structurally unique and most conserved structured regions fold into distinct RNA secondary structures. After we identified a novel functionally important RNA secondary structure specific to Sindbis virus, we revised our approach to identify regions within each virus likely to fold into a specific conformation. We identified 23 regions of the CHIKV genome that were specifically structured. The four previously known RNA secondary structures were included in the 23 regions identified, validating the approach. Further, we demonstrated that one of the uncharacterized structured regions enhanced virus replication. Lastly, we demonstrated our approach to structure identification and testing was applicable to RNA viruses beyond the alphavirus genus using Zika virus, a flavivirus responsible for a large outbreak in 2015. For each of our studies we used silent structure disrupting mutations to assess RNA structure without affecting protein coding sequence, so structures could be assessed in the context of infection. These findings improve our understanding of known pathogenic RNA viruses and provide an approach to quickly study

and assess future emerging RNA viruses. A more comprehensive knowledge of functionally important RNA structures in viruses could be used to design safer live attenuated vaccines or develop new RNA-binding small molecule therapies.

ACKNOWLEDGEMENTS

I am an incredibly fortunate individual to not only complete this journey but have had the guidance and mentorship to begin it in 2015. My undergraduate advisors at Centre College, notably Drs. Peggy Richey and Kerry Paumi, introduced me to the field of virology and first fostered the necessary resiliency that research requires. After a cold email at the start of a summer, Dr. Binks Wattenberg graciously accepted me into his lab and then found funding to pay me so I could work full time. The Wattenberg lab was where I learned how quickly a workday could pass when it was spent alongside collaborative and supportive colleagues. I don't think I can ever thank Dr. Keith Jerome enough for his and his lab's part in my journey here. He provided an opportunity to work at the Hutch, but encouraged and advocated for my participation in undergraduate programming that helped me develop my application for graduate school. Without the guidance from him and his lab I likely would not have found my way to UNC. I am so fortunate to have had two amazing bench mentors at both these labs and will forever model my mentor style after Drs. Deanna Davis and Martine Aubert.

I owe a lot of gratitude to my dissertation advisor Dr. Mark Heise. Mark provided the structure I needed to feel comfortable working in a new field but quickly gave me independence when I was ready to run. Not only did Mark let me pursue science that interested me, but he also supported my desire to share it with scientists all over the world at conferences. Thank you, Mark, for trusting me to represent the lab but also the opportunity to experience the world outside of the US. Thank you, Dr. Sharon Taft-Benz, for not only making sure the lab ran smoothly and safely, but for the countless pick-me-ups (both food and wisdom) you shared with me and others. I am so very fortunate to have had the chance to work with you at the start and end of my time at UNC.

I could not have gotten through graduate school without the support and fellowship of my friends and peers. My cohort was an invaluable source of support over the last 6 years, but there were a few I leaned on more heavily than others. Thank you, Brea Hampton, Marta Cruz Cisneros, and Becca Casazza, for being amazing friends and colleagues. Thank you for supporting me when times were rough

and celebrating with me when times were good. You each serve as a role model for me and inspire me daily to be a better scientist and person.

Friends outside of graduate school are vital, so thank you Allison Eisner for keeping me connected with home and who I am outside of academia. Thank you for supporting me and loving me despite how consumed I was with completing this degree.

Thank you, James and Melinda Sears, for accepting me as family over the last six years. I am so grateful that you not only opened your home to me during a pandemic but created a wonderful space for me to write this dissertation.

I owe so much thanks to my family, who didn't always know what I was doing but supported me unconditionally anyways. Thank you for your patience and support even through missed holidays. I know no matter what I choose to do, you all will always love me and be proud of me.

Finally, I don't think I would have gotten through this without my partner, Jerod. It was not always a smooth ride to navigate graduate school in separate cities, but whenever I really needed you, you were there. Thank you for all the times you did my chores when I was in lab late. Thank you for juggling our weekend visits to fit around experiments. Thank you for making sure I took time to enjoy living in North Carolina and take a vacation. Thank you for reminding me that life is about more than the current degree I am chasing. Thank you.

PREFACE

The work described in Chapter 1 represents a collaborative effort between several contributors: Katrina M. Kutchko, Emily A. Madden, Clayton Morrison, Kenneth S. Plante, Wes Sanders, Heather A. Vincent, Marta C. Cruz Cisneros, Kristin M. Long, Nathaniel J. Moorman, Mark T. Heise, and Alain Laederach. Nathaniel J. Moorman, Mark T. Heise, and Alain Laederach are all corresponding authors on the original publication. At the time of publication, these authors were affiliated with the University of North Carolina at Chapel Hill. This study would not have been possible without the additional help of Naomi Forrester for her assistance with the alphavirus multiple sequence alignment, Elena Rivas for her helpful comments, Diane Griffin for kindly providing the SINV nsP3 antibody, and the Weeks lab for assistance with plotting RNAs and analyzing large RNAs.

The work described in Chapter 2 represents a collaborative effort between: Emily A. Madden, Kenneth S. Plante, Clayton R. Morrison, Katrina M. Kutchko, Wes Sanders, Kristin M. Long, Sharon Taft-Benz, Marta C. Cruz Cisneros, Ashlyn Morgan White, Sanjay Sarkar, Grace Reynolds, Heather A. Vincent, Alain Laederach, Nathaniel J. Moorman, Mark T. Heise, the corresponding author on the original publication. At the time of publication, these authors were affiliated with the University of North Carolina at Chapel Hill. This study would not have been possible without the additional help of Michael Diamond of Washington University and Robert Tesh of the World Reference Center for Emerging Viruses and Arboviruses at the University of Texas Medical Branch for kindly sharing the CHIKV Caribbean isolate with us.

The work described in Chapter 3 was a collaborative effort between Wes Sanders, Emily A. Madden, Kenneth H. Dinnon 3rd, Ethan J. Fritch, Katrina M. Kutchko, Heather A. Vincent, Clayton R. Morrison, Helen Lazear, Alain Laederach, Ralph S. Baric, Mark T. Heise, and Nathaniel J. Moorman, all of whom are affiliated with the University of North Carolina at Chapel Hill.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ΔG	Gibb's free energy change
1M7	1-methyl-7nitroisatoic anhydride
cDNA	complimentary DNA
CHIKV	chikungunya virus
CSE	conserved sequence element
cHP	capsid region hairpin
DB	dumbbell
DENV	dengue virus
DLP	downstream stem loop
DSH	downstream hairpin
DVG	defective virus genome
EEEV	Eastern equine encephalitis virus
FFA	foci forming assay
FFU	foci forming units
HI-FBS	heat inactivated fetal bovine serum
ISG	interferon stimulated gene
kb	kilobases
m^6A	*N6*-methyladenosine
m^7G	7-methylguanosine
MBFV	mosquito-borne flaviviruses
MFE	mimimum free energy
NMIA	*N*-methylisotoic anhydride
nsP	nonstructural protein
nt	nucleotides
ONNV	o'nyong nyong virus
PPV	positive predictive value
RING-MaP	RNA interaction groups by mutational profiling

RRV	Ross River virus
RT-qPCR	reverse-transcription quantitative polymerase chain reaction
SCFG	stochastic context free grammar
sfRNA	subgenomic flavivirus RNA
SFV	Semliki Forest virus
sgRNA	subgenomic RNA
SHAPE	selective 2′-hydroxyl acylation analyzed by primer extension
SHAPE-MaP	selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling
SINV	Sindbis virus
SL3	stem loop 3 (chikungunya)
SLA	stem loop A (flaviviruses)
SLB	stem loop B (flaviviruses)
ssRNA	single stranded RNA
TCR	termination codon readthrough
TLS	transfer RNA-like structure
TRD	Trinidad donkey strain
tRNA	transfer RNA
UTR	untranslated region
VEEV	Venezuelan equine encephalitis virus
WNV	West Nile virus
WT	wildtype
xrRNA	exoribonuclease resistant RNA
ZIKV	zika virus

INTRODUCTION

0.1 Introduction

Viruses are small obligate intracellular parasites composed of genetic material, either RNA or DNA, a protein capsid, and sometimes a lipid bilayer envelope. RNA viruses range in size from 3.5 to 41.5 kilobases (kb). The theoretical size limit for RNA viruses was previously believed to be around 33 kb prior to the discovery of a new nidovirus with a 41.5kb genome (1). Despite having reduced space to encode virus specific proteins, RNA viruses are very successful pathogens and particularly adept at replicating across multiple host species. It is estimated that up to 44% of all emerging infectious diseases are caused by RNA viruses (2). In fact, over the course of time it took to complete this dissertation, the world experienced the emergence or re-emergence of three RNA viruses: Ebola virus, Zika virus (ZIKV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (3). Just prior to the start of this dissertation, infections caused by a fourth re-emerging RNA virus, chikungunya virus (CHIKV), were in decline after its introduction to the Western hemisphere ignited an explosive outbreak (4). Despite the numerous outbreaks that have occurred since the start of the century, there are few licensed vaccines to prevent the spread of viruses or specific antivirals available to treat viral disease.

0.2 Alphavirus Biology

Alphaviruses, the only genus in the family *Togaviradae*, are enveloped positive polarity non-segmented, single stranded RNA (ssRNA) viruses (5). Alphaviruses can be categorized by what region of the globe they were historically endemic, which coincidentally roughly groups them by disease as well. Old World alphaviruses, such as chikungunya virus (CHIKV), Ross River virus (RRV), and O'nyong'nyong virus (ONNV), were historically endemic to Africa, Asia, Australia, and the Indian Ocean islands. Old World alphaviruses generally cause febrile illness accompanied by arthritogenic disease in vertebrate hosts (6-8). New World alphaviruses, such as Venezuelan equine encephalitis virus (VEEV) and Eastern equine encephalitis virus (EEEV), are endemic to North and South America. As their name implies, New World alphaviruses can cause encephalitic disease in addition to general febrile illness in vertebrate hosts

(Forrester 2017, Strauss 1994). For most alphaviruses, disease in the vertebrate host is acute, lasting a few days to weeks, before the infection is cleared. In rare cases the host can succumb to fatal disease (9-11).

Alphavirus genomes are between 11 and 12 kb in length. The RNA genomes are capped with 7-methylguanosine (m^7G) and polyadenylated to mimic a host mRNA with cap 0 structure (8). The first two-thirds of the genome encodes the nonstructural polyprotein responsible for replicating the virus genome. The nonstructural polyprotein is composed of four essential nonstructural proteins: nsP1-4. Each protein preforms multiple essential functions within the replicase complex. The complex is anchored to the plasma membrane through nsP1, which is also responsible for capping the genomic RNA. The polyprotein is processed into individual subunits by nsP2. nsP2 also serves as the helicase during RNA transcription. The exact function of nsP3 remains ambiguous, but it is essential for virus replication. Finally, nsP4 is the RNA-dependent RNA-polymerase (8, 10). The structural polyprotein is encoded in the last one-third of the genome and translated from a subgenomic RNA (sgRNA) (5, 8). The full length structural polyprotein is composed of five proteins, with a sixth protein product generated by a frameshifting event (8, 12). Aside from capsid which is co-translationally cleaved from the structural polyprotein, the rest of the structural polyprotein is processed by host proteases in the host secretory pathway (8).

Infection is initiated through attachment and binding of cellular receptors. Alphaviruses use a variety of attachment factors and receptors but heparan sulfates, MXRA8, and NRAMP2 are among some of the molecules that have been identified (13-16). After receptor binding, the virion is internalized through clatherin mediated endocytosis (17). As the endosome acidifies, the envelope proteins undergo conformational changes allowing fusion of the virion envelope with the endosomal membrane releasing the capsid and genomic RNA into the host cell cytosol. Since the virion genome mimics a host mRNA, the virus genome can be immediately translated by host ribosomal machinery. Once synthesized, the nonstructural polyprotein establishes virus replication complexes called spherules at the plasma membrane (8). In spherules, the replication complex transcribes negative-sense antigenomes to serve as a template for transcription of both positive sense progeny genomes and sgRNAs (8).

The subgenomic RNA is translated to produce the structural polyprotein. The structural polyprotein is processed into the individual components necessary for an assembled virion. Capsid self-cleaves from the structural polyprotein while the remaining components are processed, both cleaved and glycosylated, by host proteins as it is trafficked through the secretory pathway (8). The final steps of virus assembly occur at the plasma membrane, where new genomes are encapsidated and bud through the plasma membrane (8). The newly formed progeny virions are then capable of initiating the next round of infection of a new host cell.

Due to the small size of alphaviruses and the limited coding capacity, the virus proteins serve multiple functions during infection, from structural to immune modulatory roles. However, the proteins are not the only multifunctional component of an alphavirus. The alphavirus RNA genome serves multiple roles during infection maximizing the efficiency of each component of the virus. The ssRNA genome of alphaviruses contains signal sequences and folds into functional RNA secondary structures that are recognized or impact the function of virus and host proteins during infection. The known RNA secondary structures that impact alphavirus replication are discussed below.

5′ IFIT stem loop. Alphaviruses genomes lack a 2′ O-methyl group at the 5′ end of their genomes often seen in higher order eukaryotes and other cytoplasmic viruses (18). This should make alphaviruses susceptible to interferon stimulated gene (ISG) IFIT1 restriction. IFIT1 binds uncapped or cap 0 mRNAs and signals to the cell that those RNAs are non-self, initiating an innate immune response (19). However, it was observed that the vaccine strain of VEEV TC-83, and not the parental Trinidad donkey strain (TRD), was more susceptible to IFIT1 restriction (20). There were only two nucleotides changes between the strains that were responsible for attenuation, and a mutation in the first 100 nucleotides of the 5′ untranslated region (UTR) was known to increase VEEV sensitivity to type 1-IFN (21). Hyde et al. determined that this point mutation in the 5′ UTR created a bubble in a stem loop formed at the start of the VEEV genome (20). Destabilization of the first stem loop of the VEEV genome was responsible for increasing VEEV susceptibility to IFIT1 recognition. Authors used computational folding algorithms and NMR spectroscopy to predict and experimentally confirm the presence of this 5′ UTR stem loop (20).

The observation observed in VEEV was expanded and tested in related alphaviruses shortly after. The RNA secondary structure of EEEV, Sindbis virus (SINV), CHIKV, Semliki Forest virus (SFV),

and VEEV 5′ UTRs was computationally predicted and the same regions were cloned into chimeric reporter viruses to assess the sensitivity of each 5′ UTR structure to type-1 IFN treatment. Increasing concentrations of IFN typically reduced the expression of the reporter gene. However, all alphavirus 5′ UTR sequences tested were more resistant to type-1 IFN treatment than the negative control reporter with a β-globin 5′ UTR (22). However, the level with which they promoted translation after type-1 IFN treatment varied. Authors showed the level of CHIKV resistance to type-1 IFN treatment could be affected by mutating the predicted 5′ UTR stem loop. This suggests that the variability in alphavirus resistance to type-1 IFN treatment seen with the reporter constructs could be due in part to differences in 5′ UTR RNA secondary structure (22).

5′ conserved sequence element. The most conserved RNA element across the alphavirus genus is the 5′ conserved sequence element (CSE). This element was first detected by sequence conservation analysis and established to be a replication enhancer in SINV. Computer folding programs predicted the element formed two stem loops with asymmetric adenosine bulges (23, 24). Further investigation of this element and the surrounding region was done using an infectious clone of SINV with structure disrupting mutations (25). Mutant SINV virus was observed to accumulate second site mutations in nsP2 and nsP3 to compensate for the mutations in the 5′ end of the genome. The mutations also affected virus replication in mosquito cells more severely than in vertebrate cells. Taken together, these data suggest the 5′ CSE is likely recognized by pro-viral host specific factors, nsP2, and nsP3 (25). Studies using a double promoter reporter virus showed that the 5′ CSE was more important for VEEV replication and began to dissect the contribution of each stem loop in the element (26). Deletion of either stem loop in the 5′ CSE were not as attenuating as deletion of both stem loops, suggesting their function was redundant. In total, the 5′ CSE is hypothesized to help position the replication complex to recognize the plus- and minus-strand promoter during genomic RNA transcription (24, 26, 27).

Packaging signal. The packaging signal is another important RNA secondary structure that allows for selective packaging of the genome. For SINV and VEEV, the packaging signal is found within the coding region of nsP2. It was discovered using chimeric viruses and identifying regions of high synonymous site conservation within the nonstructural polyprotein coding region (28). Computer RNA folding programs indicated the packaging signal was likely a series of 7-8 hairpins with a triple G motif in

the loops (28). Studies using replicon and chimeras indicated that absence of this structured region severely attenuated production of infectious virions, but did not eliminate it, suggesting there are other mechanisms by which alphavirus genomes are packaged into capsid. Chimera studies also showed that the packaging signal of VEEV could be recognized and used by heterologous alphavirus capsid proteins. Interestingly, CHIKV capsid is capable of packaging replicon with SINV or VEEV packaging signals, but the SINV and VEEV capsid proteins do not efficiently package replicon containing the putative CHIKV packaging signal (28). Recent studies looking at capsid binding sites along alphavirus genomes fail to specifically recognize the reported SINV packaging signal or putative SFV packaging signal, which was believed to be similarly located to that of CHIKV (29, 30). Instead, it was shown that mutation of multiple capsid binding sites across SFV was sufficient to reduce genome packaging and virus production, suggesting a multi-site packaging model for SFV and closely related alphaviruses (30). While a synthetic packaging signal structure was used to demonstrate the necessity of multiple stem loops for efficient packaging of the VEEV genome (28), the secondary structure of the VEEV or other predicted alphavirus packaging sites had yet to be experimentally supported.

Termination codon readthrough element. Full translation of the alphavirus nonstructural polyprotein requires readthrough of an opal stop codon following the nsP3 coding sequence. It was previously believed the only contextual requirement for readthrough of the stop codon was a cytidine residue 3′ of the stop codon (31). However, in 2011 Firth et al. reported a conserved stem loop following the opal stop codon in alphaviruses discovered by analyzing the conservation of wobble position nucleotides across multiple alphaviruses (32). The sequence following the opal stop codon of multiple alphaviruses was folded using computer programs and suggested a similar base stem for each. The sequences were also compared to sequences of other organisms with stem loops known to modulate read through of leaky stop codons. These data suggested that alphaviruses may have a termination codon readthrough element (TCR) that regulated readthrough of the opal stop codon. Firth et al. used dual luciferase reporters to test readthrough efficiency of the predicted VEEV and SINV stem loops along with a number of mutant stem loops. They observed reduced read though of the opal stop codon when mutations were introduced that were predicted to disrupt the TCR. Mutations were primarily focused on the first 12 base pairs of the stem that were modeled, but Firth et al did test a VEEV TCR construct that

deleted the nucleotides between the two halves of the base. This construct had increased readthrough compared to that observed for WT sequence suggesting the base of the TCR was the most important component of the structure for proper stop codon readthrough, at least in vitro (32). The TCR element of CHIKV was recently modeled from an in vitro transcribed fragment of the genome using SHAPE chemical probing data (33). The SHAPE informed model generally agreed with that predicted by comparative sequence alignment and RNA folding programs. While studies have looked at the impact of mutating the opal stop codon preceding this element in multiple alphaviruses (33-36), the impact of disrupting the TCR structure has not been assessed in the context of a virus.

Downstream hairpin. Alphaviruses infection causes host cell translation shut off by preventing the dephosphorylation of eIF2α through PKR dependent and independent signaling. This happens early in infection, prior to translation of the sgRNA, yet the virus structural proteins are still synthesized (37, 38). In SINV, it was discovered that retaining the first 275 nucleotides of the capsid sequence enhanced sgRNA translation in reporter viruses during infection (39). RNA secondary structure modeling predicted a stable hairpin in this region. The presence of this RNA secondary structure and the importance of the structure, not the sequence, was demonstrated using mutations predicted to disrupt or reconstruct this hairpin in the context of both reporter sgRNAs and full-length virus (39-41). The predicted RNA structure in SINV was further supported by chemical probing and comparative sequence analysis of a aligned sequences (38, 41, 42). The downstream hairpin (DSH), or downstream stem loop (DLP), when properly positioned within the sgRNA, traps scanning 40S ribosomal subunit allowing full ribosome assembly and translation of the sgRNA in the absence of canonical host factors and independent of a canonical initiating AUG (40, 43, 44). This structure is notably absent in CHIKV, VEEV, and ONNV but sgRNA translation still occurs during host-translation shut-off suggesting alternative translation requirements for these viruses (41, 45).

Transframe stem loop. In 2008 the TransFrame protein (TF) was discovered. Prior to 2008, the doublet observed in Western blots for 6k had puzzled scientist. The larger band was actually a novel protein translated due to a ribosome frameshifting event at the end of the 6k open reading frame. The first 2/3 of the TF protein matches the N-terminus of 6k but the last 1/3 of TF is translated after a -1 slip of the ribosome at the slippery site resulting in a roughly 8 kilodalton protein (12). This element was discovered

by systematically searching for frameshifting motifs in RNA virus alignments. A conserved stem loop following the conserved slippery sequence was identified using RNA folding programs and manual sequence alignment inspection. The presence of the TF protein was confirmed through mass spectrometry, immunoprecipitation, immunohistochemistry, and pulse-chase experiments. The necessity of the predicted stem loop following the slippery site was demonstrated using dual luciferase reporters where WT virus sequence and mutant sequences were cloned between different luciferase open reading frames (46). The reporter assays were carried out with multiple alphavirus sequences and all except SFV were predicted to fold into a hairpin or pseudoknot structure to promote frameshifting (46). The TF stem loop of CHIKV has since been modeled using SHAPE chemical probing data (33). While the necessity of the TF protein has been shown for SFV infection, the overlapping reading frames makes assessing the importance of the stem loop 3′ of the slippery sequence difficult in the context of infection.

Repeated 3′ UTR Structures. Alphaviruses 3′ UTRs are composed of repeat sequence elements that are unique to each alphavirus (47). In CHIKV, there is a range of 3′ UTR lengths due to deletion or duplication of these repeat elements (48-50). The first predictions that the repeated sequences may be forming secondary structures important for replication was in SINV. Garcia-Moreno et al. predicted that the repeat sequence elements could fold into distinct repeated structural elements. They further predicted that sequences in the loops were complimentary to sequences in the 5′ UTR and brought the 5′ and 3′ UTRs together to enhance translation in the context of mosquito infection (51). However, there were no studies directly testing the importance of the RNA secondary structures themselves. In CHIKV, secondary structures formed by 3′ UTR repeat sequences and a terminal sequence have also been modeled using predictive software (52). While alignments of closely related CHIKV 3′UTRs supports the structure models predicted, no experiments were reported directly linking the secondary structures to observed mosquito replication enhancement (52). Other RNA viruses contain 3′ UTR structures that enhance or modulate virus infection in a host dependent manner (53, 54). However, further investigation is needed in alphaviruses to both confirm the presence of repeated RNA secondary structures and determine if they are functionally involved in the host dependent phenotypes associated with 3′ UTR variants.

While there are six defined functionally important RNA secondary structures and some putative 3′ UTR structures described for alphaviruses, not all of them are conserved across the genus. For the

alphaviruses that lack a specific functional structure, it is unknown if there is an alternative RNA element that serves a similar function or if this function is provided by a virus or host protein. Finally, while these structures have some experimental or sequence conservation evidence to support the proposed RNA secondary structure models, the secondary structure landscape of the rest of the alphavirus genome remains undefined.

0.3 Modeling and Finding Functional RNA Secondary Structures

The most popular way to predict RNA secondary structure of a single RNA sequence of interest is free energy minimization. Minimum free energy (MFE) predictions identify the structure with the largest Gibbs free energy change, ΔG, between the structured and linear state of the sequence. The ΔG is calculated at a given temperature and the structure with the lowest ΔG will likely be the most prevalent structure in a solution at equilibrium between the folded and unfolded state. Computer programs have been designed to calculate the MFE structure and are freely available on the internet (55).

A common way of estimating free energy change is using the nearest neighbor model (56). This model assumes the free energy change for a given structure is a sum of the structure's individual structural motifs (e.g., single stranded loops or stacked base pairs in a helix). The free energy change for each base pair in a motif like a helix is dependent on the identity of that pair and the identity of the neighboring base pair. Similarly, the free energy change for a loop motif is dependent on the loop identity and the identity of the bounding base pairs. The parameters for these models, like the given ΔG for a Watson-Crick pairing between a cytosine and guanine residue in a helix, were largely determined experimentally by optical melting studies (57).

A major assumption to MFE structures is that the nearest neighbor model is without error, which is untrue. A complete set of "rules" for folding RNA secondary structures is still being determined and as such all current modeling algorithms are limited. However, all MFE structures on average will have correctly and incorrectly predicted base pairs so it is useful to determine the probability of each predicted pair forming in reality. To do this, a partition function is often incorporated into MFE calculations. The partition function is a sum of all the equilibrium constants for all possible structures of a given sequence. This partition function can then be used to calculate the probability of a given base pair forming. Significantly, base pairs with high probability of forming using the partition function are also the base pairs

most likely to be accurately predicted (58). Partition function calculations are incredibly useful and often automatically incorporated in popular RNA secondary structure prediction packages (59).

MFE structure predictions are useful for generating hypotheses about the structure of a given RNA, particularly if that RNA is fewer than 800 nucleotides. A given RNA structure prediction is evaluated in two ways: the positive predictive value (PPV) and sensitivity. The PPV is a measure of correctly predicted base pairs out of the total number of base pairs predicted. The sensitivity is a measure of how many base pairs in the known structure were predicted. On average MFE predictions for sequences less than 800 nucleotides have an average sensitivity of 75% and a PPV of 66%, though this accuracy diminishes significantly with increasing length of the RNA beyond 800 nucleotides (58). These measures can be improved if additional information, like data from experimental mapping techniques, is provided to guide the prediction.

The earliest experiments to probe RNA secondary structure were those investigating the transfer RNA (tRNA) structure using double-strand specific RNases (60). This technique was expanded to include single-strand specific RNases so that these enzyme probing techniques could help accurately place nucleotides in base pairs or not (61). Spectroscopy data, like nuclear magnetic resonance (NMR) spectroscopy experiments, could also be used to experimentally identify helices in small RNAs. These experiments were accomplished by labeling RNA with ^{15}N, which created detectable chemical shifts unique to specific nucleotides. These unique shifts provided information on what type of pairing the nucleotide was involved in (62). Finally, chemical modification experiments are also used to probe structure. The first reagents used to probe structure reacted with the exposed Watson-Crick faces of nucleotides. The chemically modified nucleotides caused chain termination events when the RNA was reverse transcribed into complimentary DNA (cDNA). The nucleotide at the end of the cDNA fragments could then be inferred as unpaired. Multiple chemicals were needed to probe a single RNA to obtain information about each nucleotide species, because each chemical reacted with a specific nucleotide or class of nucleotides. These techniques were useful for validating and refining RNA secondary structure models developed using MFE algorithms, but much like the algorithms, their usefulness was and is still limited to short RNAs.

In 2005, a new class of chemical was used that reacted with the more exposed ribose backbone. These chemicals more uniformly interrogate RNAs and provided information on the flexibility of every nucleotide (63, 64). This new technique is named after the type of probing chemical reaction and subsequent analysis method: selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) (63). Early versions of this technique still relied on chain termination events to determine which nucleotides were reactive with the chemical and therefore flexible and unpaired. However, the technique was modified so that reactive nucleotides are identified using high-throughput sequencing. This new technique was coined selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling, or SHAPE-MaP (65). SHAPE-MaP has been used to generate data on RNA flexibility for full cell transcriptomes and single RNAs as long as 30 kb (66, 67).

SHAPE-MaP uses *N*-methylisotoic anhydride (NMIA) or 1-methyl-7nitroisatoic anhydride (1M7) to probe conformationally flexible nucleotides. NMIA and 1M7 react with the 2′-hydroxyl of the RNA ribose backbone. Conformationally flexible nucleotides react with the SHAPE reagents quickly and form adducts. These more reactive nucleotides tend to be single-stranded, though the relationship is not linear (68, 69). Manganese ion concentrations are adjusted so that the reverse transcriptase is capable of reading through the SHAPE adduct nucleotides, introducing mutations at reactive positions, instead of stalling and resulting in chain termination. The cDNA is then prepared into a library and subjected to high-throughput sequencing. Software developed to complement this chemical probing technique is used to count the number of mutations observed at each nucleotide, compare it to negative and denatured control groups, and generate a reactivity score for each nucleotide of the RNA. The reactivity score for a nucleotide indicates its relative flexibility, with more flexible nucleotides having high reactivity scores, and generally less flexible or more protected nucleotides having low reactivity scores.

The reactivity scores for a given RNA can be used for a variety of analysis depending on the context in which the RNA was probed (in vivo, refolded in vitro, etc.). Transcriptome wide SHAPE reactivities can be used to gauge relative structuredness or flexibility of the given transcriptome. SHAPE reactivities for multiple related RNAs can also be used to inform alignments and supplement signals of evolutionary conservation (70). For secondary structure modeling, SHAPE reactivity was first included as an additional pseudo-free energy term in the nearest neighbor model (57, 71). However, incorporation of

SHAPE data as a free energy term, a prior probability, or a likelihood of pairing all improve the prediction accuracies (both sensitivity and PPV) to greater than 90% (69, 70).

Improvements to computational RNA structure modeling and experimental techniques to validate and support those models have greatly improved are understanding of the functional importance of RNA in biology. However, while we can define the secondary structure of a known biologically important RNA and connect the structure to the RNA's functional role, there is still room for improvement. There are no defined characteristics or rules that set functionally important RNA secondary structures apart from other RNA sequence and structure. Therefore, identifying which RNAs and what RNAs may be functionally important remains challenging.

Sometimes functionally important RNA secondary structures are conserved across closely related RNAs. As such, predicting or identifying conserved RNA secondary structures has been approached in three ways: folding and aligning related RNAs at the same time (72, 73), folding then aligning related RNAs by structure (74), or aligning RNAs and then folding to a structure that is common to all or most of the sequences (75). Each method has its benefits and pitfalls. The most accurate method for identifying functional RNA secondary structures and curating an accurate RNA secondary structure prediction is manual comparative sequence analysis, as was done to solve the secondary structure of tRNA (76). This manual undertaking aligns a few related RNAs and identifies regions with coordinated mutations that would preserve base pairs, not nucleotide identities. These coordinated mutations, or signals of covariation can suggest preservation of a stem loop. This method is tedious to undertake by hand and requires skill to identify covarying nucleotides that may be quite distant from each other in a linear sequence. Recently, programs have been introduced to identify covarying nucleotides and assign statistical significance for covariation support (77). These programs have been improved to indicate if low statistical significance is a sign of lack of conserved structure or simply a lack of sequence diversity to detect a conserved structure (78). However, without closely related RNAs to employ these techniques, it is difficult to identify a novel functionally important RNA secondary structure from a single sequence model. Developing methods and schemes to identify the functional RNA structures from the large sets of predicted RNA structures is an active area of investigation.

0.4 Overview of Chapters

This dissertation focuses on using SHAPE-MaP technology to identify functionally important RNA secondary structures in arboviruses, with a focus on alphaviruses. *Chapter 1* explores the utility of SHAPE-MaP to identify conserved RNA secondary structures across the alphavirus genus. We hypothesized that functionally important RNA secondary structures would be conserved across the alphavirus genus. SINV and VEEV were analyzed using SHAPE-MaP and regions of the genome with highly correlative SHAPE reactivities were identified. We determined that SINV and VEEV are structurally divergent, with most of the genomes correlating not much better than would be expected at random. Regions with the highest correlation coefficients between SINV and VEEV do not look structurally similar aside from the 5′ CSE and the TF stem loop. Using a structure disrupting mutation strategy, we confirm that RNA secondary structure of the SINV packaging signal and the 5′ CSE are important for virus replication and extend this strategy to two uncharacterized regions of the SINV genome. This study identified a novel functional RNA secondary structure in SINV located at the end of nsP1 with moderate conservation. The nsP1 structured region is likely important for regulating early stages of virus replication. These data led us to conclude that identification of functional RNA secondary structures in alphaviruses must be virus strain specific.

Chapter 2 focuses on identifying functional RNA secondary structures of CHIKV specifically. We revised our previous hypothesis and posited that functional RNA secondary structures would be highly structured and likely to adopt a single, specific conformation. We used SHAPE-MaP to inform the secondary structure model of the CHIKV genome using a 2013 Caribbean outbreak isolate of CHIKV. We then identified 23 regions of the CHIKV genome with very low SHAPE reactivity that are predicted to form a single specific structure, satisfying our criteria for potentially functional RNA secondary structures. Of these 23 specifically structured regions, we identified the four previously known functionally important RNA secondary structures in CHIKV, validating our new approach. Furthermore, we show that the third stem loop of the genome, SL3, is functionally important independent of the 5′ CSE. We also show that 3′ UTR duplication and deletion events that arose when CHIKV was introduced to the Western hemisphere resulted in duplication and deletion of specific RNA secondary structures as well. These 3′ UTR variations

impact virus replication in mosquito cells, but have no effect on virus replication or pathogenesis in mammals.

In *Chapter 3*, we extend our hypothesis that functionally important virus RNA secondary structures have low SHAPE and are specifically structured to Zika virus (ZIKV), another ssRNA arbovirus transmitted by mosquitoes belonging to the *Flaviviridae* family. We generated a SHAPE-MaP informed RNA secondary structure model of the ZIKV genomic RNA and identified regions with low SHAPE that likely fold into a single specific conformation. We identified 19 regions of the ZIKV genome that met our criteria, and again, included in those regions were the previously known functionally important RNA secondary structures. We confirmed the importance of RNA structures in the ZIKV 5′ UTR previously shown to be essential in dengue virus (DENV), a related flavivirus. We also interrogated a novel structure located within the E protein coding region for importance during virus replication but found none. We show that disruption of RNA structures in the 3′ UTR of ZIKV do not impact virus replication in vitro but do reduce morbidity and mortality in a mouse model of ZIKV pathogenesis.

Alphaviruses and other ssRNA viruses remain a major public health threat, as demonstrated by the emergence and re-emergence of four in just the past decade; CHIKV in 2013, ZIKV in 2015, Ebola intermittently, and SARS-CoV-2 in 2019 (3). There remain few to no specific antivirals to treat emerging viruses and there are only a few approved vaccines to prevent infection and disease by a select number of re-emerging viruses. Disruption of RNA secondary structures have been shown to attenuate virus replication and pathogenesis for multiple RNA viruses (YFV, ZIKV, VEEV) yet we do not have a full understanding of the RNA secondary structures for most of these viruses' genomes. Identification of functional RNA secondary structures remains a challenge for the field in general, but is particularly challenging in small RNA viruses where traditional tools, like covariation analysis, cannot be efficiently applied. Further investigation of virus RNA secondary structures is needed to identify novel targets of RNA-binding small molecule drugs or contribute to design of live-attenuated vaccines.

CHAPTER 1: STRUCTURAL DIVERGENCE CREATES NEW FUNCTIONAL FEATURES IN ALPHAVIRUS GENOMES[1]

1.1 Overview

Alphaviruses are mosquito-borne pathogens that cause human diseases ranging from debilitating arthritis to lethal encephalitis. Studies with Sindbis virus (SINV), which causes fever, rash, and arthralgia in humans, and Venezuelan equine encephalitis virus (VEEV), which causes encephalitis, have identified RNA structural elements that play key roles in replication and pathogenesis. However, a complete genomic structural profile has not been established for these viruses. We used the structural probing technique SHAPE-MaP to identify structured elements within the SINV and VEEV genomes. Our SHAPE-directed structural models recapitulate known RNA structures, while also identifying novel structural elements, including a new functional element in the nsP1 region of SINV whose disruption causes a defect in infectivity. Although RNA structural elements are important for multiple aspects of alphavirus biology, we found the majority of RNA structures were not conserved between SINV and VEEV. Our data suggest that alphavirus RNA genomes are highly divergent structurally despite similar genomic architecture and sequence conservation; still, RNA structural elements are critical to the viral life cycle. These findings reframe traditional assumptions about RNA structure and evolution: rather than structures being conserved, alphaviruses frequently evolve new structures that may shape interactions with host immune systems or co-evolve with viral proteins.

1.2 Introduction

Alphaviruses are mosquito-borne positive sense RNA viruses that infect a wide range of vertebrate and invertebrate hosts. Sindbis virus (SINV) is the prototype virus of the alphavirus genus and generally infects avian species (8, 79). However, SINV can spill over into humans, where it causes symptoms such as fever, rash, myalgia, and arthralgia (80). In contrast, Venezuelan equine encephalitis virus (VEEV), which is normally spread in an enzootic cycle be- tween rodents and mosquito vectors, can

[1] First published in *NAR* 2018 46(7):3657-3670. https://doi.org/10.1093/nar/gky012

emerge in an epizootic form leading to large-scale epidemics associated with high mortality in equine species, and symptoms ranging from flu-like symptoms to severe encephalitis in humans (81). As such, these viruses can survive and replicate in a variety of vertebrate hosts and arthropod vectors, and both protein sequences and the structure of the RNA genome itself are important for the virus life cycle. After alphavirus entry, the positive sense genome acts as a messenger RNA, leading to the translation of the viral nonstructural proteins, which mediate viral RNA synthesis and are encoded by the first two thirds of the viral genome (8). The positive sense genome is then transcribed to pro- duce the viral negative strand RNA, which in turn serves as a template for the synthesis of both the positive sense viral genome, as well as a shorter 26S RNA encompassing the last third of the genome that encodes the viral structural proteins (8). Viral genomes contain a large amount of information in a limited amount of space; consequently, alphavirus RNAs contain important regulatory structures in addition to the protein-coding sequence. These structures occur both in non-coding portions of the genome, such as the 5′ and 3′ UTRs (20, 24, 82), and in coding regions of the genome, such as the 51-nt conserved sequence element (5′ CSE) in nsP1, the packaging signal, and a frameshift signal in the structural polyprotein (25, 26, 28, 83).

Despite the importance of these RNA structures to the virus life cycle, structures from one alphavirus are not necessarily compatible with other alphaviruses (24, 28, 37). This phenomenon raises that possibility that differences in viral RNA structure may contribute to the variation in host range, cell tropism, or disease pathogenesis between alphaviruses. However, to this point, the level of RNA structural conservation among alphaviruses has not been systematically characterized and in particular not at the whole genome level, with the most extensive analysis of structural conservation focused on the 5′ UTR and the 51-nt 5′ CSE (24-26).

Historically, covariation in a multiple sequence alignment of related RNAs served as the ‘gold standard’ method of identifying conserved (and, by extension, functional) RNA secondary structures (84-88). Base pairs that covary, or evolve together to preserve base pairing but not sequence identity, reveal the conservation of secondary structure within an RNA element. Covariation is most evident for RNAs having strong and highly conserved structures, such as transfer RNAs and ribosomal RNAs (85, 88, 89). For less-conserved RNAs, including those that are specific to multicellular organisms, the structural covariation signal is much weaker or non-existent (90, 91). Many cellular RNAs including long non-coding

RNAs do not exhibit any covariation (77). This finding raises the question of whether non-conserved RNA structural elements are functional, or whether their functions are derived solely from their sequence and not their structure.

In order to understand whether novel RNA elements have a functional role in the alphavirus life cycle, and to determine whether RNA structural elements are conserved within the alphavirus family, we used SHAPE-MaP (selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling) to determine the structural profile of the SINV and VEEV genomes. SHAPE-MaP has previously been used to identify structures in the HIV and hepatitis C virus genomes (92-94). Here, we used SHAPE-MaP to obtain high quality, high-resolution structural data for the complete SINV and VEEV genomic RNAs. This analysis found that aside from a few previously defined functional structures in SINV—in particular, the 5′ conserved sequence element and the frameshift element in the 6K coding region (12, 23, 24, 83)—there seems to be little overall conservation of structured elements throughout the alphavirus family. Instead, the SINV and VEEV genomes contain a large number of highly structured regions that are unique to each virus and not shared between different alphavirus family members. Further analysis of one of these novel structures found that it affects SINV replication, indicating that these novel RNA structures are likely to play an important role in the viral lifecycle. These results suggest that alphaviruses utilize mutational space to evolve novel RNA structural elements specific to their individual biology, rather than preserving structures throughout virus evolution. Consequently, lack of conservation of structure does not indicate that a specific conformation lacks a function; instead, lack of structural conservation suggests that the role of RNA structure is highly context dependent.

1.3 Results

Structure conservation and divergence identified by high-resolution SHAPE profiling

RNA structural elements play essential roles in many aspects of the alphavirus lifecycle, including regulation of viral RNA synthesis, viral translation, and evasion of the host innate immune system. However, most of this analysis has focused on a fairly limited number of RNA structures, such as the 51-nt conserved sequence element (5′ CSE) and the 414-nt packaging signal located in the nsP1 coding sequence (23, 28). To locate additional functional structures within the SINV genome, we used SHAPE-MaP (92) to obtain a high- resolution structural profile for the entire SINV genome (Figure 1.1 A). We

performed these experiments on refolded viral RNA in the absence of viral and cellular proteins. Highly structured regions that are likely to fold into a single, unique conformation have below-average median SHAPE reactivities (92, 95).

