Genome structure in RNA viruses

Michael Liam Knight

Thesis submitted for the degree of Doctor of Philosophy

Sir William Dunn School of Pathology

St Edmund Hall

University of Oxford

Trinity term 2021

Contents

Table of Figures

Table of Tables

Abstract

RNA structures can have important and diverse functions. Influenza A viruses have eight negative sense RNA genome segments. When two influenza virus strains infect the same cell their progeny can package segments from both strains. This process, termed reassortment, can lead to rapid genetic shifts that have previously generated pandemic strains.

In this investigation the RNA-RNA interaction maps for several influenza viruses are generated using a high-throughput sequencing approach. This reveals extensive, redundant networks of RNA-RNA interactions between the genomic viral RNA segments. This analysis includes H1N1 and H3N2 reassortants, where these interactions can explain preferential co-segregation of segments during reassortment. It is also demonstrated, using chemical probing techniques, that there is extensive local RNA structure within the influenza genomic segments. Several structures conserved between the H1N1 and H3N2 viruses are identified, which may provide targets for future functional investigations. The structure of an influenza nucleoprotein, which binds to the influenza genomic RNA, is investigated. The first structure of a H3N2 virus NP is presented which shows high structural conservation with other influenza nucleoproteins.

In contrast to influenza viruses, the Severe Acute Respiratory Syndrome Coronavirus-2 genome is made up of a single positive sense strand of RNA. Sequencing approaches reveal that *in virio* the genome contains many structures formed mostly through short-range base paring (bases separated by <100 bp). A large number of structures are observed that are conserved amongst other coronaviruses and may provide targets for drug development. Overall, this work reveals the presence of diverse and abundant RNA structures in viruses from two families.

Declaration

I declare that the work presented in this thesis was performed by the author. Any contributors to the specific parts of this work are acknowledged in the text.

Michael Knight

August 2021

Acknowledgments

I would firstly like to thank my supervisors Ervin Fodor and David Bauer for their input and support. I would also like to thank Jeremy Keown for his excellent supervision of the structural part of my project. I am thankful to Phil Becker for being interested in my project and for his (occasionally) useful advice. I am extremely grateful to Jane Sharps for her advice and help with the logistics of performing experiments. I am also thankful to all members of the Fodor lab, past and present, for making the lab a fun place to come to work.

Thank you to William James and Rebecca Moore for giving me the opportunity and support to work on SARS-CoV-2. Finally, thank you to Javier Gilbert-Jaramillo and Adam Harding for making it a pleasure to come to work during a pandemic.

Abbreviations

1M7 - 1-Methyl-7-nitroisatoic anhydride
3P – The trimeric influenza polymerase complex
ACE2 - Angiotensin-Converting Enzyme 2
AMT - 4'-Aminomethyltrioxsalen
ASO – Anti Sense Oligonucleotide
BCoV - Bovine Coronavirus
BSL/PK - Bulged Stem-Loop/Pseudoknot
COMRADES - Crosslinking Of Matched RNAs and Deep Sequencing
cRNA – The positive sense replication intermediate of the influenza genomic RNA
Cryo-EM - Cryo-Electron Microscopy
CTF - Contrast Transfer Function
DI – Defective Interfering
DMS – Dimethyl Sulphate
DNA – Deoxyribonucleic acid
DTT – Dithiothreitol
E – The SARS-CoV-2 Envelope protein
EDC - 1-Ethyl-3-(3-Dimethylaminopropyl) Carbodiimide
EDTA - Ethylenediaminetetraacetic Acid
EMSA - Electrophoretic Mobility Shift Assay
FCS - Fetal Calf Serum
FFU – Focus Forming Units
FISH - Fluorescence In Situ Hybridisation
HA - Hemagglutinin
HEK - Human Embryonic Kidney
HIV - Human immunodeficiency Virus
HVR - Hypervariable Region
IAV – Influenza A virus
IRES – Internal Ribosome Entry Site
LIGR-seq - Ligation of Interacting RNA followed by high-throughput sequencing
M – The SARS-CoV-2 Membrane Protein
M1 – The influenza Matrix protein
MaP - Mutational Profiling
MEM - Minimum Essential Medium
MERS-CoV – Middle East Respiratory Syndrome Coronavirus
MDCK - Madin-Darby Canine Kidney
MFE – Minimum Free energy
MHV - Mouse Hepatitis Virus
mRNA – Messenger RNA
N – The SARS-CoV-2 Nucleoprotein
NA – Neuraminidase
NAI - 2-methylnicotinic acid imidazolide
NP – The Nucleoprotein of influenza virus
NSP - Non-Structural Protein
NT60 - A/Northern Territory/60/1968 (H3N2) virus
ORF – Open Reading Frame
OTG - Octyl β-D-1-thioglucopyranoside
PA – Polymerase Acidic protein
PARIS - Psoralen Analysis of RNA Interactions and Structures
PB1 – Polymerase Basic protein 1
PB2 – Polymerase Basic protein 2
PBS - Phosphate Buffered Saline
PCR – Polymerase Chain Reaction

PEG - Polyethylene Glycol
PFU – Plaque Forming Units
PR8 - A/Puerto Rico/8/1934 (H1N1) virus
PTAz - 4,5',8-Trimethylpsoralen-4'-methylaminoTEG Azide
RMSD - Root-Mean-Square Deviation
RNA – Ribonucleic acid
S – Spike protein
S2M - Stem–Loop II-like Motif
SARS-CoV-1 - Severe Acute Respiratory Syndrome Coronavirus 1
SARS-CoV-2 - Severe Acute Respiratory Syndrome Coronavirus 2
SDS - Sodium Dodecyl Sulphate
SEC - Size Exclusion Chromatography
sgRNA - subgenomic RNA
SHAPE - Selective 2' Hydroxyl Acylation analysed by Primer Extension
SPLASH – Sequencing of Psoralen Ligated and Selected Hybrids
Udorn - A/Udorn/307/72 (H3N2) virus
Tm – Melting Temperature
TRS - Transcription Regulatory Sequence
TRS-B - Transcription Regulatory Body Sequence
TRS-L - Transcription Regulatory Leader Sequence
UTR – Untranslated Region
UV – Ultraviolet
vRNA – The influenza genomic RNA
vRNP – The influenza viral Ribonucleoprotein complex
WSN - A/WSN/33 (H1N1) virus
Wyoming - A/Wyoming/3/03 (H3N2) virus

1. Introduction

1.1 Influenza classification, epidemiology, and societal impacts

Influenza is a respiratory disease that each year is estimated to cause 290,000 to 650,000 deaths worldwide (Iuliano et al., 2018) and cost the US economy $11.2 billion (Putri et al., 2018). Infection tends to exhibit seasonality, with annual epidemics occurring during the winter months (or the rainy season in equatorial regions). Influenza viruses have been responsible for four pandemics since the start of the 20th century, in 1918, 1959, 1968, and 2009. The most severe of these, the 1918 'Spanish flu', is estimated to have killed more than 50 million people (Taubenberger and Morens, 2006). Influenza viruses can also be pathogenic to many other species of mammals and birds. This includes a number of livestock species, with ~40 million poultry culled each year to curb the spread of highly pathogenic avian influenza viruses (Boni et al., 2013).