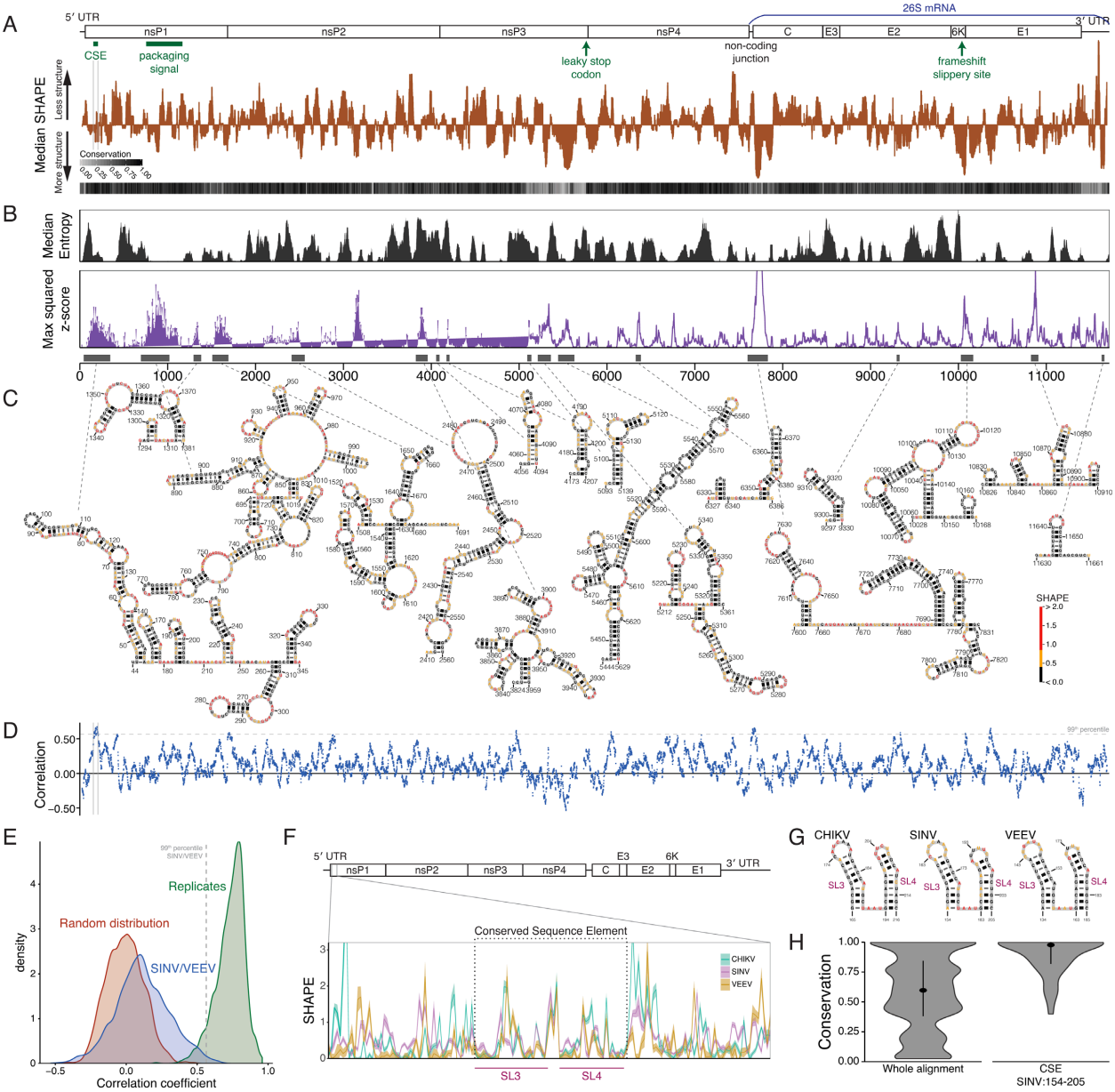


Figure 1.1: The Sindbis virus genome contains a multitude of diverse RNA structures.
(A) Top: schematic of the virus genome organization, with annotated elements. Middle: SHAPE data for the Sindbis virus genome, represented by the local median (55-nt window) compared with the global median. Reactivities below the x-axis indicate a region more structured than average. Gray lines denote the conserved sequence element (5′ CSE), which has low SHAPE reactivities and is highly structured. Bottom: sequence conservation at each position, based on sequence identity and gappiness, from a multiple sequence alignment of 37 alphaviruses. The protein-coding sequence contains both well-conserved (black and dark gray) and less-conserved (light gray) regions. (B) Top: median (55-nt window) Shannon entropies of base pairing across the SINV genome. Middle: Maximum squared *z*-score at each

position in the genome, with higher values corresponding to greater structural significance. Bottom: structured regions in the SINV genome, based on the intersection of regions with low SHAPE and low *z*-scores. (C) SHAPE-directed structural models of SINV structured regions. Nucleotide color indicates low, medium, or high SHAPE reactivity. (D) Windowed correlation coefficients of SHAPE data between the SINV and VEEV genomes. The dashed line indicates the top 1% of correlation coefficients. SHAPE data within the 5′ CSE are among the most correlated within the genome, indicating high structural conservation within that region. (E) Distribution of windowed correlation coefficients of SHAPE data. Red: a background distribution, blue: correlation coefficients between SINV and VEEV, green: correlation coefficients of two biological replicates of a virus. Although SINV and VEEV are more correlated than expected at random, there is little overlap with the correlations of the same virus, indicating little widespread correlation. Dashed line indicates top 1% of SHAPE correlations between SINV and VEEV. (F) SHAPE data of the 5′ CSE in CHIKV, SINV, and VEEV. Within the 5′ CSE, the SHAPE profiles are very similar, representing conservation of structure, but the correlation immediately disappears outside of the 5′ CSE. (G) SHAPE-directed structural models of the CSE in CHIKV, SINV and VEEV. The 5′ CSE structure is compatible with the SHAPE data and conserved in all three viruses. (H) Distribution of alignment-derived sequence conservation scores in the entire alignment (left) and 5′ CSE only (right). Dot indicates the median, with the line extending from the 25th to 75th percentile.

We also determined the sequence conservation score at each position using a multiple sequence alignment containing 37 alphaviruses representing the entire alphavirus phylogeny (7). The sequence conservation score uses sequence identity and gappiness to calculate the conservation at each position (96) (Equation 1). Although most of the protein- coding portion of the genome is highly conserved, highly divergent regions occur in both protein-coding and non- coding sections (Figure 1.1 A; light gray at the 5′ end, end of nsP3, non-coding junction, and 3′ end).

We also used the intersection of sequence features and SHAPE reactivities to find regions of highly stable structure within the SINV genome, as well as average structural entropy across the genome (Figure 1.1 B). We identified 17 ‘structured regions’ with low median SHAPE reactivities and high structural significance based on the z-score using RNAsurface (97), which compares the free energy for a region to what would be expected if the sequence were shuffled at random (98). For each of these structured regions, we used SHAPE reactivities to guide secondary structure prediction to derive a structural model (71, 99) (Figure 1.1 C). Our method successfully recapitulates known or previously predicted structured regions, including regions overlapping the 5′ CSE, the packaging signal, the non-coding junction, and the frameshift signal (83, 100), confirming the utility of the SHAPE-MaP method for accurately predicting RNA secondary structures within viral RNA genomes. In addition to identifying several previously characterized areas of RNA secondary structure, SHAPE-MaP analysis also identified several regions of the genome that contain previously uncharacterized stable RNA structures (Figure 1.1 C). Therefore, novel areas of RNA secondary structure are broadly distributed throughout the SINV

genome, which raises the possibility that these structures play as yet undefined roles in the alphavirus lifecycle.

Previous work has found that structures such as the 5′ CSE and the RNA packaging signal are conserved be- tween two or more alphavirus family members. Given the large number of stable RNA structures present in the SINV genome, we wanted to determine whether any of these novel structured regions were conserved between alphaviruses. A subset of RNA structural elements, such as the 5′ CSE and the RNA packaging signal, are known to have structural conservation between SINV and Venezuelan equine encephalitis virus (VEEV) (25, 28). Therefore, we also per- formed SHAPE-MaP on the ZPC738 strain of VEEV (Figure A.1 A).

To identify stable structures within the VEEV genome, we applied the same analysis used for SINV (Figure A.1 B). Similar to SINV, this analysis identified several highly structured regions within the VEEV genome that overlap with the 5′ CSE, packaging signal, and frameshift signal, as well as several regions with high structural stability that have not previously been characterized in VEEV (Figure A.1 C). Therefore, similar to SINV, this data suggests that the VEEV genome contains previously uncharacterized RNA structural motifs that may play an important role in VEEV replication and pathogenesis, while also providing us with an opportunity to directly assess whether stable structures are conserved between two different alphaviruses.

We used the SINV and VEEV SHAPE data to look at the correlation of SHAPE reactivities between the related genomes to assess conservation of RNA structure (101) (Figure 1.1 D). This analysis identified nine regions in the genome with correlation coefficients that surpassed the 99th per- centile, including the region covering the 51 nucleotide (nt) 5′ CSE. How- ever, when compared with correlations between biological replicates, the correlation distribution between SINV and VEEV SHAPE data overlaps minimally. The distribution is more similar, but not identical, to a random distribution, representing a limited amount of structural conservation (Figure 1.1 E). Therefore, while a small number of regions are correlated in their SHAPE signal, which could indicate structural conservation, the vast majority of highly structured regions in both the SINV and VEEV genome are unique and not shared between the two viruses. We also compared the structures of SINV and VEEV that overlap these highest-correlated regions in the 99th percentile (Figure 1.2 A). While the patterns of SHAPE data are similar between SINV

and VEEV in these regions, that effect is the result of similar but not conserved base-pairing patterns; only the 5′ CSE and the frameshift region adopt similar structures between the viruses (Figure A.2 B). Therefore, even within regions of relatively high SHAPE correlation, structures generally diverge between SINV and VEEV.

Despite the general lack of structure conservation between SINV and VEEV, one of the most correlated regions is the CSE, which shows sequence conservation across multiple alphavirus family members (25). We therefore compared the SHAPE reactivities in that region between SINV, VEEV, and chikungunya virus (CHIKV) (Figure 1.1 F). The 5′ CSE is highly correlated within the two conserved stem loops 3 and 4 (25), but the SHAPE reactivities are not similar outside of the conserved region. The similarity in SHAPE derives from the identical structures of stem loops 3 and 4 (Figure 1.1 G). In addition, sequence conservation within the 5′ CSE is remarkably high, especially when compared with the alphavirus family as a whole, based on sequence conservation scores using the multiple sequence alignment (Figure 1.1 H). Outside the 5′ CSE, however, slight divergence in sequence results in little conservation of structure. Thus, only very local structural motifs like the two hairpins in the 5′ CSE are structurally conserved.

RNA structure plays a critical role in SINV replication

Given the highly structured nature of both the SINV and VEEV RNA genomes, we set out to test the functional impact of a subset of these structures. To determine whether an RNA structure is functional, it is necessary to disrupt the structure without changing other aspects of the sequence, such as the encoded amino acid sequence. We used the program *CodonShuffle* to create mutant RNA sequences that preserve the amino acid sequence (102). The algorithm we used shuffles sets of trinucleotides (not in reading frame) in which the first and third bases remain identical, and the second base only changes when it would not affect the protein sequence (Figure 1.2 A). This method also preserves sequence composition and dinucleotide frequency. For most RNA sequences, this method generates hundreds of possible sequence mutants, so we chose mutation strategies designed to maximize disruption of the structural model by changing base pairing in a particular region with only very minor changes in codon usage (Table A.1). The frequencies of each codon in the mutant viruses remain nearly identical to WT, differing in usage by one or two in- stances at most for each codon. In addition,

although the effects of small changes in codon usage have not been studied in alphaviruses, slight changes in codon usage have not been shown to affect viral growth in polioviruses (103, 104).

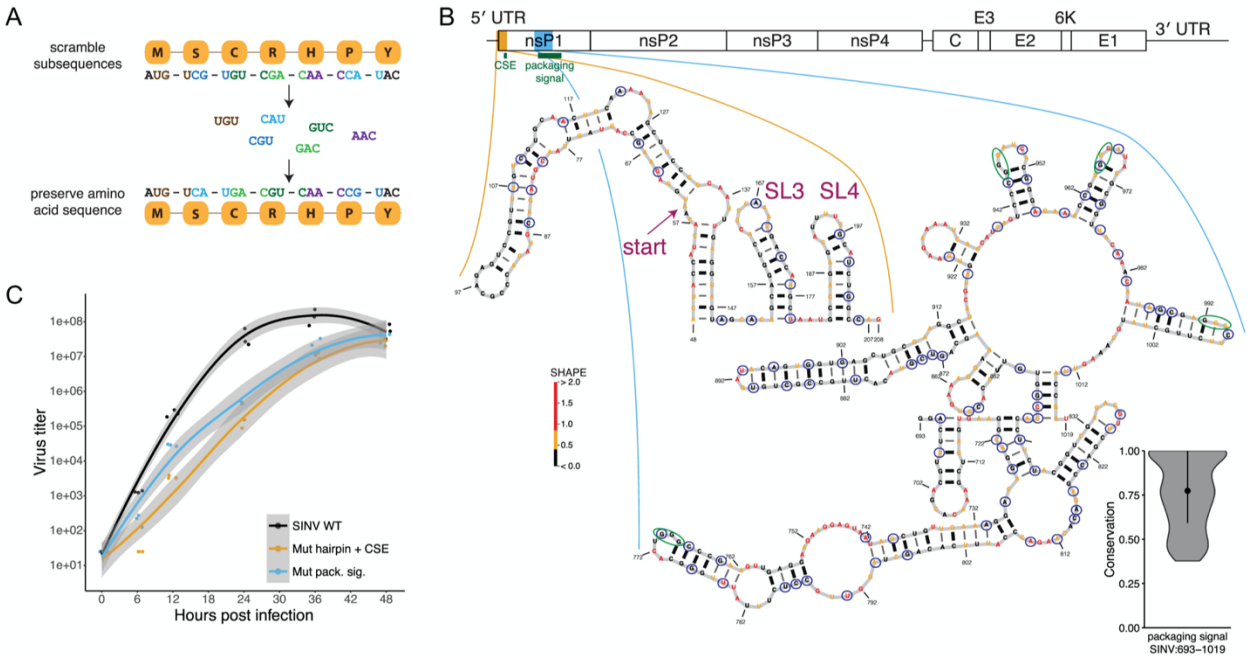


Figure 1.2: Structure-disrupting mutations successfully confirm function by impeding virus growth.
(A) Method used to disrupt RNA structure. Trinucleotide sets are shuffled, changing the nucleotide sequence while the amino acid sequence is preserved. (B) Left: structure of the hairpin + CSE element. Start codon and stem loops 3 and 4 are indicated. Right: structure of region overlapping packaging signal. Blue circles indicate positions that are mutated to disrupt the structure. Green circles indicate previously observed GGG motifs within packaging signal structure (9). Nucleotide color represents SHAPE reactivity. Violin plot displays conservation scores within the packaging signal region. (C) Growth curves for SINV WT (black), mutated hairpin + CSE (gold), and packaging signal (blue) in Vero81 cells at a MOI of 0.01. Shading indicates standard error. Both structures are necessary for optimal virus growth.

To validate our structure-disrupting method, we mutated two known SINV RNA structures, the 5′ CSE and the packaging signal (Figure 1.2 B). To disrupt the 5′ CSE, we created mutations both within the element and within the long hair- pin immediately 5′ of the element, creating twenty mutations throughout the region. Prior studies have suggested that the 5′ CSE structure has a mild impact on growth in mammalian cells but necessary in mosquito cells (25). Consistent with these prior results, we found that disruption of the 5′ CSE with the 5′ hairpin resulted in decreased viral growth in C6/36 mosquito cells compared to the wildtype virus (Figure A.2). We also found that disruption of the 5′ CSE and the 5′ hairpin resulted in a significant growth defect in mammalian cells compared to the wildtype virus (Figure 1.2 C). Therefore, these results suggest that the 5′ hairpin and 5′ CSE broadly impact viral replication in both mammalian and mosquito cells.

The viral RNA packaging signal, which falls within a region spanning nucleotides 613 to 1019, is less sequence con- served and longer than the 5′ CSE. Therefore, we introduced a total of sixty-nine mutations into the packaging signal that were designed to maximize disruption of the RNA secondary structure, while avoiding impact on the protein sequence (Figure 1.2 B). Consistent with prior results, disrupting the RNA packaging signal resulted in a significant de- crease in virus yield (Figure 1.2 C), thereby validating both the SHAPE-MaP-derived structural predictions, while demonstrating our ability to directly test the functional importance of stable RNA secondary structures within the SINV genome.

Novel RNA structures in the SINV genome

Next, we applied our method to two structured regions with low SHAPE reactivity: one downstream of the pack- aging signal in nsP1, and one extremely low-SHAPE region in the non-conserved domain of nsP3 (Figure 1.3 A). The nsP1 structured region (nsP1 SR) is conserved, but less so than the packaging signal. The novel nsP3 structured region (nsP3 SR) has little sequence conservation, and is located 133 nucleotides upstream of the leaky stop codon. We disrupted the nsP1 SR and nsP3 SR with 6 and 36 point mutations, respectively.

As with the previously tested regions, we infected Vero cells, a mammalian cell line, with the structural mutants (Figure 3B, left). In Vero cells, the mutants grew at the same rate as the wildtype virus. We also infected NIH/3T3 cells which, unlike Vero cells, have a competent interferon system (105) (Figure 1.3 B, right). Neither structural mutant had a change in phenotype in NIH/3T3 cells, indicating that these structures are not necessary for viral growth in mammalian cells. Additionally, however, we measured the specific infectivity of in vitro transcribed genomic RNA of the packaging signal, nsP1 SR and nsP3 SR mutants by electroporating the RNA into BHK-21 cells and plating serial dilutions of cells over a Vero cell monolayer (Figure 1.3 C, Figure A.3 A), an assay which measures the gross ability of naked viral RNA to produce infectious virus and detects early defects in viral fitness. We used equal amounts of RNA for the specific infectivity assay (Figure A.3 B). The packaging signal and nsP3 SR mutants had the same specific infectivity as wildtype virus, whereas the specific infectivity of nsP1 SR was reduced by three to four orders of magnitude, indicating that nsP1 SR is critical for the virus. The absence of phenotypic change between the nsP3 SR mutant, which spans almost 200 nt of coding

sequence, and WT virus confirms that merely changing the RNA sequence alone is not enough to disrupt the virus life cycle.

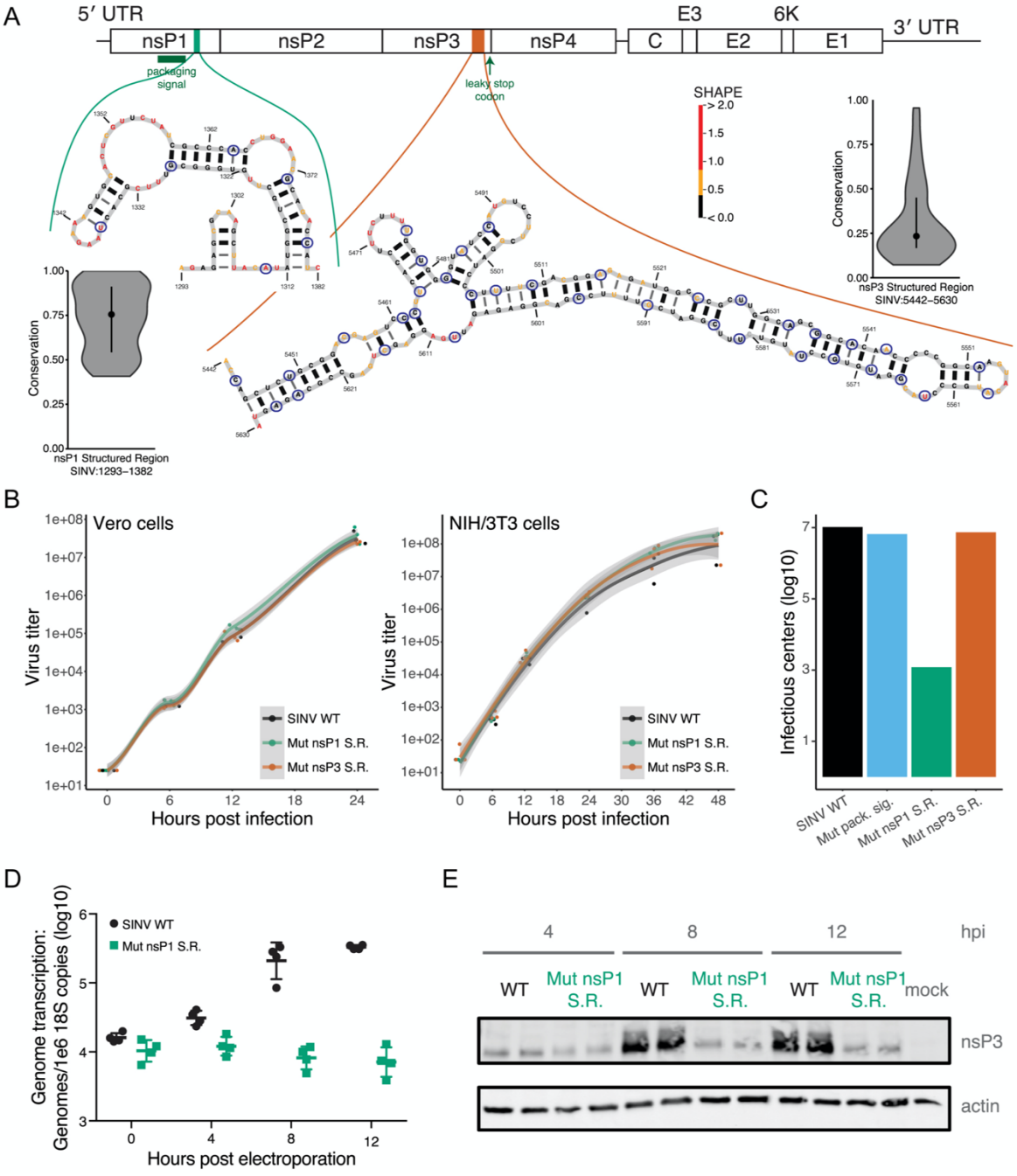


Figure 1.3: Novel virus structures tune SINV growth.
(A) Structures for the new nsP1 structured region (nsP1 SR; left) and the new nsP3 structured region (nsP3 SR; right). Nucleotide color represents SHAPE reactivity. Violin plots display sequence conservation scores within each region. Blue circles indicate positions that are mutated. (B) Growth curves for SINV WT (black), the nsP1 SR (green), and the nsP3 SR (red). Mutant growth is nearly identical to WT in both Vero cells (left) and NIH/3T3 cells (right). (C) Specific infectivity of mutant viruses. The nsP1 SR mutant has a large defect in infectivity. Graph is a representative experiment of three or

more replicates. (D) Genome transcription levels of WT and nsP1 SR mutants, measured by qRT-PCR. The nsP1 SR mutant has a defect for genome transcription. (E) RNA translation of WT and nsP1 SR. Expression of the nonstructural proteins is impaired in nsP1 SR compared to WT as measured by probing for nsP3.

Curiously, mutated nsP1 SR overcame its infectivity defect in the viral growth assays. Sequencing of the rescued virus did not reveal any compensatory mutations to restore structure or otherwise regain function.

We considered the possibility that disruption of nsP1 SR results in destabilization of the viral RNA, so we assessed transcription and stability of the input RNA after electroporation (Figure A.3 C). We saw no differences in the stability of nsP1 SR mutant RNA compared to WT RNA. This finding, along with the nsP1 SR mutant's ability to recover from its initial defect, suggests that the infectivity defect of the nsP1 SR mutant is caused by disruptions of events associated with genome replication early during infection. We found that levels of viral RNA synthesis (Figure 1.3 D) and nonstructural protein expression (Figure 1.3 E) were reduced over time compared to the WT virus. These results could be due to either defects in early non- structural protein synthesis or transcription of the genome itself. Overall, our findings suggest that nsP1 SR plays an important role in regulating early stages of the viral replication process.

Functional RNA structures are not conserved

Having confirmed that at least two known RNA structures and one novel RNA structure in the SINV genome are functional, we wanted to assess their level of conservation in related alphaviruses. For each structure, we compared the SINV model to the sequences in the thirty-six related alphaviruses (Figure 1.4 A). The structure compatibility score represents the fraction of structural model base pairs that can form at homologous positions in each virus (Equation 2).

The 5′ CSE has a high structural compatibility score in nearly every other alphavirus, indicating that the two 5′ CSE stem–loops are conserved throughout the alphavirus family. The packaging signal has less structural conservation com- pared with the 5′ CSE, with strong conservation only within a small branch of the alphavirus phylogeny. The nsP1 SR follows a similar pattern, but it is even less conserved than the packaging signal. The nsP3 SR, in a non-conserved region of the SINV genome, has no structure conservation outside of immediate relatives. These results indicate that functional structures are not necessarily conserved, and, in fact, they are in this case unique to an individual virus.

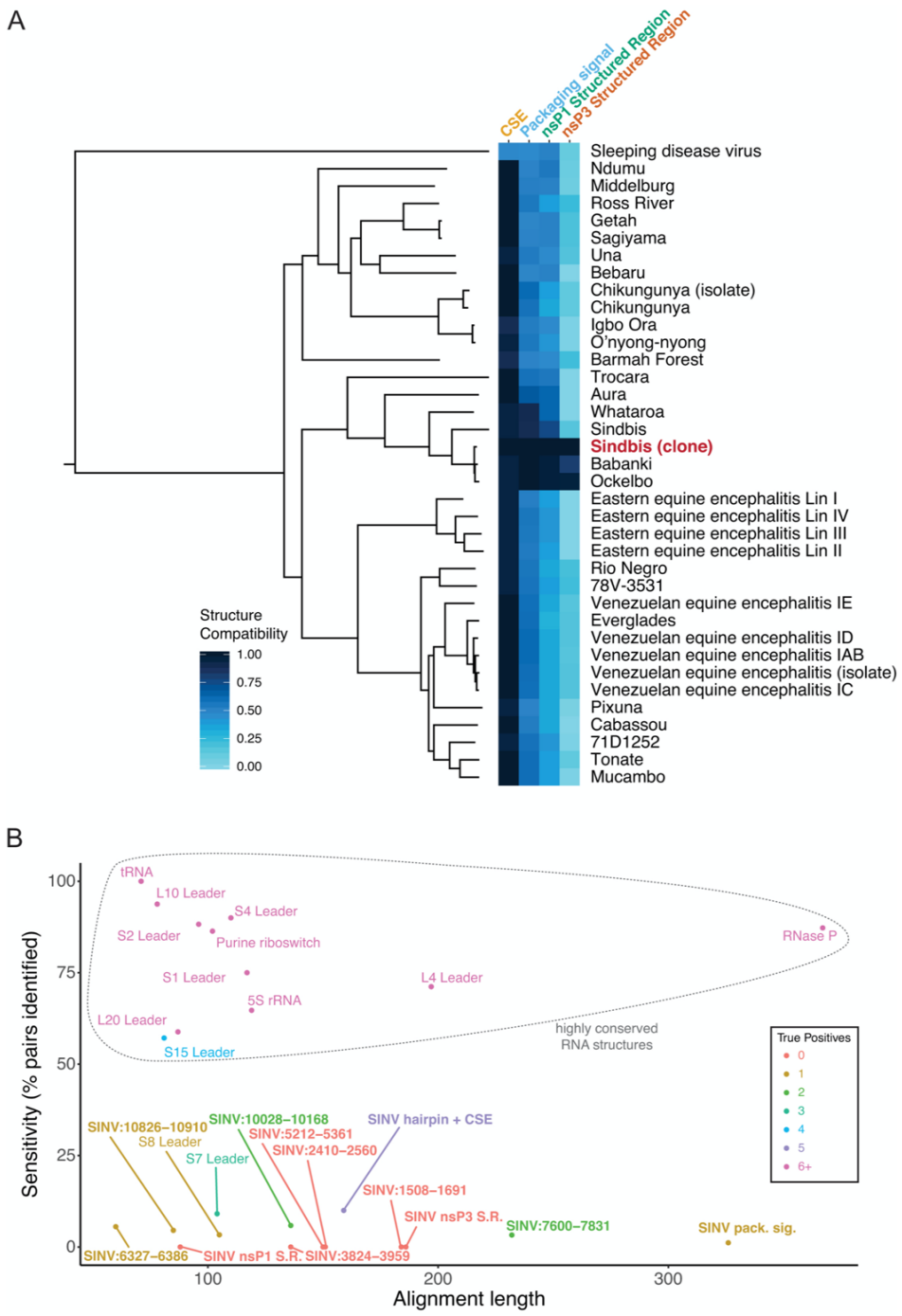


Figure 1.4: Outside of the conserved sequence element, SINV functional structures are not conserved.

(A) Structure compatibility scores for the four tested sequences, with the phylogenetic tree on the left. The SINV strain used for the reference sequence and structure is bolded and in red. The structural model for the 5′ CSE is highly compatible with other alphavirus sequences, but the other functional structures are less conserved. For the new region in nsP3, the structure essentially does not exist outside of closely related strains. (B) R-scape results for structure-informed alignments. X-axis: alignment length; y-axis: percentage of base pairs in structure found by R-scape; color: number of true positives found by R-scape. Bold typeface indicates R-scape results from SINV structure-informed alignments, whereas regular typeface indicates R-scape results for alignments from (77). While no SINV region has anywhere near the

amount of covariation of highly-conserved RNA structures, the region containing the 5′ CSE is notable for the highest number of covarying base pairs within SINV.

We also considered the possibility that functional structures may have shifted locations in other alphaviruses and would therefore not appear to be conserved based on the multiple sequence alignment. For each of the structured regions we identified (Figure 1.1 C), we constructed a stochastic context free grammar (SCFG) model from the predicted SHAPE-informed secondary structure and corresponding sequence (86, 106). We used this model to search through the related alphavirus sequences and refined our alignments in an attempt to build a covariance models for the SINV structures (107, 108). Only three regions, including the hairpin + 5′ CSE and the packaging signal, exist in almost all related alphaviruses (Table A.2).

We quantified structural covariation of these structure- informed alignments with the new program R-scape (77), which identifies base pairs having significant covariation. Because R-scape applies a significance threshold to an alignment, it filters base pairs that may appear to be con- served but do not covary more than expected by chance. We calculated the percentage of base pairs in each covariance model that R-scape found to be significant, both in the structure-informed alignments and in known conserved RNA structures (Figure 1.4 B; Table A.3). Although the structure-informed alignment of the long hairpin + 5′ CSE in nsP1 contains the highest number of significantly covarying base pairs among SINV regions, the sensitivity of these covariation models is well below classic structured RNAs such as riboswitches, tRNA, and Rnase P (77) (Figure 1.4 B; Table A.3). Importantly, the nsP1 SR, whose disruption dramatically decreases specific infectivity (Figure 1.3 C) has no covarying base-pairs.

The only other regions in SINV where R-scape found more than one significantly covarying base pair are the regions overlapping the 26S promoter (SINV:7600–7831) and the region containing the frameshift element (SINV:10028– 10168). The two covarying base pairs in the frameshift element are contiguous and part of a stem loop found in the New World clade of alphaviruses, including VEEV (83) (Figure A.5). Elements similar to this hair- pin exist in 30 out of 37 alphavirus sequences in our alignment, so although it is highly conserved, it lacks the complete structural conservation of the 51-nt 5′ CSE (Table A.2).

Consequently, although there is slight covariation in SINV within the 5′ CSE, most other regions including functional RNA structural elements have little to no covariation. The lack of structural covariation

indicates that sequence, not structure, is the primary driver of similarity in our covariance models. Despite RNA structural elements being important for the growth of SINV, there is no covariation evidence to indicate that in general these structures are conserved among alphaviruses. Thus, these results suggest that alphaviruses evolve idiosyncratic, functional RNA structures specific to their individual biology. Furthermore, these results demonstrate the difficulty in using covariation as a signal to identify functional motifs in these viruses.

1.4 Discussion

Through a combination of sequence analysis and experimental probing data, we generated whole-genome structural models for the previously uncharacterized SINV and VEEV RNA genomes. Using these models, we identified previously known and novel structures in each genome, and we found that both non-coding and coding regions of the genome contain highly structured RNA elements. We applied a systematic mutational method to disrupt RNA structures while preserving amino acid sequence, nucleotide composition, and dinucleotide frequencies. With this method, we confirmed that disrupting two known functional structures—the 5′ CSE and the packaging signal—decreases virus growth. Also, we identified a new functional RNA element in nsP1 whose disruption greatly diminishes viral RNA specific infectivity. The mutant viruses have distinct phenotypes: 5′ SL/5′ CSE and packaging signal mutants have a sharp decrease in growth in Vero cells (and the former also has greatly decreased growth in C6/36 cells), mutated nsP3 SR has no change in phenotype compared with wildtype virus, and mutated nsP1 SR has drastically impaired specific infectivity. These phenotypic differences indicate different mechanisms by which RNA structure regulates the infectivity and growth of SINV.

The data presented in Figure 1.2 recapitulates known aspects of SINV biology, that the 5′ CSE and packaging signal are critical to viral replication (23, 25, 28). These data are nonetheless important as they serve as a positive control for our automatic codon usage optimized shuffling strategy (Figure 1.2 A), suggesting that we can disrupt known RNA structures using thus approach. Furthermore, both the 5′ CSE and packaging signal have low-median SHAPE (Figure 1.1 A), confirming that our structural data is predictive of likely important RNA structures in the coding region of the virus.

We also examined the conservation of these structured regions among related alphaviruses and found that most structured regions, aside from the 5′ CSE, are highly divergent. Despite high sequence

conservation within most coding regions, there is little evidence of structural conservation. Instead, alphaviruses seem to quickly discard existing structures and evolve new ones, likely a result of their own particular environmental requirements. These viruses must survive in at least two organisms, the arthropod vector and the vertebrate host, and among the alphavirus family there is great diversity in which organisms these viruses infect (8, 80, 109). The diversity of these viruses is underscored by the discovery of Eilat virus, an alphavirus that cannot survive in vertebrates (110). The environmental diversity of these viruses is mirrored in the diversity of their RNA structures: common elements but individual uniqueness.

Structured regions in the SINV genome

RNA elements around the 5′ end of the SINV genome are critical for virus growth (23-27). We specifically investigated the structured region at the beginning of nsP1 including both the 51-nt 5′ CSE and its preceding 5′ hairpin. It is important to note that the two most highly conserved hairpins we report here in the 5′ CSE are 3′ of the start codon, indicating that specific structures can be con- served in a coding region. In contrast with (but not in contradiction to) previous research that found the 5′ CSE to be functional in mosquito cells but not in vertebrate cells (25), when we disrupted the 5′ CSE and hairpin, the virus grew poorly in both vertebrate and arthropod cells (Figure 1.2 C, Figure A.4). Although we do not yet know the mechanism by which these structures function, we conclude that the combination of the 5′ hairpin and the 5′ CSE is important for the SINV life cycle in both vector and vertebrate host.

We also confirmed that disrupting the packaging signal, which is also located in a coding region of the virus, disturbs the virus growth cycle in Vero cells, confirming the importance of the structure. Our SHAPE-directed structural model for the packaging signal region found repeated GGG-motifs in stem-loops, as previously suggested (28) (Figure 1.2 B). Because disrupting these stem-loops interferes with growth, the RNA structure throughout this region is critical for viral proteins to recognize genomic RNA. Although we have made every effort to minimize the impact of our coding mutations on codon optimality (Table A.1), we cannot exclude that our coding mutations may also have an effect on translation by altering codon usage.

Other structured regions include the non-coding junction, which overlaps with the subgenomic promoter, the frameshift element, and a highly structured hairpin ~100 nucleotides upstream of the leaky stop codon. Although that hairpin was not found to be functional in Vero cells, another study suggested

that structure formation downstream of the leaky stop codon plays a role in stop codon read-through (32). It is possible that structural elements, in conjunction with virus or host proteins, regulate the read- through frequency and translation of nsP4.

The frameshift element in the 6K coding region is particularly interesting because its two covarying base pairs are next to each other, indicating that the hairpin is conserved. This hairpin also exists in equine encephalitis viruses, but even among those related equine encephalitis viruses the structure diverges outside of that hairpin (83). Here, we find the hairpin also exists in SINV. Although this hairpin in the frameshift element is conserved in most alphaviruses that we examined, homologs were not found in every alphavirus, again indicating some degree of structural divergence within the virus family.

Lack of structural conservation among alphaviruses

Covariation, in which base pairs evolve together, is a useful indicator to identify conserved RNA structural elements (77, 86, 89, 111). We used the new program R-scape (77) to quantify the number of significantly covarying base pairs as a metric for conservation of structure. The region with the hairpin and 5′ CSE had the most conservation, with five covarying base pairs. Compared with highly structured, highly conserved RNAs, however, this number of covarying base pairs is quite limited (77, 86, 89).

In Figure 1.4 B and Table A.3, we report the percentage of pairs identified by R-scape, or sensitivity of the covariation model, for the SINV structured regions we identified experimentally, compared to classic conserved RNA structures such as tRNA and the L4 leader (77). Sensitivity, as measured by the percentage of pairs identified (Figure 1.4 B), enables us to evaluate the covariation and conservation support for RNA structures in SINV. It is striking that not a single region within the alphavirus genomes, including the highly conserved and functional 5′ CSE, has as much covariation as known conserved elements. The R- scape analysis used to compute this sensitivity is tailored to detect very high levels of conservation, mostly in non- coding RNA, such as those observed in tRNAs and ribosomal RNA (77). We are applying it here to coding regions of the genome, where protein coding sequence is likely the main driver of sequence conservation. As such, the result that all of our structures have far lower sensitivity than known structured RNAs is not necessarily surprising. How- ever, it is essential to note that none of the structures here are supported by covariation evidence, yet we were still able to identify one novel structure with a clear viral phenotype (Figure 1.3).

Experimental structure probing combined with detailed functional characterizations is likely necessary for identifying novel structured regions in alphaviruses, and it remains to be seen whether this is true for other single stranded RNA viruses. Fortunately, recent technological advances leveraging next-generation sequencing to obtain SHAPE and other forms of chemical and enzymatic probing data will facilitate this approach (92, 112-116). Furthermore, these data sets will enable further development of hybrid sequence/experiment approaches for reconciling conservation and experimental data.

Comparing the SHAPE profiles between related sequences is a useful, model-free approach to finding similarities in RNA structure (70, 90, 101). In this instance, the divergence in SHAPE data between SINV and VEEV supports the conclusion that, outside of highly conserved elements, these viruses are mostly structurally divergent (Figure 1.1 D and E, Figure A.1). We also measured the structural compatibility of related alphavirus sequences with the SHAPE-derived structures for the 5′ CSE, packaging signal, and the nsP1 and nsP3 structured regions (Figure 1.4 A). It is evident from this analysis that only the 5′ CSE has broad structural conservation across the family. What these data suggest is that specific structural elements are generally not conserved; nonetheless overall patterns of structure vs. unstructured, as evidenced by median SHAPE fluctuations, appear more conserved. These findings are consistent with the idea that in many cases, specific structures are not as important to function as the presence, or absence, of RNA structure in a particular region.

New considerations for RNA structure and evolution

It is often dogmatically suggested that functional structural elements are conserved in related organisms, with the converse being that non-conserved elements are not functional (77). However, we found a new functional, albeit non-conserved, SINV RNA element and a large amount of structural divergence within the SINV packaging signal. Consequently, traditional methods for identifying structure in certain RNAs do not apply adequately to alphaviruses, and may also be problematic with other RNA viruses. Viruses are highly divergent structurally, yet they preserve particular elements such as a single hairpin. Therefore, to adequately study RNA structure in the context of RNA viruses, new computational methods are necessary to integrate high-throughput experimental techniques such as SHAPE-MaP, and to allow for flexibility of structure out- side of the most conserved elements.

1.5 Methods

SHAPE data collection. Sindbis virus (Girdwood strain; accession #MF459683) and VEEV (ZPC783 strain; accession #MF459684) virions were concentrated by ultracentrifugation over a 20% sucrose cushion. Concentrated virions were lysed with TRIzol (Ambion) and full-length genomic RNA was purified following the manufacturer's protocol.

Modified RNA was obtained by incubation of 2µgof total RNA at 37°C for 15 min in the presence of 10 mM MgCl2 and 111 mM KCl, then treated with 100 nM of 1- methyl-7-nitroisatoicanhydride (1M7) for 5 min at 37°C. Negative control RNA was obtained by incubation of 2µgtotal RNA at 37°C for 15 min, then incubated with 5µl DMSO for 5 min at 37°C. Denatured control RNA was obtained by incubation of 2µgtotal RNA at 95°C for 2 min, then treated with 100 nM 1M7 for 2 min at 95°C. Following treatment, RNA was purified using illustra MicroSpin G-50 columns (GE Healthcare). Total purified RNAs were then incubated with 500 ng Random Primer 9 (NEB) at 65°C for 5 min, cooled to 0°C, and mixed with 10 mM dNTPs, 0.1 M DTT, 500 mM Tris pH 8.0, 750 KCl and 500 mM $MnCl_2$. The mix was then incubated at 42°C for 2 min, followed by the addition of 200 units of SuperScript II (Thermo Fisher Scientific) and a further incubation at 42°C for 180 min, heat inactivated at 70°C for 15 min, and then purified using illustra MicroSpin G-50 columns. Double stranded cDNA was created by NEBNext mRNA Second Strand Synthesis Module (NEB) using the manufacturer's protocol. The double stranded cDNA was then fragmented, tagged, amplified and barcoded using Nextera XT DNA Library Preparation Kit (Illumina) following the manufacturer's directions. Libraries were cleaned with Agencourt AMPure XP beads (BeckmanCoulter) at a DNA to bead ratio of 0.6:1, library size was determined by 2100 Bioanalyzer (Agilent Technologies) and quantified with a Qubit Flourometer using Qubit dsDNA HS Assay Kit (ThermoFisher Scientific). Sequencing was performed on a MiSeq Desktop Sequencer (Illumina).

Sequencing reads from the background SHAPE-MaP condition were used to assemble the correct virus sequences. SHAPE reactivities for the CHIKV, SINV and VEEV viruses were derived using the ShapeMapper pipeline (92). Because the background mutation rate for VEEV was higher than the CHIKV background mutation rate, SHAPE reactivities for VEEV were re-calculated using a scaled background mutation rate and 2%-8% normalization (117). The SHAPE data for SINV and VEEV are available as supplementary data in SNRNASM format online at

https://docs.google.com/spreadsheets/d/1BHORlaXPbC0npQK-zy4YEE938qzZOrsp9L6FmouWN3U/edit?usp=sharing (118). Windowed SHAPE values were calculated by finding the median SHAPE reactivity over a rolling 55-nt window and comparing those values to the global median SHAPE (70, 92, 101).

Multiple sequence alignment. The alignment was built on the conserved protein-coding sequence alignment from (7). The full nucleotide sequence for each virus in the alignment was downloaded from GenBank. Only viruses with complete genome sequences were included in the final alignment, and the assembled sequences for SINV, CHIKV and VEEV were added to the alignment. Non-conserved portions of the genome (5′ UTR, 3′ UTR, C-terminus of nsP3, and non-coding junction and N-terminus of the capsid sequence) were aligned with MAFFT (119, 120) (v7.221) and manually refined. The non-conserved and non-coding alignments were concatenated to create the final multiple sequence alignment. The phylogenetic tree was created using PhyML (121) (v3.0) with default parameters and midpoint-rooted.

Sequence conservation. The sequence conservation score *C(x)*, ranging from 0 to 1, at each alignment position *x* was computed using the following equation, adapted from (96):

$$C(x) = (1 - t(x))^{0.5} \cdot (1 - g(x)0^1 \qquad (1)$$

where *g(x)* is the frequency of gaps at position *x*, and *t(x)* is the Shannon entropy (97) at position *x*. The sequences were weighted using the algorithm in (122), and the weights were incorporated as in (96).

Correlations in SHAPE data. To compare the correlation between two sets of SHAPE data, thereby enabling comparisons across the entire genome while avoiding distortions by outliers, all gaps and missing and negative values were set to zero before calculating the Pearson correlation coefficient over a 55-nt rolling window. For the correlation between SINV and VEEV, the SHAPE reactivities were aligned according to the multiple sequence alignment (with all-gap positions removed). To generate the background distribution, the SINV and VEEV reactivities from the previous analysis were each scrambled prior to the rolling correlation coefficient calculation. The SHAPE correlation distribution for biological replicates was generated using the SHAPE data for the first 11,400 nt of the SHAPE-MaP data for two biological replicates of CHIKV. Regions in SINV and VEEV with correlation coefficients in the top 99th

percentile were expanded 27 nt on each side to incorporate all nucleotides within the window, for a total of nine highly correlated regions.

Identification of structured regions. RNASurface (123) (v1.0) was used to find regions of significant structure in the SINV genome, with a minimum z- score threshold of –2.5. Overlapping regions were merged, leading to 20 distinct predicted structured regions. In 17 of those regions, the majority of positions had below average windowed median SHAPE reactivities. These 17 regions are the final set of structured regions in the SINV genome, supported by both prediction and experimental data.