Influenza viruses are divided amongst four *genera* within the *Orthomyxoviridae* family, *alphainfluenzavirus* (Influenza A viruses), *betainfluenzavirus* (Influenza B viruses), *gammainfluenzavirus* (Influenza C viruses), and *deltainfluenzavirus* (Influenza D viruses). Influenza D viruses have been isolated from swine and cattle, with the latter believed to represent the species' main reservoir (Mazzetto et al., 2020). Influenza C viruses infect swine and humans, but tend to be associated with only mild disease (Sederdahl and Williams, 2020). Influenza B viruses have been detected in humans and seals and are a frequent cause of influenza epidemics. The focus of this thesis is on Influenza A Viruses (IAVs). The natural reservoir for IAVs is birds, however they infect a wide range of mammals as diverse as humans, bats, and whales. IAVs are divided into subtypes based on the taxonomy of their antigenic Hemagglutinin (HA) and Neuraminidase (NA) proteins. These subtypes are denoted by H and N numbers (e.g. H1N1, H7N9, or H18N11), currently there are 18 H and 11 N groups.

IAVs make the largest contribution to annual influenza epidemics and have been responsible for all recorded influenza pandemics (Houser and Subbarao, 2015).

1.2 Influenza replication

IAVs have a genome composed of eight segments of negative sense genomic Ribonucleic Acid (vRNA), each encoding at least one essential protein. These vRNAs fold back on themselves with both termini bound by a single copy of the trimeric influenza polymerase (3P), which is composed of the viral Polymerase Basic 1 (PB1), Polymerase Basic 2 (PB2), and Polymerase Acidic (PA) proteins. The rest of the vRNA is coated by many copies of the viral Nucleoprotein (NP) (Fig. 1). These viral Ribonucleoprotein (vRNP) complexes are encased within a coat formed by the viral Matrix protein (M1), which sits inside a host-derived membranous layer. Three viral proteins protrude from the virion surface: HA, NA, and the ion channel M2.

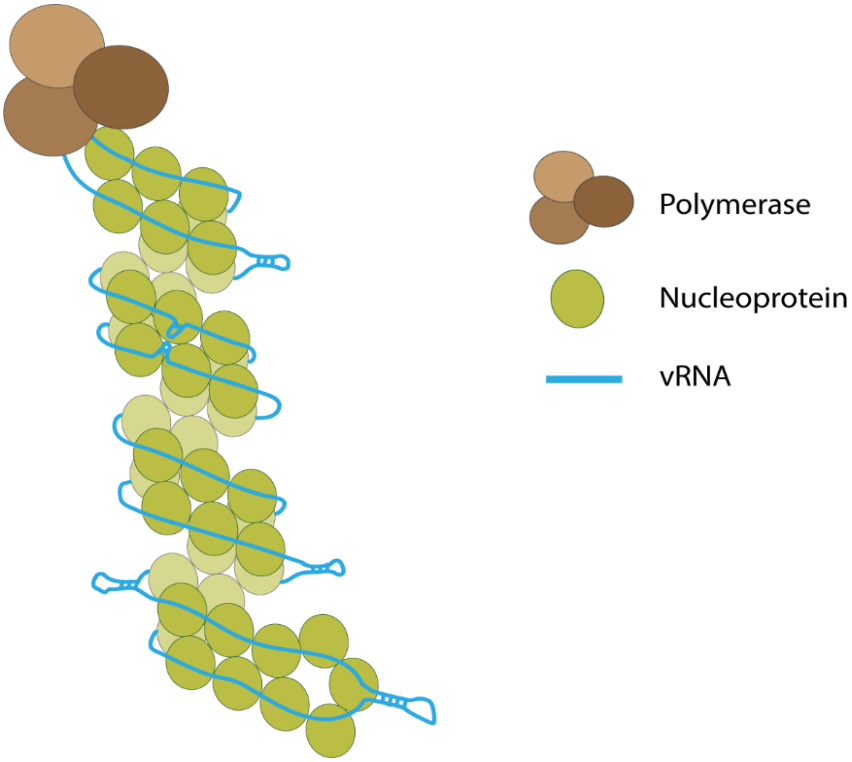


Figure 1: A diagram of a vRNP.
The influenza genome is composed of eight negative sense RNAs (vRNA shown by blue line). Both termini of the genomic RNA are bound by the trimeric influenza polymerase (shown by three brown circles) and the rest of the vRNA is bound by the influenza nucleoprotein (shown by green circles) to make a vRNP complex.

Homo-trimers of the HA protein can bind to the terminal sialic acids of the glycan chains of glycoproteins and glycolipids located on the cell surface (with the exception of the H17 and H18 HA proteins, which bind to major histocompatibility complex class II proteins) (Karakus et al., 2019). Sialic acids can be connected to their glycan chains via α-2,3 or α-2,6 linkages. The proportions of these linkages vary by species and tissue, with the former abundant in the avian intestine and the latter in the upper respiratory tract in humans (Matrosovich et al., 2004). These represent the respective sites of influenza replication in the two species. As such, the affinity of different HA proteins for the two forms of linkage is thought to contribute to the host tropism displayed by different IAVs (de Graaf and Fouchier, 2014). There is also evidence that binding may occur to some classes of phosphorylated, but non-sialyated glycans, though the relative contribution this makes during natural infection is undetermined (Byrd-Leotis et al., 2019).

Upon cellular attachment, the virus is internalised by receptor-mediated endocytosis (Matlin et al., 1981) or micropinocytosis (De Vries et al., 2011). A number of host proteins have been suggested to act as entry receptors, including the epidermal growth factor receptor (Eierhoff et al., 2010), nucleolin (Chan et al., 2016), and the voltage-dependant calcium channel CaV1.2 (Fujioka et al., 2018). During endosomal acidification, the HA protein undergoes a conformational change, exposing its fusion peptide from its pre-fusion position in a hydrophobic pocket (Gao et al., 2020a). This change facilitates fusion of the viral and endosomal membranes, allowing the virus to release its contents into the cytoplasm (Fig. 2).