Structure modeling. Minimum free energy models for each structured region were generated using RNAstructure's Fold program (124) (v5.8.1), incorporating SHAPE reactivities as a pseudo-free energy term (71), with the maximum base pairing distance set at 200 nt and standard parameters (intercept = −0.6 kcal/mol, slope = 1.8 kcal/mol, temperature = 310.15 K) otherwise.

To model the whole genome structure, the Superfold program was used with SHAPE reactivities incorporated as a pseudo-free energy term (92), with a maximum base pairing distance of 500 nt and standard parameters otherwise. This whole-genome structural model was used to obtain the Shannon entropies at each position from the base pairing probabilities (125, 126). Structures within regions with highly correlated SHAPE data were extracted from this whole- genome structural model, with long-range base pairs removed.

Creation of mutant clones. To generate mutations for each region while keeping the amino acid sequence unchanged, the program *CodonShuffle* (102) was used with the dn231 algorithm, which scrambles sets of trinucleotides while ensuring that the first and third bases of each trinucleotide set are preserved. This method also preserves sequence composition and dinucleotide frequency. For each region, 1000 shuffled sequences were randomly generated, in most cases representing hundreds of unique sequences.

Out of these shuffled sequences, mutant sequences were selected to maximize structural disruption while also avoiding large changes in codon usage. Because the virus must survive in multiple hosts, organism-specific measures such as the codon adaptive index (127) are not useful to quantify change in codon usage. Instead, codon usage change was calculated using the sum of square differences of codon frequencies within the virus transcript. Structural disruption was determined by

calculating the percentage of base pairs in the SHAPE-directed structural model that could no longer form Watson–Crick or wobble base pairs in the mutant sequence, as well as confirming that the predicted structure of the mutant was not similar to the structural model for wildtype.

Structure mutants were designed from the Girdwood S.A. cDNA clone (pg100) of SINV and created by Gibson assembly (New England BioLabs). Fragments of the Girdwood genome containing structure disrupting mutations were purchased from Integrated DNA Technologies (IDT, Iowa, USA) with ~23 bp of overhang beyond restriction endonuclease cut sites. Clones were confirmed by Sanger sequencing through UNC sequencing core.

Infectious RNA was transcribed from the cDNA clones after linearization by NotI using mMESSAGE mMACHINE SP6 Transcription Kit (Invitrogen). RNA was introduced to BHK-21 cells by electroporation. Supernatants were collected 24–48 h after electroporation based on observed cytopathic effects and aliquoted into single use aliquots stored at –80°C. Virus titer was quantified by plaque assay on Vero81 cells.

Cell culture. BHK-21 cells were maintained in 1x αMEM (Gibco) supplemented with 10% heat inactivated fetal bovine serum (FBS) (Sigma) and 0.2 mM L-glutamine (Gibco). Vero81 cells were maintained in 1x DMEM (Gibco) supplemented with 10% FBS and 0.2 mM L-glutamine. NIH-3T3 cells were maintained in 1x DMEM supplemented with 10% bovine calf serum (Colorado Serum Co., Denver, USA). Mosquito C6/36 cells were maintained in 1x Leibovitz L-15 (Corning/Cellgro) supplemented with 10% FBS, 10% tryptose phosphate broth (Sigma) and 0.2 mM L-glutamine.

Viral growth and plaque assays. Multistep growth curves were conducted by infecting cells at a multiplicity of infection (MOI) equal to 0.01 in biological triplicate. Sample of cell culture supernatants were taken at indicated times after infection and stored at – 80°C. Viral titer was quantified by plaque assay. During plaque assays, Vero81 monolayers were infected with virus samples titrated in 1x PBS (Gibco) with 1% FBS and Ca2+/Mg2+ and overlaid with 1× αMEM with 10% FBS, 0.2 mM L-glutamine, 1 mM HEPES (Corning), 1% penicillin streptomycin (Gibco) and 1.25% carboxymethylcellulose sodium (CMC) (Sigma). Virus was allowed to plaque for 40 h before cells were fixed with 4% paraformaldehyde (PFA) (Sigma), washed, and stained with 0.25% crystal violet (Fisher Chemical).

Specific infectivity assays. Wildtype and mutant RNA were electroporated into BHK- 21 cells in parallel. An aliquot of electroporated cells were titrated and plated overtop subconfluent Vero81 cell monolayer. BHK-21 cells were allowed to attach for 1.5 h, at which point the monolayers were overlaid with CMC overlay detailed previously and incubated for 40 h. After incubation, cells were fixed, washed and stained as detailed for plaque assays.

RNA stability and transcription assays. Full length genomic RNA from the wildtype and mutant viruses was produced using the SP6 DNA dependent RNA polymerase (Ambion) as described above. Wildtype and mutant RNA were electroporated into BHK-21 cells in biological duplicate or quadruplicate as described above, at which point the cells were washed 1x with media to remove remaining extracellular RNAs. To test genomic RNA stability, half the cells were treated with cyclohexamide to prevent translation of the viral nonstructural proteins, which mediate viral RNA synthesis, thereby preventing replication of virus RNA. Cells used to analyze viral genome transcription kinetics were not treated with cyclohexamide. For both experiments, cells were plated and RNA harvested at the indicated times. Monolayers were washed 1x with PBS before cells were lysed in TRIzol reagent. Total RNA was purified following manufacturer's protocol, and treated with DNAse (Promega) to remove any input plasmid from the initial transcription reaction. Virus genome and 18S copy number was quantified by qRT-PCR using iTaq Universal Probes One-Step Kit (Bio-Rad) alongside a standard for absolute quantitation. SINV primer-probe sets were used as described previously, and 18S primer probe was purchased ThermoFisher (Catalog #4331182) (128). We also performed PCR amplification on samples without reverse transcriptase to test for carryover plasmid contamination within the electroporated RNA stocks. After quantitation of genome and 18S copies, SINV genomes were normalized to 1 x 10^6 18S copies and log transformed.

Western blots. Wildtype and mutant RNA were electroporated into BHK- 21 cells in biological duplicate as described above. Cells were washed 1x with media to remove remaining extracellular RNAs and plated. Cell lysates were harvested at indicated time points with RIPA buffer containing protease inhibitor at indicated times. Lysates were prepared by incubation on ice for 30 min followed by centrifugation at 12,000 rpm for 15 min. Total protein was quantified by BCA Pierce assay (ThermoFisher) using a BSA standard curve. Equal amounts of protein were boiled in SDS loading buffer for 5 min.

Samples were run on a 4–15% gradient Mini- PROTEAN TGX Precast Protein gel (BioRad) and transferred to PVDF membrane (Li-Cor) with the BioRad Wet Transfer system. Membranes were blocked in 1x TBS–0.1% tween-20 (TBST) and 5% milk overnight at 4°C. Primary antibodies were diluted in TBST + 5% milk (1:2000 rabbit polyclonal anti-nsP3; 1:1000 goat anti-actin, Santa Cruz) (129). Membranes were washed 3× in 1× TBST while rocking at room temperature for 10 min before incubation with secondary antibody (1:10 000) IRDye conjugated mouse anti-rabbit for nsP3 and IRDye conjugated anti-goat for actin (Li-Cor). Secondary antibodies were incubated for 1 h at room temperature in 5% milk with 0.01% SDS in 1x TBST on a rocker. Membranes were washed 3x with 1x TBST for ten minutes each wash. The membranes were then washed 3x in 1× TBS for 10 min each time. The membranes were visualized with the Odyssey infrared Imaging system (Li-Cor).

Structural conservation. The structure compatibility (*SC*) score of a structure within a given, related sequence (*x*) was defined as follows:

$$SC = \frac{bp_x}{bp_{orig}} \qquad (2)$$

specifically, the fraction of base pairs (*bp*) in the SINV structure (*orig*) that can form in the related sequence, using the multiple sequence alignment to identify the locations of homologous base pairs.

To search for homologous structures, we used the Infer- nal software suite (v1.1.1) (106, 130). For each SINV structured region, the sequence and the minimum free energy structure were used to create an alignment in Stockholm format (with long-range base pairs in the packaging signal removed). A covariance model was built and calibrated using *cmbuild* and *cmcalibrate* for each region. Hits in homologous alphaviruses were found using *cmsearch* with the model on the sequences in the multiple sequence alignment, and those hits were assembled into a new alignment with *cmalign*, to create a structure-informed alignment for each region of interest.

The R-scape program (v0.3.2) was used to identify base pairs with significant covariance in each structure-informed alignment >50 nucleotides (77), and applied to conserved RNA alignments from the same report. Average percent sequence identities were calculated with the *alistat* program, part of the HMMER package (131).

CHAPTER 2: USING SHAPE-MaP TO MODEL RNA SECONDARY STRUCTURE AND IDENTIFY 3′UTR VARIATION IN CHIKUNGUNYA VIRUS[2]

2.1 Overview

Chikungunya virus (CHIKV) is a mosquito-borne alphavirus associated with debilitating arthralgia in humans. RNA secondary structure in the viral genome plays an important role in the lifecycle of alphaviruses; however, the specific role of RNA structure in regulating CHIKV replication is poorly understood. Our previous studies found little conservation in RNA secondary structure between alphaviruses, and this structural divergence creates unique functional structures in specific alpha- virus genomes. Therefore, to understand the impact of RNA structure on CHIKV biology, we used SHAPE-MaP to inform the modeling of RNA secondary structure throughout the genome of a CHIKV isolate from the 2013 Caribbean outbreak. We then analyzed regions of the genome with high levels of structural specificity to identify potentially functional RNA secondary structures and identified 23 regions within the CHIKV genome with higher-than-average structural stability, including four previously identified, functionally important CHIKV RNA structures. We also analyzed the RNA flexibility and secondary structures of multiple 3′ UTR variants of CHIKV that are known to affect virus replication in mosquito cells. This analysis found several novel RNA structures within these 3′ UTR variants. A duplication in the 3′ UTR that enhances viral replication in mosquito cells led to an overall increase in the amount of unstructured RNA in the 3′ UTR. This analysis demonstrates that the CHIKV genome contains a number of unique, specific RNA secondary structures and provides a strategy for testing these secondary structures for functional importance in CHIKV replication and pathogenesis.

2.2 Importance

Chikungunya virus (CHIKV) is a mosquito-borne RNA virus that causes febrile illness and debilitating arthralgia in humans. CHIKV causes explosive outbreaks but there are no approved therapies to treat or prevent CHIKV infection. The CHIKV genome contains functional RNA secondary structures

[2] First published in *JVI* 2020 94:e00701-20. https://doi.org/10.1128/JVI.00701-20.

that are essential for proper virus replication. Since RNA secondary structures have only been defined for a small portion of the CHIKV genome, we used a chemical probing method to define the RNA secondary structures of CHIKV genomic RNA. We identified 23 highly specific structured regions of the genome, and confirmed the functional importance of one structure using mutagenesis. Furthermore, we defined the RNA secondary structure of three CHIKV 3′ UTR variants that differ in their ability to replicate in mosquito cells. Our study highlights the complexity of the CHIKV genome and describes new systems for designing compensatory mutations to test the functional relevance of viral RNA secondary structures.

2.3 Introduction

Chikungunya virus (CHIKV) is an arthropod-borne alphavirus that causes febrile illness associated with severe acute and persistent arthralgia. Since its identification in 1952, CHIKV has caused sporadic outbreaks in Africa, Asia, and the Indian subcontinent. However, recent outbreaks in the countries surrounding the Indian Ocean in 2005, as well as the 2013 introduction of the virus into the Americas, illustrate CHIKV's reemergence as a global threat to public health (4, 132). Despite its status as a significant emerging disease threat, there are currently no approved vaccines or virus-specific therapies for treating acute or chronic CHIKV disease. Therefore, it is important to understand the factors that contribute to CHIKV pathogenesis, since this information may inform the development of safe and effective vaccines and therapies.

The alphavirus genome is a positive sense, single-stranded RNA that encodes two polyproteins. The first polyprotein encodes the four nonstructural proteins (nsP1 to nsP4), which together comprise the RNA replication machinery. The second, an internally encoded polyprotein encompassing the 3′ third of the viral genome, encodes the virion structural proteins from a subgenomic RNA. While the role of viral proteins in the alphavirus life cycle has been extensively studied, a growing body of evidence suggests that coding and noncoding RNA structural elements (e.g., stem loops) are also critical determinants of alphavirus replication and pathogenesis. These include structures in the 5′ UTR that prevent host innate immune recognition, RNA packaging signals, and RNA elements that regulate viral transcription and translation (25, 26, 28, 33, 82). However, the full complement of RNA secondary structures in the CHIKV genomic RNA has not been determined. Given the importance of RNA secondary structure in alphavirus

biology, a better understanding of CHIKV RNA secondary structures is likely to provide new insights into the viral factors that contribute to the CHIKV life cycle and CHIKV disease pathogenesis.

We attempted to identify conserved alphavirus RNA secondary structures using Sindbis virus (SINV), Venezuelan equine encephalitis virus (VEEV), and CHIKV but found little of the RNA secondary structure landscape is conserved across these viruses (133). Retrospectively, this is not entirely surprising, since some recognized functional RNA secondary structures, such as RNA packaging signals, are found at different locations of the genomes of different alphaviruses, while other structures are only found in a subset of viruses in the genus (41, 134). Furthermore, alphaviruses have very low signals of nucleotide covariation, the traditional “gold standard” for identifying conserved, functional RNA secondary structures (65, 99, 102, 123, 133, 135).

Because the RNA secondary structure landscape of alphaviruses is not highly conserved (133), we used selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) to perform de novo RNA structural analysis on the CHIKV genome to identify potentially functional RNA secondary structures. SHAPE- MaP is an RNA structure probing technique that combines chemical probing of unpaired nucleotides of the genome with next generation sequencing to identify highly flexible, or unstructured, regions in long RNAs (92). The mutational profiling is combined with rigorous thermodynamic free energy modeling to generate experimentally derived, high-confidence models of RNA secondary structure (65, 92). This method has been applied to several other RNA viruses to identify important RNA structural features (93, 94, 133, 136, 137). We hypothesized that functionally important RNA secondary structures specific to CHIKV would fold into a single, specific conformation relative to the rest of the genome. With this approach, we identified the four known functional RNA elements in the CHIKV coding sequence, as well as 19 previously unidentified elements. We confirmed the functional importance of one element through structure disrupting mutagenesis strategies. Furthermore, three variants of the CHIKV 3′ UTR have been reported (49, 138), and our studies defined the RNA structure of each variant. We further characterized the impact of each of these variants on CHIKV host range. Together, these studies provide important new information on the location and stability of RNA structures distributed throughout the CHIKV genome.

2.4 Results

SHAPE-MaP analysis of the CHIKV genomic RNA

RNA structure plays an important role in alphavirus biology and contributes to functions ranging from regulation of RNA and protein synthesis to immune evasion (25, 26, 28, 33, 82). However, despite CHIKV's importance as an emerging pathogen, our understanding of how viral RNA structure impacts the CHIKV life cycle has largely been inferred from analysis of the RNA secondary structure in other alphaviruses (25, 26). Extensive analysis of functional RNA secondary structures has been performed on Sindbis virus (SINV), Semliki Forest virus (SFV), and Venezuelan equine encephalitis virus (VEEV). These analyses identified specific regions within alphavirus genomes where RNA secondary structure plays important functional roles in RNA packaging, RNA and protein synthesis, and immune evasion (12, 28, 42, 82). We recently determined the full genome RNA structure of the Girdwood S.A. strain of SINV and the ZPC738 strain of VEEV using SHAPE-MaP (selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling). This analysis found that the genomes of both SINV and VEEV are highly structured beyond the 5′ UTRs of both the genomic and subgenomic RNAs. This finding suggests that in addition to the previously defined functional RNA structures, additional RNA secondary structures may play a role in SINV or VEEV replication. However, the structure profiles are highly divergent between the two viruses, with little correlation between SHAPE-MaP profiles, few conserved RNA secondary structures, and low structure compatibility among conserved structures (133). We used structural conservation as a method to identify functionally important RNA structures in SINV, but this method failed to identify all the known RNA secondary structures important for SINV replication. Therefore, the field needs alternative methods to identify potentially important structured regions in alphaviruses.

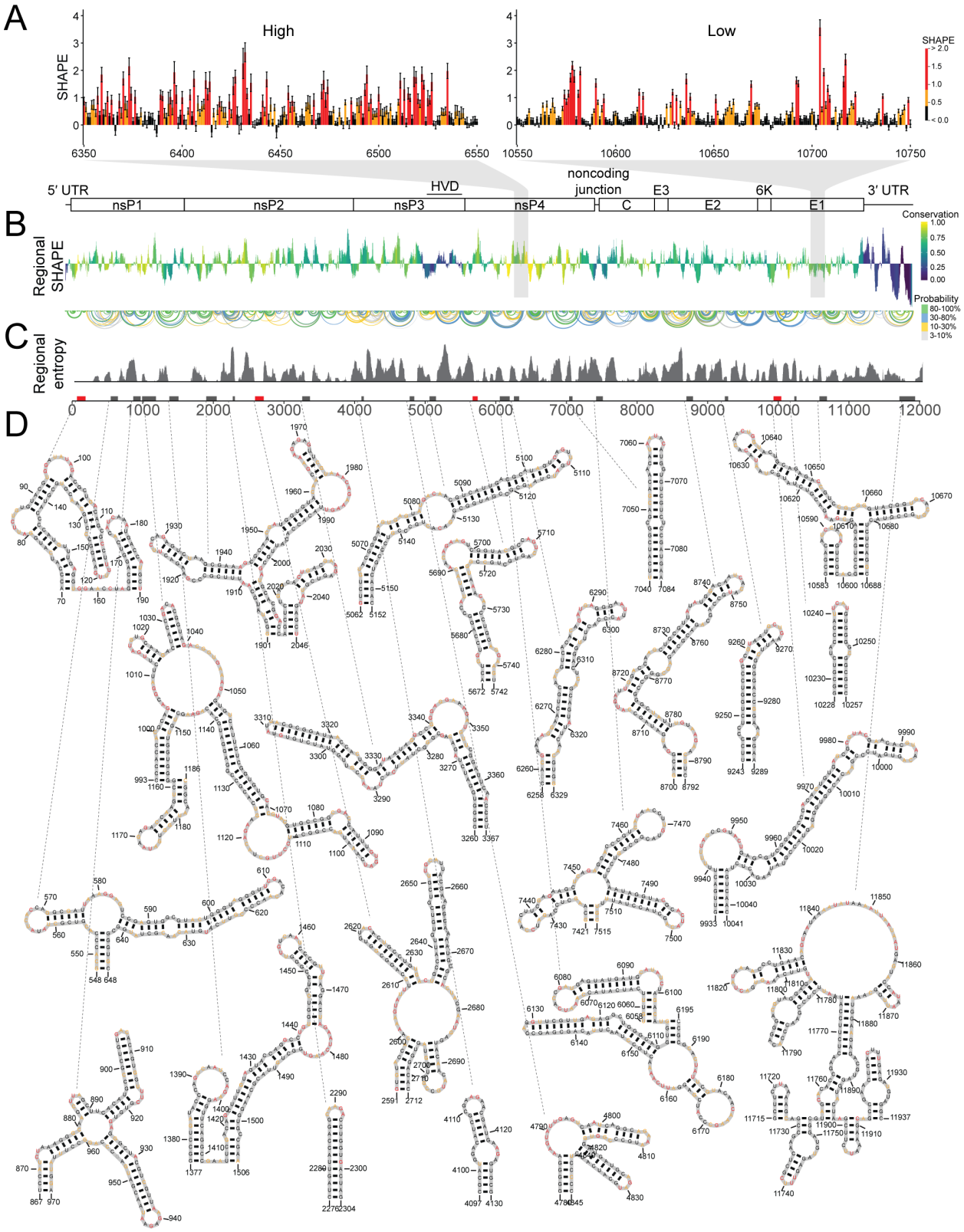


Figure 2.1: SHAPE-MaP indicates specific RNA secondary structures are found throughout the CHIKV genome.

(A, left) Example of SHAPE reactivities in a high-SHAPE, unstructured region. (Right) Example of SHAPE reactivities in a low-SHAPE, very structured region. Although individual reactivities do not reveal local structure, the pattern of reactivities in a region provides information about structuredness within that region. (B, top) Cartoon of the CHIKV genome. (Bottom) Regional median SHAPE reactivities across the CHIKV genome compared to the global median SHAPE. Regions below the *x* axis indicate more structure than average, while regions above the *x* axis indicate less structure than average. The histogram is colored according to the regional conservation score using representative full-genome sequences across the alphavirus genus (7, 133). Yellow indicates highly conserved regions, and purple indicates less conserved regions. HVD, hypervariable domain. (C, top) Base pairs within CHIKV genome. The color indicates base pairing probability. (Bottom) Windowed Shannon entropy across the genome. Low Shannon entropy values correspond to regions that form a single structure. (D, top) Boxes along the genome indicate highly structured regions, as determined by both computational prediction and experimental reactivities. Red boxes indicate structured regions with previously known functional importance. Black boxes indicate novel structured regions. (Bottom) SHAPE-MaP informed secondary structure models of highly structured regions. Nucleotide color corresponds to SHAPE reactivity scale in panel A.

We hypothesized that RNA structures likely to fold into a single specific conformation had an increased likelihood of being functionally important. Therefore, to test this idea, we set out to determine the RNA secondary structure of the genomic RNA of a human CHIKV isolate from the 2013 outbreak on the Caribbean island of St. Martinique. We treated purified CHIKV genomic RNA isolated from cell-free virions with the 1M7 SHAPE reagent, and SHAPE-MaP was performed to generate a SHAPE reactivity profile for the entire CHIKV genome (Figure 2.1). SHAPE reactivity indicates the relative flexibility of a nucleotide, and nucleotide flexibility correlates with base-pairing likelihood. The SHAPE-MaP technique measures SHAPE reactivities with single nucleotide precision (70). SHAPE reactivities above 0.8 indicate likely unpaired bases (shown in red), as illustrated by an unstructured region spanning nucleotides 6350 to 6550 (Figure 2.1 A, left). SHAPE reactivities below 0.4 indicate likely paired and therefore unreactive bases (colored black), as illustrated by a representative structured region spanning nucleotides 10550 to 10750 (Figure 2.1 A, right). There are large-scale fluctuations in the SHAPE reactivity across the genome, which are best visualized as a median windowed SHAPE reactivity, as shown in Figure 2.1 B, where we plotted the median windowed SHAPE (called regional SHAPE). These data indicate there are specific regions in the CHIKV genome that have low median SHAPE or are more likely to form RNA secondary structures.

While RNA secondary structures are not highly conserved across alphaviruses, it is unknown if overall viral RNA structured-ness correlates with sequence conservation. Figure 2.1 B shows the regional SHAPE of CHIKV colored by the sequence conservation scores from our previous work (133). We observed very little pattern for highly conserved regions and their overall structured-ness within CHIKV.

Regions with high sequence conservation scores (Figure 2.1 B, in yellow) are found with low- and above-average SHAPE reactivity values. Likewise, the least conserved regions of the genome (dark purple and blue) also fluctuate between very low and very high SHAPE reactivity values. For example, the 3′ UTR of CHIKV, which differs within CHIKV strains, contains some of the least and most reactive nucleotides in the genome (48). Interestingly, the hypervariable domain near the end of nsP3, which is often omitted from alignments of alphavirus genomes due to drastic divergence in sequence, has very low SHAPE reactivity in CHIKV, SINV, and VEEV (7, 133). However, this shared low SHAPE reactivity does not result in similar secondary structures (133).

In order to identify the likely RNA secondary structures across the CHIKV genome, we computed base-pairing probabilities for the entire genome from the SHAPE reactivity data (Figure 2.1 C, top arches). From these base-pairing probabilities, we computed the Shannon entropy of base pairing for each nucleotide using a 55-nucleotide window (Figure 2.1 C, bottom), where low-entropy suggests a single, well-defined conformation (71, 92, 99, 123, 125). We used these base pairing probabilities and entropy values to generate SHAPE-MaP-derived structural predictions, which represent the most likely structural conformation for a specific region of the genome. We included the SHAPE-MaP derived RNA secondary structures for the entire CHIKV genome, with nucleotides colored according to SHAPE reactivity in Figure 1 (https://drive.google.com/file/d/1ZlrFGYxsrUeRE0QcUB_rWPirpLki4C6F/view?usp=sharing).

Identification of specific RNA secondary structures within the CHIKV genomic RNA

SHAPE-MaP analysis of the CHIKV genome indicates that RNA secondary structure is distributed throughout the viral genome (Figure 2.1). While SHAPE-MaP- derived structural predictions represent the most likely structural conformation for any region of the genome, it is likely that many regions of the genome are capable of adopting several different conformations. We hypothesized that functionally important RNA secondary structures would be both highly structured (SHAPE reactivity of <0.3) and highly specific (Shannon entropy of <0.04) compared to the rest of the genome and therefore likely to adopt a single RNA secondary structure. Low SHAPE reactivity indicates the region is often involved in base pairing interactions. While low SHAPE indicates likely structure, the RNA sequence can be involved in many different conformations to achieve this low SHAPE reactivity (65). Regions with low Shannon

entropy indicate that there are few possible conformations; therefore, regions with low SHAPE reactivity and low Shannon entropy are likely to have a single well-determined secondary structure (92). By using cutoffs that were shown to yield specific structures in prior studies (133), we identified 23 regions that meet these criteria (Figure 2.1 D). Four of these 23 regions contain RNA structures known to be functionally important in CHIKV replication or pathogenesis: a stem-loop of the 5′ conserved sequence element (CSE), a region within the CHIKV packaging signal, a stem-loop just 3′ of the opal termination codon that is involved in opal termination codon readthrough, and an RNA structure involved in ribosome frameshifting to produce the viral TF protein (12, 25, 28, 32, 33).

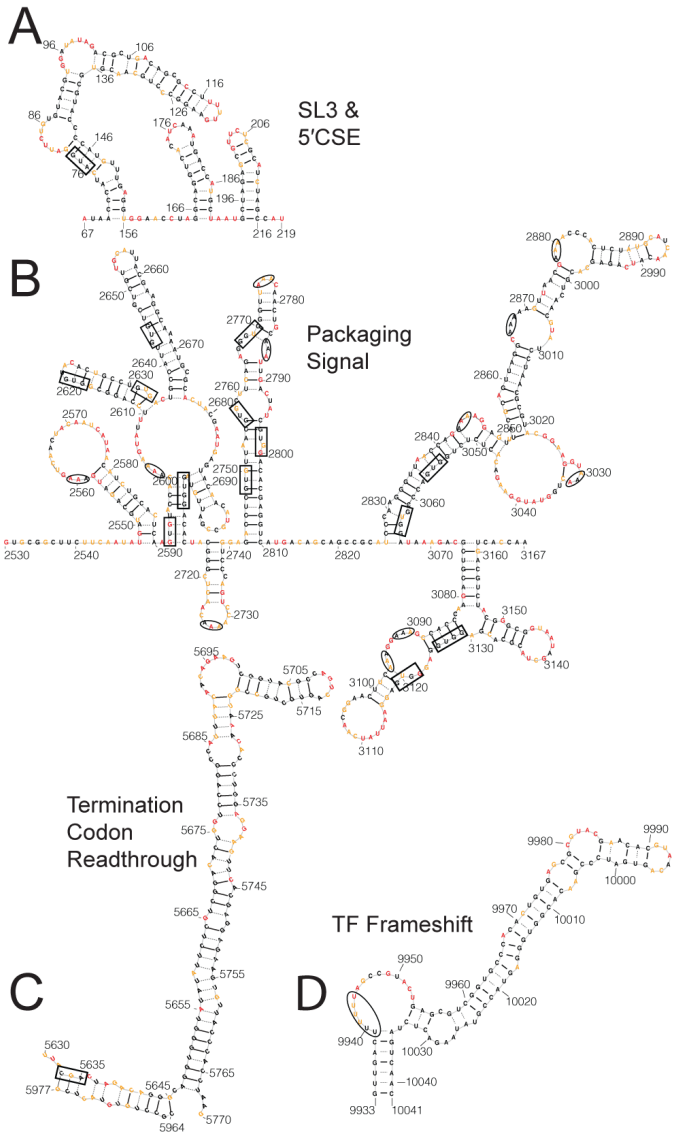


Figure 2.2: SHAPE-MaP analysis identifies previously known functional RNA secondary structures.

(A) SL3 of the CHIKV genome and the 5= conserved sequence element modeled using SHAPE-MaP data. The nonstructural polyprotein start codon is boxed. (B) Putative CHIKV packaging signal, as identified by Kim et al. (28). Triple adenosine motifs within loops and bulges are circled; the predicted GUG(G) motifs from Kim et al. are boxed (28). (C) The CHIKV TCR with the canonical opal stop codon position boxed. The CHIKV strain used had predominantly Arg at this position and is modeled as such. (D) The CHIKV TF frameshift element is plotted with the slippery U motif circled. Nucleotide color corresponds to the SHAPE reactivity key in Figure 2.1 A.

5′ conserved sequence element. Our analysis identified nucleotides 70 to 195 as being highly specific and highly structured. This region includes stem-loop 3 (SL3) of the genome, which contains the nonstructural polyprotein start codon, and SL4, the first stem-loop of the 5′ CSE. The 5′ CSE (nucleotides 165 to 216 in CHIKV) is one of the few structurally conserved motifs across alphaviruses and is important for proper alphavirus genome replication during infection (25, 26, 133). Figure 2.2 A expands this structured region in CHIKV to include the entire 5′ CSE. Both stem loops of the 5′ CSE are predicted to be composed of 9 bp each, with the first stem-loop having seven unpaired nucleotides in the terminal loop and the second stem-loop having four unpaired nucleotides composing the terminal loop. The SINV 5′ CSE second stem-loop has more than 4 unpaired nucleotides in the apical loop and only 8 bp making up the stem (23, 25, 133). The CHIKV 5′ CSE secondary structure is more similar to that of VEEV, which is also predicted to have 9 bp making up each stem-loop and only four unpaired nucleotides of the second stem-loop (26, 133).

Packaging signal. The putative packaging signal for CHIKV is in nsP2 (nucleotides 2501 to 3079) (28), but an RNA structure had yet to be determined for this region (Fig. 2.2 B). Our analysis identified a portion of this region (nucleotides 2590 to 2713) as one of the 23 highly specifically structured regions. For SINV and VEEV, the packaging signal is composed of four to six stem loops with triple G motifs in the loops. The CHIKV packaging signal was predicted to have a similar multistem motif but instead of GGG, the stems would be topped with a GUG(G) motif (28). The SHAPE-MaP model indicates that the region encompassing the CHIKV packaging signal is composed of eight stem loops with the predicted GUG(G) motifs being predominantly nonreactive nucleotides contained entirely within stems (Figure 2.2 B, boxed). Instead we identified multiple in- stances of a triple A motif located in loops and bulges following a similar reactivity pattern to that of the triple G motif in SINV and VEEV (133).

Opal termination readthrough element. The third known functional RNA second- ary structure is a stem-loop found at the start of nsP4, or just 3′ of the canonical opal stop codon (Figure 2.2 C). We

found a highly specific stem-loop at nucleotides 5672 to 5742 that is part of the termination codon readthrough (TCR) element, which increases readthrough of the opal stop codon in order to translate the full-length nonstructural polyprotein (nsP1 to nsP4) (32, 33). It should be noted that our virus contained a mixed population at the opal stop codon itself (nsP3 aa520), with the majority coding for an arginine (codon CGA) at this position and a minor population containing the canonical opal stop codon (UGA) (34). While this region was modeled with the majority Arg residue codon (520R), it should be noted that the structure of the TCR is not dependent on the codon at nsP3 aa520. Our SHAPE-MaP derived structural prediction indicates that the 520R codon is contained in a stem structure adjacent to the TCR, where the two stems are separated by a one nucleotide space (Figure 2.2 C). However, many of these nucleotides are moderately reactive and therefore flexible during treatment. This suggests these nucleotides likely adopt an open conformation which would create a spacer of 11 nucleotides between the 520R codon and the base of the TCR, which is consistent with the TCR model previously proposed by Firth et al. (32).

TF frameshift element. The final known functionally important RNA secondary structure identified was the TF frameshift element (nucleotides 9933 to 10041) (Figure 2.2 D). This element is located in the 6K coding region and causes a -1 frameshift due to a slippery UUUUUU motif followed by a hairpin. The new reading frame encodes the TF protein. The UUUUUU element is present in other alphavirus genomes but secondary structure following this motif is not predicted to be present in all alphaviruses (12, 33, 133). The general motif predicted for this element in CHIKV was the UUUUUU motif, followed by a spacer of five to nine nucleotides, and then a structured region, which was based largely on comparison to other viruses (12). Our model, generated from data gathered from the full-length genomic RNA, indicates the UUUUUU motif is followed by nine nucleotides, the majority of which are highly reactive supporting the prediction that these nucleotides are unpaired. A long stem follows the unpaired nucleotides, which agrees with the general motif predicted for other frameshift elements (12). Further- more, all nucleotides predicted to participate in base pairing are supported by very low reactivity scores.

Overall, comparisons between our SHAPE-MaP data and prior computational RNA structure predictions are largely consistent (12, 33, 133), though we did identify subtle, but potentially functional important differences. This illustrates the utility of combining structure probing techniques combined with

high-throughput sequencing like SHAPE- MaP for both identifying and providing more refined information of RNA secondary structures in RNA virus genomes.

CHIKV SL3 enhances genome transcription

As noted above, identification of the 5′ CSE by SHAPE-MaP validated the accuracy of our approach. However, we were intrigued by the identification of SL3, the stem-loop immediately upstream of the 5′ CSE that contains the initiating AUG, as a highly specific structure. Previous studies of the 5′ CSE in other alphaviruses predicted this stem-loop and disrupted it when probing the functional importance of the 5′ CSE itself. However, the function of SL3 alone has never been studied (25, 26). Disruption of the stem loops in the 5′ CSE in combination with the upstream stem-loop results in a decrease in SINV replication in mammalian cells and severe defects in replication within mosquito cells, while analogous mutations in VEEV are lethal to the virus (25, 26). However, during serial passaging of the disrupted VEEV, Michel et al. reported that compensatory mutations were generated in the large stem-loop 5′ of the 5′ CSE, analogous to SL3 of CHIKV (26), and these compensatory mutations were predicted to stabilize the large stem-loop. This suggests that SL3 is functionally important, at least in the context of a structurally disrupted 5′ CSE.

To assess the impact of disruption of this region on CHIKV replication, we mutated nucleotides 67 to 216, which encompasses SL3 and the 5′ CSE (∂SL3-5). We also designed two mutants that disrupted SL3 alone (∂SL3) and the 5′ CSE alone (∂5′ CSE) (Figure 2.3 A, red stars; see also Appendix B). To avoid affecting coding capacity, our mutagenesis strategy used wobble-base codon shuffling to maximally disrupt base pairing in the RNA secondary structure, while maintaining both the coding capacity and dinucleotide frequency (133). In vitro-transcribed genomic CHIKV RNA was electroporated into BHK-21 cells and successful infection was measured by the number of resulting infectious centers. Disrupting SL3 alone had no impact on infectious center production compared to the wild-type (WT) control (Figure 2.3 B). In contrast, the ∂5′ CSE mutant produced significantly fewer infectious centers, confirming the importance of this region for alphavirus replication (25, 26). The ∂SL3-5 mutant was nonviable, yielding no infectious centers. This suggests that both SL3 and the 5′ CSE are necessary for optimal CHIKV RNA infectivity.

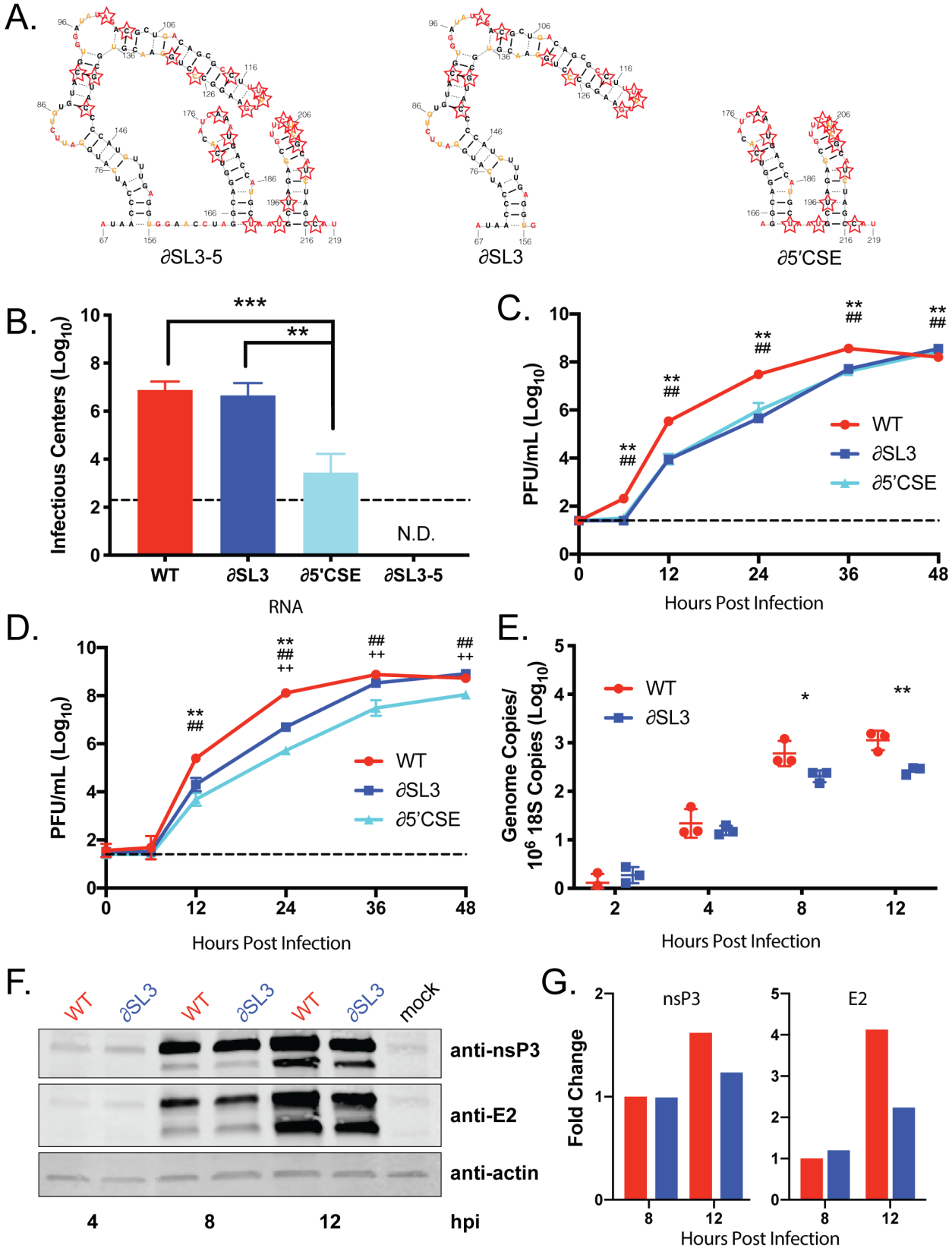


Figure 2.3: CHIKV SL3 enhances RNA transcription.

(A) Secondary structure models of CHIKV SL3-5, SL3, and the 5′ CSE alone, SL4-5. Starred nucleotides indicate nucleotides mutated to disrupt RNA secondary structure and sequence using *CodonShuffle* (102). (B) Infectious centers assay of mutant viruses. The data represent aggregates of three independent experiments. (C and D) Mutant virus growth in mammalian Vero81 cells (C) and mosquito C6/36 cells (D). The data are means of nine biological replicates across three independent experiments. (E) SL3 genome transcription was assessed by qRT-PCR in mammalian Vero81 cells. The data shown represent one of three independent experiments, each performed with three biological replicates. (F) Viral protein synthesis was assessed by Western blotting. The blot is representative of three independent experiments. (G) Densitometry was performed for nsP3 and E2 using ImageJ software. The data are representative of two independent experiments analyzed. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. The symbols in panels C and D indicate the *P* value for the following comparisons: *, WT versus ∂SL3; #, WT versus ∂5′ CSE; and +, ∂SL3 versus ∂5′ CSE.

We next tested whether ∂SL3 and ∂5′ CSE were impaired for replication in mammalian and mosquito cells. Both mutants exhibited slower replication kinetics compared to WT in mammalian cells (Figure 2.3 C). In mosquito cells, the ∂SL3 mutant had an intermediate phenotype between that of WT and the ∂5′ CSE mutant (Figure 2.3 E). This suggested that SL3 is required for efficient virus replication in both mosquito and mammalian cells, but that the magnitude of effect of SL3 is host dependent, while the 5′ CSE is important for CHIKV replication for both host cell types.

We next defined the stage in the CHIKV replication cycle that requires SL3. Genome and protein accumulation were measured early in infection by qRT-PCR and Western blotting, respectively (Figure 2.3 D and F). Disrupting SL3 resulted in a delay in accumulation of genomic RNA compared to WT. However, the protein nsP3, a component of the viral replicase complex, and E2 glycoprotein accumulated to similar levels in cells infected with the WT or mutant viruses. Densitometry analysis indicated that viral proteins were slightly more abundant in WT infected cells than in ∂SL3 infected cells by 12 h post-infection (Figure 2.3 G). This is likely due to a greater abundance of WT RNA present at 8 and 12 h post-infection and not due to impaired translation of the ∂SL3 RNA since there are similar levels of viral protein accumulation at 8 h post-infection. This suggests that SL3 functions to enhance genomic RNA replication and is not necessary for proper viral protein synthesis in mammalian cells.

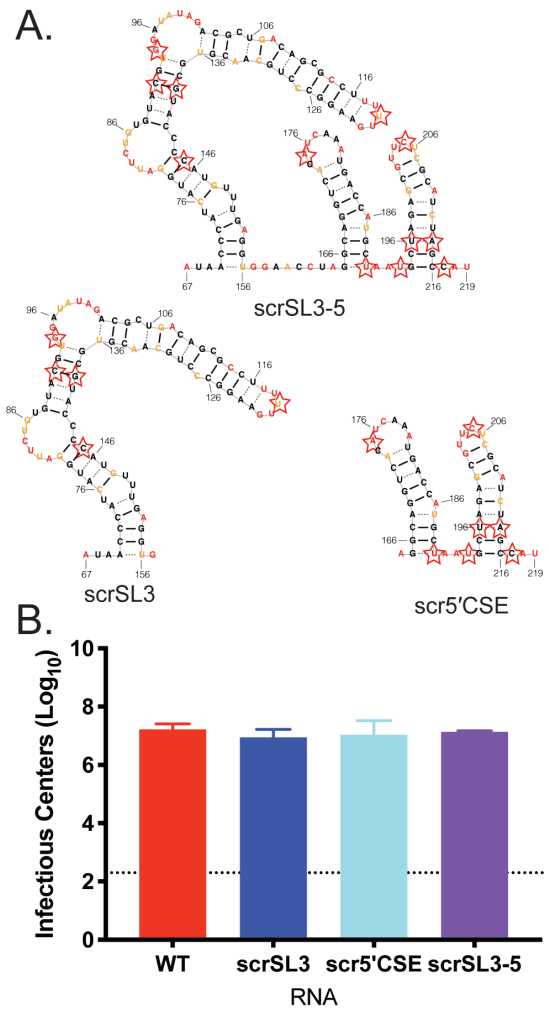


Figure 2.4: Preservation of RNA secondary structure when primary sequence is disrupted complements full structure disruption phenotypes.
(A) Secondary structure models of CHIKV SL3-5, SL3, and the 5′ CSE alone, SL4-5. Starred nucleotides indicate nucleotides mutated to disrupt primary sequence and maintain RNA secondary structure using *CodonShuffle* (102). (B) Infectious centers assay of mutant viruses. The data are aggregated from three independent experiments.