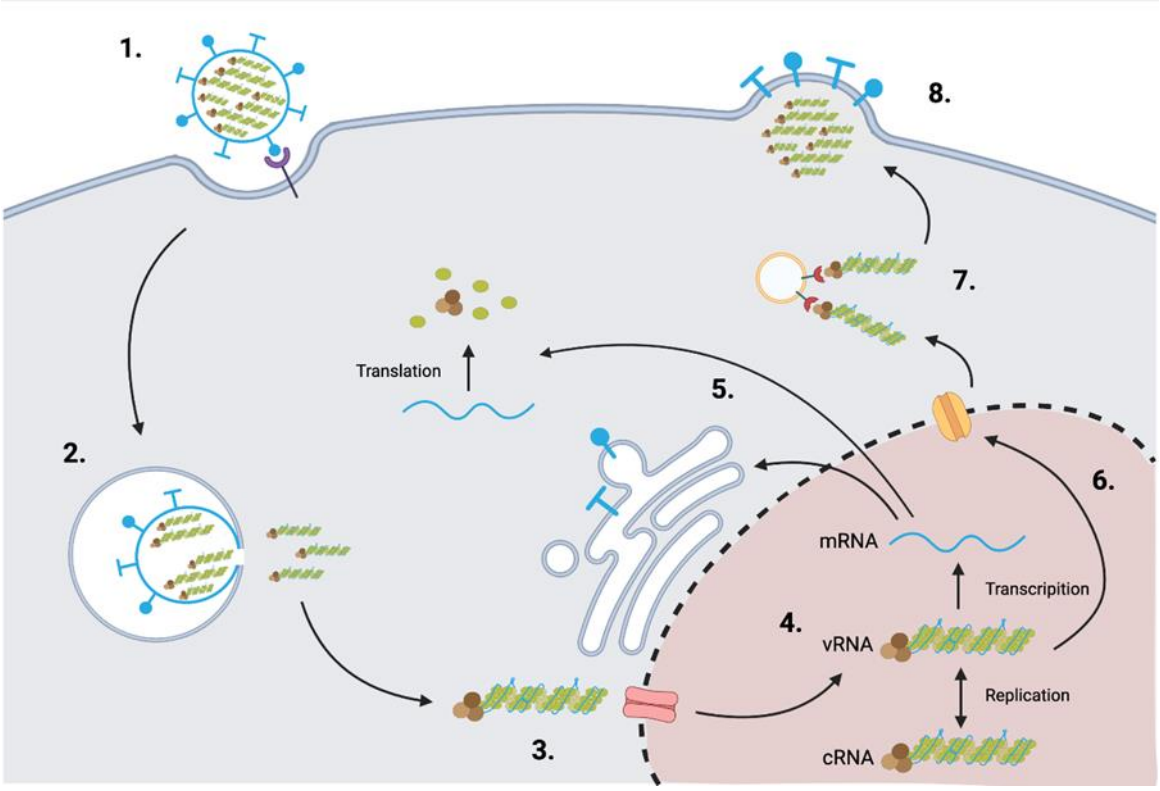


Figure 2: The influenza replication cycle.
1. Influenza virions bind to sialic acids on the cell surface and are endocytosed. 2. Acidification of the endosome leads to HA mediated fusion of the viral and endosomal membranes allowing release of the vRNPs into the cytoplasm. 3. The vRNPs are imported into the nucleus via importin-β. 4. Inside the nucleus the negative sense vRNP is replicated via a full length positive sense intermediate (cRNA). The vRNA is also transcribed to produce mRNA. (5) The mRNA is exported from the nucleus to allow production of viral proteins. (6) Newly made vRNPs are exported from the nucleus by the CRM1 pathway. (7) Rab-11 coated vesicles traffic the vRNPs to the cell membrane. (8) Virions bud from the cell membrane.

Inside virions, the M1 protein is associated with the vRNPs via interaction with NP (Noton et al., 2007). This association must be broken to allow release of the vRNPs into the cytoplasm. Destabilisation of the M-NP interaction is thought to occur through a mixture of the M2 mediated transport of potassium ions into the virion during endosomal acidification and possible interactions with host proteins (Borau and Stertz, 2021). After the vRNPs have been released the nuclear localisation signal of the NP proteins recruits importin-α (O'Neill et al., 1995). Importin-α can in turn interact with the importin-β transporter that facilitates transport of the vRNPs into the nucleus, via the nuclear pore complex (Dou et al., 2018).

The nucleus is the site of both transcription and replication of the influenza virus genome. The viral polymerase transcribes the negative sense genomic RNA into positive sense, capped and polyadenylated, messenger RNA (mRNA) that can be transported from the nucleus for translation. The influenza polymerase steals methylated 5′ cap structures for the influenza mRNAs using a process called cap-snatching (Walker and Fodor, 2019). This first requires 3P to interact with the host RNA polymerase II, bringing the vRNP to be transcribed into close proximity with a capped host mRNA transcript. The cap binding domain of PB2 binds to the cap of the host RNA, which is then cleaved by the PA endonuclease domain (Fodor and Te Velthuis, 2020). This cleavage results in the production of a 10-14 nucleotide capped RNA, which is transferred to the active site of the viral polymerase. This short host-derived RNA has complementarity to the 3′ sequence of the vRNA and is used to prime initiation of transcription (Fodor and Te Velthuis, 2020). The influenza mRNAs are exported to the cytoplasm or endoplasmic reticulum (HA, NA, and M2) to allow translation of the viral proteins.

Replication of the influenza genome requires the production of full-length positive sense copies of the vRNAs, termed complementary RNAs (cRNAs). These act as replication intermediates that can be transcribed to produce more copies of the negative sense vRNAs. The synthesis of cRNA from vRNA begins with *de novo* initiation at the 3′ end of the vRNA to produce a pppApG dinucleotide, a process dependent upon the viral polymerase's priming loop (Te Velthuis et al., 2016). This dinucleotide is then extended by the viral polymerase to

produce the full length cRNA. Much like vRNA, the cRNA is folded back on itself with the 5′ and 3′ termini both bound by 3P and the rest of the RNA bound with NP to produce a complementary ribonucleoprotein complex (York et al., 2013). The production of vRNA from cRNA begins with the synthesis of pppApG at the 4th and 5th bases from the 3′ terminus of the cRNA. The cRNA template then slides to position the first two bases of the template on the pppApG dinucleotide, priming cRNA to vRNA replication from position 1. This process is called the 'prime and realign' step of replication and is dependent upon the formation of a dimeric complex, formed from two 3P complexes (i.e. a dimer of heterotrimers) (Fan et al., 2019).

1.3 Packaging of the influenza genome

The exact mechanism that triggers packaging of influenza vRNPs into virions (as opposed to host RNAs) is not known. It has been shown that the inclusion of the 5′ and 3′ regions of vRNA is sufficient to achieve packaging of foreign RNA sequences (Liang et al., 2005) or of naturally occurring defective interfering (DI) RNAs (Duhaut and Dimmock, 2002). These regions are of ~50-250 nucleotides in length (depending on the segment) with successive deletions from the non-terminal ends leading to a gradual decrease in packaging efficiency. Synonymous mutations in these regions have been found to reduce packaging efficiency by as much as 90% (Marsh et al., 2008). This reduction suggests that these terminal regions contain some sort of packaging signal.

As a segmented virus, influenza faces the additional challenge of packaging at least one of each of its eight different genome segments in order to produce an infectious particle. The vast majority of influenza virions have been shown to package exactly eight segments (Noda et al., 2006) (Noda et al., 2012) (Fournier et al., 2012). If segments were selected randomly for packaging, only one in ~400 particles would be infectious due to most not receiving all eight different vRNAs. However, a much higher proportion of particles are observed to be infectious (Noton et al., 2009) (Wei et al., 2007). Fluorescence In Situ Hybridisation (FISH)

analysis (Chou et al., 2012) and studies using reporter genes (Inagaki et al., 2012) also indicate that the majority of particles contain all 8 different vRNAs. This observation suggests that there is a selective mechanism governing the assembly of the different vRNAs into bundles ready for packaging (this process is referred to as 'bundling').

The most well-supported explanation for selective bundling is that the process is mediated by RNA-RNA interactions between segments. *In vitro* studies using native Electrophoretic Mobility Shift Assays (EMSAs) have indicated that the eight segments form an interaction network in which each segment interacts with at least one other segment (Fournier et al., 2012). Another study using EMSAs showed that these interaction networks appear to be different for different viruses and are not limited to the terminal regions that make up the minimal requirements for packaging of a segment (Gavazzi et al., 2013). More recently Psoralen Crosslinking of Ligated, and Selected Hybrids (SPLASH) has been performed on IAVs allowing direct capture of inter-segment interactions on a genome wide scale (Dadonaite et al., 2019). This study indicated that there are extensive, redundant networks of RNA interactions between the vRNAs that are not limited to terminal regions.

The vRNPs are exported from the nucleus by the CRM1 pathway and transported to the cell membrane by interaction with Rab11 coated vesicles (Bruce et al., 2010) (de Castro Martin et al., 2017). It is not clear if the process of bundling begins in the nucleus or takes place entirely in cytoplasm (Chou et al., 2013) (Lakdawala et al., 2014). One study using FISH found that not all possible sub-bundles (bundles of fewer than eight segments) are seen during the genome assembly process and that the sub-bundles rarely contain multiple copies of the same segment (Haralampiev et al., 2020).

1.4 Reassortment

Frequent changes in the amino acid sequence of the HA and NA proteins prevent universal immunity against IAVs and ensure the annual occurrence of influenza outbreaks. This diversity of sequence is accomplished through a combination of antigenic drift, due to the error prone influenza polymerase, and antigenic shift, due to reassortment. Reassortment occurs when two (or more) different influenza viruses infect the same cell (Fig. 3). Their progeny can then package a mixture of segments from both parental strains. This can produce large genetic shifts, particularly where it occurs between viruses circulating predominantly in different species (Mena et al., 2016).

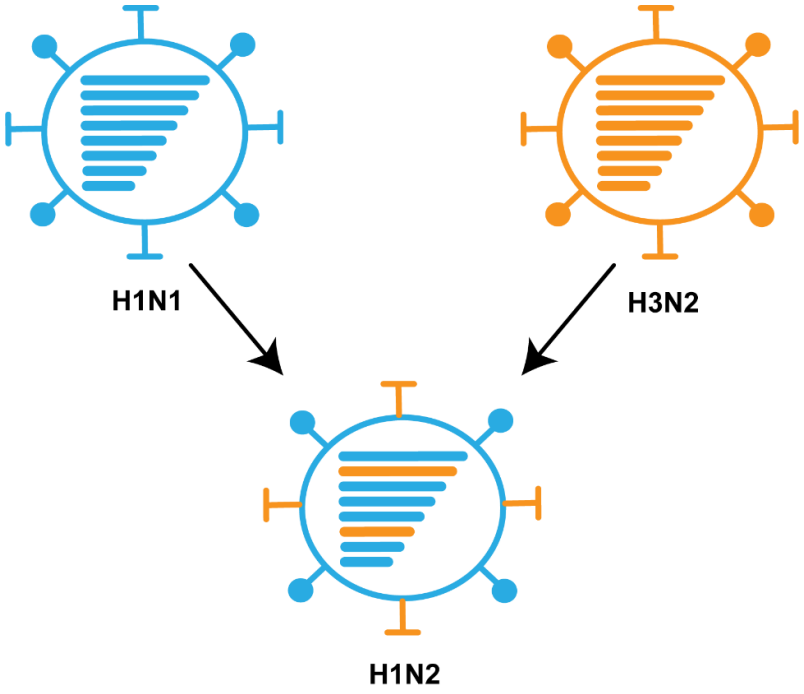


Figure 3: The process of reassortment.
When two IAVs infect the same cell their offspring may package segments from both viruses. In this example, the offspring packages a H1 HA and a N2 NA segment meaning it would be classified in a different sub-type to both of the parental strains.

In vitro reassortment experiments show that it is often not possible to generate progeny containing all of the theoretically possible combinations of gene segments from two parental strains (Jackson et al., 2009) (Li et al., 2008a). In addition, certain vRNAs have a tendency to co-segregate during reassortment (Essere et al., 2013) (Cobbin et al., 2014) and more closely related viruses tend to reassort more efficiently than those that are more distantly related (Gerber et al., 2014), suggesting reassortment is not an entirely random process. If inter-segment RNA-RNA interactions govern segment bundling, it seems likely that these same interactions may influence the propensity for segments from different IAVs to be packaged into the same virion during reassortment events.

1.5 RNA structure in viruses

Aside from their role in influenza segment bundling, structures formed through RNA-RNA interactions can have important and diverse functional roles. This includes the aminoacyl-transferase activity of ribosomal RNA and regulatory protein recruitment to the Untranslated Regions (UTRs) of mRNAs. The genomes of many RNA viruses have been found to contain structured regions essential to virus function. An example of this are internal ribosome entry sites (IRES), which were first identified in picornaviruses (Pelletier and Sonenberg, 1988). IRES allow viral RNAs to be translated in a cap-independent manner. The poliovirus IRES is a 450-nucleotide region containing five stem-loop structures, which recruit a number of IRES-transacting factors. The 3′-most stem-loop is bound by the eukaryotic initiation factors 4A and 4G, which then indirectly recruit the ribosome through interactions with other initiation factors (Beckham et al., 2020).

Viruses have also developed other RNA structures designed to facilitate translation. An example of this are the 3′ located cap-independent translation elements seen in a number of plant viruses. This includes the multi-partite stem-loop structure of the barley yellow dwarf virus. This structure is able to directly recruit eukaryotic initiation factor 4G and form interaction

with the 5′ region of the genome (effectively circularising the genome). This allows recruitment of the ribosome to the 5′ region and has been suggested to help recycle it as it reaches the 3′ end (Jaafar and Kieft, 2019). Another example of functional RNA structure in a virus is the 'kissing loop' required for packaging of the Human immunodeficiency Virus (HIV) genome (Paillart et al., 1996). This structure is formed from a 9-nucleotide loop which contains a 6-nucleotide long palindromic sequence. HIV packages two copies of its genome per viral particle and this is mediated by an inter-segment interaction between these loops, located in the 5′ UTR of the genome (Mundigala et al., 2014).

1.6 RNA structure prediction

Attempts at computationally predicting the structures of RNA based on their sequence have been ongoing for over half a century (Tinoco et al., 1971) and mainly use covariance or Minimum Free Energy (MFE) based approaches. Covariance based approaches perform comparative analysis of homologous RNA sequences. As sequences diverge from a common ancestor, it is assumed that important RNA structures will be maintained. These approaches assume that when a mutation occurs in a residue involved in a region of base pairing, its partner will also mutate to maintain sequence complementarity. However, such techniques suffer from high false positive rates, in part due to difficulties in establishing appropriate negative controls (Eddy, 2014).

MFE approaches to structure prediction assume that at equilibrium RNA will form the most stable structure available to it, corresponding to that with the lowest free energy. At the most basic level, this could be thought of as trying to fold the sequence into the structure which gives the highest number of paired bases (Fig. 4). More sophisticated approaches utilise data from melting experiments that have been performed on a large range of RNA sequences in the context of different structures (Freier et al., 1986) (Turner and Mathews, 2010). The energy values determined in these experiments can be used to assign free energy contributions to

particular base pairs. To simplify structure prediction, the energy contribution of a base pair is usually predicted based only in the context of the two base pairs immediately either side of it. This is termed the next nearest neighbour energy model. The energy contribution of loops is determined primarily by their length and degree (which is a function of the number of enclosed bases and the identity of the base pair enclosing the loop) (Lorenz et al., 2016). MFE prediction seeks to generate the structure in which the sum of the free energy scores for all of the base pairs (negative contribution) and loops (positive contribution) present is lowest.