While our mutations were designed to disrupt RNA secondary structure in SL3 and the 5′ CSE, this method also disrupts sequence. In order to test whether ∂SL3 and ∂5′ CSE mutants were attenuated due to structure disruption or sequence disruption, we generated three additional mutants. Using the same wobble-base codon shuffle algorithm, we chose sequences that maintained the secondary structure of the region but used a different sequence to maintain coding capacity. The *CodonShuffle* program predicts a minimum free energy structure for each codon shuffled sequence generated (102). We mutated nearly all available nucleotides possible that would also maintain the predicted secondary structure. These “fixed” structure mutants (scrSL3, scr5′CSE, and scrSL3-5) differ in sequence from WT

but are predicted to be structurally the same (Figure 2.4 A; see also Appendix B). When these fixed structure mutants are assessed for infectivity as before, all mutants produce the same number of infectious centers as WT RNA (Figure 2.4 B). These data indicate that structure within SL3 plays an important role in promoting efficient CHIKV replication in combination with the 5′ CSE stem loops.

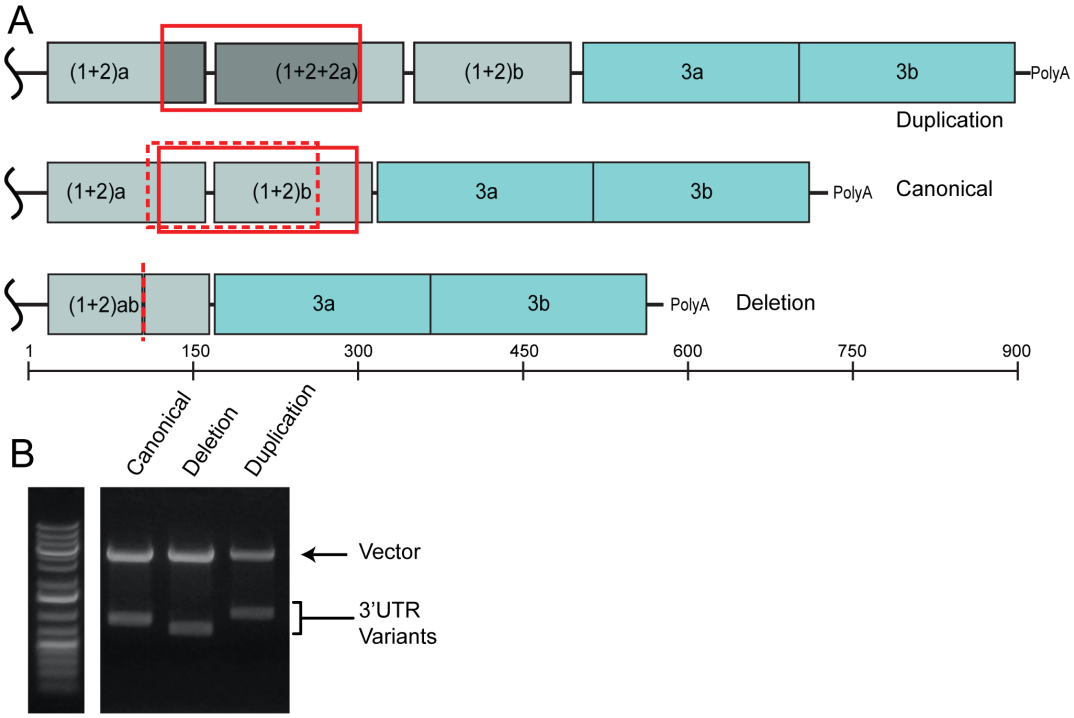


Figure 2.5: St. Martinique CHIKV isolate contained three 3′ UTR variants.
(A) Cartoon representations of the CHIKV 3′ UTR. Colored boxes indicate unique repeat elements (RE) that are labeled within the diagrams. (Top) Duplication variant 3′ UTR. (Middle) Canonical Asian genotype 3′ UTR. (Middle bottom) Deletion variant 3′ UTR and nucleotide length ruler. Solid red boxes indicate the location of the duplicated sequence. The additional sequence is both boxed and shaded in the duplication 3′ UTR cartoon. A dashed red box indicates the sequence that was deleted in reference to the canonical 3′ UTR, and the dashed red line indicates where this sequence would be in the deletion 3′ UTR. (B) 3′ RACE products were subcloned into blunt vectors for sequencing and clarification of 3′ UTR sizes. Clones containing one of each 3′ UTR were digested and separated by gel electrophoresis.

Identification of 3′ UTR variants in CHIKV

SHAPE-MaP analysis requires high- throughput sequencing of RNA treated with a SHAPE chemical probe to detect SHAPE adduct induced mutations. These sequencing results are compared to an untreated control, therefore providing deep sequencing results of the viral genomic RNA. The sequencing results for the negative-control portion of our SHAPE-MaP analysis found that the 3′ UTR of the CHIKV isolate used in our study was 738 nucleotides in length and had repeat element nucleotide sequences and pattern consistent with those found in Asian CHIKV strains (20, 48). However, we also

noted the read depth increased in the (1+2) repeat element regions of the 3′ UTR (Figure 2.5 A), similar to a previously described duplication in this region within Caribbean CHIKV isolates (49, 138). To further define the 3′ UTR of our isolate, we performed 3′ RACE on RNA isolated from the virus stock. This analysis revealed three distinct isoforms of the viral 3′ UTR (Figure 2.5 B): (1) the 738- nucleotide canonical 3′ UTR (Figure 2.5 A, top), (2) a 912-nucleotide variant with a partial 3′ UTR duplication (Figure 2.5 A, middle, solid red box) which has been previously identified (49, 138), and (3) a novel 583-nucleotide variant containing a 152-nucleotide deletion that removes the 3′ end of the first copy of the (1+2) repeat element and the majority of the second (1+2) repeat element (Figure 2.5 A, bottom, dotted red box and line).

We constructed three CHIKV infectious clones, each containing one of the three 3′ UTR variants. Since the 3′ UTR duplication has been shown to enhance CHIKV replication in mosquito cells and deletion mutants were observed to be attenuated, we initially tested the three viruses for their ability to replicate in C6/36 mosquito cells (49, 50, 52, 138). We observed three distinct phenotypes from the three 3′ UTRs (Figure 2.6 A). The 3′ UTR duplication clone (3′ dup) exhibited faster kinetics and achieved an overall higher peak titer than the canonical (3′ canon) or deletion (3′ del) 3′ UTR variant. The 3′ del virus was severely attenuated for growth in mosquito cells, achieving a peak titer 100- to 1,000-fold lower than the other 3′ UTR variants. However, all three viruses exhibited similar growth kinetics in Vero81 mammalian cells (Figure 2.6 B). These data show that duplications in the CHIKV 3′ UTR are beneficial for virus replication in mosquito cells and are unimportant for virus replication in mammalian cells.

To determine how 3′ UTR variation impacts virus replication and CHIKV-induced pathology in vivo, we used the C57BL/6J mouse model of CHIKV pathogenesis. Previous studies of CHIKV 3′ UTR variation and its impact on replication in vivo have used infant mice and artificial 3′ UTR constructs (48). The impact of naturally occurring 3′ UTR variants on CHIKV-induced disease has not yet been assessed. We infected 6-week-old C57BL/6J mice with 100 PFU of virus in the left hind footpad and monitored CHIKV-induced footpad swelling using our established methods (34). As shown in Figure 2.6 C, each of the three variants induced a similar degree of swelling in the footpad. We also found no differences in viral replication between the three viruses at 3 days post-infection in the left foot (inoculation site), right foot, sera, or spleen (Figure 2.7 D). Therefore, we did not detect a role for the 3′ UTR variation in

modulating CHIKV replication or disease in a mouse model of CHIKV disease. Rather, consistent with prior results, our studies suggest that variants in this region of the CHIKV 3′ UTR are primarily affecting viral fitness in mosquito cells (49, 50, 52, 138).

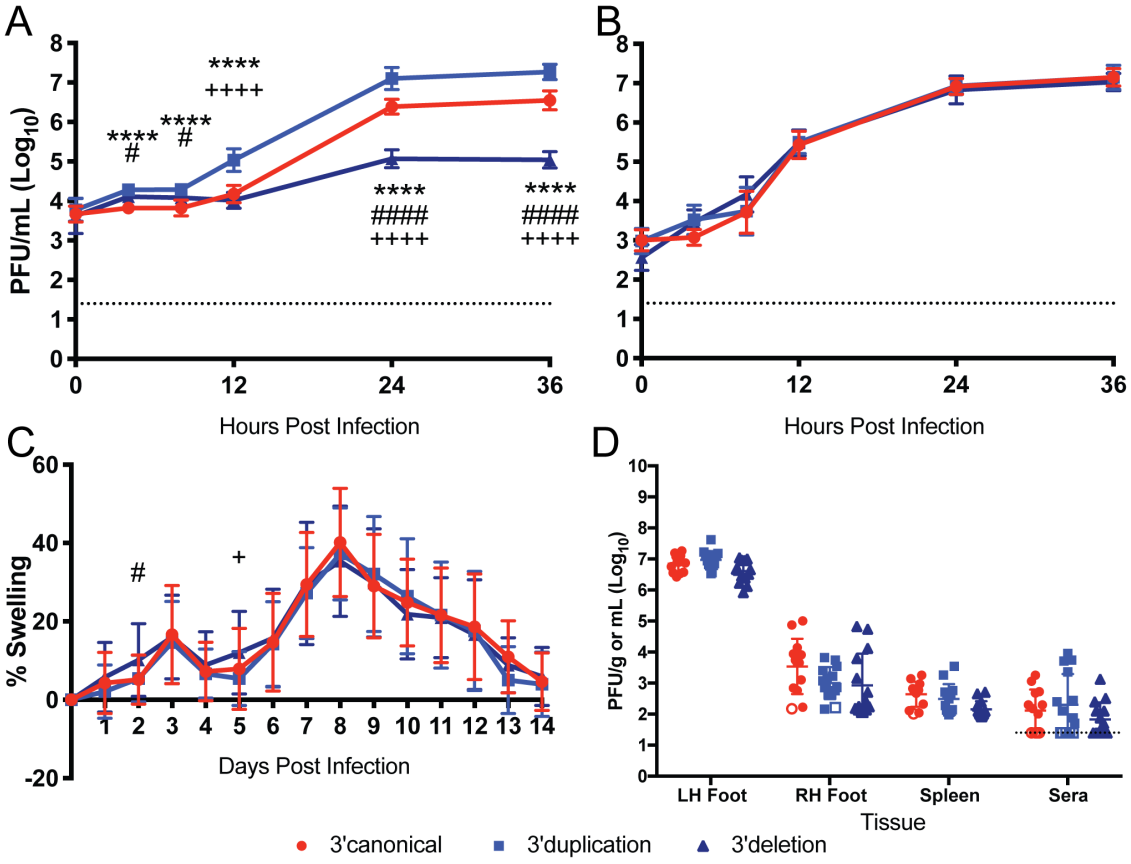


Figure 2.6: Variation in CHIKV 3′ UTR impacts virus replication in mosquito cells but not the vertebrate host.
Growth curves of 3′ UTR variant infectious clones in mosquito C6/36 cells (A) and mammalian Vero81 cells (B) are shown. The data are aggregated from three independent experiments with nine total biological replicates. (C) Inoculated footpad swelling of C57BL/6J mice after infection with 3′ UTR variants. Data for days 0 to 7: *n* = 30 for 3′ canon, *n*=30 for 3′ dup, and *n*=31 for 3′ del. Data for days 8 to 14: *n*=16 for 3′ canon, *n*=15 for 3′ dup, and *n*=16 for 3′ del. The data are aggregated from four independent experiments. Male and female mice were used. (D) Infectious virus load of C57BL/6J mice at 3 days postinfection. Open symbols indicate samples with undetectable virus and are plotted at the limit of detection for that tissue (dictated by tissue weight). Dashed lines indicate limit of detection for liquid samples. *n* = 13 for 3′canon, *n* = 14 for 3′dup, and *n* = 14 for 3′del. Male and female mice were used. Symbols indicate the following comparisons: *, 3′ canonical versus 3′ duplication; #, 3′ canonical versus 3′ deletion; and +, 3′ duplication versus 3′ deletion. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

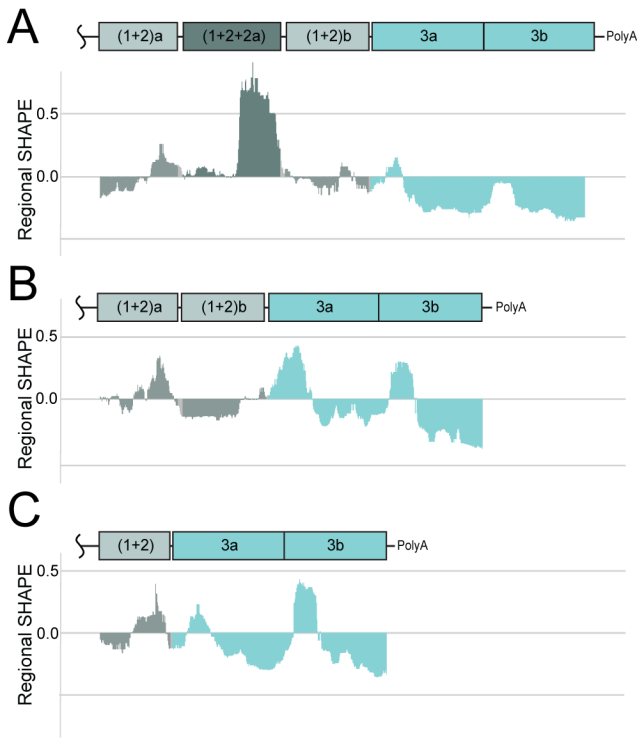


Figure 2.7: 3′ UTR variants are sequence similar but distinct in reactivity.
Median SHAPE reactivity profiles for the duplication (A), canonical (B), and deletion (C) 3′ UTRs are shown beneath cartoons depicting the repeat regions of the 3′ UTR. Histograms are colored according to repeat element. Gray, (1+2) repeat elements; dark gray, hybrid (1+2) repeat element; light turquoise, repeat element 3a and 3b.

RNA structure differences between CHIKV 3′ UTR variants

The differences in sequence of the 3′ UTR variants suggested that each 3′ UTR might assume a different RNA secondary structure. The 3′ UTR variants primarily differ in sequence at the start of the 3′ UTR where the duplication and deletion occurred, creating one, two, or two and a hybrid copy of a repeat element. The 3′ ends of the 3′ UTR variants are identical in sequence and repeat composition. Therefore, we hypothesized that if the RNA structure of the three 3′ UTRs were to differ, it would be at the start of the 3′ UTR, while the 3′ end of the UTR would have a similar SHAPE reactivity pattern and RNA secondary structure model.

The duplication, deletion, and canonical 3′ UTRs were in vitro transcribed from the respective infectious clone. SHAPE-MaP was performed on the in vitro transcribed 3′ UTR RNA using the same conditions applied to the full-length RNA genome. Figure 2.7 shows the median SHAPE reactivity profiles for each 3′ UTR. The color of the reactivity profiles corresponds to the repeat elements in the 3′ UTR. The SHAPE reactivity profile of the duplication variant contains highly reactive nucleotides that correspond to

the hybrid sequence repeat element unique to the duplication 3′ UTR (Figure 2.7 A, dark gray). These nucleotides reach much higher reactivity levels than any of the nucleotides in the canonical or deletion 3′ UTR (Figure 2.7 B and C, respectively). This suggests that the additional sequence in the duplication 3′ UTR is less structured than any portion of the canonical or deletion 3′ UTRs.

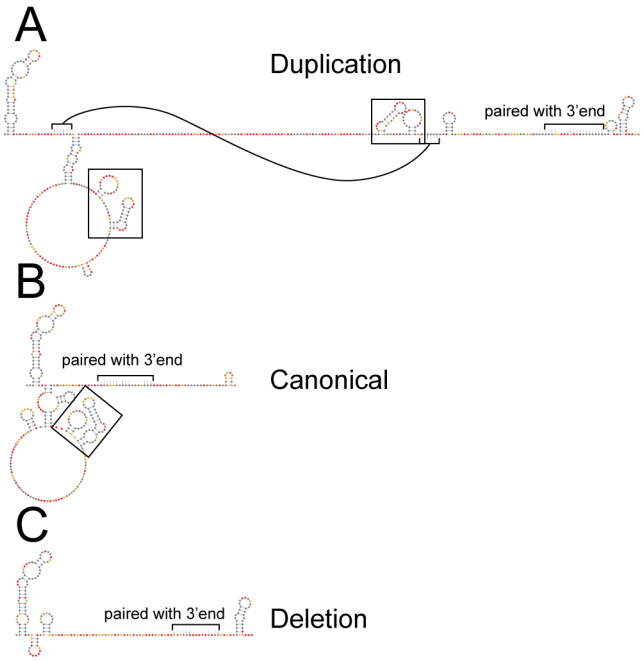


Figure 2.8: Variation in 3′ UTR reactivity corresponds to distinct models of secondary structure. SHAPE informed secondary structure models of the 5′ ends of the (A) duplication, (B) canonical, and (C) deletion 3′ UTRs are shown. Secondary structure models are for the (1+2) repeat element region for each 3′ UTR where the duplication and deletion events occurred. Nucleotides that pair with sequence not shown have been indicated by brackets and a note. In the duplication 3′ UTR one stem-loop is indicated with connected brackets for clarity. Black boxes indicate two stem-loop motif that was duplicated. Nucleotides are colored according to SHAPE reactivity key in Figure 2.1 A.

This highly reactive extra sequence supported our hypothesis that RNA secondary structure at the start of the 3′ UTR would be different among the three variants, but the end of each 3′ UTR would be similar. We modeled the RNA secondary structure for each 3′ UTR based on SHAPE data and found all three 3′ UTRs share a secondary structure at the end of the 3′ UTR (see online supplemental figure, end), while the major differences in secondary structure occur at the start of the 3′ UTR (Figure 2.8). Each 3′ UTR begins with the same hairpin, but the structures that follow are slightly different. The duplication 3′ UTR is predicted to be largely single stranded beyond some shared secondary structure at the start and a few other hairpins. The start of the deletion 3′ UTR also lacks significant structure compared to the canonical and duplication 3′ UTR. Interestingly, we noticed a two-hairpin motif that occurs once in the

canonical 3′ UTR and twice in the duplication 3′ UTR and is absent in the deletion 3′ UTR (Figure 2.8, boxes). Of note, the duplication virus has the most copies of this structure motif and replicates faster and to higher titers in mosquito cells, while the deletion virus lacks this structure motif and replicates slower and to lower titers (Figure 2.6 A and 2.8).

2.5 Discussion

RNA secondary structure plays a major role in multiple aspects of RNA virus biology. Given the importance of CHIKV as a re-emerging pathogen, we generated a whole-genome RNA secondary structure model for a human isolate of CHIKV to identify potentially functional novel RNA secondary structures. Our single-nucleotide resolution model both confirmed and refined past structural analyses of CHIKV RNA motifs and identified novel RNA structures in the CHIKV genome. We found RNA secondary structures are distributed throughout the genome and identified 23 regions across the genome that are predicted to adopt a single RNA conformation due to their high structural stability. These 23 regions include four previously identified functional RNA elements (5′ CSE, packaging signal, TCR, and the TF slippery site) (12, 25, 26, 28, 33). This suggests that high structural stability may provide some predictive value for identifying additional functionally relevant RNA structures in alphavirus genomes. We also demonstrate that a stem-loop adjacent to the 5′ CSE was functionally important in combination with the 5′ CSE. In the process of performing SHAPE-MaP analysis of the CHIKV genome, we analyzed several previously identified variants of the viral 3′ UTR. SHAPE-MaP analysis of these 3′ UTR variants indicates a duplication in the 3′ UTR results in more unstructured RNA at the start of the 3′ UTR as well as the duplication of a two-stem-loop motif. Using infectious clones, we showed that the variation in the CHIKV 3′ UTR had host-specific effects. The duplication 3′ UTR enhances virus replication and the deletion 3′ UTR inhibits virus replication in mosquito cells. However, variation in the CHIKV 3′ UTR had no effect on virus replication or swelling in our mouse model of pathogenesis.

Validation and discovery of highly specific structures.

Several RNA secondary structures were previously shown to be functionally important for CHIKV replication. These structures provided us an opportunity to both validate our SHAPE-MaP results and to test our new method for identifying important RNA secondary structures for CHIKV replication and pathogenesis. The 5′ CSE is the most conserved RNA secondary structure element across the alphavirus

genus. It can be found by sequence conservation or structure conservation analysis and unsurprisingly is one of the most specifically structured regions in CHIKV (Figure 2.2A) (133). Most prior analyses of this region focused on SINV and VEEV showing variable necessity depending on the virus (25, 26). It is interesting that our analysis also identified the large stem loop 5′ of the 5′ CSE, SL3, as being highly specific and structured. SL3, and its homolog across alphaviruses, contains the start codon for the nonstructural polyprotein. Early structure disruption studies of the 5′ CSE included the SL3 homolog in their disruption strategies, but the role of this specific stem-loop had not been studied (25, 26).

Recently, Kendall et al. released a SHAPE-informed RNA secondary structure model for the CHIKV 5′ end of the genome (139). Our model of the 5′ CSE and the base of SL3 agree with the Kendall model. Our models diverge after the base of SL3. The top of the Kendall et al. model diverges into a Y-shaped structure, while the model we propose continues from the base as a single stem with bulges. The 5′ arm of the “Y” has a number of reactive nucleotides placed in a stem with many unreactive nucleotides placed in a single-stranded loop. These reactivity data are somewhat contradictory to the model. They propose the shorter arm of the “Y” may actually form a pseudoknot with the start codon and surrounding nucleotides to explain this discrepancy, but attempts to confirm this structure were inconclusive (139). While we cannot rule out a pseudoknot, the Kendall et al. reactivity data corresponds well with our reactivity data and would support our model of a single long stem-loop. Regardless of this difference in modeling, our functional analyses and the Kendall et al. functional analyses of this stem-loop demonstrate the importance of SL3 in regulating CHIKV replication (Figure 2.3). Our data and Kendall et al. data support that RNA secondary structure is necessary for SL3 and 5′ CSE enhancement of virus replication (Figure 2.4). Therefore, these data refine and support recent advancements in understanding the role RNA structure plays in the CHIKV life cycle. Furthermore, our identification of SL3 as one of the 23 most specifically structured regions of the genome, along with the previously known functional RNA elements, suggests the remaining 19 specifically structured regions may play functional roles as well.

Our analysis also identified a region of the putative packaging signal in CHIKV and provides an RNA secondary structure model for this region (28). The packaging signal of CHIKV, and likely very closely related alphaviruses, is located in the coding region of nsP2, while SINV and other New World alphaviruses have a packaging signal further upstream in nsP1 (28). Our data suggest the CHIKV

packaging signal has multiple stem loops in close proximity and is more specifically structured compared to the rest of the genome, which is consistent with the packaging signal motifs identified for SINV and VEEV. However, it does not have the triple G motif reported in other alphaviruses or the hypothesized GUG(G) motif located in the loops (28). Instead, our secondary structure model of the region contains a triple A motif in five of the loops with the same reactivity pattern observed for the triple G motif in SINV (Figure 2.2 B, circled) (133). The lack of a triple G motif in the loops of the CHIKV packaging signal, combined with the data that CHIKV capsid can recognize and use the CHIKV, VEEV, and SINV packaging signals suggest that CHIKV capsid is capable of recognizing a triple pyrimidine motif in order to package genomic RNA (28).

The third known structure our analysis identified was the top of the TCR, a large stem-loop 3′ of the nsP3 coding region, canonically following an opal stop codon. This RNA secondary structure was predicted to be conserved across alphaviruses and can be found in other viruses and organisms in general that contain an opal stop codon (32). The CHIKV TCR was modeled recently by Kendra et al. from SHAPE data generated from in vitro transcribed RNA segments (33). This model differed slightly from past predictions that used sequence analysis of closely related alphaviruses (12, 32). Models generated in silico for the TCR of VEEV, EEEV, and SINV clade viruses predicted an 8- to 12- nucleotide spacer between the opal stop codon and the base of the TCR stem-loop followed by an 11- to 12-bp stem with a one nucleotide asymmetric bulge (32). The Kendra et al. CHIKV TCR model contains a spacer of five nucleotides between the opal stop codon and the base of the TCR element with only a 3-bp stem before a large bulge, followed by a 9-bp stem (33). Our model, generated from full-length genomic RNA probed with the SHAPE-MaP technique, places the opal stop codon in a stem 11 nucleotides away from the base of the TCR, which itself contains a one-nucleotide asymmetric bulge in the first 12 bp of the stem. We propose the true base of the CHIKV TCR be with the first A:U, as shown in our model, which corresponds to the base of the second helix in the Kendra et al. model, with the 11 intervening nucleotides between the opal codon and the TCR base either single-stranded and flexible or in transitory base pairs with neighboring sequence. We believe this would bring the three CHIKV TCR models into concordance.

The final known functional RNA secondary structure in CHIKV identified by our analysis was the frameshift element in the 6K coding region to generate the TF protein. This structure has been validated as functional in CHIKV and SFV (12, 33). Our frameshift element model differs from that proposed by Kendra et al. but is in accordance with the general model proposed previously for secondary structures following slippery sites (12, 33). The Kendra et al. SHAPE data were generated from a CHIKV genome fragment and predicts the spacer region involved in a helix at the base of a stem-loop. The discrepancies in this model and the previously discussed TCR model are likely a reflection of the methods used (SHAPE versus SHAPE-MaP), the sample RNA probed (in vitro- transcribed RNA fragments versus genomic RNA), and the length of the sequence modeled (78-nucleotide fragment versus the 109-nucleotide fragment composite generated from multiple windows) (33).

There are two RNA secondary structures found in other alphaviruses we either did not identify or which failed to meet our criteria for highly structured and specific in our analysis: the 5′ IFIT stem-loop and the subgenomic downstream loop (DLP). The 5′ IFIT stem-loop was first shown to be functionally important for evading detection by IFIT1 in VEEV (82). Later studies showed that this stem-loop is likely present and serves the same purpose, albeit with varied efficacy, in other alphaviruses including CHIKV (22). Our SHAPE-MaP data support the presence of this stem-loop as the first stem-loop of the CHIKV genome. Since this structure appears in the first 28 nucleotides of the genome, and our analysis relied on calculations over a rolling window of 55 nucleotides of the genome it was not captured in our analysis of highly structured and specific regions. While the subgenomic DLP has been shown to aid in replication of the SINV sgRNA during infection, this structure was not predicted to be present in CHIKV (38, 140). We also do not find an RNA secondary structure similar to the SINV DLP within the first 200 nucleotides of the CHIKV capsid coding region. This supports the hypotheses that CHIKV has an alternative mechanism for translating the sgRNA during infection.

Cocirculation of 3′ UTR variants

When CHIKV was introduced to the Western hemisphere in 2013, a bottle-necking event occurred in which a duplication in the 3′ UTR of CHIKV became fixed in the population (49, 138). The additional duplication in the 3′ UTR set the American strain of CHIKV apart from the parental Asian strain, which itself harbored multiple duplications from the predicted ECSA parental strain (48). We confirmed by

3′ RACE that our virus stock consisted of a mixed population of viruses that possessed the duplication 3′ UTR, as well as the canonical and a deletion 3′ UTR variant (Figure 2.5).

We confirmed previous reports that the 3′ UTR duplication event enhanced replication in mosquito cells but found that, conversely, the deletion inhibited replication compared to the canonical and duplication 3′ UTR (Figure 6 A) (48-50). It was known that the 3′ UTR variants had no effect on virus replication in mammalian tissue culture, but the impact of these specific 3′ UTRs on in vivo replication and pathogenesis had not been assessed. We saw no effect on virus replication, dissemination, or pathogenesis in our mouse model of CHIKV disease (Figure 2.6 C and D). This differs from previous studies that suggested deletions of the 3′ UTR increased virus fitness compared to the WT 3′ UTR isoform. However, those studies evaluated virus RNA persistence in a 12-day-old CD1 mouse model (48), while our studies evaluated acute viral replication and pathogenesis in 6-week-old adult C57BL/6J mice.

3′ UTR variant structure

Given the impact of the 3′ UTR variants on replication in mosquito cells, we used SHAPE-MaP to analyze the RNA secondary structure of each of these variants. Of note, the extra sequence found in the duplication 3′ UTR is composed of highly reactive nucleotides. Instead of creating an additional 177 nt with similar reactivity pattern, these nucleotides are more reactive, or flexible, than any other nucleotides found in the other two 3′ UTRs or elsewhere in the duplication 3′ UTR. Differing reactivities of individual repeat elements, which reflect different levels of structured versus unstructured RNA in the 3′ UTR (Figure 2.7 and 2.8) may affect RNA accessibility to host factors in the context of infection. Importantly, the additional sequence in the duplication 3′ UTR created a second copy of a two-hairpin motif found only once in the canonical 3′ UTR. This same motif is completely absent in the deletion 3′ UTR. The copy number of this secondary structure corresponds with the replicative fitness observed in mosquito cells by each 3′ UTR variant virus. The repeat of this two-stem-loop structure is reminiscent of the 3′ UTR structures found in SINV that have been implicated in enhanced SINV replication in insect cells (51). However, future work will need to be done to assess whether the mosquito cell replication phenotypes in CHIKV are due to nucleotide sequence, copy number of novel RNA secondary structures, or the length and flexibility of unstructured RNA present in the 3′ UTR. This may influence accessibility to host factors interacting with the primary sequence motifs within the duplicated 3′ UTR.

The variability in 3′ UTR structure and the consequences of this variation on virus replication in one host but not the other suggests that the 3′ UTR is a flexible part of the genome used to aid in host switching. A similar phenomenon has been reported in dengue virus and related flaviviruses (141). The 3′ UTR of mosquito-borne flaviviruses contain stem loops resistant to Xrn1 degradation, termed xrRNAs (141-143). Flaviviruses produced from mosquito cells have highly heterogenous 3′ UTR sequences, with often mutated or deleted xrRNAs. These mutated variants exhibit replication advantages in mosquito cells. However, when these viruses are transmitted to the vertebrate host, the 3′ UTR diversity collapses to a nearly singular 3′ UTR variant with multiple stable xrRNAs. This is because at least one functional xrRNA significantly enhances virus replication in vertebrate cells (53, 54).

The opposite appears to be true for CHIKV. Replication in vertebrate cells generates 3′ UTR diversity, often through deletions of repeat elements by copy-choice recombination, but these deletions are deleterious for replication in mosquito cells (52). Our analysis confirms these observations and provides experimental evidence of previously predicted RNA secondary structures associated with specific repeat elements in the 3′ UTR: SL-a, SL-b, and SL-Y (52, 144). Our chemical probing data support that presence of SL-a, a large stem-loop at the start of the 3′ UTR present in all CHIKV 3′ UTR variants, and SL-b, the second stem-loop of the two-hairpin motif found to be deleted or duplicated, and the forked stem loops predicted for SL-Y (Figure 2.8; see also nucleotides 11715 to 11753 and 11899 to 11937 in online supplemental data). Finally, our data also suggests these structures are separated by large spans of unstructured regions (52, 144). Our data in combination with prior studies of the 3′ UTR, strengthens the hypothesis that the 3′ UTR contains RNA secondary structures and sequences that are functionally important for efficient host switching. Duplicated RNA secondary structures and repeat sequence elements are found in other alphavirus 3′ UTRs, and there is some evidence that other alphaviruses generate heterogenous 3′ UTRs in a host-dependent manner like CHIKV (51, 145). Future studies should focus on the host specific function of each structure and begin to tease apart the relationship between the RNA secondary structure and underlying sequence.

CHIKV RNA structure considerations

We now understand that RNA secondary structures are not conserved across the alphavirus family aside from a few important structures (133), and it remains difficult to identify novel functional RNA

secondary structures in the broader context of RNA viruses generally (93, 94, 137, 146). We proposed identification of low Shannon entropy regions within a specific virus genome as a method to identify novel functional RNA elements. Our study supports the idea that functionally important RNA secondary structures can be identified by determining the most stably structured regions of a genome after SHAPE-MaP analysis, as we identified the four known important secondary structures and 19 novel structures in CHIKV with this strategy. Future studies will focus on assessing the 19 uncharacterized structured regions for functional importance using structure disruption strategies similar to those done in this study to assess the 5′ end of the genome (Figure 2.3). However, these types of analyses will require multiple mutagenesis and phenotyping strategies for each structure and are beyond the scope of this study.

New methods assessing the tertiary structure of RNAs, such as RING-MaP and SPLASH, offer additional strategies when looking for functional RNA elements (137, 147). These strategies can more accurately identify longer range nucleotide interactions than SHAPE-MaP, where our analysis was limited to structures with a maximum pairing distance of 500 nucleotides. Tertiary structure and long-distance RNA-RNA interactions within a single RNA molecule are also likely to be more transient and, in terms of RNA viruses, highly dependent on the stage of replication and interacting proteins (148-150). These considerations also illustrate the need for further refinement of methods to assess RNA secondary structure in cells. Therefore, while our models of RNA secondary structures for virion-derived genomic RNA provide an important resource, there are likely other viral RNA conformations that occur within the infected cell (66, 137). Using sequence-based strategies like synonymous site conservation, which was designed to identify functional elements in RNA with coding constraints, in combination with experimentally informed RNA secondary structure models may also aid in identification of functional RNA elements (151). However, sequence-based methods often require dozens of sequences of the same RNA to be reliable. This would not be useful for newly emerged viruses, where few sequences are available, or for regions of viral genomes with little to no coding capacity, or with extensive overlapping reading frames.

The ability to generate experimentally informed RNA secondary structure models of long RNAs is advantageous for known and recently emerged viruses. Future work should be done to identify characteristics of known functional secondary structures so that they can be used to predict functional

importance in novel structures. These characteristics can help prioritize novel structures for experimental testing. This will be especially helpful among structurally divergent but related RNAs, like virus genomes, and newly emerged viruses for which few genome sequences are available.

2.6 Methods

SHAPE-MaP of CHIKV genomic and 3' UTR RNA. CHIKV genomic RNA was extracted from sucrose-purified virions produced from Vero81 cells with TRIzol according to the manufacturer's protocol. Individual 3′ UTRs were transcribed in vitro. For SHAPE modification, 2µgof virion RNA was incubated at 37°C for 15 min in folding buffer (110 mM HEPES [pH 8], 10 mM MgCl2, 111 mM KCl) and then treated with 100 nM 1-methyl-7-nitroisatoicanhydride (1M7) for 5 min at 37°C. Negative-control RNA was incubated with 5µlof dimethyl sulfoxide in place of 1M7. The denatured control RNA sample was incubated at 95°C for 2 min and then treated with 100 nM 1M7 for 2 min at 95°C. The treated RNA was purified over Zymo RNA Clean and Concentrator-5 columns (Zymo). A 500-ng portion of Random Primer 9 (NEB) was added, followed by incubation at 65°C for 5 min, and then the sample was placed on ice. Reverse transcriptase buffer (10 mM deoxynucleoside triphosphates, 0.1 M dithiothreitol, 500 mM Tris [pH 8.0], 750 KCl, and 500 mM $MnCl_2$) was added, and the reaction mixture was incubated at 42°C for 2 min before the addition of 200 U of SuperScript II (Thermo Fisher Scientific), followed by incubation at 42°C for 180 min. The reaction was heat inactivated at 70°C for 15 min, and RNA was purified over G-50 columns (GE Healthcare). The NEBNext mRNA second strand synthesis module (NEB) was used to generate double-stranded cDNA according to the manufacturer's protocol. The double-stranded cDNA was fragmented, tagged, amplified, and barcoded using a Nextera XT DNA library preparation kit (Illumina) according to the manufacturer's directions. Excess oligonucleotides, primer dimers, small library fragments, and nucleotides were removed with Agencourt AMPure XP beads (Beckman Coulter) at a DNA-to-bead ratio of 0.6:1. Library size was determined by using a 2100 Bioanalyzer (Agilent Technologies) and quantified with a Qubit fluorimeter (Thermo Fisher Scientific). Sequencing was performed on a MiSeq desktop sequencer (Illumina).

SHAPE data processing. Reads from the untreated RNA control were aligned to CHIKV reference sequence AF369024.2 using bowtie2 (v2.2.3) to generate a reference sequence (135). The ShapeMapper pipeline (v1.2) was used to map SHAPE reactivities to the CHIKV genome (92). Because

of low coverage for nucleotides 11,400 to 12,012 with the CHIKV genomic RNA, the SHAPE reactivities for this region were calculated with additional data from *in vitro*-transcribed RNA. Default parameters for the ShapeMapper pipeline were used except for a maximum insert size of 1,000, and for the 3′ end, a minimum map quality of 20. The mean SHAPE reactivity and standard error for each nucleotide of the genomic RNA and individual 3′ UTRs tested are reported in online supplemental material. Median SHAPE values were calculated over a rolling 55-nucleotide window. Generally, SHAPE reactivities below 0.4 indicate likely paired bases, and SHAPE reactivities above 0.8 indicate likely unpaired bases. We were unable to determine a quantitative odds ratio for paired and unpaired nucleotides at these specific SHAPE reactivity thresholds (68, 69). Odds ratio calculations require a known RNA secondary structure, such as rRNA which was not present due to using purified virion particles for our genomic RNA source. Base pairing probabilities for the whole genome and each individual 3′ UTR were obtained with Superfold v1.0 using the RNA structure software suite (v5.8.1), with a maximum pairing distance of 500 nucleotides (92, 99). SHAPE data were used as a folding restraint. Superfold was also used to find Shannon entropies of base pairing at each position. Regional entropies were generated by finding the median Shannon entropy over a 55-nucleotide rolling window. Highly structured regions were defined as regions with low median Shannon entropy and low SHAPE, as in Siegfried et al. and Smola et al. (65, 92). In total, 23 structured regions were found within the CHIKV genome. Structures for these regions were extracted from the whole-genome structure obtained with Superfold.

To compare the SHAPE reactivity among the 3′ UTR variants in Figure 2.7, the rolling median SHAPE values were calculated in reference to the median SHAPE reactivity of the whole-genome CHIKV SHAPE excluding the 3′ UTR to create a common reactivity scale. Consequently, the denominator when calculating the median SHAPE reactivity for each 3′ UTR was the average SHAPE reactivity of the CHIKV genome from positions 1 to 11,301 (the 3′ UTR begins at position 11,302).

Sequence conservation. Sequence conservation analysis used the multiple sequence alignment generated in Kutchko and Madden et al., as well as the same method of calculating conservation scores (133). Data were smoothed by calculating the median conservation score over a rolling 55-nucleotide window.

3′ RACE analysis of CHIKV genomic RNA. 3′ RACE of the CHIKV genome was performed on purified genomic RNA using the RLM-RACE kit (Ambion) according to the manufacturer's directions. Briefly, RNA was reverse transcribed using the 3′ RACE adapter (5′ -GCGAGCACAGAATTAATACGACTCACTATAGGT12V N-3′). 3′ UTR-specific PCR was performed with the 3=RACE adapter outer primer (5=-GCGAGCACAGAATT AATACGACT-3′) and 3′ UTR gene-specific forward primer (5=-CTTGACAACTAGGTATGAAG-3=) recognizing CHIKV genome position 11,302. PCR products were resolved on a 1% agarose gel and gel purified before cloning into the pCR-Blunt vector (Thermo Fisher). Multiple transformants of each PCR amplicon were sequenced, and the 3′ UTR variants were cloned into the infectious clone of the Caribbean CHIKV isolate using Gibson Assembly. In each case, the infectious clone was re-sequenced to ensure the absence of unintended mutations.

Generation of an infectious clone of the CHIKV Caribbean isolate. The full-length cDNA clone of the early outbreak Caribbean CHIKV isolate virus was assembled from the consensus nucleotide sequence of purified genomic viral RNA (MG208125) (34). Michael Diamond (Washington University) provided the clinical isolate sequenced for construction of the infectious clone. The isolate was originally obtained on the island of St. Martin during the 2013 outbreak and was banked at the World Reference Center for Emerging Viruses and Arboviruses (University of Texas Medical Branch). This virus was amplified on Vero cells three times prior to receipt by our lab and amplified once on C6/36 cells before sequencing. Briefly, 14 dsDNA gBlocks were synthesized by IDT to span the entire 12,012nt genome. gBlocks were assembled into 5 overlapping genomic fragments, and each cloned into the pCR-Blunt vector (Thermo Fisher). Each fragment was PCR amplified and assembled by ligation using unique restriction sites in the CHIKV genome. A unique SacI restriction site was included upstream of the SP6 promoter in fragment 1, and a unique NotI site was included in the fragment downstream of a poly(A) sequence. The assembled CHIKV genome was inserted into the SacI and NotI sites of plasmid pSinRep5 (Invitrogen). The sequence of the full-length clone and each 3′ UTR variant clone was confirmed by Sanger sequencing (GenBank accession no. MT228631, MT228632, and MT228633).

Cells and viruses. Vero81 cells were cultured in DMEM (Gibco) supplemented with 10% heat-inactivated fetal bovine serum (FBS) and 0.2 mM L-glutamine (Gibco). BHK-21 cells were cultured in αMEM (Gibco) supplemented with 10% FBS and 0.2 mM L-glutamine. The mosquito cell line C6/36 was

cultured in Leibovitz L-15 media (Corning/Cellgro) supplemented with 10% FBS, 10% tryptose phosphate broth (Sigma), and 0.2 mM L-glutamine. A low-passage-number isolate of Caribbean CHIKV was propagated in C6/36 cells, and infectious supernatants were collected and purified over a 20% sucrose cushion.

The St. Martin CHIKV infectious clone was used for 3′ UTR replication and pathogenesis studies (GenBank accession no. MG208125) (34), and the 181/25 infectious clone was used for 5′ end structure studies (GenBank accession no. EF452494) (76). Clonal virus pools made from infectious clones were generated by linearizing the infectious clone plasmid and in vitro transcribing full-length capped genomic RNA using mMessage mMachine SP6 transcription kits (Ambion).

1.0 x 10^7 BHK-21 cells were electroporated (850 V, 25 µF, three pulses) in a 4-mm gap cuvette (Bio-Rad) with 10 µg of RNA after being washed three times with PBS lacking Ca2+ and Mg2+. Cells were recovered in maintenance media. Supernatants with virus were harvested 24 h later at peak titer. Cell debris were pelleted by centrifugation at 1,000 rpm for 10 min at 4°C. Single-use aliquots were made and stored at –80°C.