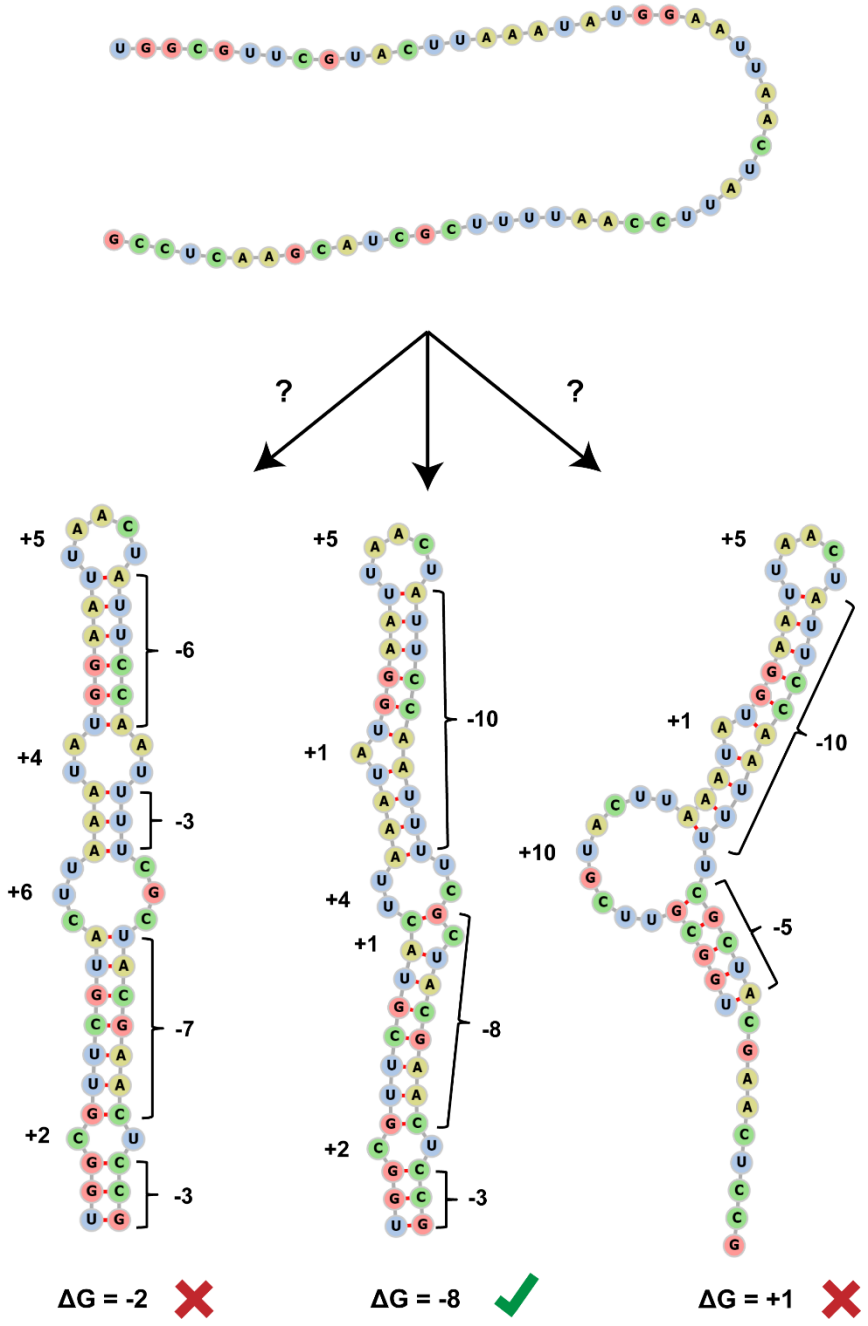


Figure 4: Illustrating MFE based RNA structure prediction.
The RNA shown could fold into a number of different conformations (3 of which are shown). It is likely that the RNA will form the most stable structure (the one with the lowest free energy). Paired regions give a negative contribution (favoured) to the free energy and loops a positive contribution (disfavoured). Note this figure is meant as an illustration of the process and the values given are not accurate. Adapted from (Tinoco et al., 1971).

Predicting structures using MFE can be reasonably effective for short RNAs with accuracies of ~70% seen for some sequences shorter than 500 nucleotides (Mathews et al., 1999).

However, accuracy declines considerably as RNA length increases (Doshi et al., 2004) and MFE prediction cannot account for factors that arise when RNA is folded *in vivo*, such as protein binding, that may alter RNA structure. It also ignores the effect of tertiary interactions, which can have a substantial impact on RNA folding (base pairs contribute −0.9 to −3.4 kcal/mol whilst the contribution of tertiary interactions ranges from −0.3 to −1.5 kcal/mol) (Schroeder, 2018). This means that experimental data is often required to achieve accurate predictions of RNA structure. This most commonly entails the use of chemical probes which can selectively modify RNA.

Chemical probes that will preferentially modify single stranded RNA bases have been developed since the 1960s (Metz and Brown, 1969). Examples of these reagents include Dimethyl Sulphate (DMS), 1-Cyclohexyl-3′-(2-Morpholinoethyl) Carbodiimide and 1-Ethyl-3-(3-Dimethylaminopropyl) Carbodiimide (EDC) (Fig. 5) (Ehresmann et al., 1987) (Wang et al., 2019). Modified bases terminate reverse transcription. Thus, the position of a modified base can be determined by performing primer extension assays and then determining the size of the reverse transcription products (by gel electrophoresis). As the chemical probes used preferentially modify single stranded nucleotides, those nucleotides that are frequently seen to be modified can be inferred to be single stranded.

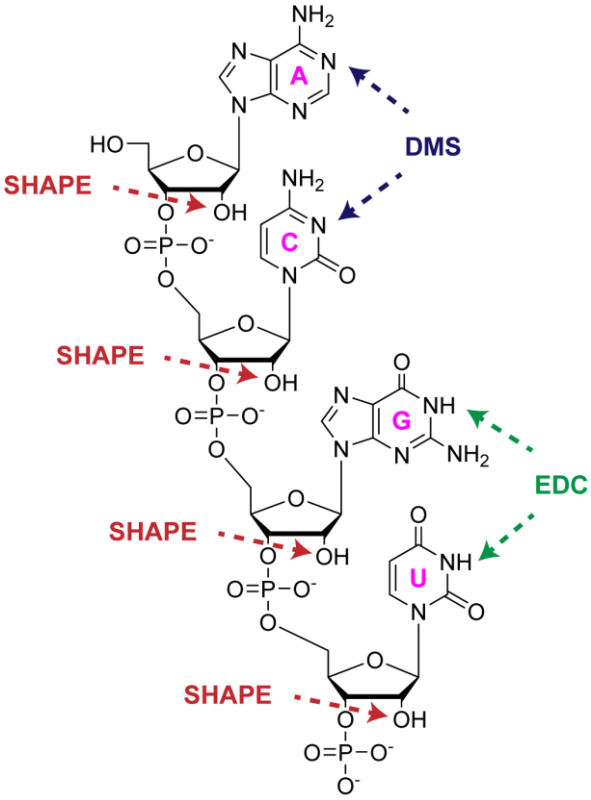


Figure 5: Modification of RNA by SHAPE reagents, EDC, and DMS.
A number of chemical probes are able to preferentially modify unpaired nucleotides. The information gained from probing experiments can thus give information on the structure of a given RNA. Adapted from (Mitchell III et al., 2019).

Today the same principles are used in combination with high-throughput sequencing technologies to probe much longer RNAs and achieve greater depths of information. In addition, under certain conditions reverse transcriptase's can be used that will introduce mutations into the complimentary Deoxyribonucleic Acid (cDNA) at chemically modified bases (either a mismatch, insertion, or deletion) rather than terminating. These mutations can be identified when mapping the sequencing reads back to the reference genome. These Mutational Profiling (MaP) approaches can further increase sequencing depth as they allow for multiple modifications to be detected per piece of RNA (Smola et al., 2015).

One of the most popular techniques to emerge for probing RNA secondary structure has been Selective 2′ Hydroxyl Acylation analysed by Primer Extension (SHAPE) (Fig. 5) (Wilkinson et al., 2006). Hydrogen bonds can form between the 2′ hydroxyl and 3′ oxygen in the sugar

phosphate backbone. The greater rigidity of double stranded RNA will stabilise these interactions. However, in single stranded RNA this hydrogen bonding will be more transient as the greater flexibility allows the 2′ and 3′ positions to become separated by greater distances. When not involved in hydrogen bonding the 2′ hydroxyl is more nucleophilic allowing it to interact with SHAPE reagents. Thus regions of single stranded RNA are more likely to see high modification frequencies. The most popular SHAPE reagents are currently 2-methylnicotinic acid imidazolide (NAI) and 1-Methyl-7-nitroisatoic anhydride (1M7), due to their fast reactivity time, ability to cross lipid membranes, and unbiased modification of all 4 bases (Busan et al., 2019).

Regardless of the reagent used structural probing data is used to assign a reactivity value to each nucleotide in the probed RNA. This is done by comparing the mutation rate of a given nucleotide in the chemically modified sample to its mutation rate in a control sample, where the reagent is not added. The reactivity rates have no units and are normalised to arbitrary values. Normalisation can be performed in a number of different ways, but usually involves dividing all of the reactivity values in a dataset by the reactivity value of a relatively highly reactive base in that dataset (Deigan et al., 2009) (Zarringhalam et al., 2012) (Incarnato et al., 2018). For example, the reactivity values of all bases could be divided by the reactivity value of the base that is the 90th percentile when bases are ordered by reactivity. This results in the vast majority of bases having a reactivity value in the range 0 to 2. A low reactivity value indicates that a particular base is likely to be structurally constrained and a high value indicates that it is likely to be single stranded.

It is important to note that reactivity values from probing experiments do not give an unambiguous indication as to whether or not a base is paired, but rather confer probabilistic information on the likelihood of this being the case. Attempts to quantify this for SHAPE data suggest that a very low reactivity value (i.e. normalised SHAPE value below 0.1) indicates that a base is ~5 times more likely to be paired than unpaired whilst a high reactivity value (i.e. normalised SHAPE value above 1) indicates that a base is ~5 times more likely to be unpaired

than paired (Sükösd et al., 2013) (Eddy, 2014). This shows that there is substantial overlap in the reactivity value distributions of paired and unpaired bases. The exact relationship between reactivity value and the probability of base pairing is likely to vary depending upon the RNA (e.g. an RNA with lots of bases structurally constrained by non-base pairing interactions would give lower quality information in SHAPE probing) and the experimental set up (e.g. chemical probe used) (Kutchko and Laederach, 2017).

The probabilistic relationship between base pairing and reactivity values obtained from chemical probing experiments means that it is advisable to use reactivity values as 'soft constraints' to RNA structure prediction (as opposed to hard constraints where a lowly reactive base would be forced to base pair). Deigan *et al* proposed to incorporate chemical probing data into structure prediction by applying the reactivity values from SHAPE data as a pseudo-free energy term to MFE based prediction (Deigan et al., 2009).

Equation 1: $\Delta G'_i = m \log(\alpha_i + 1) + b$

In equation 1 $\Delta G'_i$ represents the free energy contribution that will be assigned to a particular base, i, when the prediction program is attempting to determine the structure with the lowest free energy. The term α_i is the relative SHAPE reactivity assigned to a base while m and b are free parameters. In the original paper by Deigan, values of m = 2.6 and b = −0.8 kcal mol^{-1} were found to give the most accurate structure prediction, based on chemical probing of RNA structures previously determined by crystallography (Deigan et al., 2009). When using these default parameters, the equation has the effect of applying an extra −0.8 kcal mol^{-1} to the free energy contribution of a base pair with a relative reactivity value of 0 (i.e. the probing data is in agreement with that residue being base paired). Conversely the program will be discouraged from predicting structures where bases exhibiting high relative reactivity are base paired (a relative reactivity value of 2 would lead to +2.1 kcal mol^{-1} being added to the free

energy contribution of that base pair). In this manner structure prediction was able to be achieved with ~90% accuracy against structures determined by crystallography (Deigan et al., 2009).

A number of slightly different approaches have been proposed for integrating SHAPE reactivity values as pseudo-free energy terms for structure prediction (Zarringhalam et al., 2012) (Washietl et al., 2012). All claim to increase structure prediction accuracy in comparison to predictions made without incorporating SHAPE reactivities. However, there is no consensus that any one of these approaches is superior to the others. This is in part due to the lack of complete RNA structures that have been determined by X-ray crystallography or Cryo-Electron Microscopy (Cryo-EM) against which to compare the different folding approaches.

A given RNA can be folded to form many different structures, with the number of possible structures increasing exponentially with the length of the RNA. This collection of possible structures is termed the structural 'ensemble'. By generating the ensemble of possible structures, along with the free energies of these structures, it is possible to calculate the partition function for a particular RNA (McCaskill, 1990). The partition function is a dimensionless figure encoding statistical information on a system in thermodynamic equilibrium. It, in theory, represents the ratio between the number of particles that will adopt the most stable structure and the number of particles that will adopt alternative structures, at a given temperature. Only a limited number of sub-optimal structures, within a certain free energy range from the lowest free energy structure, are generally considered in this calculation (Wuchty et al., 1999). Individual structures in an RNA can be assigned pairing probabilities based on their free energy relative to the partition function (Mathews, 2004).

The partition function also allows the calculation of Shannon entropy – a measure of the uncertainty associated with the formation of a structure in a given region of an RNA. Shannon entropy is calculated by determining the probability of a base being paired with its partner relative to the possibility of it pairing with all of its potential partners (Huynen et al., 1997) (Siegfried et al., 2014). This means that if a region can form two (or more) structures that have

similar free energies the region will have high Shannon entropy (note the structure may still have high pairing probability if these structures have low free energy). This is partly a measure of uncertainty around the pairing status of a base and partially attempts to address the fact that an RNA may not form a single stable structure. RNA may exist in an equilibrium, adopting one or more alternative structures. Dealing with structural ensembles is a particular limitation of SHAPE experiments, where the reactivity data is averaged out over many copies of an RNA, which may form different structures.

1.7 Determination of RNA structure by interaction capture

Alternative methods for determining RNA structure seek to directly capture RNA-RNA interactions by using chemical crosslinking. The most commonly used reagents are derivatives of psoralen. Psoralen preferentially intercalates into double stranded regions of RNA where there are adjacent, opposite pyrimidines (with a preference for uracil over cytosine) (Fig. 6A) (Cimino et al., 1985). Upon exposure to long wave UV radiation, the interacting RNAs will be crosslinked together through reversible covalent bonds. Multiple techniques use psoralen derivatives to analyse RNA-RNA interactions including SPLASH (Sharma et al., 2016), Ligation of Interacting RNA followed by high-throughput sequencing (LIGR-seq) (Lu et al., 2016), Psoralen Analysis of RNA Interactions and Structures (PARIS) (Aw et al., 2016), and Crosslinking Of Matched RNAs and Deep Sequencing (COMRADES) (Ziv et al., 2018).