Structure-disrupting mutations. *CodonShuffle* was used to generate mutant sequences within the coding portion of the virus genome that maintain amino acid sequence and nucleotide composition (102). The dn231 algorithm was used, which also preserves dinucleotide frequency. Because CodonShuffle generates many possible mutant sequences, the final mutant sequence for a region was selected to maximize structural disruption in that region while maintaining similar codon usage frequencies within the virus. Synthetic DNA fragments (IDT) containing selected mutations and two unique restriction sites were incorporated into an infectious clone of the 181/25 (TSI-GSD-218) vaccine strain of CHIKV (76) by Gibson assembly (NEB) (see Appendix B).

***In vitro* analysis of virus replication.** C6/36 mosquito cells and Vero81 monkey kidney cells were infected at a multiplicity of infection (MOI) equal to 0.01 in biological triplicate. Supernatant samples were collected at indicated times and stored at –80°C until titering. Virus titers of cell culture supernatants were quantified by plaque assay on Vero81 cells after samples were diluted in 1x PBS (Gibco) with 1% FBS and Ca^2+/Mg^2+. Cells were overlaid with 1x αMEM with 5% FBS, 0.2 mM L-glutamine (Gibco), 1 mM HEPES (Corning), 1% penicillin-streptomycin (Gibco), and 1.25% carboxymethylcellulose sodium

(Sigma). Virus was allowed to plaque for 48 h (SM CHIKV) or 72 h (181/25) before monolayers were fixed with 4% paraformaldehyde, rinsed, and stained with crystal violet (0.25%; VWR). Data were analyzed by two-way analysis of variance (ANOVA) with Tukey's multiple-comparison test in Prism8 (GraphPad Software).

Infectious centers assay. A total of 1.0 x 10^7 BHK-21 cells were electroporated (850 V, 25 μF, three pulses) in a 4-mm gap cuvette (Bio-Rad) with 10 μg of virus genomic RNA after being washed three times with PBS lacking Ca^2+ and Mg^2+. Cells were recovered in maintenance media and serially diluted in maintenance media. Cell dilutions were plated over Vero81 monolayers overlaid with 1x αMEM with 5% FBS, 0.2 mM L-glutamine (Gibco), 1 mM HEPES (Corning), 1% penicillin-streptomycin (Gibco), and 1.25% carboxymethylcellulose sodium (Sigma). Plaques were allowed to form for 72 h before monolayers were fixed with 4% paraformaldehyde, rinsed, and stained with crystal violet (0.25%; VWR). Data were analyzed by one-way ANOVA with Tukey's multiple-comparison test in Prism8 (GraphPad Software).

Western blotting for viral proteins during infection. Vero81 cells were infected at an MOI of 5 with either 181/25 or mutant virus and incubated for 1 h with rocking every 15 min. After the incubation, cells were washed three times with PBS, and the medium was replaced. Cellular lysates were made at indicated times in radioimmunoprecipitation assay buffer. First, 5 μg of total protein was loaded and separated on a 4 to 20% gradient TGX precast protein gel and transferred to a polyvinylidene difluoride membrane. The membrane was then blocked overnight with 5% milk in PBST and probed with primary antibodies for 1 h to overnight (mouse anti-nsP3 1:1,000 and mouse anti-E2 1:500 in 5% milk PBS-T). Secondary antibodies were incubated for 1 h at room temperature in 5% milk with 0.01% SDS in 1x TBST on a rocker. Membranes were washed three times with 1x TBST for 10 min each wash. The membranes were then washed three times in 1x TBS for 10 min each time. The membranes were visualized with the Odyssey infrared Imaging system (Li-Cor).

Densitometry analysis. Bands were quantified for E2, nsP3, and actin using ImageJ software (National Institutes of Health; v1.53a). Band densities for individual proteins were normalized to actin loading control densities. Fold change in expression was measured relative to WT expression of E2 or nsP3 at 8 h post-infection.

qRT-PCR for detecting virus genome during infection. Vero81 cells were infected at an MOI of 5 with either 181/25 or mutant virus for 1 h with rocking every 15 min. Following the incubation, cells were washed three times with PBS, and the medium was replaced. At the indicated times post-infection, the medium was removed, and the cells were washed once in PBS before being lysed in TRIzol (Life Technologies) for total RNA isolation. RNA was purified according to the manufacturer's protocol. qPCR was performed on RNA using an iTaq Universal Probes one-step kit (Bio-Rad) and primers and probe specific to either the 18S rRNA gene or the CHIKV nsP1 gene. Standard curves of both mammalian 18S cDNA and 181/25 infectious clone were run in parallel with samples for absolute quantification of the gene copy number. All reactions were run in 96-well plates on an ABI 7300 real-time PCR machine in technical duplicate. Data were analyzed by multiple *t* tests with Holm-Sidak correction in Prism8.

In vivo analysis of virus replication and pathogenesis. All animal studies were done following IACUC-approved protocols under the supervision and scrutiny of University of North Carolina veterinarians. The C57BL/6J mice that were utilized in this study were bred at UNC after breeding pairs were purchased from Jackson Laboratories. Animals were allowed to age to 6 weeks before use. Animals were inoculated with 100 PFU of virus in 10 µl of vehicle (PBS with 1% FBS and Ca^2+/Mg^2+). The inoculation was given as a subcutaneous injection in the left hind footpad. For analysis of footpad swelling, the footpad width was measured daily for 1 week using calipers (152). Data were analyzed by two-way ANOVA with Tukey's correction for multiple comparisons using Prism8 (GraphPad Software). To quantify infectious virus levels in tissues, infected animals were sacrificed on day three after infection, and tissues were harvested into Vero81 media containing sterile glass beads. After weighing, the tissues were homogenized, and infectious virus was quantified by plaque assay on Vero81 cells as done for analysis of virus replication in vitro. Data within each tissue were analyzed by one-way ANOVA with Tukey's correction for multiple comparisons using Prism8.

Data availability. Viruses and materials used in this study will be provided upon request. Nucleotide sequences for the CHIKV 3′ UTR variant infectious clones have been deposited in GenBank under accession numbers MT228631, MT228632, and MT228633. SHAPE-MaP data are available online in SNRNASM format (https://docs.google.com/spreadsheets/d/1OrnU4lmvytfHhv-nh47PHyUxMdiHc -

hDe0OS0r4o9dA/edit?usp=sharing). Figure S1 is available online in a single PDF document (https://drive.google.com/file/d/1ZlrFGYxsrUeRE0QcUB_rWPirpLki4C6F/view?usp=sharing).

CHAPTER 3: APPLICATION OF SHAPE-MaP TO OTHER RNA VIRUSES

3.1 Overview

Zika virus (ZIKV) is a mosquito-borne flavivirus that has recently been associated with severe birth defects, and therefore represents a significant emerging threat to human health. Secondary structure within positive sense RNA genome of ZIKV and other flaviviruses is known to be important for multiple stages in the viral lifecycle. However, most of this analysis has focused on the viral 5′ and 3′UTRs while RNA structure within other regions of the genome has largely gone unstudied. Therefore, to define the RNA structure landscape of the ZIKV genome, we used selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) to identify secondary structures throughout the ZIKV genomic RNA. This analysis found high levels of RNA structure distributed throughout the ZIKV genome, including 19 structured regions that exhibited significantly higher than average levels of structural specificity. Structures in the 5′ and 3′UTR that have previously been identified in ZIKV and other flaviviruses were among the 19 regions with significant structural specificity. Mutational analysis of a selection of these structures demonstrated some are essential for virus replication and others impact induction of virus-induced disease. Importantly, we also determined that the coding region of the ZIKV genome is highly structured, including 17 novel, highly structured RNA regions. Therefore, this analysis demonstrates that the ZIKV genome is highly structured and provides the field with a resource for understanding how RNA structure impacts ZIKV replication and pathogenesis.

3.2 Introduction

Zika virus (ZIKV) was identified in Uganda in 1947, and since it's discovery, ZIKV had only been associated with mild human disease, including symptoms such as fever, rash, and mild arthralgia. However, ZIKV demonstrated its capacity to emerge outside of Africa in 2007, when it caused a significant outbreak in Micronesia, followed by its introduction into Brazil in 2015 and subsequent spread throughout the Americas (153, 154). The Brazil outbreak was associated with a new constellation of ZIKV disease outcomes, including an alarming capacity to cause microcephaly and other severe

neurologic developmental defects in infants born to ZIKV infected mothers. These defects are collectively referred to as congenital Zika syndrome (155, 156).

ZIKV is a positive sense single stranded RNA virus whose genome encodes a single polyprotein that is then post-translationally processed to produce the mature viral replicase and structural proteins. The majority of studies on the biology of ZIKV and other flaviviruses focus on the role of viral proteins in mediating functions such as viral RNA synthesis, viral assembly, or immune evasion. However, the viral RNA genome also plays important roles in the viral life cycle independently of its protein coding capacity. For example, while the ZIKV 5′UTR has not been functionally evaluated, a stem loop structure (SLA) within the 5′UTR is highly conserved among flaviviruses. SLA was shown to be required for proper virus replication in DENV due to its direct interaction with virus replication machinery (150, 157, 158). This structure interacts with the viral RNA-dependent RNA polymerase, NS5A, to promote viral RNA synthesis (159-161). Complimentary sequences sequestered in conserved stem loops in the 5′ and 3′UTRs are also necessary for genome cyclization prior to RNA synthesis (162, 163). Many flaviviruses have a conserved 3′ UTR pseudoknot structures capable of inhibiting degradation of the 3′ UTR by exoribonucleases, termed exoribonuclease resistant RNAs (xrRNAs). These xrRNAs produce subgenomic flavivirus RNAs (sfRNAs) which play important roles in regulating flavivirus replication, host range, and immune evasion (141, 143, 164). However, while these analyses of the flavivirus 5′ and 3′ UTR illustrate the importance of viral RNA secondary structure in flavivirus replication and pathogenesis, relatively little is known about whether RNA structures exist in other parts of the flavivirus genome, or whether these structured RNA regions are important for the replication or pathogenesis of ZIKV.

In order identify structured and unstructured regions throughout the ZIKV genome, we used selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) to analyze the full-length genome of ZIKV. SHAPE-MaP uses both chemical structure probing and bioinformatics to model experimentally tested RNA secondary structures (65, 92). This technique has advantages over other RNA analysis techniques because it allows for probing of long RNAs, such as virus genome RNAs, and the chemical probing aspects of SHAPE-MaP provide experimental support of the modeled RNA structures (146).

SHAPE-MaP analysis of ZIKV confirmed the presence of SLA within the 5′UTR of the viral genome, while also identifying xrRNA structures within the 3′UTR. The xrRNA structure models for ZIKV agree with the crystal structure of xrRNAs of related flaviviruses (142, 143). We also demonstrate that these structures are functionally important for ZIKV replication and pathogenesis, demonstrating the utility of SHAPE-MaP for identifying biologically relevant RNA structures within RNA virus genomes. Furthermore, our analysis identified structured and unstructured regions distributed throughout the genome, including 17 novel, highly specific, RNA structures distributed throughout the viral polyprotein coding region. These results demonstrate the importance of RNA secondary structure during ZIKV replication and infection and provide a resource of all RNA secondary structures in the ZIKV genome for the field to reference.

3.3 Results

Stable secondary structures of the ZIKV genome. Analysis of RNA secondary structure within the genome of ZIKV specifically has focused on the viral 3′UTR and the sfRNAs, while studies in other flaviviruses have identified highly conserved stem loops within the 5′UTR which are essential for viral RNA synthesis (150, 157, 158, 165). However, relatively little is known about RNA structure throughout the rest of the genome. Therefore, to identify structures and unstructured regions within the ZIKV genome, we treated virion-derived genomic RNA from the H/PF/2013 strain of ZIKV with 1M7 SHAPE reagent, which reacts readily with flexible, and likely unpaired, nucleotides of RNA. Modification by 1M7 induces mutations during reverse transcription of cDNA, which are then detected by high throughput RNA sequencing and used to determine the SHAPE reactivity profile across the entire genome (Figure 3.1 A). The SHAPE reactivity data can then be assessed in combination with regional Shannon entropy to identify structured regions of interest (Figure 3.1 B) (92, 123). Regions with below average SHAPE reactivity are indicative of structured regions of RNA, and regions with below average entropy suggests the RNA primarily adopts a single specific conformation. Therefore, we evaluated the ZIKV genome for regions with low SHAPE reactivity and low entropy to identify regions with higher than average secondary structural specificity when compared to the rest of the genome (Figure 3.1 C).

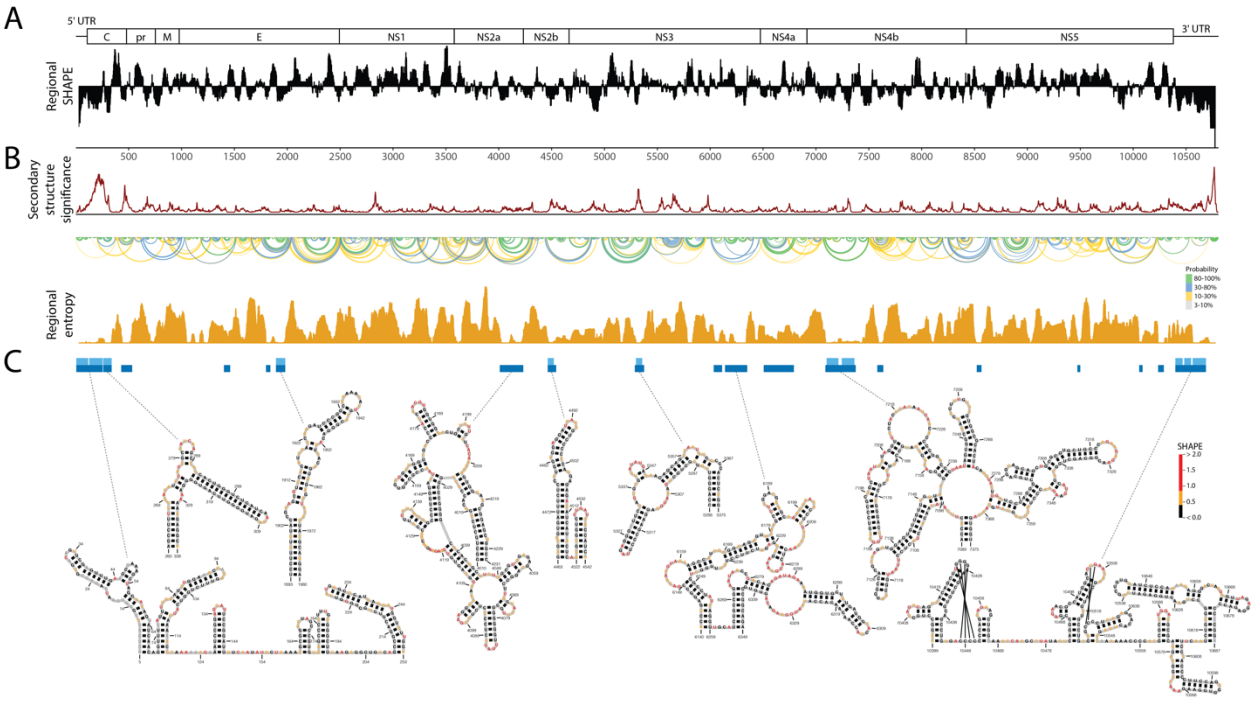


Figure 3.1: High resolution structural profile of the ZIKV genome.
(A) Windowed SHAPE reactivities across the ZIKV genome. Parts of the genome with regional SHAPE below the x-axis are more structured, while regions above the x-axis are less structured. Schematic on top represents the organization of the ZIKV genome. (B) Top: Secondary structure significance across the genome, based on the maximum squared z-score at each genomic position. Middle: Pairing probabilities between genomic positions indicated by arcs, with color representing the likelihood of pairing. Bottom: Windowed regional Shannon entropy across the ZIKV genome. Regions with low entropy adopt a single structure. (C) Secondary structure models of predicted RNA structures within the ZIKV genome. Nucleotides are colored by SHAPE reactivities. Regions were selected by having both low-entropy and low-SHAPE, represented by the blue boxes, with the light blue boxes having the highest confidence. Structures previously predicted or identified are boxed in red.

Our analysis identified 19 specifically structured regions distributed across the ZIKV genome, where 11 of these regions met our highest confidence intervals (Figure 3.1 C). These high confidence structures included RNA secondary structures within the 5′ end of the viral genome that have been previously identified in ZIKV and other flaviviruses, validating our SHAPE-MaP results (Figure 3.2 A). These include stem loop A (SLA, nucleotides 5-69) which contains a distinct Y shaped stem loop that is highly conserved across the flavivirus genus, stem loop B (SLB, nucleotides 71-117), and the capsid region hairpin (cHP, nucleotides 127-146). SLA acts as a promoter for ZIKV replication and SLB contains complimentary sequences to the 3′UTR necessary for long range interactions that promote genome replication. The cHP, a stable stem loop immediately following the start codon was found to be necessary for virus replication in DENV (150, 157, 161, 162, 166) (Figure 3.2 A). Of note, SHAPE-MaP data

suggests that the poly U sequence (73-76) often modeled as a “spacer” between SLA and SLB, is involved in the formation of the SLB stem loop (Figure 3.2 A).

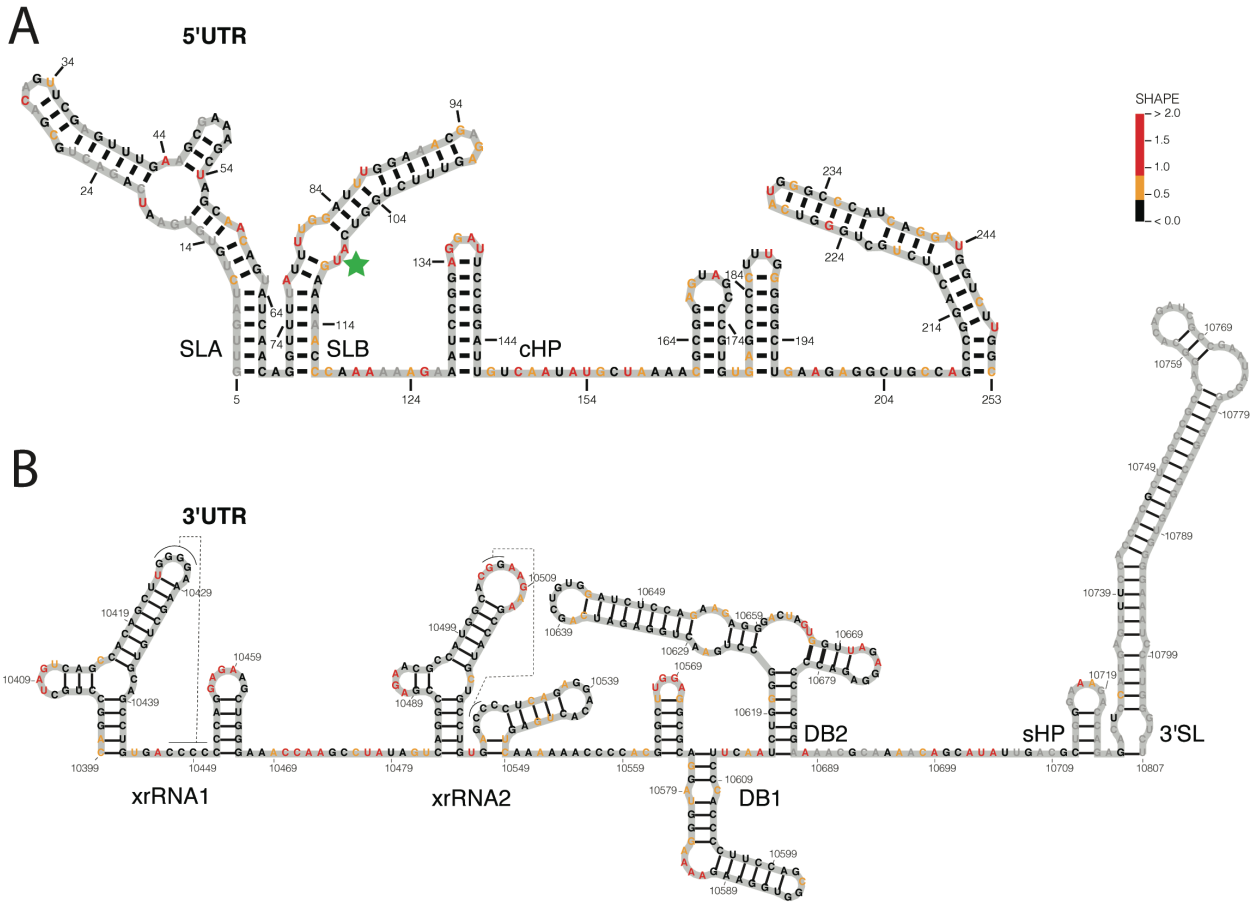


Figure 3.2: Identification of known important RNA structures validates SHAPE-MaP analysis. (A) Known and conserved functionally important RNA secondary structures at the 5′end of ZIKV genome. Green star indicates start codon for ORF. (B) Functionally important RNA secondary structures of ZIKV 3’UTR. Reactivity of nucleotides is colored according to the scale.

Analysis of RNA structures within the ZIKV 3′UTR found that much of the ZIKV 3′UTR was specifically structured, with three distinct structures falling within the category of high confidence SHAPE-MaP structures (light blue boxes, Figure 3.1 C). These are the two xrRNAs (nucleotides 10399-10466, 10483-10549), Xrn1 nuclease resistant stem loops that lead to the generation of sfRNAs during flavivirus infection (Figure 3.2 B). The third highly structured region includes the dumbbell structures (DB1 and DB2), which are also putative Xrn1 resistant RNA secondary structures (142, 164, 167).

To determine if known flavivirus RNA structures function similarly in the context of ZIKV, we mutated SLA in the 5′UTR. Three different mutation strategies were used to test the function of SLA

during ZIKV infection: overall sequence and structure (mutant “A”), the sequence alone (mutant “B”), and the loops only (mutant “C”) of SLA (Appendix B). Mutations were made in the ZIKV infectious clone used for SHAPE-MaP analysis (Figure 3.3 A). The SLA mutant RNAs failed to produce infectious virus after multiple electroporation attempts using the mosquito cell line C6/36. Wildtype (WT) ZIKV RNA was electroporated in parallel and was recovered 4 days post infection between 10^7 and 10^8 FFU/mL (Figure 3.3 B). This suggests that the sequence and exposed loops of SLA are essential for ZIKV replication similar to other flaviviruses (166).

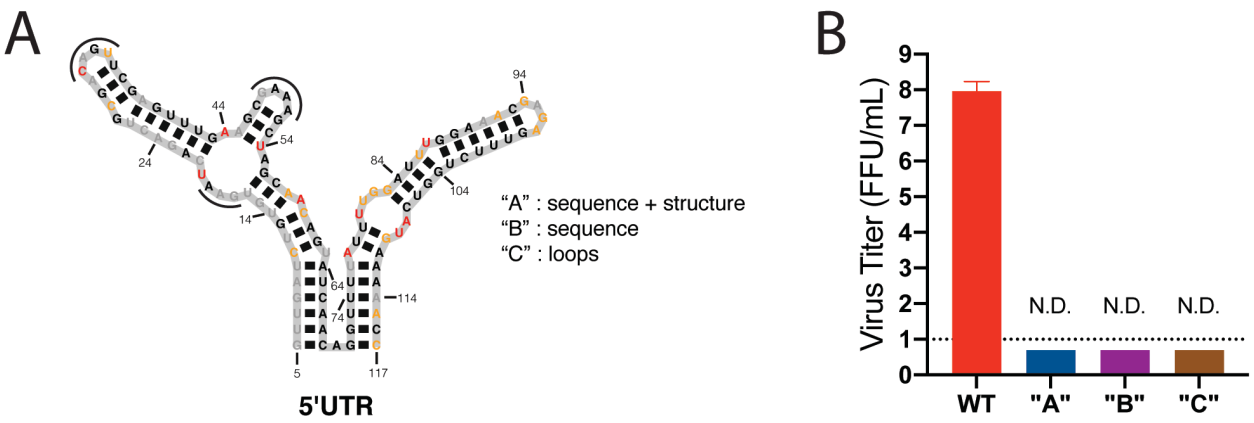


Figure 3.3: 5’UTR structure and sequence are necessary for ZIKV RNA infectivity.
(A) Mutations were introduced to the 5′UTR SLA to assess the importance of sequence and structure for virus replication. Mutant “A” disrupted sequence and structure by substituting scrambled sequence for the first 70 nucleotides of the genome, mutant “B” contained mutant sequence in SLA that maintained structure but disrupted sequence, and mutant “C” contained mutations to the loops highlighted by the black arcs. (B) WT and mutant genomic RNAs were introduced into C6/36 cells by electroporation. Virus titer in the supernatant 5 days post electroporation was quantified by foci forming assay. Data is from three independent experiments. N.D. indicates “not detected.”

Disruption of uncharacterized structure in E protein coding region. Beyond the confirmation of known structures in the 5′ and 3′UTRs, our analysis identified multiple significant structures within the protein coding region of ZIKV (n=17). We chose to disrupt one of the previously uncharacterized structures, a large specifically structured region containing a stem loop within the E coding region (1893-1980) (Figure 3.4 C). The RNA secondary structure of this region was disrupted using the program *CodonShuffle*. We chose a new primary sequence predicted to disrupt RNA secondary structure but maintain amino acid sequence and dinucleotide frequency as in previous studies (Figure 3.4 C) (102). We were careful to not alter the normal ZIKV dinucleotide frequency since flaviviruses specifically repress certain dinucleotide combinations to evade detection by innate immune sensors (168). The E structure

disruption mutant, E^{mut}, was successfully recovered after electroporation of transcribed mutant RNA into C6/36 cells. In order to test if disruption of this region impacted general virus replication or host range, we compared virus growth of E^{mut} to WT virus in Vero81 and C6/36 cells (Figure 3.4 A and B). We observed no differences in replication kinetics or peak titer output as determined by foci forming assay. We also assessed replication of the E^{mut} in the interferon competent MRC-5 cell line (Figure 3.4 D). Again, we observed no differences between WT virus and the E^{mut} virus replication kinetics or peak titer output. These data indicate the E structured region from nucleotides 1893 to 1980 is not functionally important for virus replication in vitro.

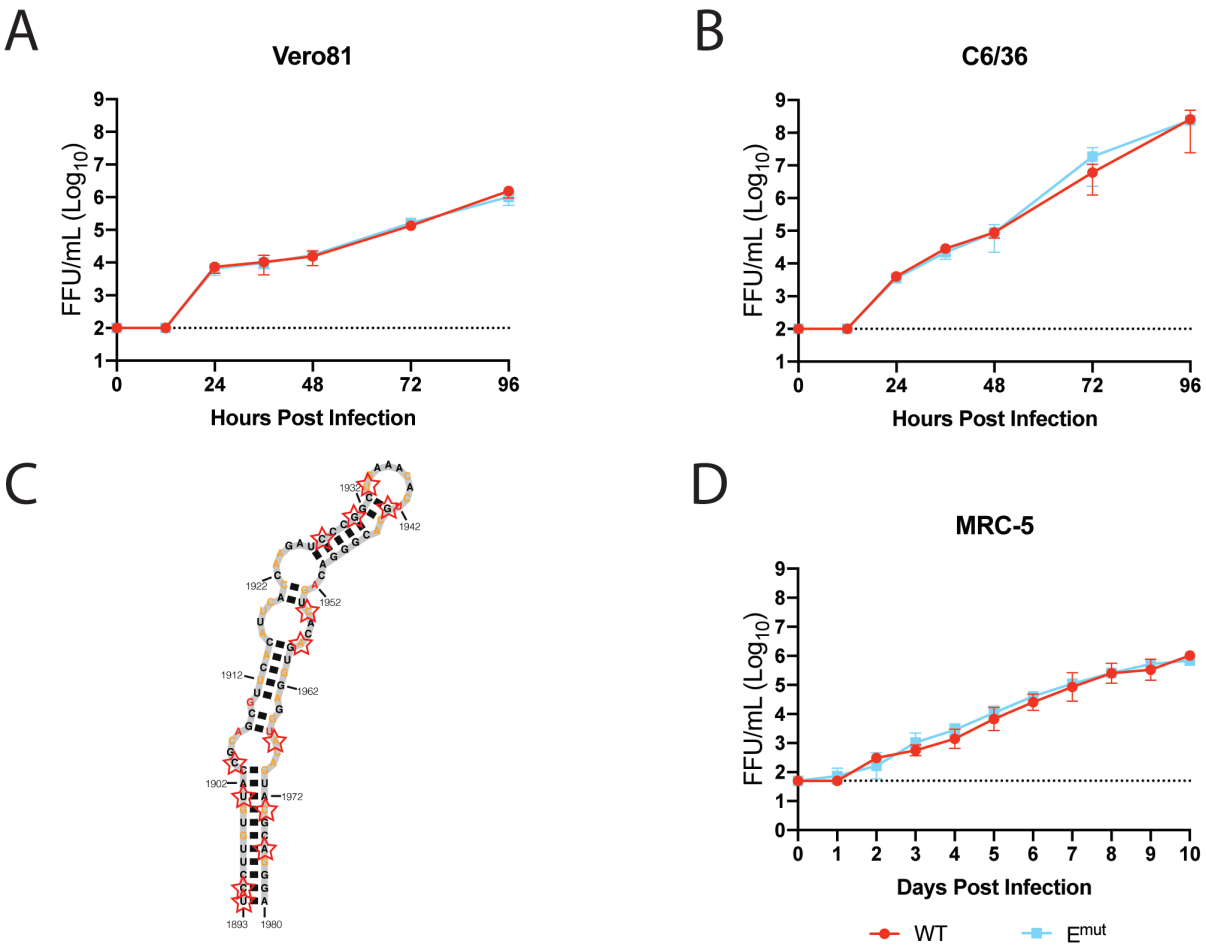


Figure 3.4: Disruption of E structured region has no effect on ZIKV replication in vitro.
(A and B) Vero81 and C6/36 cells were infected with WT and E^{mut} virus at an MOI = 0.01. Supernatants were sampled at indicated time points and infectious virus was quantified by foci forming assay (FFA). Data are from four independent experiments. (C) RNA secondary structure found from nucleotides 1893 to 1980 in the E protein coding region of ZIKV. This structure was disrupted with silent point mutations to generate the E^{mut} virus. Nucleotides that were disrupted at starred. There are 7 additional mutations made to surrounding sequence not shown. (D) MRC-5 cells were infected with WT and E^{mut} virus at an MOI =

0.01. Supernatants were samples at indicated time points and infectious virus was quantified by FFA. Data represents 3 biological replicates of one experiment.

Disruption of ZIKV 3′UTR xrRNA1. Structures in the 3′UTR of DENV and West Nile virus (WNV) have been extensively studied for their role in producing sfRNAs. Production of sfRNAs is caused by preventing the degradation of the genomic RNA by stalling the exoribonuclease Xrn1 at pseudoknot structures called xrRNAs. The 3′UTR of ZIKV contains two xrRNA structures and a third putative DB xrRNA (Figure 3.2 B). To assess the functional importance of xrRNA1 and xrRNA2, we introduced mutations to disrupt xrRNA1 (xrRNA1^{mut}) or both xrRNA1 and xrRNA2 (xrRNA1/2^{mut}), but maintain dinucleotide frequency in the 3′UTR (Figure 3.5 A) (102). These mutations were predicted to disrupt RNA secondary structure according to our SHAPE data informed RNA secondary structure model.

Mutant RNAs were introduced into C6/36 cells by electroporation. The xrRNA1^{mut} virus was recovered at a slightly lower titer than the WT control and had smaller foci than WT when quantified by FFA (Figure 3.5 B and C). We were unable to recover the xrRNA1/2^{mut} virus above the limit of detection. This suggests at least one or both xrRNAs are necessary for virus viability in mosquito cells.

We assessed the replication of xrRNA1^{mut} in both Vero81 cells and C6/36 cells (Figure 3.5 D and E). These cell lines are highly permissible to virus replication because they lack key innate immune responses to infection (169, 170). There was no difference in replication between WT ZIKV and xrRNA1^{mut}, indicating that disruption of xrRNA1 does not impact general ZIKV virus replication. To assess virus replication in a cell line with functioning innate immune responses we use the human lung fibroblast cell line MRC-5 (171). We observed lower titers of the xrRNA1^{mut} virus compared to WT virus after 5 days post infection, but these differences were not significant (Figure 3.5 F).

Past studies of sfRNA production and xrRNAs of WNV demonstrated that disruption or deletion of these structures caused less pathology during infection of neonatal mice (Pijlman 2008). To assess the role xrRNA1 structure plays during ZIKV pathogenesis, we infected C57BL/6J αβ/γ receptor knockout (αβ/γ$^{-/-}$) mice with 1,000 FFU of WT or xrRNA1^{mut} virus (Figure 3.6) (172). Due to concerns of low-level reversion of the xrRNA1^{mut} at the time of infection, we used a passage 0 (p0) and a passage 1 (p1) stock of the xrRNA1^{mut} virus. Sanger sequencing of the p0 stock 3′UTR later determined there was no reversion occurring in the 3′UTR. Sanger sequencing of separate p1 stocks generated after passaging on Vero81 or C6/36 cells indicated no reversion in the 3′UTR as well. Full genome sequencing has not been done for

the p0 or p1 stock xrRNA1^{mut} viruses. Mice were monitored for signs of disease and weight loss for 35 days after infection. Mice that reached 80% of their starting weight, became moribund, or developed dual hind limb paralysis were humanely euthanized according to UNC IUCUC protocols.

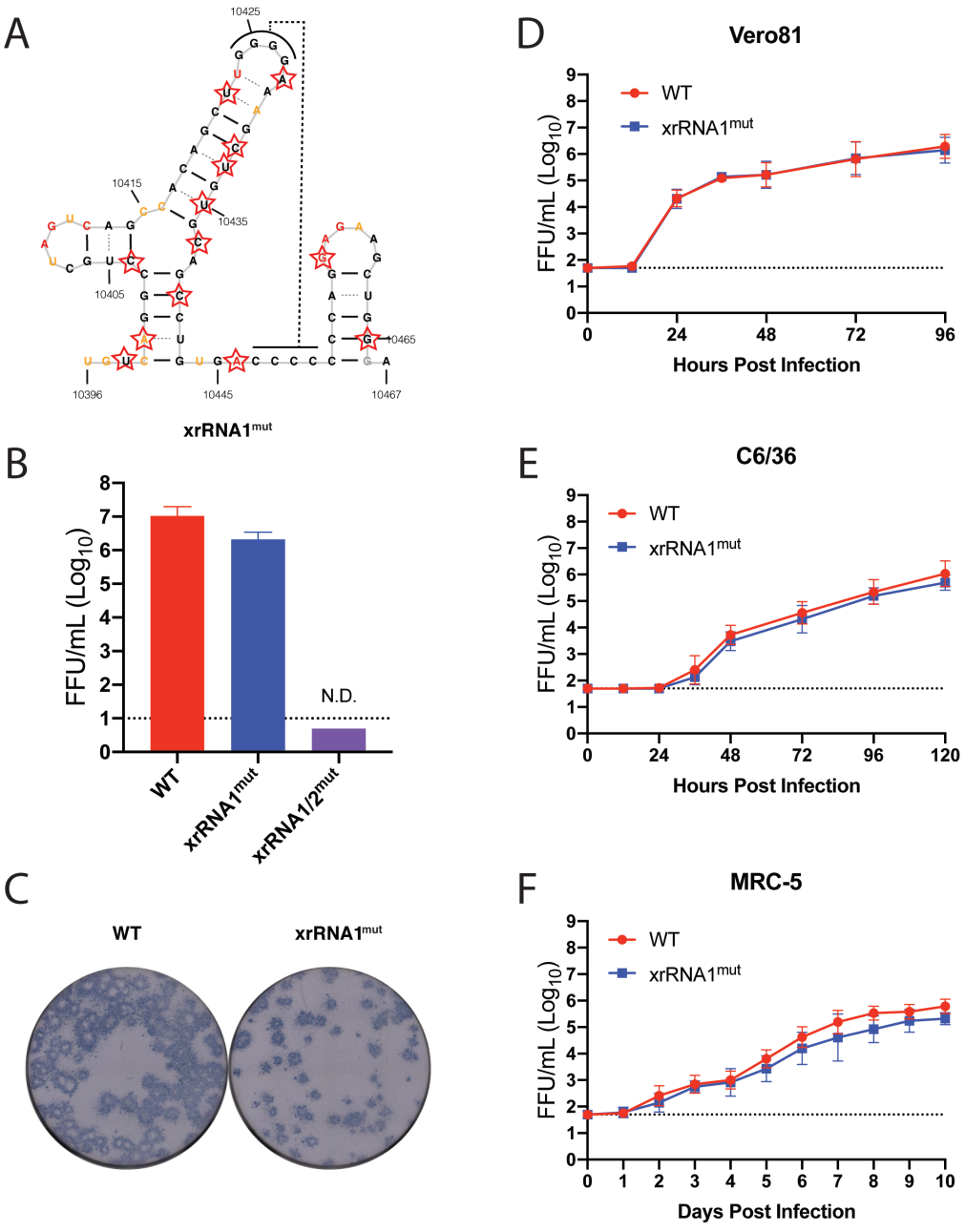


Figure 3.5: Disruption of ZIKV 3'UTR xrRNA1 structure does not impact virus replication in vitro. (A) xrRNA1 structure was disrupted by mutating starred nucleotides. (B) WT and mutant genomic RNAs were introduced into C6/36 cells by electroporation and titered by FFA. Data are from two independent experiments. (C) Representative foci staining and plaque morphology differences between WT and xrRNA1^{mut} 4 days post infection. (D-F) Vero81 (three independent experiments), C6/36 (four independent

experiments), or MRC-5 cells (two independent experiments) were infected with WT or xrRNA1^{mut} virus at an MOI = 0.01. Supernatant samples were collected at indicated time points and virus was quantified by foci forming assay. Dashed lines indicate the limit of detection. N.D. indicates “not detected.”

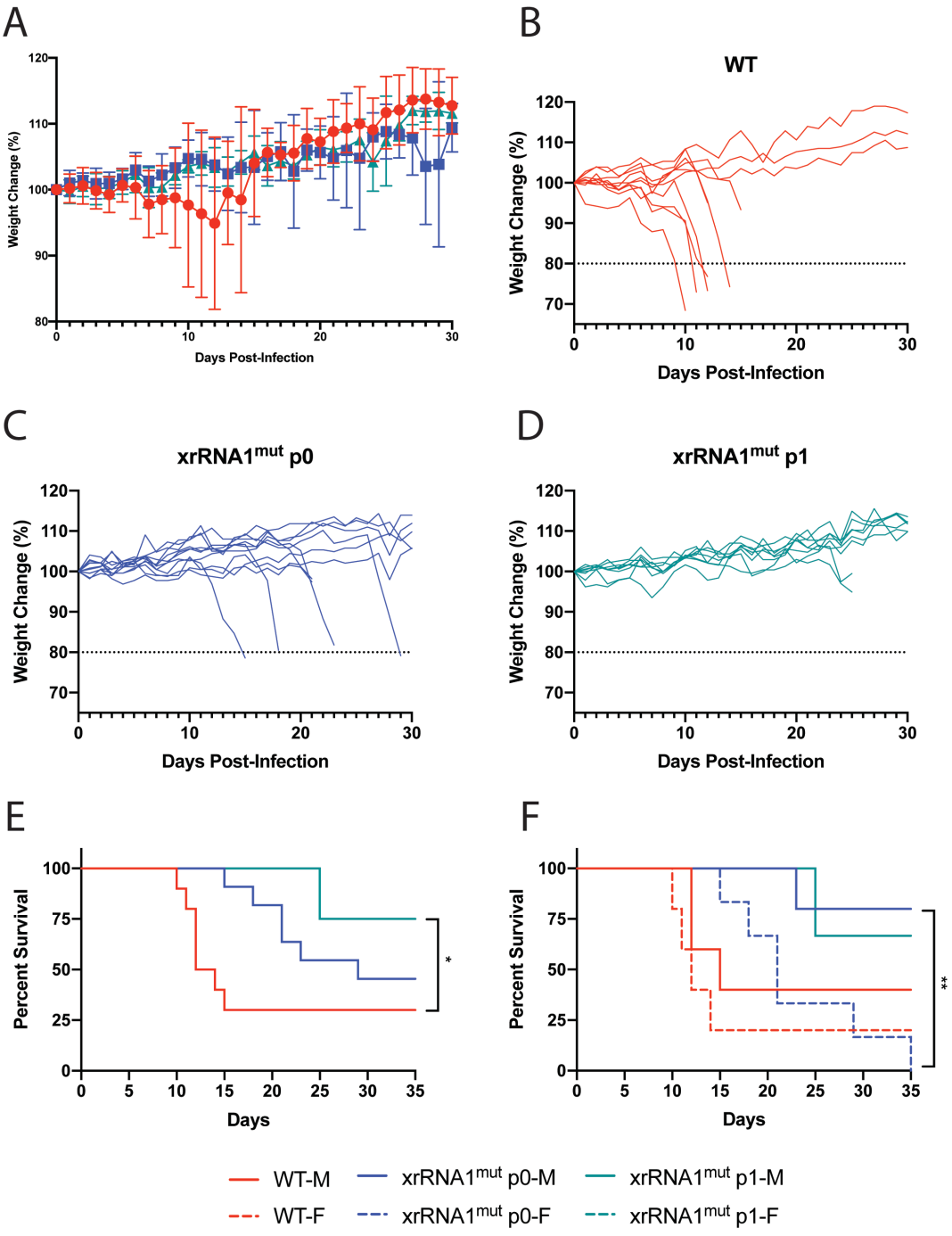


Figure 3.6: Disruption of xrRNA1 differentially impacts pathogenicity based on sex in vivo. C57BL/6J αβ/γ$^{-/-}$ mice were inoculated in the left-hind footpad with 1,000 FFU of WT, xrRNA1^{mut} p0, or xrRNA1^{mut} p1 virus in 10µlof PBS. Mice were weighed and monitored for signs of disease daily. Mice were euthanized when they dropped below 80% of their starting body weight, developed dual hind-limb paralysis, or become moribund. (A) Aggregate group weight loss for mice infected. (B). Weight loss of individual mice infected with WT virus (n = 5 male and 5 female). (C) Weight loss of individual mice

infected with xrRNA1^{mut} p0 virus (n = 5 male and 6 female). (D) Weight loss of individual mice infected with xrRNA1^{mut} p1 virus (n = 3 male and 5 female). Survival curves are plotted as group totals (E) or by sex within infection groups (F). Significance was determined by log-rank tests between indicated groups. * $p < 0.05$. ** $p < 0.01$.

About 50% of mice infected with WT ZIKV reached a humane endpoint due to weight loss and 20% of WT ZIKV infected mice reached a humane endpoint other than weight loss (Figure 3.6 A, B, and E). Mice infected with WT ZIKV succumbed to disease between 10 and 15 days post infection, while mice infected with the xrRNA1^{mut} p0 had no lethality until after 15 days post infection. Four of the five mice that reached a humane endpoint in the xrRNA1^{mut} p0 group exhibited weight loss but only two reached 80% of their starting weight (Figure 3.6 A, C, and E). Finally, the xrRNA1^{mut} p1 infected mice that succumbed to infection exhibited little to no weight loss prior to reaching a humane endpoint unrelated to weight loss (Figure 3.6 A, D , and E). There was no significant difference in overall survival between WT and xrRNA1^{mut} p0 infected mice but there was a significant difference between WT and xrRNA1^{mut} p1 infected mice (Figure 3.6 E). The differences in mortality between the xrRNA1^{mut} p0 virus and the xrRNA1^{mut} p1 virus may be due to second site mutations in the virus coding region accumulated during passaging.

Interestingly, there was a significant difference in survival based on sex for mice infected with the xrRNA1^{mut} p0 virus. No female mice in this group survived beyond 35 days after infection, while 80% of the male mice survived to the study end point (Figure 3.7 F). There were more male mice that survived the WT virus infection, though this difference was not significant. We observed no differences in survival between sexes for the mice infected with xrRNA1^{mut} p1 virus (Figure 3.7 F). Overall, this data suggests that disruption of xrRNA1 and passaging attenuates virus pathogenesis. It also indicates that disruption of xrRNA1 has sex dependent effects on pathogenesis in C57BL/6J αβ/γ$^{-/-}$ mice.

3.4 Discussion

Interest in identifying the functional roles RNA secondary structures play in flavivirus replication and pathogenesis has increased since the 2013 ZIKV outbreak. Many of the studies on flavivirus RNA structure prior to the outbreak focused on the 5′ and 3′UTRs of DENV or WNV (142, 150, 157, 162, 164, 166). These studies identify functionally important structures in the non-coding regions of flavivirus genomes, but neglect to look at internal RNA secondary structures found within the polyprotein coding region. This is due in part to limitations of past RNA secondary structure analysis techniques. Computer modeling of RNA secondary structures becomes less accurate with longer RNAs, such as a virus

genomes, and require experimental validation. Techniques like SHAPE-MaP provide experimental data and have the throughput capacity to make analysis of long RNAs feasible (92).

We used SHAPE-MaP to identify the genomic RNA secondary structures of the 2013 French Polynesian strain of ZIKV. Our analysis identified 19 specifically structured regions of the ZIKV genome. Two of these regions lie within the 5′ and 3′UTRs, but the remaining 17 fall within the polyprotein coding sequence. We confirmed the presence of SLA and SLB in the 5′UTR of ZIKV (Figure 3.2 A). The poly U sequence between SLA and SLB was previously predicted to be a “spacer” sequence in other flaviviruses and not involved in stem loop formation (148, 150, 173). Our analysis indicates this sequence forms the base of SLB in ZIKV causing SLB to be longer when compared to other flaviviruses. This also puts SLA and SLB in closer proximity. The presence of two xrRNAs in the 3’UTR followed by a dumbbell was also confirmed (147, 165). These structures are found in the 3′UTR of most mosquito borne flaviviruses (MBFVs), but the organization and number can vary (141). ZIKV differs from its closest relative, DENV, in that it only contains a single DB in the 3′UTR but has two xrRNAs similar to DENV 1-3. Viruses belonging to the YFV group of MBFVs are also predicted to contain only a single DB in their 3′UTRs (141).

In the 3′UTR, we found that disrupting xrRNA1 was not attenuating in either vertebrate or invertebrate cells in vitro. However, we could not harvest a double mutant in C6/36 cells suggesting both are needed in invertebrate cells. This contrasts with what was observed for DENV. Both are dispensable in mosquitos while necessary in vertebrate cells (53, 54). The xrRNA structures are resistant to Xrn1 degradation, leaving behind subgenomic RNAs known as sfRNAs (142, 164, 165). These sfRNAs are involved in regulating host innate immune responses during infection and pathogenesis. Previous studies that disrupted xrRNA structures in WNV showed that the mutant viruses with disrupted xrRNA structure were less pathogenic in neonatal mice (164). Similarly, xrRNA1^{mut} caused less mortality and delayed disease onset in our C57BL/6J αβ/γ$^{-/-}$ mouse model (Figure 3.6). These data provided further support that RNA secondary structures can be determinants of pathogenesis. These data also contribute to the existing body of data on other RNA viruses like YFV and VEEV that RNA secondary structure disruption is a viable attenuation strategy for future live-attenuated vaccines (82, 174, 175).

Of particular interest was our observation that disruption of the xrRNA1 structure caused differential mortality in C57BL/6J αβ/γ$^{-/-}$ mice based on biological sex. Studies have been done looking at

ZIKV replication and pathogenesis in reproductive tract tissues due to congenital Zika syndrome in infants born to ZIKV infected mothers and the ability of ZIKV to spread through sexual contact (176). However, there have been no studies directly comparing ZIKV induced pathology between male and female animals. This is likely due to WT infection having no significant difference in mortality, though we did observe that more female mice died compared to males in the same group (P = 0.284). When xrRNA1 is disrupted in ZIKV, there is a significant difference in mortality between male and female mice (Figure 3.6 F). This is of particular interest because biological sex impacts immune responses to both pathogens and vaccinations with females having a more severe response to vaccines and virus infection, in particular (177, 178).

The disruption of xrRNA1 should induce less pathology due to reduced accumulation of sfRNAs (164). On the whole, this is what we observe however the reduced pathogenesis is much more pronounced for male mice than it is for female mice when comparing WT and xrRNA1 p0 infected groups (Figure 3.6 F). In vertebrates, sfRNAs increase stability of host cell mRNAs through inhibition of Xrn1 and suppress type-I interferon (IFN) responses through mechanisms not fully defined (179-181). While C57BL/6J $\alpha\beta/\gamma^{-/-}$ mice lack type-I IFN receptors, this difference in pathology could be mediated through other signaling pathways and subsequent activation of innate immune cells. It was recently shown that DENV sfRNA interacts with TRIM25 to inhibit downstream signaling through RIG-I (182). There are known differences between males and females in gene induction during antiviral responses and in innate immune cell abundances and activation (183, 184). Female animals generally show a stronger response in all areas when compared to males (184). This may explain why the difference in mortality between female mice infected with WT or xrRNA1^{mut} viruses is less pronounced than the mortality difference between male mice infected with WT or xrRNA1^{mut} virus (Figure 3.6 F). Future work should assess if these differences in pathogenesis between male and female mice infected with xrRNA1^{mut} are due to differences in sfRNA accumulation or another mechanism. Further, studies of ZIKV pathogenesis, particularly concerning xrRNAs and sfRNA production, should assess pathology in a sex-disaggregated manner.

Our study of ZIKV RNA secondary structure adds to the growing body of RNA secondary structure data for ZIKV and closely related flaviviruses. Two other studies have assessed RNA secondary

structure of full-length ZIKV or DENV genomes (137, 185). Dethoff et al. assessed DENV RNA secondary and tertiary structures using SHAPE-MaP and RING-MaP (RNA interaction groups by mutational profiling) techniques (137). Huber et al. probed local RNA secondary structure and tertiary structures of four ZIKV strains and four DENV serotypes using SHAPE-like chemical probing techniques (185). Both studies used slightly different approaches to identify structured regions of interest, but both their studies and ours identified the known functional RNA secondary structures in flaviviruses, demonstrating that multiple approaches can be taken to investigate virus RNA structure. Dethoff et al. used a similar strategy of identifying low SHAPE and low Shannon entropy regions of DENV, resulting in 24 regions of interest (137). Some of these regions were also identified by Huber et al., who also considered sequence and structure conservation when identifying structured regions of interest in DENV and ZIKV. Huber et al. identified 12 conserved and potentially important structured regions in ZIKV, four of which overlap with regions we report in this study (nt 114-413, 6114-6438, 7016-7397, and 9732-10807) (185). Both studies in ZIKV and DENV identified functionally important long range RNA interactions for virus replication. However, Huber et al. observed fewer long range interactions of virus RNA probed in cells than in virions, suggesting that short range interactions, or secondary structures, may be more important during infection (185). Further, authors also noted that secondary structure was similar between RNA probed in cells and in virions, and proposed that many RNA secondary structure may be stable across multiple stages of replication (185). Overall there is a high level of concordance in predicted RNA secondary structures in ZIKV and regions identified by both Huber et al. and us should be prioritized for further characterization.

In conclusion, SHAPE-MaP analysis provided full genome secondary structures for ZIKV. These secondary structures are useful targets for mutation to create new live attenuated vaccines. This study contributes to a growing body of RNA secondary structure models and data for ZIKV and related flaviviruses (137, 146, 185). This study also expands the functional roles of RNA secondary structures in pathogenesis by showing that RNA secondary structures can play sex-dependent roles in pathogenesis.

3.5 Methods

SHAPE analysis. The ShapeMapper pipeline (v1.2) (92) was used to align sequencing reads to the ZIKV genome using bowtie2 (v2.2.3) (135) with a maximum insert size of 1000. Mutation rates found by ShapeMapper were used to derive the SHAPE reactivity at each position, by the difference between

the reagent and background conditions divided by the denatured control (92). SHAPE reactivities were normalized by scaling by a factor so that the distribution matched the distribution of SHAPE reactivities of other RNA viruses.

Structure modeling. The Superfold pipeline (v1.1) (92) was used to find base pairing probabilities, entropies, and a structural model for the whole ZIKV genome. A partition window size of 1500 nucleotides and a maximum pairing distance of 500 nucleotides were used. Normalized SHAPE reactivities were incorporated as a pseudo-free energy term to inform structural modeling with experimental data (71). Phylogenetic-based pairing constraints were used within the 3′ UTR to ensure correct folding of the xrRNA and dumbbell regions.

The pseudoknots within the xrRNA regions were forced as single stranded for structure predictions, which cannot predict pseudoknots, and then added to the structural model. The RNAstructure suite (v5.8.1) was used for calculating base pairing probabilities with the *partition* program and minimum free energy structures with the *Fold* program (99). Structure significance was found using the maximum squared z-score from RNASurface (v1.0), with default parameters (123). Median SHAPE reactivities and Shannon entropies were obtained using a 55-nucleotide rolling window.

Highly structured regions. Structured regions were defined as regions with both low Shannon entropies and low SHAPE. Structured regions with high confidence had median Shannon entropies lower than 0.04 and median SHAPE reactivities lower than 0.3. Expanded structured regions were found with a median SHAPE cutoff of 0.4 and expanded up to 150 nucleotides to incorporate well-defined encapsulating base pairs. Structural models for each region were found from the Superfold SHAPE-informed structural model.

Mutant ZIKV generation. ZIKV structure mutants were designed by hand or using the *CodonShuffle* program (102), with the dn231 algorithm, which conserves dinucleotide frequency (Appendix B). Synthetic DNA fragments containing the designed mutations were ordered from Integrated DNA Technologies as gBlock Gene Fragments. Fragments were inserted into the quadripartite unidirectional H/PF/2013 ZIKV infectious clone system (186) using Gibson Assembly (NEB).

WT and mutant viruses were produced as described previously (186). Briefly, plasmids were digested, genomic fragments were purified and subsequently ligated to create a full-length genomic

template for in vitro RNA synthesis. Genomic ZIKV RNA was synthesized using an mMessage mMachine T7 transcription kit. Genomic RNA was introduced by electroporation to 8 x 10^7 C6/36 or Vero cells after being washed 3x with Gibco 1x PBS. Supernatants from electroporated cells were harvested after 6 to 7 days, aliquoted in single use volumes, and stored at -80°C. Single passage virus (p1) was generated by taking p0 stocks and infecting C6/36 cells. Supernatants were harvested 4-5 days post infection, aliquoted in single volumes, and stored at -80°C.

Cell culture. Media for C6/36 cells was Leibovitz L-15 media (Corning/Cellgro) supplemented with 10% FBS, 10% tryptose phosphate broth (Sigma), and 0.2 mM L-glutamine. C6/36 cells were maintained at 28°C with no CO_2. Media for Vero-81 cells was DMEM (Gibco) supplemented with 10% heat inactivated fetal bovine serum (HI-FBS) and 0.2 mM L-glutamine (Gibco). Vero81 cells were maintained at 37°C with 5% CO_2. Media for MRC-5 cells was 1x DMEM (Gibco) supplemented with 10% HI-FBS. MRC-5 cells were maintained at 37°C with 5% CO_2.

Mutant virus characterization. Cells were infected with Wild-type or mutant ZIKV at MOI=0.01. Supernatant samples were collected at indicated times post infection. Virus was quantified by focus forming assay on Vero81 monolayers.

Samples to be titered were serially diluted in Vero81 maintenance media before addition to Vero81 monolayers. Cells were overlayed with 1x αMEM with 5% FBS, 0.2 mM L-glutamine (Gibco), 1mM HEPES (Corning), 1% penicillin streptomycin (Gibco), and 1.25% carboxymethylcellulose sodium (Sigma). Virus was allowed to plaque for 72 hours before plates were fixed with 4% PFA for 20 min. Monolayers were washed with 1x PBS, permeabilized (Invitrogen Perm Buffer) and blocked with 5% non-fat dry milk in 1x PBS. Plates were incubated with 4G2 antibody (hybridoma supernatants obtained from the Baric lab at UNC) for 1h at 37°C while rocking. Plates were washed 2x with 1x PBS and incubated at 37°C for 1h with horseradish peroxidase conjugated goat anti-mouse secondary antibody (KPL) while rocking. Plates were washed 2x with 1x PBS and finally incubated with TrueBlue peroxidase substrate (KPL) for 15 min to visualize foci. TrueBlue was rinsed from plates with distilled H_2O and plates were allowed to dry before foci were counted to determine sample titer.

In vivo analysis of pathogenesis. All animal studies were done following IACUC approved protocols under the supervision and scrutiny of University of North Carolina veterinarians. The C57BL/6

αβ/γ$^{-/-}$ mice used were bred at UNC and originally obtained from the Whitmire Lab at the University of North Carolina at Chapel Hill. Nine week old animals were inoculated with 1,000 FFU of virus in 10µlof vehicle (1x PBS with 1% FBS and Ca^{2+}/Mg^{2+}). Inoculum was administered as a subcutaneous injection in the left hind footpad. Male and female mice were used with 10 mice infected with WT virus and 11 mice infected with xrRNA1^{mut} virus. Infected mice were weighed and monitored daily for signs of disease. Mice that developed dual hind-limb paralysis, became moribund, or fell below 80% of starting weight were euthanized.

CONCLUSION

4.1 Key Findings

Since the turn of the century, there have been multiple epidemics and pandemics caused by RNA viruses, including more than a dozen caused by alphaviruses (3, 154, 187, 188). Alphaviruses, which are positive sense arboviruses, are a particularly important pathogen to address because they will almost certainly cause future outbreaks and there are no approved therapeutics or vaccines to treat or prevent alphavirus infection (10, 188). CHIKV alone caused nearly 2 million infections when it was first introduced to the Western hemisphere in 2013 and continues to cause outbreaks in Central and South America (4, 188). While many resources have been devoted to study the several virus proteins encoded in the ~12 kb genome, prior to this study there was relatively little information about the secondary structures formed by the RNA genome.

Using SHAPE-MaP we were able to develop experimentally informed RNA secondary structure models of three alphaviruses (SINV, VEEV, CHIKV) and one flavivirus (ZIKV). The models themselves are a valuable resource to the molecular arbovirology field. Previously, only a few regions of each virus had experimentally informed RNA secondary structure models, and there were no structure models that accounted for the full-length of the RNA genome. These RNA secondary structure data can inform future studies assessing protein binding specificity, RNA localization, post-transcriptional modification, translation regulation and transcription regulation since RNA structure is known to play important roles in these and other areas of biology (82, 139, 162, 164, 189-192). Further, synonymous and non-synonymous mutations, whether naturally occurring or synthetically introduced, can now be assessed with respect to the impact they may have on local RNA secondary structure. If a specific RNA structure or structured regions is suspected of functional importance, we also demonstrated a new method for disrupting local RNA secondary structure in the context of replication competent virus without impacting protein coding sequences. These data and techniques advance the strategies available to study viruses in biologically meaningful ways.

Alphaviruses are Uniquely Structured

In *Chapter 1*, we sought to identify conserved RNA secondary structures by comparing SHAPE reactivity profiles between distinct but related alphaviruses. We hypothesized the conserved RNA secondary structures would be functionally important and may play roles in replication or pathogenesis expanding our fundamental understanding of virus biology. This analysis demonstrated that alphaviruses genomes from different clades are actually structurally unique and that further RNA secondary structure analysis would have to be virus specific. However, despite the structural diversity between SINV and VEEV, we did identify and preliminarily characterize a novel functional structure at the end of nsP1 in SINV (Figure 1.3 A, structured region 1293-1382). This structured region was chosen because the region was somewhat conserved at the sequence level and moderately structured when considered in the context of the SINV genome. In vitro transcribed genomes with silent mutations disrupting this structured region were less infectious than WT genomes, however mutant infectious virions replicated as efficiently as WT virions (Figure 1.3 B &C). Since assembled infectious particles had no overt replication defect, but naked genomes did, one explanation could be that virus proteins compensate for the lost function of the disrupted structure. The same region we structurally disrupted was also identified as a SINV capsid binding site (29), suggesting that defects early during infection may be due to impaired capsid binding at this location.

An alternative, but not mutually exclusive, hypothesis is disruption of the structured region delayed deposition of a post-transcriptional modification that enhances virus replication. The nsP1 structured region contains a potential signal sequence for the post-transcriptional modification *N6*-methyladenosine (m^6A) (193, 194). Online m^6A prediction tools identify the bulged adenosine of the nsP1 S.R., nt 1377, as highly likely to be an m^6A nt in certain cell types (unpublished data). Adenosine 1377 is not mutated in our structure disruption mutant, but silent mutations around this adenosine do reduce the likelihood it is modified according to the same m^6A prediction tools (unpublished data). An additional hypothesis that would explain the difference in infectivity between naked in vitro transcribed RNA and virions is that the RNA found in virions is post-transcriptionally modified in some way.

To assess if virion derived RNA was modified, we purified virion RNA from WT and nsP1 S.R. mutant virus stocks collected at 24 and 48 hours post electroporation. Equal quantities of virion derived

RNA were electroporated into BHK-21 cells and infectious centers were quantified in the same manner as in *Chapter 1.5 Methods*. Mutant virion derived RNA collected at 24 hours post electroporation is only ~4 fold less infectious than WT virion derived RNA, while in vitro transcribed mutant RNA was ~4 log less infectious than in vitro transcribed WT RNA (Figure 4.1 and Figure 1.3 C). Further, mutant virion derived RNA collected at 48 hours post electroporation is equally as infectious as WT virion derived RNA collected at 24 hours post electroporation. This indicates that mutant virion derived RNA becomes more infectious over time. Sequencing of virus collected at 48 hours post electroporation did not reveal any reversion or second site mutations that could explain an increase in infectivity. This suggests that virions produced later after electroporation are modified in some way and modified RNAs are more abundant at later times. Flavivirus particle assembly and production has been shown to depend on the m^6A modification status of virus genomes (195). It is possible that m^6A modification is beneficial for alphaviruses and efficient deposition of this modification depends on RNA secondary structure and sequence. Future work will need to determine if disruption of the nsP1 structured region does affect post-transcriptional modification of the genomic RNA and if this impairs capsid binding to this site.

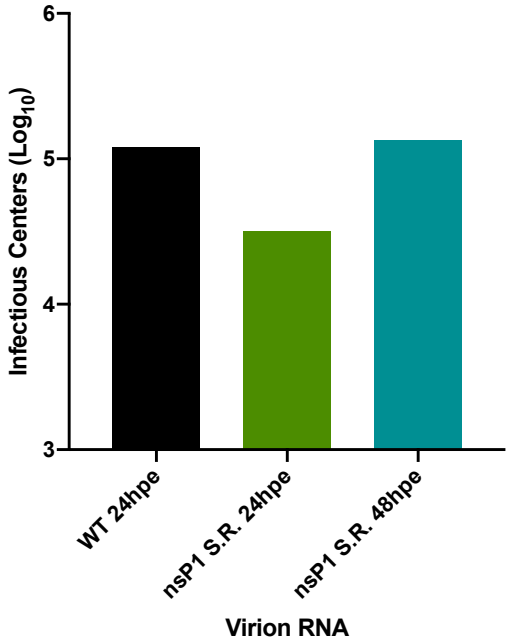


Figure 4.1: Virion derived RNA infectivity depends on time of virus collection.
Virion RNA was purified from p0 stocks harvested at 24 or 48 hours post electroporation. Purified RNA was reverse-transcribed to create cDNA, and the mutated region was sequenced for each virus to confirm there was no contamination or reverting mutations. Once confirmed, 1µgof purified virion RNA was used in an infectious centers assay as described in *Chapter 2.6*. Data is of one independent experiment.

New Insights of Known Structures

Given that alphavirus genome secondary structure was highly divergent, we modified our approach in *Chapter 2* to identify specifically structured regions of a single virus genome. We focused on identifying specifically structured regions of CHIKV because CHIKV currently poses the greatest threat to human health of the alphaviruses (10, 188). With this strategy, we identified 23 specifically structured regions of CHIKV, including the previously known functionally important RNA secondary structures. Identifying all the previously known functionally important RNA secondary structures supported our hypothesis that yet discovered functional structures would be specifically structured.

One of the most well-known secondary structures in the alphavirus genome is the 5′ CSE. This 51-nt element was identified by both approaches and used as a positive control for our structure disruption strategy in both *Chapter 1* for SINV and *Chapter 2* for CHIKV. This element was first identified and assessed in SINV and later VEEV, but it's importance had not been demonstrated for CHIKV. Further, studies done in SINV and VEEV disrupted the large stem loop 5′ of the CSE and found the degree of replication inhibition was different for each virus (25, 26). We demonstrated that disruption of the 5′ CSE alone was attenuating for CHIKV replication, but disruption of SL3-5 was lethal, indicating the importance of the region in CHIKV is similar to that of VEEV (Figure 2.3). However, our structure disrupting strategies and past studies of the element also disrupted the sequence of the 5′ CSE. Therefore it was unclear if the structure or sequence of the region was more important for proper function. Using the same program that provided new primary sequences to disrupt structure but maintain protein coding sequence, we created mutants with a different primary sequence that was predicted to maintain the WT RNA secondary structure of the region. These mutants were just as infectious as WT RNA indicating that structure was important for virus replication (Figure 2.4). These data have been recently supported by the work of Kendall et al. who used compensatory mutations to also demonstrate the importance of RNA secondary structure in this region (139).

One of the specifically structured regions identified in the coding region of CHIKV was the apical portion of the CHIKV termination codon readthrough (TCR) element. Our model of the TCR element refined and supported previous work that assessed this translational recoding element (32, 33). The TCR element increases the rate of ribosomal readthrough of the opal stop codon preceding it (32). Mutating

the opal stop codon that precedes the TCR element in related alphaviruses impacts accumulation of nsP34, virus fitness in mosquitos, and virus pathogenicity (35, 36, 196, 197). In CHIKV, mutation of the opal stop codon to an arginine attenuates virus pathogenicity and increases the production of the full nonstructural polyprotein (33, 34).

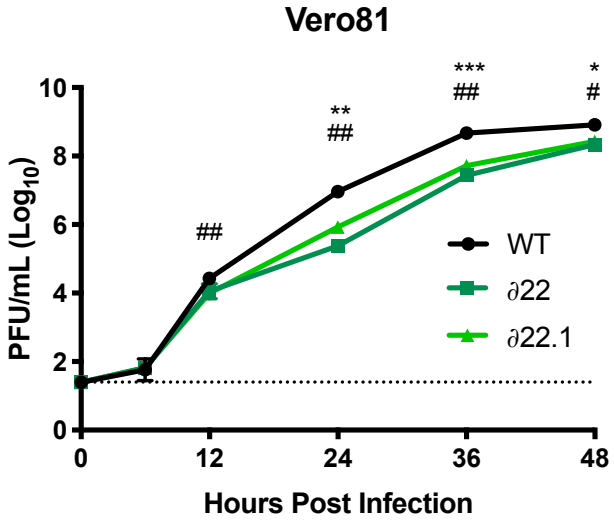


Figure 4.2: Disruption of the full TCR element in 181/25 CHIKV is attenuating.
Vero81 cells were infected with WT, ∂22 or ∂22.1 virus at an MOI = 0.01. Supernatants were sampled at indicated time points and infectious virus was quantified by plaque assay as in Madden 2020. Data represent three biological replicates of one experiment. * $P < 0.05$; ** $P < 0.01$. The symbols in panel B represent the P value of the following comparisons: * WT versus ∂22; # WT versus ∂22.1.

The impact of the opal stop codon was addressed by multiple groups, but the importance of the structure of the TCR element following the stop codon had yet to be assessed in the context of infection. Preliminary data using structure disruption strategies similar to previous studies indicates that disruption of the TCR element does reduce virus replication compared to WT (Figure 4.2). The structure was disrupted in two ways: ∂22 disrupts the full TCR element, ∂22.1 disrupts just the apical portion of the TCR element identified as specifically structured (Figure 2.2 C and Figure 2.1 D nt 5672-5742). The background virus strain encodes the opal stop codon preceding the TCR element (198), therefore attenuation of the mutant viruses is likely due to reduced readthrough of the opal stop codon. This would reduce production of mature nsP4, the RNA-dependent RNA polymerase, compared to WT (32).

However, previous analysis of this element using in vitro translation read through assays indicated that the base of this stem loop was the only portion necessary for this element to function properly; the top of the stem loop could be varied or deleted. Our data indicate that the full structure of the

TCR is important for proper virus replication. WT and structure disruption mutant viruses with the opal stop codon mutated to an arginine residue (O524R) have been cloned and recovered. These mutants will be used to test if ablation of the opal stop codon can rescue the growth defect seen in the TCR disruption mutants. The O524R variants were recovered at a somewhat higher titer than the 524Opal counterparts, but the viruses were not recovered in parallel so a direct comparison cannot be made. Future studies will need to be done directly comparing growth kinetics and nsP4 accumulation between these viruses.

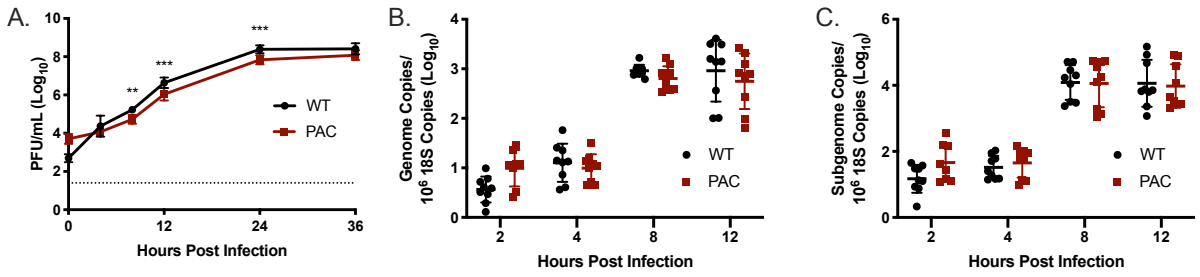


Figure 4.3: Disruption of the putative packaging signal is mildly attenuating.
(A) Vero81 cells were infected with WT or PAC mutant viruses at an MOI = 5. Supernatants were sampled at indicated time points and infectious virus was quantified by plaque assay as in Madden 2020. (B & C) Vero81 cells were infected with WT or PAC mutant viruses at an MOI = 5. Cell lysates were harvested for quantification of genome (B) and subgenome (C) RNA copy number as in *Chapter 2.6 Methods*. Data shown are aggregate of nine biological replicates over three independent experiments. ** $P < 0.01$; *** $P < 0.001$.

A functional element that previously lacked a structure model was the putative packaging signal of CHIKV (Figure 2.2 B, nucleotides 2530-3167). The putative CHIKV packaging signal region was originally identified in 2011 by Kim et al. and located in a different area of the genome than that of SINV and VEEV (28). While authors hypothesized what the RNA secondary structure of this region might be, there were no studies done to directly test if RNA secondary structure was important for CHIKV packaging. We created a structure disruption mutant by introducing silent mutations at every possible wobble site position (PAC, Appendix B). PAC mutant virus replication was mildly attenuated compared to WT virus (Figure 4.3 A). Mutations did not impact virus genome or subgenome accumulation suggesting attenuation was due to a genome packaging defect (Figure 4.3 B & C). However, given the relatively minor effect on replication, it is likely that the region disrupted is not the only signal required for efficient genome packaging. For comparison, disruption of the SINV packaging signal was significantly more attenuating for virus replication (Figure 1.2 C).

A recent publication by Brown et al., also found that disruption of the putative packaging signal of CHIKV did not significantly affect virus replication (30). Further, authors demonstrated that the homologous region in SFV was not largely responsible for genome packaging. The authors show that capsid binding sites are distributed throughout the SFV genome and proposed a multisite genome packaging model for SFV and related viruses like CHIKV (30). The top capsid binding site in SFV maps to one of the highly structured regions identified in our analysis of CHIKV RNA secondary structure (Figure 2.1, nt 6058-6195). Overall, an alignment of SFV and CHIKV indicate that four of the top 10 SFV capsid binding sites correspond to sequences in highly structured regions for CHIKV. This suggests that RNA secondary structures throughout the genome may be important for proper capsid binding and genome packaging. However, since RNA secondary structure is divergent between alphaviruses, further investigation of the SFV RNA secondary structure of these regions should be done or specific capsid binding site should be mapped in CHIKV.

Finally, we contributed to a growing body of studies identifying host determinants of replication and pathogenesis in the 3′ UTRs of arboviruses. In *Chapter 2*, we showed that variation in CHIKV 3′ UTR sequence and structure was associated with increased replication in invertebrate cells without increasing replication fitness or pathogenicity in a vertebrate animal model (Figure 2.6). Flaviviruses like ZIKV have also been shown to modulate duplicated 3′UTRs sequence and structure as a host adaptation strategy (53, 54, 141). In *Chapter 3*, our studies of ZIKV 3′UTR structure support prior work showing that there is no replication cost in invertebrate cells when xrRNA1 is disrupted (53). However disruption of xrRNA1 is less pathogenetic than WT ZIKV in a mouse model of disease (Figures 3.5 and 3.6 E). Previous studies disrupted another putative xrRNA structure in the ZIKV 3′ UTR and the mutant proved to be a promising live attenuated vaccine candidate (174). Our work supports attenuation strategies of flaviviruses through 3′ UTR disruption but highlights that this attenuation may be most significant in male subjects, as male mice were less likely to succumb to infection when infected with the xrRNA1^{mut} ZIKV (Figure 3.6 F). Structure disruption is a promising strategy for developing live-attenuated vaccines but future work with live attenuated vaccines should assess virus pathogenesis in a sex-disaggregated manner.

4.2 Improving RNA Structure Discovery in Viruses

This study of functional RNA secondary structures in RNA viruses contributes to a larger field of RNA structure discovery using chemical structure probing. We improved structure discovery when we took a virus specific approach, focusing on structures in a single virus strain, as opposed to an evolutionary approach, focusing on conserved structures across multiple viruses in the same genus. This method seems to hold true when assessing flavivirus RNA structure as well, since we also identified all known functional structures in ZIKV when we identified the most specifically structured regions, or regions likely to fold into a single conformation, of the ZIKV genome as well. Identifying specifically structured regions has also been supported by other groups assessing RNA secondary structure as well or was found to be a common characteristic of putative functional structures identified using a combination of approaches (137, 185). Identifying the specifically structured regions in a virus provides a starting list of putatively functional structures to be tested. The validity of the starting list is strengthened when known functionally important structures are included. Further refinement of this list in an unbiased manner is difficult, though, because application of other tools does not consistently narrow the pool of candidates while maintaining all the known functional structures. Additionally, you cannot infer an RNA structure's function from structure alone. Therefore, in order to prioritize structure candidates, we must turn back to sequence conservation analyses. Tools like covariation and synonymous site conservation can be used to guide prioritization of the candidate list instead of as requirements for inclusion on the list.

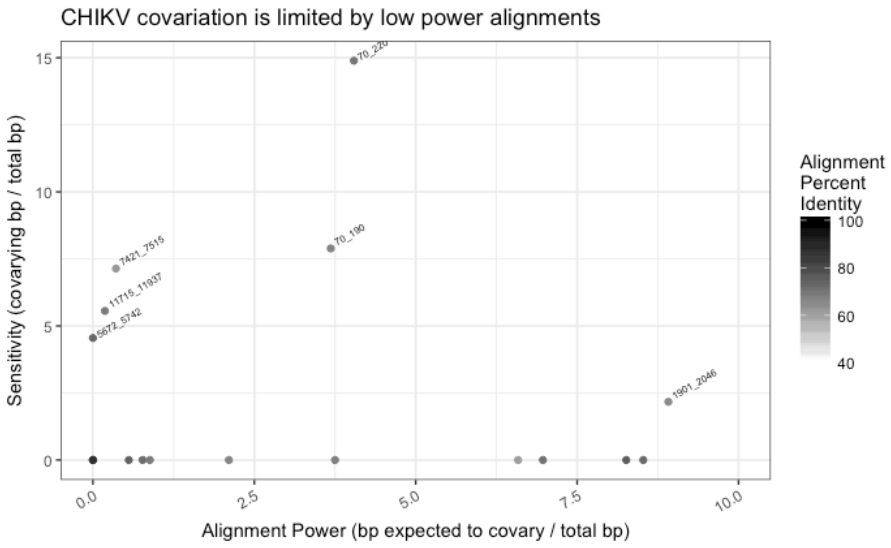


Figure 4.4: CHIKV covariation is limited by low power alignments.

Structured regions of CHIKV were used to search for homologous structures using Infernal software suite 1.1.2 (199). Homologous structures found using sequences assembled by Forrester et al. were assembled into a structure informed alignment (7). The R-scape program v1.5.16 was used to identify base pairs with significant covariance (77, 78). The sensitivity of the covariation analysis is reported as the number of base pairs found to covary significantly out of all base pairs predicted by SHAPE-MaP informed structure modeling. The alignment power is the number of base pairs expected to covary, calculated by R-scape v1.5.16 out of the total base pairs predicted in a structured region. The percent sequence identity of the alignments used during the R-scape analysis are also reported. A summary of this data can be found in Table 4.1. Regions with any covarying base pairs are labeled. Six regions are not plotted due to constant covariation observed and termination of calculations early (n = 4) or no other homologous structures were found and covariation could not be done (n = 2). bp, base pairs.

Covariation analysis for small RNA viruses like alphaviruses and flaviviruses is not very sensitive, particularly within coding regions (133, 137, 185). The null hypothesis for covariation analysis assumes the RNA sequence is not constrained and each position is free to evolve independently. For the coding regions of RNA viruses, this assumption is false from the beginning since the sequence is already being constrained by the coding sequence. Originally, if structures were found to have no evidence of covariation, it was unclear if this was evidence against structural conservation or lack of sufficiently variable sequences to detect a significant covariation signal. The R-scape program has recently been updated to report the power of the alignment supplied to clarify this point. Now sufficiently powered alignments that result in no significant covariation are suggestive of a lack of conserved structure (78). This is useful when trying to prioritize multiple candidate RNAs or RNA structured regions of interest for experimental testing. A sufficiently powered alignment of a structured region with no covariation could be de-prioritized. Unfortunately, alignments generated for CHIKV structured regions of interest from sequences used in *Chapter 1* are low power and cannot be used to de-prioritize any candidate regions (Figure 4.4 and Table 4.1). Studies with flaviviruses that have used all available flavivirus sequences have had better luck using covariation as a prioritization guide but the power of the alignments were not reported (185). These studies indicate that perhaps the covariation analysis can be improved for alphaviruses with improved alignments and inclusion of more sequences in the alignments. A limitation of this prioritization scheme is that it would really only identify structures conserved across the genus and therefore may demote structures with virus specific functions.

The program *synplot2*, was designed to look for elements being conserved within coding regions by focusing on the wobble base position of synonymous codons. The program compares the number of observed mutations in a codon-informed alignment to the number of mutations expected by neutral

evolution (151). Similar to SHAPE-MaP analysis, regions with lower than expected synonymous site substitution are identified over a rolling window whose size is determined by alignment depth and the size of RNA element you are searching for (RNA structures tend to be smaller than overlapping ORFs and would therefore require smaller windows). However, both covariation and synonymous site conservation analysis are dependent on the quality of alignments used. Unlike covariation analysis, synonymous site conservation can be successfully applied over alignments of specific viruses, like only CHIKV sequences, or very closely related virus strains if it the virus of interest is newly discovered and there are few quality sequences available (151). While *synplot2* can be applied to alignments as shallow as two sequences, alignments that are deeper and contain more varied, high quality sequences improve the power to detect features that are constraining the RNA sequence (151). Further, *synplot2* requires all sequences be compared to a specific reference sequence. While this makes some aspects of the analysis easier, it can also bias the analysis if an inappropriate reference sequence is chosen. The program only compares two sequences at any given time and does not consider any amino acid positions that are not synonymous to the reference sequence. No information about the alignment as a whole is considered, for example if a position is nearly always a single specific amino acid or if the position tolerates a specific type of amino acid more broadly (e.g. acidic residues). This type of calculation would likely increase the number of expected mutations at a given position perhaps turning a moderately conserved position into a highly conserved position.

It would be useful to develop a program that combined features of both these tools, covariation analysis and synonymous site conservation. The covariation analysis tool could be improved if you could force consideration of a coding constraint on the RNA alignments. Overall coding capacity severely limits a nucleotides ability to evolve neutrally, but the cost of mutating is not equal at all positions in a codon. This would likely severely limit the number of nucleotides that could be assessed for covariation but may improve the power of the alignments and improve sensitivity. It is useful to know if a structure has only four pairs that could be assessed for covariation and all four pairs covaried significantly, while another structure had five pairs that could be assessed for covariation and none of them were observed to covary significantly despite a high-powered alignment.

The program *synplot2* is advantageous because position in a codon is considered when calculating the expected number of substitutions for a sequence. However, it does not consider the full flexibility of a virus genome. Instead, it only considers positions that are synonymous between two sequences. This does not take advantage of all the information provided in the protein-informed alignment the user must assemble prior to running *synplot2*. Instead of using one specific virus as a reference sequence, it would be advantageous to create a modified consensus sequence based on information from the protein informed alignment. For example, if at any given position there is no single consensus amino acid, but the position is nearly always a negatively charged amino acid, it could be assumed that codons for both aspartate (GAU and GAC) and glutamate (GAA and GAG) were tolerated. Therefore, when calculating the expected number of substitutions, all four nucleotides could be expected to substitute in the third position instead of just two possible nucleotides.

These modified analyses, preferably with statistics for each nucleotide position, could then be combined with reactivity data. We could look for unreactive nucleotides with lower-than-expected mutation rates, or covariation signal, and assess to identify potentially functional RNA helices from helices formed as a byproduct of the given coding sequence. Sequences with nucleotides mutating at an expected rate, with a powerful enough alignment, could then be deprioritized for follow-up regardless of the SHAPE data, while regions with low SHAPE reactivity, low Shannon entropy, and lower than expected mutation rates in stems could be prioritized for experimental follow-up.

STRUCTURED REGION (INCLUSIVE)	**PREDICTED BASEPAIRS**	**COVARIATION SENSITIVITY**	**COVARIATION ALIGNMENT POWER**	**SYNONYMOUS SITE CONSERVATION SIGNIFICANCE**		
				15-codon sliding window	25-codon sliding window	45-codon sliding window
70 -190^	38	7.89	3.68	214.94	709.11	*42877.87
(70-220) ^	47	14.89	4.04	216.02	728.57	*52604.56
548-648	33	0	6.97	4.23	3.95	9.17
867-970	40	0	3.75	2.82	2.88	1.95
993-1186	61	0	8.52	4.58	4.63	2.66
1377-1506	41	0	0	2.84	2.94	2.63
1901-2046	46	2.17	8.91	3.92	2.53	1.63
2276-2304	12	!	!	5.83	4.88	7.71
2591-2712	38	0	2.11	50.61	138.90	250.98
3260-3367	39	0	0.77	2.28	2.26	1.96
4097-4130	10	!	!	6.38	4.54	2.55
4780-4845	22	0	0	2.02	1.54	1.26
5062-5152	32	#	#	1.46	1.21	1.09

5672-5742	22	4.55	0	2.48	2.81	5.78
6058-6195	41	0	6.59	1.25	1.14	1.12
6258-6329	23	0	8.26	18.82	19.90	23.05
7040-7084	18	0	0.56	10.82	1.66	1.12
7421-7515^	28	7.14	0.36	76.92	48.42	13.84
8700-8792	28	!	!	1.68	1.41	1.15
9243-9289	18	#	#	1.11	1.04	1.04
9933-10041	34	0	0.88	*2.05 x 10^6	*4.00 x 10^7	*3.09 x 10^6
10228-10257	11	0	0	10.05	3.27	1.96
10583-10688	36	!	!	2.85	2.82	2.35
11715-11937	54	5.56	0.19	NA	NA	NA

Table 4.1: Summary covariation and synonymous site conservation analysis of specifically structured regions in CHIKV.
Specifically structured regions determined by SHAPE-MaP were assessed for significantly covarying bases using R-scape v1.5.16 (78). Alignments used for the analysis were created as in *Chapter 1* and the power of each alignment is reported. Synonymous site conservation data for CHIKV reported in (151) was re-analyzed in 15, 25, and 45-codon sliding windows. The average *P*-value for a structured region in each window is reported. Region 70-220 is the expanded region of 70-190 to include both 5′ CSE stem loops. Structured regions in red overlap with previously reported functionally important RNA secondary structures. Significance of synonymous site conservation for the whole genome can be found in Appendix C.
* Indicates a *P*-value < 0.05 for that window after a Bonferonni-like correction as reported in (151).
! Covariation analysis was not completed due to constant covariation observed and termination of calculations early.
No other homologous structures were found to create an alignment and covariation analysis could not be done.
^ Synonymous site conservation data reported is for the coding sequence only of these structured regions.

4.3 Future Directions

Having identified a list of 19 potentially functional uncharacterized RNA secondary structures of CHIKV that have been preliminarily prioritized using available covariation data and synonymous site conservation analysis (Table 4.1), future work needs to empirically test these structured regions for functional importance to assess the accuracy of this structure discovery approach. A limitation of past experimental approaches to assess structure for functional importance was assessing structured regions in isolation. In reality, RNA viruses exist as a quasispecies with multiple selective pressures being imposed by not just the host, but also by the component virus genomes of the quasispecies (200). Specific RNA secondary structures may not be a requirement for virus replication but may provide an advantage for one virus genome over another that lacked that structure. In this instance, the only way to identify an RNA secondary structure that provides an advantage is to assess it in competition.