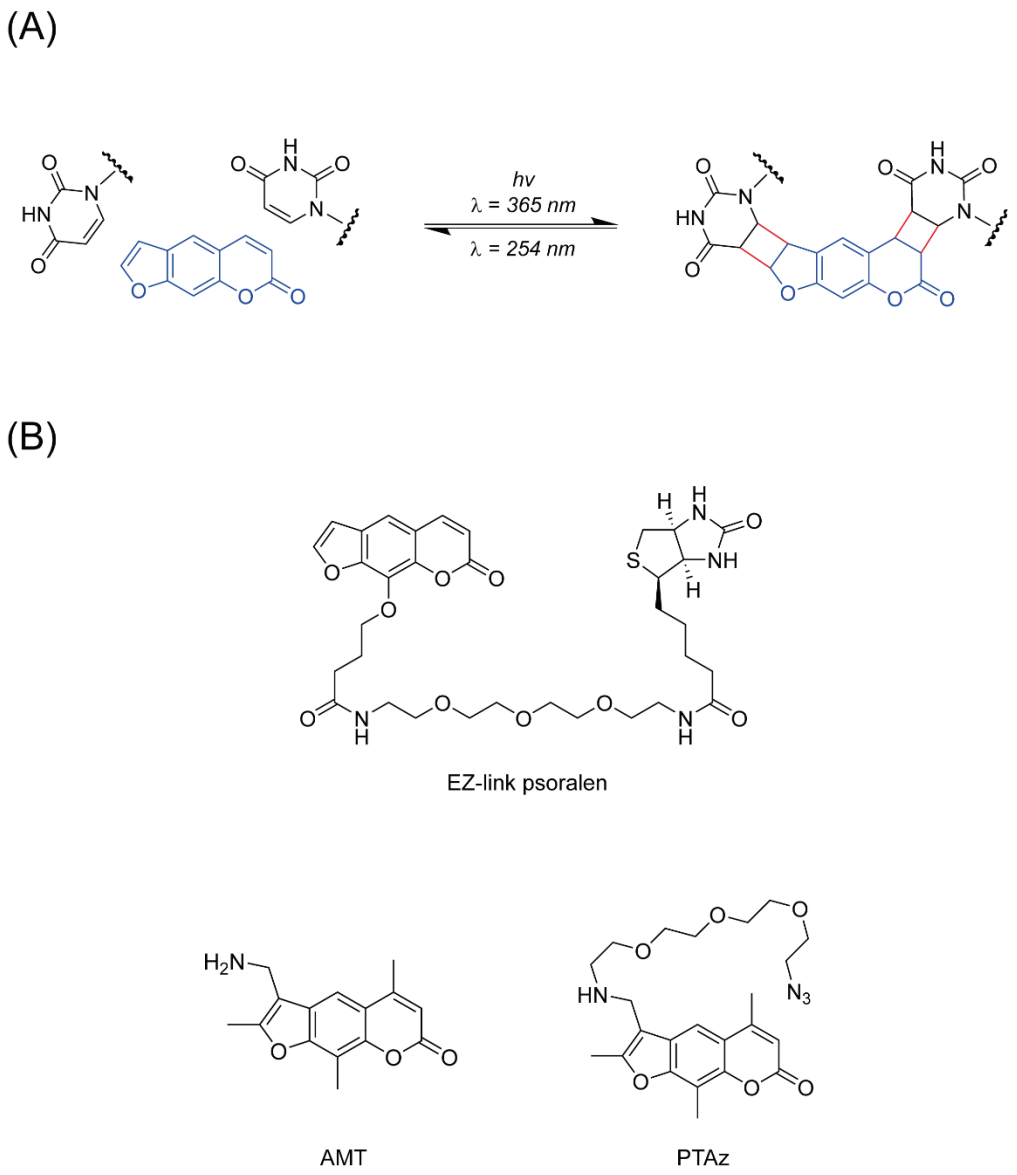


Figure 6: Chemical crosslinking for RNA interaction capture
(A) The mechanism by which psoralen crosslinks pyrimidines which are in close proximity due to RNA-RNA interactions. (B) The structures of different psoralen derivatives used in chemical crosslinking experiments.

All of the psoralen-based RNA interactions techniques use a very similar methodology. After crosslinking, the RNA is usually fragmented (either chemically or by nuclease treatment). Proximity ligation is then performed to join the two interacting RNAs into one single hybrid strand. The cross-linking can then be reversed by exposure to short wave UV radiation, removing the psoralen. This leaves hybrid strands, made up of the interacting regions of RNA, which can be reverse transcribed and used to generate sequencing libraries. By mapping the

sequencing reads back to the reference genome (at the two different locations), it is possible to identify the interacting partners.

The psoralen crosslinking techniques differ mostly in the manner in which they enrich for crosslinked RNA. SPLASH makes use of the biotin tag on the psoralen-Polyethylene Glycol 3 (PEG3)-biotin (commercial name EZlink) (Fig. 6B) molecule to affinity purify the crosslinked RNA (Sharma et al., 2016). Both PARIS and LIGR-seq use 4′-Aminomethyltrioxsalen (AMT) for crosslinking and enrich for crosslinked RNA by electrophoresis or use of a single-stranded specific nuclease (RNAse R) respectively (Aw et al., 2016) (Lu et al., 2016). COMRADES uses 4,5',8-Trimethylpsoralen-4'-methylamino PEG3 Azide (PTAz) for crosslinking. Initial enrichment for the RNA of interest is performed using biotinylated capture probes. Following this, a biotin is added to the crosslinked RNA via a click chemistry reaction between the azide group on the PTAz and a strained alkyne-functionalised biotin derivative, to allow enrichment of cross-linked RNA (Ziv et al., 2018).

1.8 Thesis objectives

The objectives of this thesis were to investigate the presence and function of RNA structure in viruses. Chapter 3 seeks to determine the RNA-RNA interactions between the influenza vRNAs. Extensive, redundant networks are uncovered and their importance in reassortment is demonstrated. In chapter 4, the intra-segment RNA-RNA interactions in the influenza genome are investigated. These structures are found to differ greatly between viruses from different sub-types, though some conserved structures are identified that may provide targets for future functional studies. In chapter 5, the structure of the influenza NP protein, which coats the vRNA, is investigated. The first structure of a H3N2 virus NP is presented which shows high structural conservation with other IAV NPs. Chapter 6 investigates the presence of RNA structure in the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) genome.

A large number of structures are observed that are conserved amongst other coronaviruses and may provide targets for drug development.

2. Methods

2.1 Cell culture and determination of viral titres

2.1.1 Cell culture

Madin-Darby Canine Kidney (MDCK) cells were grown in Minimum Essential Medium (MEM) (Merck) with 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), 2 mM L-glutamine and 10% Fetal Calf Serum (FCS) (Table. 1). Human Embryonic Kidney 293T (HEK) cells were grown in Dulbecco's modified Eagle medium (Merck) with 10% FCS, 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), and 2 mM l-glutamine. Vero Ccl-81 cells (ATCC) were grown in Dulbecco's modified Eagle medium (Merck) with 10% FCS, 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), and 2 mM l-glutamine. All cell types were maintained by splitting 1 in 10 approximately every 3 days.