In order to assess two genomes in competition with one another, you must be able to accurately measure the presence of each genome in a mixed pool. Inserting fluorescent markers like GFP or mKate to mark a WT genome from a mutant genome is one method to accomplish this, however reporter viruses

with these tags often have reduced fitness already due to the additional sequence and would likely affect local RNA secondary structure confounding the results. Deep sequencing provides the best quantitative measure of genome variants in a pool of viruses but would be expensive to apply to multiple assays and multiple pooled combinations that would be required by a screen of this kind. Instead, we have taken advantage of the sequence diversity created by the structure disrupting mutations and designed sets of qRT-PCR assays that specifically target mutant structure regions or the cognate wildtype sequence of the region (Figure 4.5). These competition assays provide an absolute quantitation of mutant and WT sequence present in a sample and can be applied to both in vitro and in vivo competition studies. Further, additional assays can be designed quickly if new mutation strategies are needed to differentiate the importance of sequence vs. structure or additional regions need to be assessed.

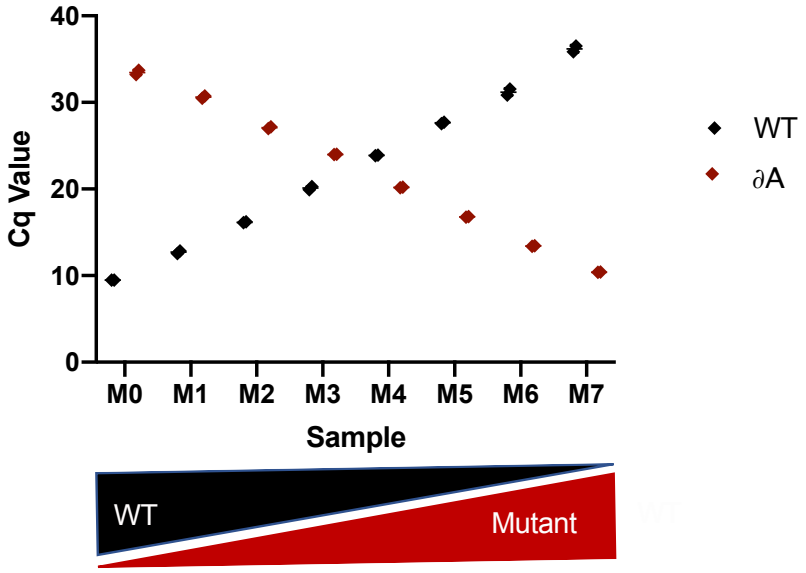


Figure 4.5: RT-PCR is highly specific for CHIKV genome target.
RT-PCR assays were designed to specifically target unmutated WT sequence or mutant sequence found in structure disruption viruses. To test the accuracy of the assays, target sequences were serially diluted from 4.08 x 10^8 copies/µl to 35 copies/µl and a final 0 copy/µl control was added. The WT dilution series was inverted and mixed 1:1 with the mutant target series. The resultant mixed samples were run in technical duplicate with either an assay designed to detect WT sequence (WT) or mutant sequence (∂A). Reported Cq values for each reaction with each probe are reported.

Past work assessing RNA structure for functional importance has been done piece-meal, assessing one structure at a time. While this method is thorough, it is also time consuming. It is unlikely that every structure identified is functionally important for each assay. In reality, a handful of the identified uncharacterized structures will likely be functionally important. To accelerate structure screening, multiple

structures have been disrupted across four mutants, and one mutant with all structured regions of interest were mutated. This limits initial screens to four mutants and a WT control (Table 4.2). If one of the grouped structure disruption mutants is attenuated in a specific assay, the component structured regions can then be prioritized for more detailed follow-up. There are multiple approaches that can then be taken to identify which structured region in a mutant caused the attenuated phenotype: Attenuated structure mutants can be passaged and sequenced for reversion mutations; new mutants can be cloned with individual structured regions disrupted and screened; or mutants with different combinations of disrupted and repaired structures can be cloned and screened.

MUTANT	REGIONS DISRUPTED	TOTAL MUTATIONS
∂A	521-673	29
	842-994	22
	1046-1198	23
	1361-1513	30
	1880-2032	26
	2213-2365	21
∂BC	3239-3388	21
	4037-4189	24
	4736-4888	18
	5030-5182	30
	5630-5782	22
	6032-6181	23
	6218-6370	18
	6986-7138	23
∂D	8671-8823	21
	9190-9342	21

Table 4.2: Mutants with multiple structured regions disrupted.
Specifically structured regions were disrupted with mutant sequences generated using *CodonShuffle* as in *Chapter 2*. The name of the mutant virus is listed with which regions are mutated with structure disrupting point mutations. The number of mutations made in each region is also listed. The specific mutant sequences can be found in Appendix B.

We also plan to assess these structure disruption mutants in the context of a vertebrate host regardless of the in vitro assay phenotypes. It is possible that some RNA structure elements may provide a fitness advantage in a cell type or tissue specific manner better observed in vivo, like the miRNA sequences restricting replication of VEEV (201). Our structure disruption mutants will be assessed in the immune competent mouse model of CHIKV pathogenesis (152). Mice will be inoculated with individual viruses or mixed inoculums of WT and mutant. Virus replication will be assessed in multiple tissue compartments early in infection by qRT-PCR to quantify which variant is most abundant.

Further, mice will be infected with individual structure mutants and compared to a WT control to determine if structure disruption affects CHIKV-induced pathology. CHIKV infection of C57BL/6J mice

induces biphasic swelling of the ipsilateral footpad. Footpad swelling severity and kinetics depends on host genetics but also depends on CHIKV genotype (202) (Sarkar and Plante et al. unpublished). This suggests there are yet discovered virus determinants modulating this disease severity. Therefore, the footpad swelling of mice infected with a single mutant will be assessed and compared to WT to see if structure mutants impact footpad swelling severity, kinetics, or duration.

Follow-up studies with deep sequencing will be necessary to see if the designed structure mutations are being maintained through multiple rounds of replication or if mutant sequences are being lost due to reversion, recombination with WT sequences, or simply being out competed. If multiple structured regions are outcompeted by WT, it is possible that different mechanisms may explain the loss of different structured regions. Complete reversion of mutants to WT has occurred with other structure disrupting mutations, but is unlikely with these mutants (203, 204). There are dozens of structure-disrupting mutations in each mutant, and these mutations are located in coding regions, limiting the number of possible mutations that would restore a structure and not interfere with the amino acid sequence. It is more likely that mutant viruses that are less fit than WT will be outcompeted or recombine with WT sequence. Recombination events would be exciting to analyze since little is known about alphavirus recombination mechanisms, though preliminary data suggests RNA secondary structure may be involved (205) (Levi and Madden et al. unpublished). Recombination events may also help identify which disrupted structure in the mutant was primarily responsible for attenuation.

After identification of an attenuated structure mutant, further investigation will be necessary to understand the mechanism behind the attenuation. RNA elements are known to be important for proper evasion of host innate immune sensing or permitting replication in some host tissue compartments (82, 201). RNA structure may also be important for generating pools of defective virus genomes (DVG) created during replication which could then impact pathology during infection (205). Furthermore, the mutations designed to disrupt structure may impact deposition of post-transcriptional modification on the virus genomes that in turn impact replication and pathogenesis (195, 206). There are many roles RNA structure can play during virus replication and pathogenesis, which makes further study of this area an exciting avenue of investigation.

APPENDIX A: STRUCTURAL DIVERGENCE CREATES NEW FUNCTIONAL FEATURES IN ALPHAVIRUS GENOMES SUPPLEMENTAL FIGURES AND TABLES

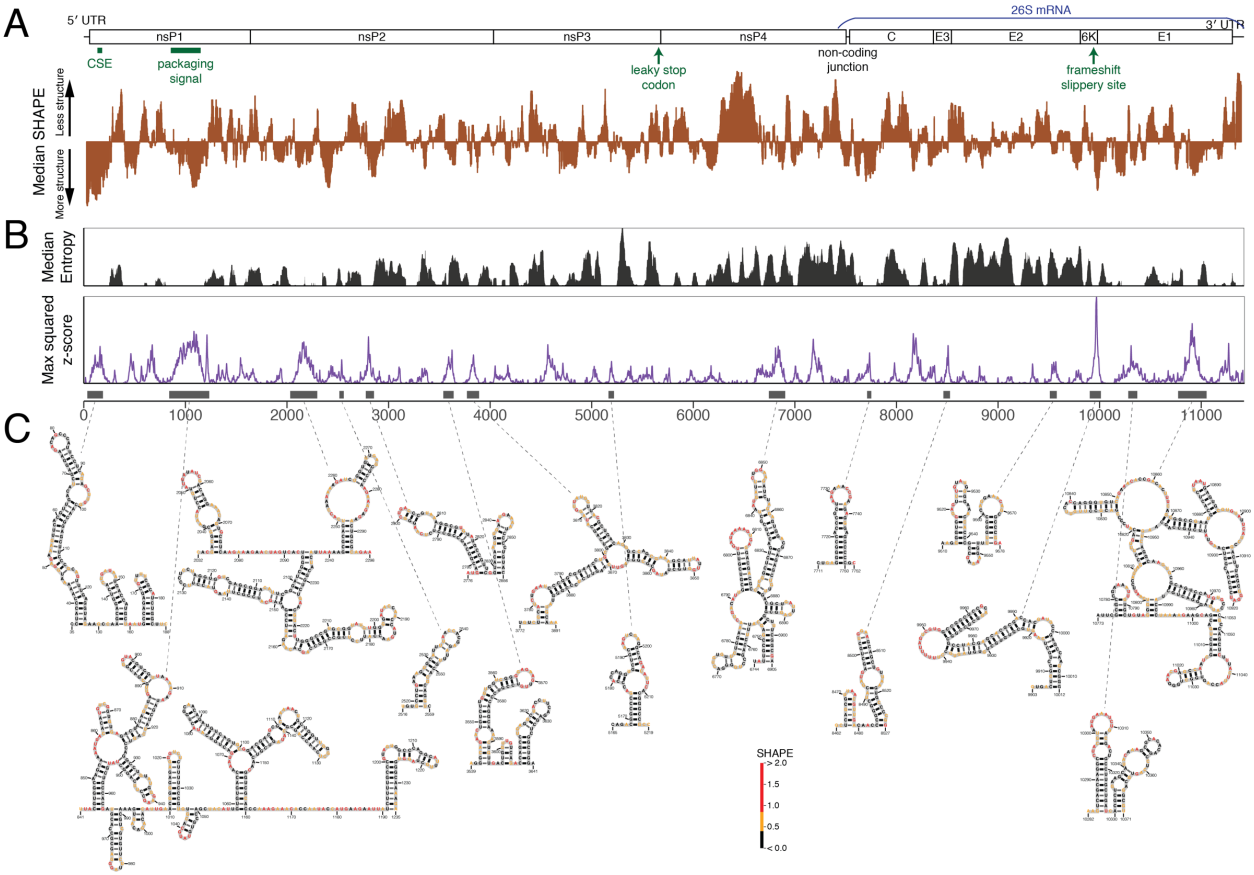


Figure A.1: Highly stable structures in the VEEV genome.
(A) Top: schematic of the virus genome organization, with annotated elements. Bottom: SHAPE data for the Venezuelan equine encephalitis virus genome, represented by the local median (55-nt window) compared with the global median. Reactivities below the x-axis indicate a region more structured than average. Gray lines denote the conserved sequence element (CSE), which has low SHAPE reactivities and is highly structured. (B) Top: median (55-nt window) Shannon entropies of base pairing across the SINV genome. Middle: Maximum squared z-score at each position in the genome, with higher values corresponding to greater structural significance. Bottom: structured regions in the SINV genome, based on the intersection of regions with low SHAPE and low z-scores. (C) SHAPE-directed structural models of SINV structured regions. Nucleotide color indicates low, medium, or high SHAPE reactivity.

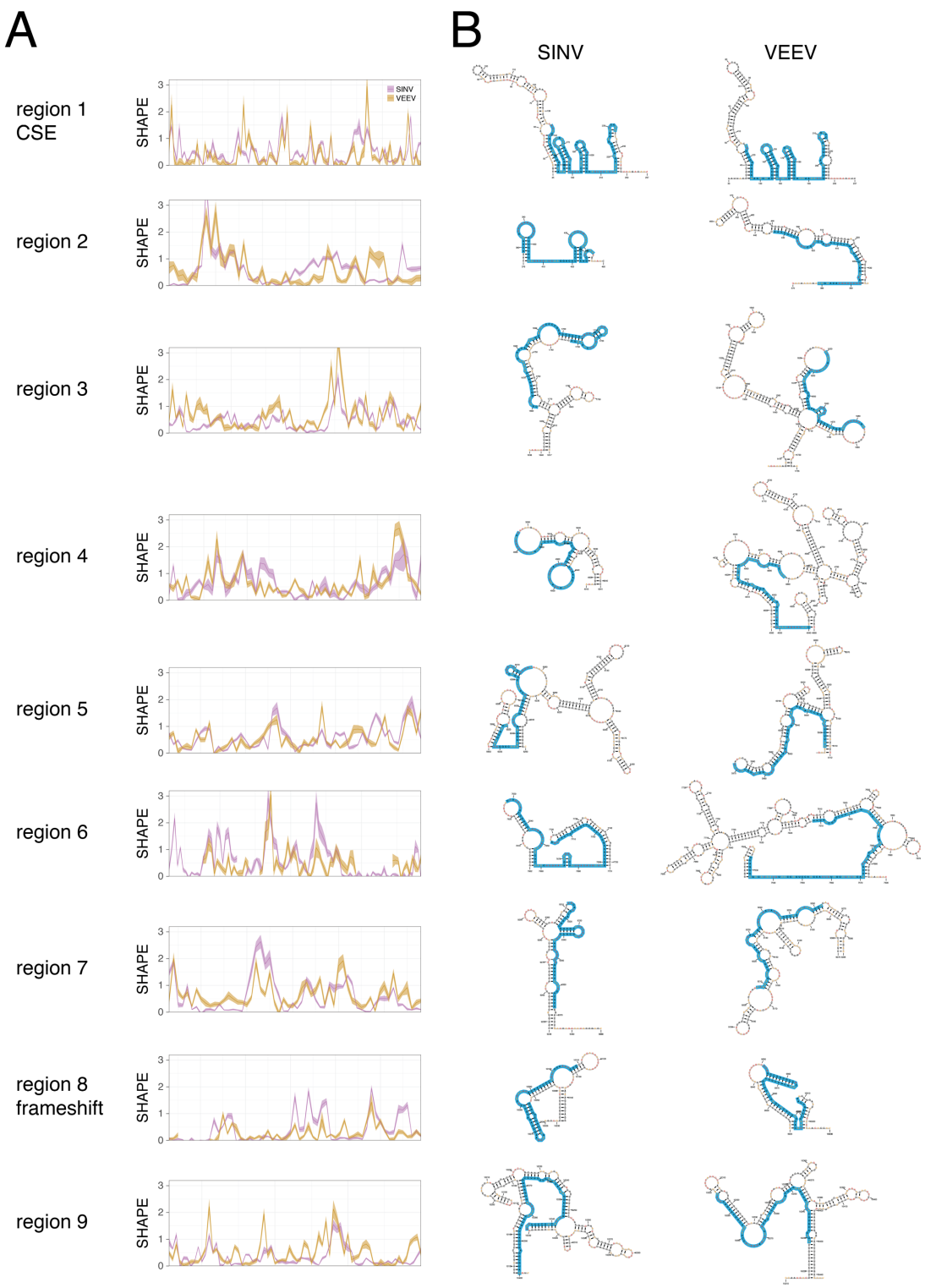


Figure A.2: Most high-correlated SHAPE regions do not adopt similar structures.
(A) Structural models for SINV and VEEV that overlap areas with highly correlated SHAPE data. Regions within the correlation window are highlighted in blue. Only the first region, which contains the 5′ CSE, and the eighth region, which contains the frameshift signal hairpin, adopt similar secondary structures between the two viruses. (B) SHAPE data of SINV and VEEV within each correlation window. Base

pairing patterns that result in similar SHAPE profiles do not necessarily correspond to similar secondary structures and are in most cases the result of chance.

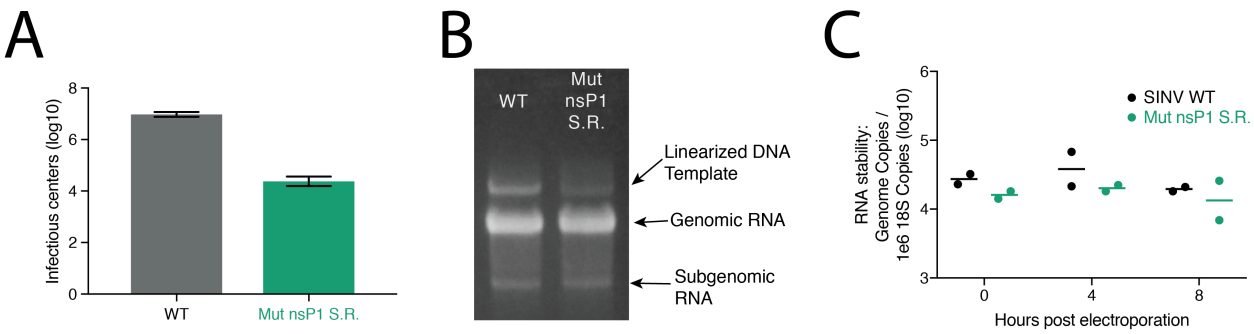


Figure A.3: Mutations to nsP1 SR does not affect RNA quality or stability.
(A) Electroporation of WT and nsP1 SR in vitro transcribed RNA result in a reproducible decrease in infectious centers of nsP1 SR compared to WT (n=4). (B) The RNA quality of the WT and nsP1 SR in vitro transcribed RNA is the same. (C) RNA stability was assessed by qRT-PCR post electroporation and no differences in decay were observed.

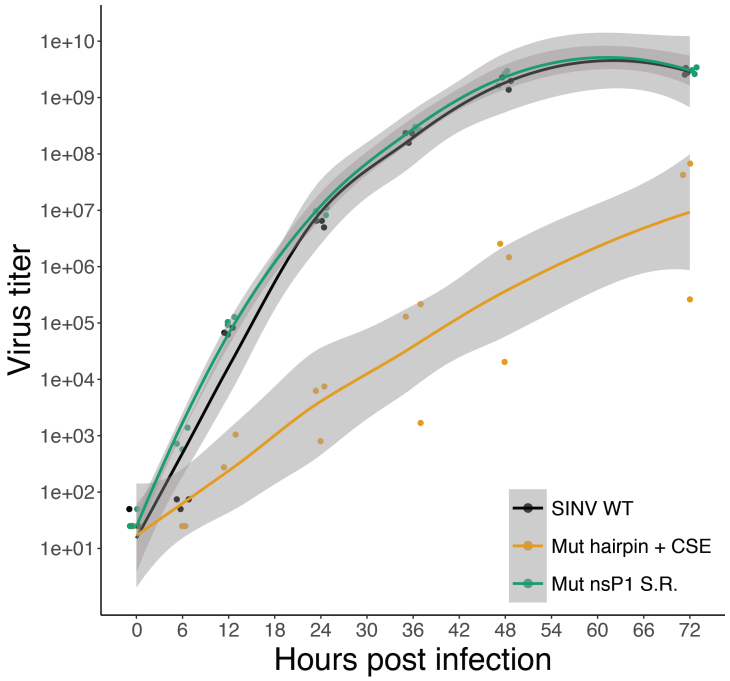


Figure A.4: The 5′ hairpin and CSE are necessary for optimal virus growth in mosquito cells.
When the mutant viruses were grown in C6/36 cells, the mutant hairpin + CSE had a titer several orders of magnitude lower compared with wildtype, whereas the nsP1 SR mutant had no change in growth.

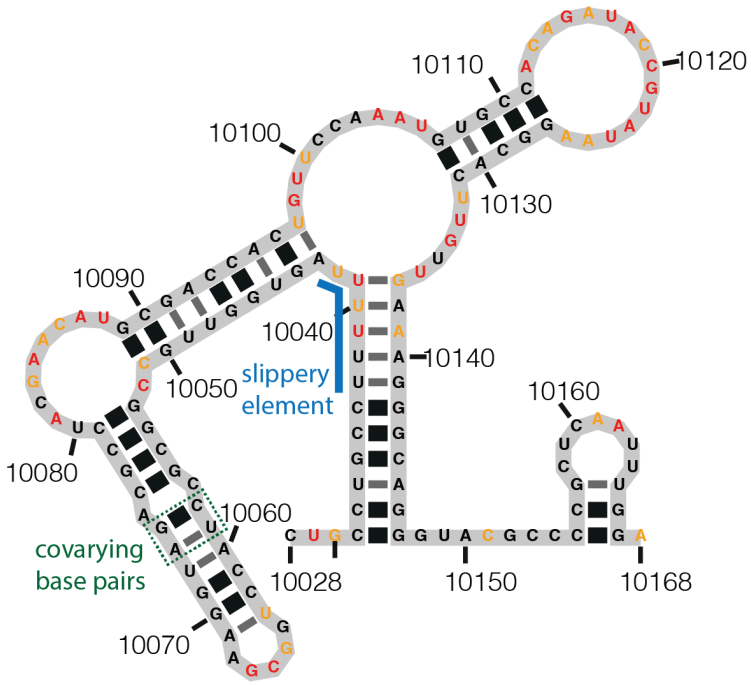


Figure A.5: Evidence of conserved structure in the frameshift element.
R-scape revealed two covarying base pairs (green) downstream of the poly-U slippery element (blue), but no covariance was found outside of that stem-loop. This covariance supports a previous study that found conservation of the same stem-loop in equine encephalitis viruses (83).

effective number of codons			SINV WT	Mut hairpin + CSE	Mut pack. Sig.	Mut nsP1 SR	Mut nsP3 SR
		ENC	54.502	54.551	54.615	54.496	54.453
codon counts	amino acid	codon	SINV WT	Mut hairpin + CSE	Mut pack. Sig.	Mut nsP1 SR	Mut nsP3 SR
	Ala	gca	98	98	99	99	96
	Ala	gcc	120	120	119	120	120
	Ala	gcg	68	69	68	67	68
	Ala	Gct	51	50	51	51	53
	Arg	Aga	62	62	62	62	61
	Arg	Agg	34	34	33	34	34
	Arg	Cga	16	16	18	16	16
	Arg	Cgc	46	46	45	46	46
	Arg	Cgg	15	15	14	15	16
	Arg	Cgt	23	23	24	23	23
	Asn	Aac	92	92	92	92	92
	Asn	Aat	44	44	44	44	44
	Asp	Gac	124	122	125	124	124
	Asp	Gat	54	56	53	54	54
	Cys	Tgc	80	80	81	80	80
	Cys	Tgt	21	21	20	21	21
	Gln	Caa	57	55	58	57	56
	Gln	Cag	75	77	74	75	76
	Glu	Gaa	114	115	112	114	115
	Glu	Gag	108	107	110	108	107
	Gly	Gga	90	90	87	90	91

	Gly	Ggc	53	53	53	53	53
	Gly	Ggg	39	39	41	39	38
	Gly	Ggt	25	25	26	25	25
	His	Cac	57	59	56	57	57
	His	Cat	43	41	44	43	43
	Ile	Ata	41	41	41	41	41
	Ile	Atc	94	94	93	94	95
	Ile	Att	45	45	46	45	44
	Leu	Cta	33	34	33	33	33
	Leu	Ctc	48	48	48	48	48
	Leu	Ctg	98	97	98	98	99
	Leu	Ctt	40	40	40	40	38
	Leu	Tta	18	18	18	18	19
	Leu	Ttg	50	50	50	50	50
	Lys	Aaa	115	116	116	115	115
	Lys	Aag	131	130	130	131	131
	Met	Atg	86	86	86	86	86
	Phe	Ttc	74	75	74	74	73
	Phe	Ttt	58	57	58	58	59
	Pro	Cca	86	85	86	85	86
	Pro	Ccc	41	41	40	41	43
	Pro	Ccg	80	80	81	81	80
	Pro	Cct	42	43	42	42	40
	Ser	Agc	59	59	60	59	59
	Ser	Agt	34	34	35	34	34
	Ser	Tca	46	47	48	46	49
	Ser	Tcc	40	40	39	40	38
	Ser	Tcg	52	51	49	52	53
	Ser	Tct	24	24	24	24	22
	Thr	Aca	78	78	76	78	78
	Thr	Acc	100	100	101	100	100
	Thr	Acg	45	45	47	45	44
	Thr	Act	52	52	51	52	53
	Trp	Tgg	37	37	37	37	37
	Tyr	Tac	90	90	90	90	90
	Tyr	Tat	35	35	35	35	35
	Val	Gta	65	64	65	65	64
	Val	Gtc	84	83	86	84	84
	Val	Gtg	76	77	75	76	75
	Val	Gtt	56	57	55	56	58

Table A.1: Effective number of codons (ENC) and codon counts for WT SINV and the four mutants. Changes in codon usage between WT and mutant viruses are of very small magnitude.

Region	**Number of *cmsearch* hits**
SINV hairpin + CSE	36
SINV packaging signal	37
SINV nsP1 SR	12
SINV nsP3 SR	4
SINV: 1508-1691	7
SINV: 2410-2560	37
SINV: 3824-3959	9
SINV: 4056-4094	6

SINV: 4173-4207	12
SINV: 5093-5139	4
SINV: 5212-5361	3
SINV: 6327-6386	28
SINV: 7600-7831 (non-coding junction)	8
SINV: 9297-9330	7
SINV: 10028-10168 (frameshift)	30
SINV: 10826-10910	29
SINV: 11630-11661	10

Table A.2: Number of alphavirus sequences found by *cmsearch*.
Only three covariance models, including the 5′ hairpin + CSE and the packaging signal, find homologs in all or almost all related alphaviruses.

Model	Sensitivity (%)	PPV (%)	Model pairs	True positives	Alignment length	Avg. % identity
SINV hairpin + CSE	10	100	50	5	159	62
SINV pack. sig.	1.16	50	86	1	326	60
SINV nsP1 SR	0	0	24	0	88	59
SINV nsP3 SR	0	0	61	0	186	81
SINV: 1508-1691	0	0	47	0	184	63
SINV: 2410-2560	0	0	37	0	151	70
SINV: 3824-3959	0	0	34	0	136	74
SINV: 5212-5361	0	0	46	0	150	96
SINV: 6327-6386	5.56	100	18	1	60	67
SINV: 7600-7831 (non-coding junction)	3.28	100	61	2	232	62
SINV: 10028-10168 (frameshift)	5.88	100	34	2	136	59
SINV: 10826-10910	4.55	100	22	1	85	58
RNase P	87.25	91.75	102	89	367	58
Purine riboswitch	86.36	100	22	19	102	55
tRNA	100	56.76	21	21	71	44
5S rRNA	64.71	73.33	34	22	119	56
L10 Leader	93.75	78.95	16	15	78	46
L1 Leader	66.67	85.71	9	6	31	53
L20 Leader	58.82	100	34	20	87	63
L4 Leader	71.19	97.67	59	42	197	58
S15 Leader	57.14	57.14	7	4	81	67
S1 Leader	75	100	24	18	117	60
S2 Leader	88.24	83.33	17	15	96	43
S4 Leader	90	100	10	9	110	73
S7 Leader	9.09	100	33	3	104	80
S8 Leader	3.33	100	30	1	105	82

Table A.3: R-scape results for covariance models of known RNA structures and structure informed alignments.
The number of covarying base pairs R-scape found is used for the sensitivity and positive predictive value (PPV) calculations. For most structures outside of SINV, the number of true base pairs found by R-scape is much higher than found for any SINV structure-informed alignment. SINV alignments have similar average percent identity to the alignments of conserved structures with significantly covarying base pairs, but the sequence diversity within SINV alignments does not result in a large number of covarying base pairs.

APPENDIX B: MUTANT VIRUS SEQUENCES

Virus Strain	Mutant Name	Sequence range	# mutations	Mutant sequence
CHIKV 181/25	∂SL3	67-220	11	ATAACCCATCATGGATTCTGTGTAGTGG ACATAGACGCTGACAGCGCCTTTTTGAA GGCCCTGCAACGTGCGTACCCCATGTT TGAGGTGGAACCTAGGCAGGTCACGTC GAATGATCATGCTAATGCCAGAGCATTC TCGCACCTAGCTATA
CHIKV 181/25	∂5′CSE	67-220	12	ATAACCCATCATGGATTCTGTGTACGTG GACATAGACGCTGACAGTGCGTTTTTGA AGGCCCTGCAACGCGCCTACCCCATGT TTGAGGTGGAACCTAGGCAGGTCACAT CGAATGACCATGCTAATGCTAGAGCGTT CTCGCATCTAGCCATA
CHIKV 181/25	∂SL3-5	80-220	23	GATTCTGTGTATGTGGACATTGATGCTG ACAGCGCGTTCCTCAAGGCGCTTCAAC GTGCCTATCCCATGTTTGAGGTGGAAC CTAGGCAGGTGACATCTAACGACCATG CGAACGCCAGAGCGTTTAGCCACCTAG CTATA
CHIKV 181/25	scrSL3	80-220	5	GATTCTGTGTATGTAGACATAGACGCTG ACAGCGCCTTTCTGAAGGCCCTGCAAC GTGCATACCCTATGTTTGAGGTGGAACC TAGGCAGGTCACATCGAATGACCATGCT AATGCTAGAGCGTTCTCGCATCTAGCCAT A
CHIKV 181/25	scr5′CSE	80-220	7	GATTCTGTGTACGTGGACATAGACGCTG ACAGCGCCTTTTTGAAGGCCCTGCAACG TGCGTACCCCATGTTTGAGGTGGAACCT AGGCAGGTCACGTCGAATGACCATGCCA ACGCCAGAGCGTTTTCGCATCTGGCTATA
CHIKV 181/25	scrSL3-5	80-220	12	GATTCTGTGTATGTAGACATAGACGCTGA CAGCGCCTTTCTGAAGGCCCTGCAACGT GCATACCCTATGTTTGAGGTGGAACCTA GGCAGGTCACGTCGAATGACCATGCCAA CGCCAGAGCGTTTTCGCATCTGGCTATA
ZIKV H/PF/2013	“A”	1-70	54	GAUUACAUUAGGGCACAGUGCACAGG UCGCGAUUUCUUCGAAAAGCGUAGCG UAAUACUAAGAUUUAAGG
ZIKV H/PF/2013	“B”	1-70	58	AGUACAACUAGAGACUAUUAGUGUGA GGCUAUAAAGCUCAUACUUCCAUUUG GAAGGAUCUCUUAGUUGU
ZIKV H/PF/2013	“C”	1-70	11	AGUUGUUGAUCUGUGUAUUUCAGACU GCGAAUAAUCGAGUUUGAAGCAUUUG CUAGCAACAGUAUCAACA
ZIKV H/PF/2013	E^{mut}	1860-2012	20	ATGGATAAACTCAGATTGAAGGGCGTG TCTTACAGCTTGTGCACTGCAGCGTTC ACATTCACCAAGATACCAGCAGAAACA CTTCACGGGACAGTTACGGTGGAGGT CCAGTATGCCGGGACAGACGGACCAT GTAAGGTCCCTGCTCAGATG
ZIKV H/PF/2013	$xrRNA1^{mut}$	10396-10467	13	TGCCTGGCTTGCTAGTCAGCCACAGC CTGGGGCAAGGAGGGAAGACTGTGT

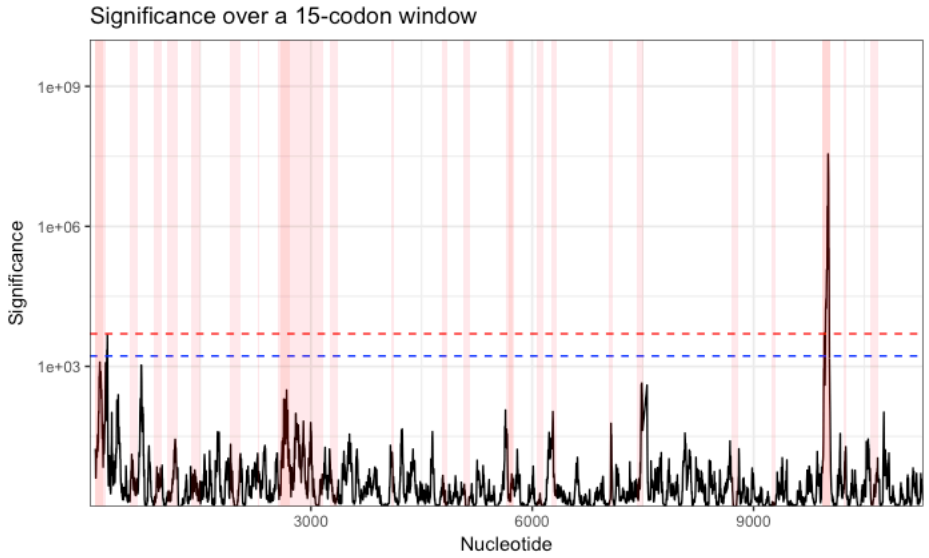
				CCCCCCCAGCAGAAGCTGTGA
ZIKV H/PF/2013	xrRNA1/2^{mut}	10396-10548	26	TGCCTGGCTTGCTAGTCAGCCACAGC CTGGGGCAAGGAGGGAAGACTGTGTC CCCCCCAGCAGAAGCTGTGAAACCAA GCCTATAGTCAGGCCGAGAACCCCCG AGCTCAGAAGAAGCCATGCTGCAGGTG GGCCGCACTGATGACACTGAGT
CHIKV 181/25	∂22	5639-5791	25	GAGCTGTGACTAGACAGGGCTGGTGGG TACATATTTTCGTCGGATACAGGCCCGG GTCACCTACAACAGAAGTCGGTCCGCCA GTCAGTATTACCAGTAAACACCCTGGAG GAGGTGCATGAAGAGAAGTGCTATCCAC CTAAGTTGGACGAA
CHIKV 181/25	∂22.1	5639-5791	12	GAGTTATGACTGGACAGGGCTGGTGGG TATATATTCTCGTCGGACACAGGTCCGG GCCATTTGCAACAGAAGTCAGTGCGCCA GTCGGTACTACCAGTAAACACCCTAGAG GAAGTCCACGAGGAGAAGTGTTACCCAC CTAAGCTGGATGAA
CHIKV 181/24	PAC	2501-3079	178	AAGGTCGTGCTCTGCGGCGATCCCAAAC AATGTGGATTTTTTAACATGATGCAGATG AAAGTTAATTATAACCACAATATATGTAC ACAGGTATATCATAAGAGCATATCAAGA CGGTGCACCCTTCCCGTAACGGCTATC GTATCCTCATTACACTATGAGGGTAAGA TGCGAACGACCAACGAATATAATATGCC AATCGTGGTTGATACCACCGGTTCCACT AAGCCCGATCCGGGGGATCTTGTATTG ACTTGTTTCAGGGGATGGGTCAAGCAG CTACAGATCGATTACCGCGGGCATGAA GTAATGACCGCGGCTGCGTCACAGGGT TTGACGAGGAAGGGCGTCTATGCGGTC AGACAGAAGGTGAATGAGAATCCGCTAT ACGCCTCCACGTCGGAACATGTTAATGT CCTACTCACACGAACTGAGGGCAAGCT AGTGTGGAAAACCCTTTCGGGGGATCC TTGGATTAAAACCCTACAAAATCCCCCA AAGGGCAATTTTAAGGCGACGATCAAA GAATGGGAAGTAGAACATGCCTCTATC ATGGCAGGTATTTGTAGCCATCAGGTG
CHIKV St. Martin	∂A	521-673	29	ATCTATCAGGACGTATATGCGGTCCAT GCACCCACCTCCCTATACCACCAAGCA ATCAAAGGTGTACGAGTAGCCTACTGG GTGGGGTTCGACACTACGCCGTTTATG TACAACGCGATGGCTGGAGCGTACCC GTCATATTCAACGAACTGG
CHIKV St. Martin	∂A	842-994	22	AAGAGCTGGCACCTACCATCAGTTTTC CATTTAAAGGGAAAGCTCAGCTTTACA TGCAGGTGTGACACGGTGGTGTCGTG TGAGGGCTATGTCGTAAAAAGAATAAC CATGAGCCCCGGCCTCTACGGCAAGA CGACGGGCTATGCGGTTACC
CHIKV St. Martin	∂A	1046-1198	23	GAGAGAGTGTCGTTTAGTGTGTGCACA TACGTCCCGGCTACGATCTGCGATCAA ATGACGGGCATCCTGGCCACAGAAGTG

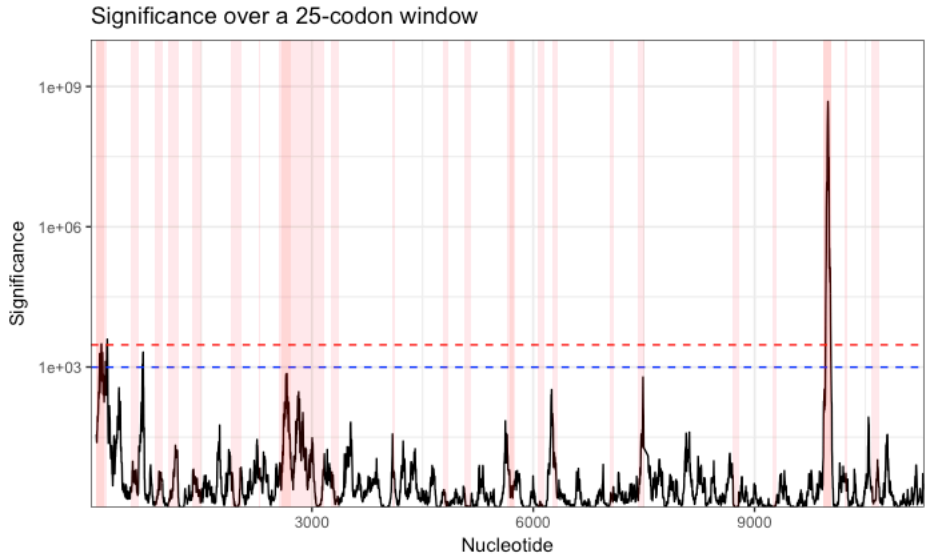
				ACGCCGGAGGATGCACAAAAGTTGCTG GTGGGGCTTAACCAGAGAATTGTTGTC AACGGCAGAACGCAACGG
CHIKV St. Martin	∂A	1361-1513	30	CATACCGTGTACAAGAGGCCAGACACC CAGTCGATCCAGAAAGTACAGGCAGAA TTCGACTCGTTTGTAGTTCCCGGGCTG TGGTCTTCAGGCCTGTCAATACCGTTG AGAACAAGGATCAAGTGGTTGCTGCGC AAGGTCCCTAAGGCGGAC
CHIKV St. Martin	∂A	1880-2032	26	GGCCGAGTCCTCGTACCGAGCGGTTA CGCCATCAGTCCTGAAGACTTCCAGA GTTTGAGCGAAAGCGCAACGATGGTC TACAACGAAAGAGAGTTTGTAAACAGA AAGCTCCACCACATAGCAATGCATGGA CCTGCGCTAAACACTGACGAA
CHIKV St. Martin	∂A	2213-2365	21	AAAATAGCAGTGATAGGGGTGTTCGG AGTACCAGGATCTGGTAAGTCCGCCA TTATCAAGAACCTAGTCACCCGGCAA GACCTAGTCACTTCAGGAAAGAAAGA GAACTGTCAAGAAATTAGCACAGACG TGATGAGACAAAGAGGCCTGGAA
CHIKV St. Martin	∂B	3239-3388	21	GAAATATGCACTAGAATGTATGGGGT GGATCTGGATAGCGGATTATTCTCCA AACCGCTAGTGTCTGTATATTACGCT GATAACCACTGGGATAACAGGCCAG GCGGCAAGATGTTCGGGTTCAACCC GGAGGCGGCGTCGATTCTAGAA
CHIKV St. Martin	∂B	4037-4189	24	AACGCCGCGTTTGTGGGACAGGCCA CCAGGGCCGGGTGTGCTCCCTCATA CCGCGTCAAGCGCATGGACATAGCG AAGAATGACGAGGAGTGCGTCGTAA ACGCAGCCAACCCACGTGGATTACC GGGAGACGGTGTTTGCAAAGCAGTA TAT
CHIKV St. Martin	∂C	4736-4888	18	CCAAAGCAAATTGAAGCCAATGAGC AGGTTTGCCTCTATGCCCTGGGGG AGAGTATAGAGTCCATCCGGCAAAA ATGCCCAGTGGATGATGCAGATGC ATCATCCCCTCCAAAAACTGTCCCG TGCCTATGCCGTTATGCGATGACTC CGGAA
CHIKV St. Martin	∂C	5030-5182	30	GTGAGTCCCCGCGAGTATAGATCAA GCCAGGAATCCGTAAGGGAAGTGA GTATGACCACGTCATTAACACACAG TCAGTTTGATCTAAGCGCTGACGG GGAGACGCTCCCAGTCCCGTCTGA CTTAGATGCCGATGCCCCTGCACT GGAACCG
CHIKV St. Martin	∂C	5630-5782	22	TTGCGACTGGACAGGGCAGGTGGG TACATATTTTCGTCAGATACGGGCC CTGGTCACCTGCAACAGAAGTCGG TACGCCAGTCGGTACTTCCAGTAAA CACACTAGAGGAAGTACACGAGGA GAAGTGTTATCCCCCTAAGCTGGAT GAATTA

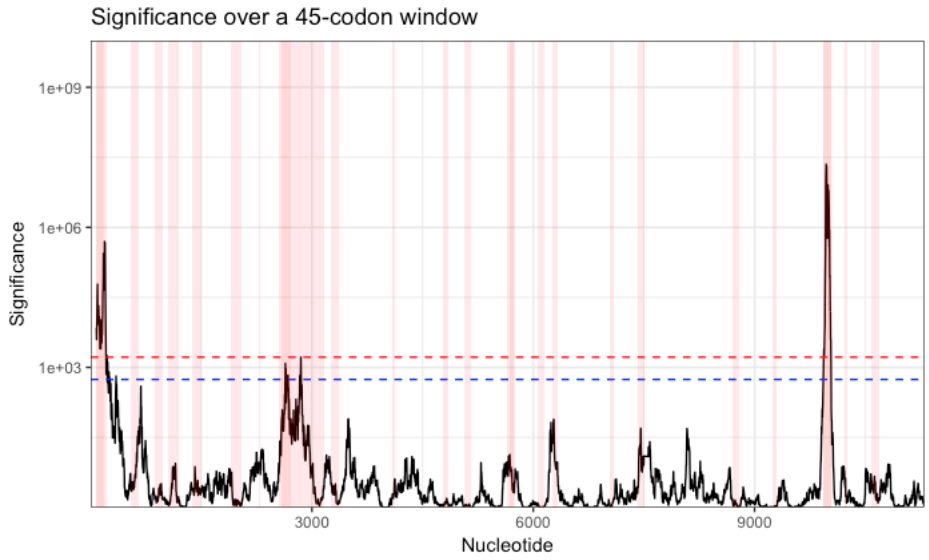
CHIKV St. Martin	∂C	6032-6181	23	GAGTTCTTGGCAAGGAACTATCCGACC GTCTCATCATACCAGATTACTGATGAG TACGATGCGTATTTAGACATGGTGGAT GGGTCGGAAAGCTGCTTAGACCGAGC TACATTCAACCCATCAAAACTCAGAAG TTACCCGAAACAACAT
CHIKV St. Martin	∂C	6218-6370	18	TCACCTTTCCAAAACACATTACAGAAC GTACTGGCCGCAGCCACCAAGAGGA ACTGCAATGTCACCCAGATGAGAGAA CTACCGACATTGGACTCAGCAGTATT CAACGTGGAATGTTTTAAAAAATTCGC ATGCAACAGGGAGTACTGGGAA
CHIKV St. Martin	∂C	6986-7138	23	ACTATCGCCAGTCGTGTCTTGGAAGA TCGCCTGACAAAATCCGCCTGCGCA GCTTTCATAGGCGATGACAACATAAT ACACGGGGTAGTGTCGGATGAATTG ATGGCTGCTCGATGCGCCACATGGA TGAACATGGAAGTGAAGATCATCGAT
CHIKV St. Martin	∂D	8671-8823	21	ATCCAGGTTTCGTTGCAAATTGGAAT AAAGACAGATGACAGCCACGATTGG ACGAAGCTGCGGTACATGGATAATC ATATGCCTGCAGATGCCGAGCGGG CAGGCTTATTCGTAAGAACGTCGGC ACCCTGCACCATTACAGGAACAATG GGA
CHIKV St. Martin	∂D	9190-9342	21	AAGGTCGATCAATGCCATGCGGCT GTGACCAATCACAAAAAATGGCAA TACAATTCGCCCCTGGTGCCTCGT AATGCCGAGTTCGGGGACAGAAAA GGGAAAGTCCATATTCCATTTCCTC TGGCTAATGTCACATGCCGGGTTC CAAAAGCA

APPENDIX C: SIGNIFICANCE OF SYNONYMOUS SITE CONSERVATION IN CHIKV

Data reported in Firth et al for the CHIKV strain NC004162 was replotted using R (151). CHIKV strain NC004162 is 92.3% identical at the nucleotide level and 97% identical at the protein level to the St. Martin CHIKV strain used in *Chapter 2*. Rolling windows of 15, 25, and 45 nucleotides were analyzed. The significance of codon conservation compared to expected was calculated using the equation reported in *synplot2* user guide. The significance thresholds for $P < 0.05$ (red dotted line) and $P < 0.15$ (blue dotted line) were also calculated for each window as recommended in Firth et al (151). Regions of the genome with known functionally important structures or identified as highly structured regions in *Chapter 2* are highlighted in red. Darker red regions indicate an overlap between known functionally important structures and highly structured regions.




Significance over a 25-codon window
1e+09
1e+06
1e+03
Significance
3000
6000
9000
Nucleotide


Significance over a 45-codon window
1e+09
1e+06
1e+03
Significance
3000
6000
9000
Nucleotide

REFERENCES

1. Saberi,A., Gulyaeva,A.A., Brubacher,J.L., Newmark,P.A. and Gorbalenya,A.E. (2018) A planarian nidovirus expands the limits of RNA genome size. *PLoS Pathog*, **14**, e1007314.

2. Carrasco-Hernandez,R., Jácome,R., López Vidal,Y. and Ponce de León,S. (2017) Are RNA Viruses Candidate Agents for the Next Global Pandemic? A Review. *ILAR Journal*, **58**, 343–358.

3. Morens,D.M. and Fauci,A.S. (2020) Emerging Pandemic Diseases: How We Got to COVID-19. *Cell*, **182**, 1077–1092.

4. Weaver,S.C. and Forrester,N.L. (2015) Chikungunya: Evolutionary history and recent epidemic spread. *ANTIVIRAL RESEARCH*, **120**, 32–39.

5. Chen,R., Mukhopadhyay,S., Merits,A., Bolling,B., Nasar,F., Coffey,L.L., Powers,A., Weaver,S.C.ICTV Report Consortium (2018) ICTV Virus Taxonomy Profile: Togaviridae. *Journal of General Virology*, **99**, 761–762.

6. Powers,A.M., Brault,A.C., Shirako,Y., Strauss,E.G., Kang,W., Strauss,J.H. and Weaver,S.C. (2001) Evolutionary relationships and systematics of the alphaviruses. *Journal of Virology*, **75**, 10118–10131.

7. Forrester,N.L., Palacios,G., Tesh,R.B., Savji,N., Guzman,H., Sherman,M., Weaver,S.C. and Lipkin,W.I. (2012) Genome-scale phylogeny of the alphavirus genus suggests a marine origin. *Journal of Virology*, **86**, 2729–2738.

8. Strauss,J.H. and Strauss,E.G. (1994) The alphaviruses: gene expression, replication, and evolution. *Microbiol Rev*, **58**, 491–562.

9. Adouchief,S., Smura,T., Sane,J., Vapalahti,O. and Kurkela,S. (2016) Sindbis virus as a human pathogen-epidemiology, clinical picture and pathogenesis. *Rev. Med. Virol.*, 10.1002/rmv.1876.

10. Silva,L.A. and Dermody,T.S. (2017) Chikungunya virus: epidemiology, replication, disease mechanisms, and prospective intervention strategies. *Journal of Clinical Investigation*, **127**, 737–749.

11. Baxter,V.K. and Heise,M.T. (2018) Genetic control of alphavirus pathogenesis. *Mamm. Genome*, **29**, 408–424.

12. Firth,A.E., Chung,B.Y., Fleeton,M.N. and Atkins,J.F. (2008) Discovery of frameshifting in Alphavirus 6K resolves a 20-year enigma. *Virol. J.*, **5**, 108.

13. Zhang,R., Kim,A.S., Fox,J.M., Nair,S., Basore,K., Klimstra,W.B., Rimkunas,R., Fong,R.H., Lin,H., Poddar,S., *et al.* (2018) Mxra8 is a receptor for multiple arthritogenic alphaviruses. *Nature*, **69**, 1.

14. Kim,A.S., Zimmerman,O., Fox,J.M., Nelson,C.A., Basore,K., Zhang,R., Durnell,L., Desai,C., Bullock,C., Deem,S.L., *et al.* (2020) An Evolutionary Insertion in the Mxra8 Receptor-Binding Site Confers Resistance to Alphavirus Infection and Pathogenesis. *Cell Host & Microbe*, 10.1016/j.chom.2020.01.008.

15. Rose,P.P., Hanna,S.L., Spiridigliozzi,A., Wannissorn,N., Beiting,D.P., Ross,S.R., Hardy,R.W., Bambina,S.A., Heise,M.T. and Cherry,S. (2011) Natural resistance-associated macrophage protein is a cellular receptor for sindbis virus in both insect and mammalian hosts. *Cell Host & Microbe*, **10**, 97–104.

16. Gardner,C.L., Ebel,G.D., Ryman,K.D. and Klimstra,W.B. (2011) Heparan sulfate binding by natural eastern equine encephalitis viruses promotes neurovirulence. *PNAS*, **108**, 16026–16031.

17. Doxsey,S.J., Brodsky,F.M., Blank,G.S. and Helenius,A. (1987) Inhibition of endocytosis by anti-clathrin antibodies. *Cell*, **50**, 453–463.

18. Ramanathan,A., Robb,G.B. and Chan,S.-H. (2016) mRNA capping: biological functions and applications. *Nucleic Acids Research*, **44**, 7511–7526.

19. Diamond,M.S. and Farzan,M. (2013) The broad-spectrum antiviral functions of IFIT and IFITM proteins. *Nature Reviews Immunology*, **13**, 46–57.

20. Hyde,J.L., Chen,R., Trobaugh,D.W., Diamond,M.S., Weaver,S.C., Klimstra,W.B. and Wilusz,J. (2015) The 5‘ and 3’ ends of alphavirus RNAs--Non-coding is not non-functional. *Virus Res.*, **206**, 99–107.

21. Kinney,R.M., Chang,G.J., Tsuchiya,K.R., Sneider,J.M., Roehrig,J.T., Woodward,T.M. and Trent,D.W. (1993) Attenuation of Venezuelan equine encephalitis virus strain TC-83 is encoded by the 5'-noncoding region and the E2 envelope glycoprotein. *Journal of Virology*, **67**, 1269–1277.

22. Reynaud,J.M., Kim,D.Y., Atasheva,S., Rasalouskaya,A., White,J.P., Diamond,M.S., Weaver,S.C., Frolova,E.I. and Frolov,I. (2015) IFIT1 Differentially Interferes with Translation and Replication of Alphavirus Genomes and Promotes Induction of Type I Interferon. *PLoS Pathog*, **11**, e1004863.

23. Niesters,H.G.M. and Strauss,J.H. (1990) Mutagenesis of the Conserved 51-Nucleotide Region of Sindbis Virus. *Journal of Virology*, **64**, 1639–1647.

24. Frolov,I., HARDY,R. and RICE,C.M. (2001) Cis-acting RNA elements at the 5′ end of Sindbis virus genome RNA regulate minus- and plus-strand RNA synthesis. *RNA*, **7**, 1638–1651.

25. Fayzulin,R. and Frolov,I. (2004) Changes of the Secondary Structure of the 5' End of the Sindbis Virus Genome Inhibit Virus Growth in Mosquito Cells and Lead to Accumulation of Adaptive Mutations. *Journal of Virology*, **78**, 4953–4964.

26. Michel,G., Petrakova,O., Atasheva,S. and Frolov,I. (2007) Adaptation of Venezuelan equine encephalitis virus lacking 51-nt conserved sequence element to replication in mammalian and mosquito cells. *Virology*, **362**, 475–487.

27. Gorchakov,R., HARDY,R., RICE,C.M. and Frolov,I. (2004) Selection of functional 5' cis-acting elements promoting efficient sindbis virus genome replication. *Journal of Virology*, **78**, 61–75.

28. Kim,D.Y., Firth,A.E., Atasheva,S., Frolova,E.I. and Frolov,I. (2011) Conservation of a packaging signal and the viral genome RNA packaging mechanism in alphavirus evolution. *Journal of Virology*, **85**, 8022–8036.

29. Sokoloski,K.J., Nease,L.M., May,N.A., Gebhart,N.N., Jones,C.E., Morrison,T.E. and Hardy,R.W. (2017) Identification of Interactions between Sindbis Virus Capsid Protein and Cytoplasmic vRNA as Novel Virulence Determinants. *PLoS Pathog*, **13**, e1006473.

30. Brown,R.S., Anastasakis,D.G., Hafner,M. and Kielian,M. (2020) Multiple capsid protein binding sites mediate selective packaging of the alphavirus genomic RNA. *Nat Commun*, **11**, 4693–16.

31. Li,G. and Rice,C.M. (1993) The signal for translational readthrough of a UGA codon in Sindbis virus RNA involves a single cytidine residue immediately downstream of the termination codon. *Journal of Virology*, **67**, 5062–5067.

32. Firth,A.E., Wills,N.M., Gesteland,R.F. and Atkins,J.F. (2011) Stimulation of stop codon readthrough: frequent presence of an extended 3' RNA structural element. *Nucleic Acids Research*, **39**, 6679–6691.

33. Kendra,J.A., Advani,V.M., Chen,B., Briggs,J.W., Zhu,J., Bress,H.J., Pathy,S.M. and Dinman,J.D. (2018) Functional and structural characterization of the chikungunya virus translational recoding signals. *J. Biol. Chem.*, **293**, 17536–17545.

34. Jones,J.E., Long,K.M., Whitmore,A.C., Sanders,W., Thurlow,L.R., Brown,J.A., Morrison,C.R., Vincent,H., Peck,K.M., Browning,C., *et al.* (2017) Disruption of the Opal Stop Codon Attenuates Chikungunya Virus-Induced Arthritis and Pathology. *mBio*, **8**.

35. Suthar,M.S., Shabman,R., Madric,K., Lambeth,C. and Heise,M.T. (2005) Identification of Adult Mouse Neurovirulence Determinants of the Sindbis Virus Strain AR86. *Journal of Virology*, **79**, 4219–4228.

36. Myles,K.M., Kelly,C.L.H., Ledermann,J.P. and Powers,A.M. (2006) Effects of an opal termination codon preceding the nsP4 gene sequence in the O'Nyong-Nyong virus genome on Anopheles gambiae infectivity. *Journal of Virology*, **80**, 4992–4997.

37. Gorchakov,R., Frolova,E., Williams,B.R.G., RICE,C.M. and Frolov,I. (2004) PKR-dependent and -independent mechanisms are involved in translational shutoff during Sindbis virus infection. *Journal of Virology*, **78**, 8455–8467.

38. Ventoso,I., Sanz,M.A., Molina,S., Berlanga,J.J., Carrasco,L. and Esteban,M. (2006) Translational resistance of late alphavirus mRNA to eIF2alpha phosphorylation: a strategy to overcome the antiviral effect of protein kinase PKR. *Genes Dev.*, **20**, 87–100.

39. Frolov,I. and Schlesinger,S. (1994) Translation of Sindbis virus mRNA: effects of sequences downstream of the initiating codon. *Journal of Virology*, **68**, 8111–8117.

40. Frolov,I. and Schlesinger,S. (1996) Translation of Sindbis virus mRNA: analysis of sequences downstream of the initiating AUG codon that enhance translation. *Journal of Virology*, **70**, 1182–1190.

41. Ventoso,I. (2012) Adaptive changes in alphavirus mRNA translation allowed colonization of vertebrate hosts. *Journal of Virology*, **86**, 9484–9494.

42. Toribio,R., Díaz-López,I., Boskovic,J. and Ventoso,I. (2018) Translation initiation of alphavirus mRNA reveals new insights into the topology of the 48S initiation complex. *Nucleic Acids Research*, **46**, 4176–4187.

43. Sanz,M.A., Almela,E.G., Garcia-Moreno,M., Marina,A.I. and Carrasco,L. (2019) A viral RNA motif involved in signaling the initiation of translation on non-AUG codons. *RNA*, **25**, 431–452.

44. Sanz,M.A., Castelló,A., Ventoso,I., Berlanga,J.J. and Carrasco,L. (2009) Dual mechanism for the translation of subgenomic mRNA from Sindbis virus in infected and uninfected cells. *PLoS ONE*, **4**, e4772–12.

45. Broeckel,R., Sarkar,S., May,N.A., Totonchy,J., Kreklywich,C.N., Smith,P., Graves,L., DeFilippis,V.R., Heise,M.T., Morrison,T.E., *et al.* (2019) Src Family Kinase Inhibitors Block Translation of Alphavirus Subgenomic mRNAs. *Antimicrob Agents Chemother*, **63**, 517–20.

46. Chung,B.Y.W., Firth,A.E. and Atkins,J.F. (2010) Frameshifting in alphaviruses: a diversity of 3' stimulatory structures. *J. Mol. Biol.*, **397**, 448–456.

47. Ou,J.H., Trent,D.W. and Strauss,J.H. (1982) The 3'-non-coding regions of alphavirus RNAs contain repeating sequences. *J. Mol. Biol.*, **156**, 719–730.

48. Chen,R., Wang,E., Tsetsarkin,K.A. and Weaver,S.C. (2013) Chikungunya virus 3' untranslated region: adaptation to mosquitoes and a population bottleneck as major evolutionary forces. *PLoS Pathog*, **9**, e1003591.

49. Stapleford,K.A., Moratorio,G., Henningsson,R., Chen,R., Matheus,S., Enfissi,A., Weissglas-Volkov,D., Isakov,O., Blanc,H., Mounce,B.C., *et al.* (2016) Whole-Genome Sequencing Analysis from the Chikungunya Virus Caribbean Outbreak Reveals Novel Evolutionary Genomic Elements. *PLOS Neglected Tropical Diseases*, **10**, e0004402.

50. Morley,V.J., Noval,M.G., Chen,R., Weaver,S.C., Vignuzzi,M., Stapleford,K.A. and Turner,P.E. (2018) Chikungunya virus evolution following a large 3'UTR deletion results in host-specific molecular changes in protein-coding regions. *Virus Evol*, **4**, vey012.

51. Garcia-Moreno,M., Sanz,M.A. and Carrasco,L. (2016) A Viral mRNA Motif at the 3'-Untranslated Region that Confers Translatability in a Cell-Specific Manner. Implications for Virus Evolution. *Sci Rep*, **6**, 19217.

52. Filomatori,C.V., Bardossy,E.S., Merwaiss,F., Suzuki,Y., Henrion,A., Saleh,M.-C. and Alvarez,D.E. (2019) RNA recombination at Chikungunya virus 3'UTR as an evolutionary mechanism that provides adaptability. *PLoS Pathog*, **15**, e1007706.

53. Villordo,S.M., Filomatori,C.V., Sánchez-Vargas,I., Blair,C.D. and Gamarnik,A.V. (2015) Dengue virus RNA structure specialization facilitates host adaptation. *PLoS Pathog*, **11**, e1004604.

54. Filomatori,C.V., Carballeda,J.M., Villordo,S.M., Aguirre,S., Pallarés,H.M., Maestre,A.M., Sánchez-Vargas,I., Blair,C.D., Fabri,C., Morales,M.A., *et al.* (2017) Dengue virus genomic variation associated with mosquito adaptation defines the pattern of viral non-coding RNAs and fitness in human cells. *PLoS Pathog*, **13**, e1006265.

55. Mathews,D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**, 526–532.

56. 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.

57. 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.

58. 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.

59. Bellaousov,S., Reuter,J.S., Seetin,M.G. and Mathews,D.H. (2013) RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Research*, **41**, W471–4.

60. Lockard,R.E. and Kumar,A. (1981) Mapping tRNA structure in solution using double-strand-specific ribonuclease V1 from cobra venom. *Nucleic Acids Research*, **9**, 5125–5140.

61. Auron,P.E., Weber,L.D. and Rich,A. (1982) Comparison of transfer ribonucleic acid structures using cobra venom and S1 endonucleases. *Biochemistry*, **21**, 4700–4706.

62. 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.

63. 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). *J. Am. Chem. Soc.*, **127**, 4223–4231.

64. Mortimer,S.A. and Weeks,K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.*, **129**, 4144–4145.

65. 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.

66. Smola,M.J. and Weeks,K.M. (2018) In-cell RNA structure probing with SHAPE-MaP. *Nat Protoc*, **13**, 1181–1195.

67. Sanders,W., Fritch,E.J., Madden,E.A., Graham,R.L., Vincent,H.A., Heise,M.T., Baric,R.S. and Moorman,N.J. (2020) Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. *bioRxiv*, 10.1101/2020.06.15.153197.

68. Sükösd,Z., Swenson,M.S., Kjems,J. and Heitsch,C.E. (2013) Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Research*, **41**, 2807–2816.

69. Eddy,S.R. (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu Rev Biophys*, **43**, 433–456.

70. Kutchko,K.M. and Laederach,A. (2017) Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *Wiley Interdiscip Rev RNA*, **8**.

71. Deigan,K.E., Li,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination. *PNAS*, **106**, 97–102.

72. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.

73. Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

74. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.

75. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474–13.

76. Levitt,M. (1969) Detailed molecular model for transfer ribonucleic acid. *Nature*, **224**, 759–763.

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

78. Rivas,E., Clements,J. and Eddy,S.R. (2020) Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, **36**, 3072–3076.

79. Laine,M., Luukkainen,R. and Toivanen,A. (2004) Sindbis viruses and other alphaviruses as cause of human arthritic disease. *J Intern Med*, **256**, 457–471.

80. Suhrbier,A., Jaffar-Bandjee,M.-C. and Gasque,P. (2012) Arthritogenic alphaviruses--an overview. *Nat Rev Rheumatol*, **8**, 420–429.

81. Go,Y.Y., Balasuriya,U.B.R. and Lee,C.-K. (2014) Zoonotic encephalitides caused by arboviruses: transmission and epidemiology of alphaviruses and flaviviruses. *Clin Exp Vaccine Res*, **3**, 58–77.

82. Hyde,J.L., Gardner,C.L., Kimura,T., White,J.P., Liu,G., Trobaugh,D.W., Huang,C., Tonelli,M., Paessler,S., Takeda,K., *et al.* (2014) A viral RNA structural element alters host recognition of nonself RNA. *Science*, **343**, 783–787.

83. Kendra,J.A., la Fuente,de,C., Brahms,A., Woodson,C., Bell,T.M., Chen,B., Khan,Y.A., Jacobs,J.L., Kehn-Hall,K. and Dinman,J.D. (2017) Ablation of Programmed -1 Ribosomal Frameshifting in Venezuelan Equine Encephalitis Virus Results in Attenuated Neuropathogenicity. *Journal of Virology*, **91**, 281–13.

84. Noller,H.F., Kop,J., Wheaton,V., Brosius,J., Gutell,R.R., Kopylov,A.M., Dohme,F., Herr,W., Stahl,D.A., Gupta,R., *et al.* (1981) Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Research*, **9**, 6167–6189.

85. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research*, **20**, 5785–5795.

86. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Research*, **22**, 2079–2088.

87. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

88. Holley,R.W., APGAR,J., EVERETT,G.A., MADISON,J.T., MARQUISEE,M., MERRILL,S.H., PENSWICK,J.R. and ZAMIR,A. (1965) STRUCTURE OF A RIBONUCLEIC ACID. *Science*, **147**, 1462–1465.

89. Gutell,R.R., Lee,J.C. and Cannone,J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Engineering and design*, **12**, 301–310.

90. Kutchko,K.M., Sanders,W., Ziehr,B., Phillips,G., Solem,A., Halvorsen,M., Weeks,K.M., Moorman,N. and Laederach,A. (2015) Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR. *RNA*, **21**, 1274–1285.

91. Somarowthu,S., Legiewicz,M., Chillón,I., Marcia,M., Liu,F. and Pyle,A.M. (2015) HOTAIR forms an intricate and modular secondary structure. *Mol. Cell*, **58**, 353–361.

92. 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 Meth*, **11**, 959–965.

93. 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. *PNAS*, 10.1073/pnas.1416266112.

94. Watts,J.M., Dang,K.K., Gorelick,R.J., Leonard,C.W., Bess,J.W., Swanstrom,R., Burch,C.L. and Weeks,K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.

95. Smola,M.J., Calabrese,J.M. and Weeks,K.M. (2015) Detection of RNA-Protein Interactions in Living Cells with SHAPE. *Biochemistry*, **54**, 6867–6875.

96. Valdar,W.S.J. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.

97. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

98. Le,S.Y. and Maizel,J.V. (1989) A method for assessing the statistical significance of RNA folding. *J Theor Biol*, **138**, 495–510.

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

100. Snyder,J.E., Kulcsar,K.A., Schultz,K.L.W., Riley,C.P., Neary,J.T., Marr,S., Jose,J., Griffin,D.E. and Kuhn,R.J. (2013) Functional characterization of the alphavirus TF protein. *Journal of Virology*, **87**, 8511–8523.

101. Pollom,E., Dang,K.K., Potter,E.L., Gorelick,R.J., Burch,C.L., Weeks,K.M. and Swanstrom,R. (2013) Comparison of SIV and HIV-1 Genomic RNA Structures Reveals Impact of Sequence Evolution on Conserved and Non-Conserved Structural Motifs. *PLoS Pathog*, **9**, e1003294–17.

102. Jorge,D.M. de M., Mills,R.E. and Lauring,A.S. (2015) CodonShuffle: a tool for generating and analyzing synonymously mutated sequences. *Virus Evol*, **1**, vev012–9.

103. Mueller,S., Papamichail,D., Coleman,J.R., Skiena,S. and Wimmer,E. (2006) Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *Journal of Virology*, **80**, 9687–9696.

104. Burns,C.C., Shaw,J., Campagnoli,R., Jorba,J., Vincent,A., Quay,J. and Kew,O. (2006) Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *Journal of Virology*, **80**, 3259–3272.

105. Desmyter,J., Melnick,J.L. and Rawls,W.E. (1968) Defectiveness of interferon production and of rubella virus interference in a line of African green monkey kidney cells (Vero). *Journal of Virology*, **2**, 955–961.

106. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

107. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J., *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, **43**, D130–7.

108. Gardner,P.P. (2009) The use of covariance models to annotate RNAs in whole genomes. *Briefings in Functional Genomics and Proteomics*, **8**, 444–450.

109. Weaver,S.C., Winegar,R., Manger,I.D. and Forrester,N.L. (2012) Alphaviruses: population genetics and determinants of emergence. *ANTIVIRAL RESEARCH*, **94**, 242–257.

110. Nasar,F., Palacios,G., Gorchakov,R.V., Guzman,H., Da Rosa,A.P.T., Savji,N., Popov,V.L., Sherman,M.B., Lipkin,W.I., Tesh,R.B., *et al.* (2012) Eilat virus, a unique alphavirus with host range restricted to insects by RNA replication. *PNAS*, **109**, 14622–14627.

111. Ritz,J., Martin,J.S. and Laederach,A. (2013) Evolutionary evidence for alternative structure in RNA sequence co-variation. *PLoS Comput Biol*, **9**, e1003152–11.

112. Loughrey,D., Watters,K.E., Settle,A.H. and Lucks,J.B. (2014) SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research*, **42**, e165–e165.

113. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

114. Ding,Y., Tang,Y., Kwok,C.K., Zhang,Y., Bevilacqua,P.C. and Assmann,S.M. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**, 696–700.

115. Wan,Y., Qu,K., Zhang,Q.C., Flynn,R.A., Manor,O., Ouyang,Z., Zhang,J., Spitale,R.C., Snyder,M.P., Segal,E., *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.

116. 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.

117. Wilkinson,K.A., Gorelick,R.J., Vasa,S.M., Guex,N., Rein,A., Mathews,D.H., Giddings,M.C. and Weeks,K.M. (2008) High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol*, **6**, e96–17.

118. Rocca-Serra,P., Bellaousov,S., Birmingham,A., Chen,C., Cordero,P., Das,R., Davis-Neulander,L., Duncan,C.D.S., Halvorsen,M., Knight,R., *et al.* (2011) Sharing and archiving nucleic acid structure mapping data. *RNA*, **17**, 1204–1212.

119. Katoh,K., Misawa,K., Kuma,K.-I. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

120. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

121. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

122. Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

123. Soldatov,R.A., Vinogradova,S.V. and Mironov,A.A. (2014) RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments. *Bioinformatics*, **30**, 457–463.

124. 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. *PNAS*, **101**, 7287–7292.

125. Huynen,M., Gutell,R. and Konings,D. (1997) Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.*, **267**, 1104–1112.

126. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

127. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**, 1281–1295.

128. Wollish,A.C., Ferris,M.T., Blevins,L.K., Loo,Y.-M., Gale,M. and Heise,M.T. (2013) An attenuating mutation in a neurovirulent Sindbis virus strain interacts with the IPS-1 signaling pathway in vivo. *Virology*, **435**, 269–280.

129. Park,E. and Griffin,D.E. (2009) The nsP3 macro domain is important for Sindbis virus replication in neurons and neurovirulence in mice. *Virology*, **388**, 305–314.

130. Nawrocki,E.P. and Eddy,S.R. (2013) Computational identification of functional RNA homologs in metagenomic data. *RNA Biol*, **10**, 1170–1179.

131. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

132. Morrison,C.R., Plante,K.S. and Heise,M.T. (2016) Chikungunya Virus: Current Perspectives on a Reemerging Virus. In *Emerging infections 10*. American Society of Microbiology, pp. 143–161.

133. 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 Research*, **46**, 3657–3670.

134. Carrasco,L., Sanz,M.A. and González-Almela,E. (2018) The Regulation of Translation in Alphavirus-Infected Cells. *Viruses*, **10**.

135. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth*, **9**, 357–359.

136. 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.

137. 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., *et al.* (2018) Pervasive tertiary structure in the dengue virus RNA genome. *PNAS*, 10.1073/pnas.1716689115.

138. Chen,R., Puri,V., Fedorova,N., Lin,D., Hari,K.L., Jain,R., Rodas,J.D., Das,S.R., Shabman,R.S. and Weaver,S.C. (2016) Comprehensive Genome Scale Phylogenetic Study Provides New Insights on the Global Expansion of Chikungunya Virus. *Journal of Virology*, **90**, 10600–10611.

139. Kendall,C., Khalid,H., Müller,M., Banda,D.H., Kohl,A., Merits,A., Stonehouse,N.J. and Tuplin,A. (2019) Structural and phenotypic analysis of Chikungunya virus RNA replication elements. *Nucleic Acids Research*, 10.1093/nar/gkz640.

140. Toribio,R., Díaz-López,I., Boskovic,J. and Ventoso,I. (2016) An RNA trapping mechanism in Alphavirus mRNA promotes ribosome stalling and translation initiation. *Nucleic Acids Research*, **44**, 4368–4380.

141. Villordo,S.M., Carballeda,J.M., Filomatori,C.V. and Gamarnik,A.V. (2016) RNA Structure Duplications and Flavivirus Host Adaptation. *Trends Microbiol.*, **24**, 270–283.

142. Chapman,E.G., Costantino,D.A., Rabe,J.L., Moon,S.L., Wilusz,J., Nix,J.C. and Kieft,J.S. (2014) The Structural Basis of Pathogenic Subgenomic Flavivirus RNA (sfRNA) Production. *Science*, **344**, 307–310.

143. Chapman,E.G., Moon,S.L., Wilusz,J. and Kieft,J.S. (2014) RNA structures that resist degradation by Xrn1 produce a pathogenic Dengue virus RNA. *eLife*, **3**, e01892.

144. Schneider,A. de B., Ochsenreiter,R., Hostager,R., Hofacker,I.L., Janies,D. and Wolfinger,M.T. (2019) Updated Phylogeny of Chikungunya Virus Suggests Lineage-Specific RNA Architecture. *Viruses*, **11**.

145. Filomatori,C.V., Merwaiss,F., Bardossy,E.S. and Alvarez,D.E. (2020) Impact of alphavirus 3'UTR plasticity on mosquito transmission. *Semin. Cell Dev. Biol.*, 10.1016/j.semcdb.2020.07.006.

146. Boerneke,M.A., Ehrhardt,J.E. and Weeks,K.M. (2019) Physical and Functional Analysis of Viral RNA Genomes by SHAPE. *Annu. Rev. Virol.*, **6**, 93–117.

147. 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.

148. Villordo,S.M. and Gamarnik,A.V. (2009) Genome cyclization as strategy for flavivirus RNA replication. *Virus Res.*, **139**, 230–239.

149. de Borba,L., Villordo,S.M., Iglesias,N.G., Filomatori,C.V., Gebhard,L.G. and Gamarnik,A.V. (2015) Overlapping local and long-range RNA-RNA interactions modulate dengue virus genome cyclization and replication. *Journal of Virology*, **89**, 3430–3437.

150. Filomatori,C.V., Lodeiro,M.F., Alvarez,D.E., Samsa,M.M., Pietrasanta,L. and Gamarnik,A.V. (2006) A 5' RNA element promotes dengue virus RNA synthesis on a circular genome. *Genes Dev.*, **20**, 2238–2249.

151. Firth,A.E. (2014) Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Research*, **42**, 12425–12439.

152. Morrison,T.E., Oko,L., Montgomery,S.A., Whitmore,A.C., Lotstein,A.R., Gunn,B.M., Elmore,S.A. and Heise,M.T. (2011) A Mouse Model of Chikungunya Virus–Induced Musculoskeletal Inflammatory Disease. *The American Journal of Pathology*, **178**, 32–40.

153. Duffy,M.R., Chen,T.-H., Hancock,W.T., Powers,A.M., Kool,J.L., Lanciotti,R.S., Pretrick,M., Marfel,M., Holzbauer,S., Dubray,C., *et al.* (2009) Zika virus outbreak on Yap Island, Federated States of Micronesia. *N. Engl. J. Med.*, **360**, 2536–2543.

154. Faria,N.R., Azevedo,R.D.S.D.S., Kraemer,M.U.G., Souza,R., Cunha,M.S., Hill,S.C., Theze,J., Bonsall,M.B., Bowden,T.A., Rissanen,I., *et al.* (2016) Zika virus in the Americas: Early epidemiological and genetic findings. *Science*, **352**, 345–349.

155. França,G.V.A., Schuler-Faccini,L., Oliveira,W.K., Henriques,C.M.P., Carmo,E.H., Pedi,V.D., Nunes,M.L., Castro,M.C., Serruya,S., Silveira,M.F., *et al.* (2016) Congenital Zika virus syndrome in Brazil: a case series of the first 1501 livebirths with complete investigation. *Lancet*, **388**, 891–897.

156. Miranda-Filho,D. de B., Martelli,C.M.T., Ximenes,R.A. de A., Araújo,T.V.B., Rocha,M.A.W., Ramos,R.C.F., Dhalia,R., França,R.F. de O., Marques Júnior,E.T. de A. and Rodrigues,L.C. (2016) Initial Description of the Presumed Congenital Zika Syndrome. *Am J Public Health*, **106**, 598–600.

157. Gebhard,L.G., Filomatori,C.V. and Gamarnik,A.V. (2011) Functional RNA Elements in the Dengue Virus Genome. *Viruses*, **3**, 1739–1756.

158. Li,X.-F., Jiang,T., Yu,X.-D., Deng,Y.-Q., Zhao,H., Zhu,Q.-Y., Qin,E.-D. and Qin,C.-F. (2010) RNA elements within the 5' untranslated region of the West Nile virus genome are critical for RNA synthesis and virus replication. *J. Gen. Virol.*, **91**, 1218–1223.

159. 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.

160. 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.

161. Bujalowski,P.J., Bujalowski,W. and Choi,K.H. (2020) Identification of the viral RNA promoter stem loop A (SLA)-binding site on Zika virus polymerase NS5. *Sci Rep*, **10**, 13306–13.

162. 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.

163. Selisko,B., Wang,C., Harris,E. and Canard,B. (2014) Regulation of Flavivirus RNA synthesis and replication. *Current Opinion in Virology*, **9**, 74–83.

164. 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.

165. Akiyama,B.M., Laurence,H.M., Massey,A.R., Costantino,D.A., Xie,X., Yang,Y., Shi,P.-Y., Nix,J.C., Beckham,J.D. and Kieft,J.S. (2016) Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science*, **354**, 1148–1152.

166. Lodeiro,M.F., Filomatori,C.V. and Gamarnik,A.V. (2008) Structural and Functional Studies of the Promoter Element for Dengue Virus RNA Replication. *Journal of Virology*, **83**, 993–1008.

167. Göertz,G.P., Abbo,S.R., Fros,J.J. and Pijlman,G.P. (2018) Functional RNA during Zika virus infection. *Virus Res.*, **254**, 41–53.

168. Kunec,D. and Osterrieder,N. (2016) Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep*, **14**, 55–67.

169. Miller,J.R., Koren,S., Dilley,K.A., Puri,V., Brown,D.M., Harkins,D.M., Thibaud-Nissen,F., Rosen,B., Chen,X.-G., Tu,Z., *et al.* (2018) Analysis of the Aedes albopictus C6/36 genome provides insight into cell line utility for viral propagation. *Gigascience*, **7**, 1–13.

170. Osada,N., Kohara,A., Yamaji,T., Hirayama,N., Kasai,F., Sekizuka,T., Kuroda,M. and Hanada,K. (2014) The genome landscape of the african green monkey kidney-derived vero cell line. *DNA Res*, **21**, 673–683.

171. Jacobs,J.P., Jones,C.M. and Baille,J.P. (1970) Characteristics of a human diploid cell designated MRC-5. *Nature*, **227**, 168–170.

172. Lazear,H.M., Govero,J., Smith,A.M., Platt,D.J., Fernandez,E., Miner,J.J. and Diamond,M.S. (2016) A Mouse Model of Zika Virus Pathogenesis. *Cell Host & Microbe*, **19**, 720–730.

173. Alvarez,D.E., Filomatori,C.V. and Gamarnik,A.V. (2008) Functional analysis of dengue virus cyclization sequences located at the 5‘ and 3’UTRs. *Virology*, **375**, 223–235.

174. Shan,C., Muruato,A.E., Nunes,B.T.D., Luo,H., Xie,X., Medeiros,D.B.A., Wakamiya,M., Tesh,R.B., Barrett,A.D., Wang,T., *et al.* (2017) A live-attenuated Zika virus vaccine candidate induces sterilizing immunity in mouse models. *Nature Medicine*, **23**, 1–7.

175. Barrett,A.D.T. and Gould,E.A. (1986) Comparison of Neurovirulence of Different Strains of Yellow Fever Virus in Mice. *J. Gen. Virol.*

176. Miner,J.J. and Diamond,M.S. (2017) Zika Virus Pathogenesis and Tissue Tropism. *Cell Host & Microbe*, **21**, 134–142.

177. Klein,S.L., Marriott,I. and Fish,E.N. (2015) Sex-based differences in immune function and responses to vaccination. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **109**, 9–15.

178. Steeg,vom,L.G. and Klein,S.L. (2016) SeXX Matters in Infectious Disease Pathogenesis. *PLoS Pathog*, **12**, e1005374–6.

179. Moon,S.L., Anderson,J.R., Kumagai,Y., Wilusz,C.J., Akira,S., Khromykh,A.A. and Wilusz,J. (2012) A noncoding RNA produced by arthropod-borne flaviviruses inhibits the cellular exoribonuclease XRN1 and alters host mRNA stability. *RNA*, **18**, 2029–2040.

180. Schuessler,A., Funk,A., Lazear,H.M., Cooper,D.A., Torres,S., Daffis,S., Jha,B.K., Kumagai,Y., Takeuchi,O., Hertzog,P., *et al.* (2012) West Nile virus noncoding subgenomic RNA contributes to viral evasion of the type I interferon-mediated antiviral response. *Journal of Virology*, **86**, 5708–5718.

181. Slonchak,A. and Khromykh,A.A. (2018) Subgenomic flaviviral RNAs: What do we know after the first decade of research. *ANTIVIRAL RESEARCH*, **159**, 13–25.

182. Manokaran,G., Finol,E., Wang,C., Gunaratne,J., Bahl,J., Ong,E.Z., Tan,H.C., Sessions,O.M., Ward,A.M., Gubler,D.J., *et al.* (2015) Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness. *Science*, **350**, 217–221.

183. Boissier,J., Chlichlia,K., Digon,Y., Ruppel,A. and Moné,H. (2003) Preliminary study on sex-related inflammatory reactions in mice infected with Schistosoma mansoni. *Parasitology Research*, **91**, 144–150.

184. Klein,S.L., Jedlicka,A. and Pekosz,A. (2010) The Xs and Y of immune responses to viral vaccines. *The Lancet Infectious Diseases*, **10**, 338–349.

185. 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**, 1408.

186. Widman,D.G., Young,E., Yount,B.L., Plante,K.S., Gallichotte,E.N., Carbaugh,D.L., Peck,K.M., Plante,J., Swanstrom,J., Heise,M.T., *et al.* (2017) A Reverse Genetics Platform That Spans the Zika Virus Family Tree. *mBio*, **8**, 509–15.

187. Zeller,H., Van Bortel,W. and Sudre,B. (2016) Chikungunya: Its History in Africa and Asia and Its Spread to New Regions in 2013-2014. *J. Infect. Dis.*, **214**, S436–S440.

188. Azar,S.R., Campos,R.K., Bergren,N.A., Camargos,V.N. and Rossi,S.L. (2020) Epidemic Alphaviruses: Ecology, Emergence and Outbreaks. *Microorganisms*, **8**, 1167–37.

189. Spitale,R.C., Flynn,R.A., Zhang,Q.C., Crisalli,P., Lee,B., Jung,J.-W., Kuchelmeister,H.Y., Batista,P.J., Torre,E.A., Kool,E.T., *et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**, 486–490.

190. Alford,R.L., Honda,S., Lawrence,C.B. and Belmont,J.W. (1991) RNA secondary structure analysis of the packaging signal for Moloney murine leukemia virus. *Virology*, **183**, 611–619.

191. Duke,G.M., Hoffman,M.A. and Palmenberg,A.C. (1992) Sequence and structural elements that contribute to efficient encephalomyocarditis virus RNA translation. *Journal of Virology*, **66**, 1602–1609.

192. McKnight,K.L. and Lemon,S.M. (1998) The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA*, **4**, 1569–1584.

193. Linder,B., Grozhik,A.V., Olarerin-George,A.O., Meydan,C., Mason,C.E. and Jaffrey,S.R. (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Meth*, **12**, 767–772.

194. Meyer,K.D. and Jaffrey,S.R. (2014) The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat. Rev. Mol. Cell Biol.*, **15**, 313–326.

195. Gokhale,N.S., McIntyre,A.B.R., McFadden,M.J., Roder,A.E., Kennedy,E.M., Gandara,J.A., Hopcraft,S.E., Quicke,K.M., Vazquez,C., Willer,J., *et al.* (2016) N6-Methyladenosine in Flaviviridae Viral RNA Genomes Regulates Infection. *Cell Host & Microbe*, **20**, 654–665.

196. Li,G.P. and Rice,C.M. (1989) Mutagenesis of the in-frame opal termination codon preceding nsP4 of Sindbis virus: studies of translational readthrough and its effect on virus replication. *Journal of Virology*, **63**, 1326–1337.

197. Levinson,R.S., Strauss,J.H. and Strauss,E.G. (1990) Complete sequence of the genomic RNA of O'nyong-nyong virus and its use in the construction of alphavirus phylogenetic trees. *Virology*, **175**, 110–123.

198. Levitt,N.H., Ramsburg,H.H., Hasty,S.E., Repik,P.M., Cole,F.E. and Lupton,H.W. (1986) Development of an attenuated strain of chikungunya virus for use in vaccine production. *Vaccine*, **4**, 157–162.

199. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

200. Lauring,A.S. and Andino,R. (2010) Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog*, **6**, e1001005–8.

201. Trobaugh,D.W., Gardner,C.L., Sun,C., Haddow,A.D., Wang,E., Chapnik,E., Mildner,A., Weaver,S.C., Ryman,K.D. and Klimstra,W.B. (2014) RNA viruses can hijack vertebrate microRNAs to suppress innate immunity. *Nature*, **506**, 245–248.

202. Langsjoen,R.M., Haller,S.L., Roy,C.J., Vinet-Oliphant,H., Bergren,N.A., Erasmus,J.H., Livengood,J.A., Powell,T.D., Weaver,S.C. and Rossi,S.L. (2018) Chikungunya Virus Strains Show Lineage-Specific Variations in Virulence and Cross-Protective Ability in Murine and Nonhuman Primate Models. *mBio*, **9**.

203. Stern,A., Yeh,M.T., Zinger,T., Smith,M., Wright,C., Ling,G., Nielsen,R., Macadam,A. and Andino,R. (2017) The Evolutionary Pathway to Virulence of an RNA Virus. *Cell*, **169**, 35–46.e19.

204. Kenney,J.L., Volk,S.M., Pandya,J., Wang,E., Liang,X. and Weaver,S.C. (2011) Stability of RNA virus attenuation approaches. *Vaccine*, **29**, 2230–2234.

205. Vignuzzi,M. and López,C.B. (2019) Defective viral genomes are key drivers of the virus-host interaction. *Nat Microbiol*, **4**, 1075–1087.

206. Price,A.M., Hayer,K.E., McIntyre,A.B.R., Gokhale,N.S., Abebe,J.S., Fera,Della,A.N., Mason,C.E., Horner,S.M., Wilson,A.C., Depledge,D.P., *et al.* (2020) Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. *Nat Commun*, **11**, 6016–17.