Media name	**Ingredients**
MDCK growth media	MEM with 10 units/mL penicillin, 10 µg/mL streptomycin, 2 mM L-glutamine and 10% FCS.
HEK cell media	Dulbecco's modified Eagle medium with 10% FCS, 10 units/mL penicillin, 10 µg/mL streptomycin, and 2 mM l-glutamine.
Vero cell media	Dulbecco's modified Eagle medium with 10% FCS, 10 units/mL penicillin, 10 µg/mL streptomycin, and 2 mM l-glutamine.
WSN infection media	MEM with 10 units/mL penicillin, 10 µg/mL streptomycin, 2 mM L-glutamine and 0.5% FCS.
Influenza infection media	MEM with 10 units/mL penicillin, 10 µg/mL streptomycin, 2 mM L-glutamine and 0.3% BSA.
SARS-CoV-2 infection media	Dulbecco's modified Eagle medium with 1% FCS, 10 units/mL penicillin, 10 µg/mL streptomycin, and 2 mM l-glutamine.

Table 1: Cell growth and infection media.

2.1.2 Influenza virus rescue

Virus rescue was performed using the 8 plasmid bi-directional system (Hoffmann et al., 2002). The pPOLI plasmids for the 8 genome segments were mixed, with 500 ng added of each plasmid. Next 8 µL of Lipofectamine 2000 Transfection Reagent (Invitrogen) was added to 242 µL of Opti-MEM (Gibco) and incubated for 5 minutes at room temperature. The plasmid mixture was then added to the Opti-MEM-Lipofectamine mixture and incubated for a further 20 minutes at room temperature. HEK cells were split and 9 x 10^5 cells added to one well of a 6 well plate in 1.5 mL of MEM with 2 mM L-glutamine, and 10% FCS. The mixture containing the plasmids was then added to the cells and incubated for 24 hours at 37°C, before washing with Phosphate Buffered Saline (PBS) and adding 2 mL of MEM with 2 mM L-glutamine and 0.5% FCS.

After a further 48 hours at 37°C, 10 µL of the media from the cells was taken and added to 70% confluent MDCKs in a T75 plate, in 10 mL of MEM with 2 mM L-glutamine and 0.5% FCS. These cells were then incubated at 37°C until the cells displayed signs of infection (usually 48 hours). The media was then harvested and centrifuged at 2000 rpm in a Rotanta 460 R centrifuge at 4°C for 10 minutes to remove cell debris. The supernatant was then aliquoted as required and stored at -80°C.

2.1.3 Influenza plaque assay

MDCK cells were seeded on 6 well plates MEM (Merck) with 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), 2 mM L-glutamine and 0.5% FCS. Cells were incubated at 37°C for 24 hours (until monolayer was ~80% confluent). Virus was serially diluted (usually in 10 fold increments) in MEM with 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), 2 mM L-glutamine, and 1% FCS. Cells were washed twice with PBS and then the serially

diluted virus was added to each well in a volume of 200 µL and left for 1 hour at room temperature. During incubation the MEM (Merck) with 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), 2 mM L-glutamine and 0.5% FCS was mixed in a 1:1 volume to volume ratio with 2% low gelling temperature agarose (dissolved in PBS). Each well of the six-well plate then had 2 mL of this mixture added as an overlay. The overlay was allowed to set for 15 minutes at room temperature before plates were placed upside down and incubated for 3 days at 37°C. Coomassie blue solution was prepared by dissolving 2 g of Coomassie Brilliant Blue R-250 (Sigma) in 375 mL water, 75 mL acetic acid, and 500 mL of ethanol. After incubation the solid overlays were removed using a spatula and 1 mL of Coomassie blue solution was added to each well. After incubating for 1 hour at room temperature the stain was removed and wells washed with water. Plaques (non-stained regions in the monolayer) could then be counted to determine viral titre.

2.1.4 Determination of viral titres for SARS-CoV-2

Plaque assays were performed in 24 well plates with 0.5 mL of Vero cells, resuspended in infection media at 5 x 10^5 cells/mL, added to each well. The cells were incubated with dilutions of the virus stocks or purified virus for 2 hours at 37°C. A semi-solid overlay of 0.5 mL of infection media containing 1.5% carboxymethyl cellulose was then added to each well. After 4 days of incubation at 37°C the overlays were removed and plaques were revealed by staining with Amido Black.

Focus forming assays were performed as described previously (Skelly et al., 2021). In brief 100 µL of Vero cells, resuspended in infection media at 4.5 x 10^5 cells/mL, were added to each well. The cells and diluted virus were incubated together for 2 hours at 37°C, before addition of 100 µL of semi-solid overlay. After a further 20 hours of incubation at 37°C, overlays were removed and cells were washed with PBS before being fixed with 4% paraformaldehyde in PBS for 30 minutes. After fixation cells were washed in 1% ethanolamine in PBS and then

permeabilised for 30 minutes at 37°C with 2% Triton X100 in PBS. Cells were incubated for 1 hour at room temperature with a primary antibody against the SARS-CoV-2 N protein (EY-2A from Arthur Huang via Alain Townsend and Jack Tan). Following washes in PBS containing 1% Tween-20, a goat anti-human IgG antibody conjugated to HRP (Sigma A0170-1ML) was added and infection foci revealed by addition of TrueBlue peroxidase substrate (Insight Biotechnology Ltd). Following washes with water, foci were counted using an ELISPOT plate reader.

2.2 Investigating inter-segment RNA-RNA interactions in influenza

2.2.1 Influenza virus production and purification

MDCK cells were gown until 70% confluent. Six T175 flasks per virus were infected at a multiplicity of infection of 0.01 in MEM containing 2 mM L-glutamine, 0.3% BSA and 0.8 µg/mL of TPCK-trypsin (Merck). Viruses used for SPLASH experiments were; A/Udorn/307/72 (H3N2) (Udorn), A/Puerto Rico/8/1934 (H1N1) (PR8), A/Wyoming/3/03 (H3N2) (Wyoming), and reassortants of these strains (provided by Steven Rockman (Seqirus)).

After 48 hours the media was harvested and clarified by centrifugation at 4000 rpm in a Rotanta 460 R centrifuge at 4°C. This was followed by a further clarification step of 15 min at 10,000 rpm in a SW32 rotor (Beckman Coulter) in an Optima XPN 80K Ultracentrifuge (Beckman Coulter) at 4°C. The virus was then pelleted through a 30% sucrose cushion for 90 min at 25,000 rpm in an SW32 rotor in a Optima XPN 80K Ultracentrifuge (Beckman Coulter) at 4°C before being resuspended in NTC buffer (100 mM NaCl, 20 mM Tris-HCl pH 7.4 and 5 mM $CaCl_2$) (480 µL per virus).

2.2.2 SPLASH of influenza viruses

SPLASH was performed as described previously (Dadonaite et al., 2019). In brief, Digitonin (Merck) was added to purified virus to a final concentration of 0.01% and EZ-link™ Psoralen-PEG3-Biotin (Thermo Fisher) to 200 µM. The mixture was irradiated for 45 minutes on the long wave (365 nm) setting of an Ultra Violet Product™ Handheld UV Lamp (Fisher).

The samples were then subject to Proteinase K (Merck) digestion in PK buffer (0.5% Sodium Dodecyl Sulphate (SDS), 100 mM Tris-Cl at pH 7.5, 50 mM NaCl, 10 mM Ethylenediaminetetraacetic Acid (EDTA)) with 0.250 mg/mL Proteinase K. RNA was extracted using TRI Reagent LS® (Sigma) according to manufacturer's protocol. RNA was fragmented using the NEBNext Magnesium RNA fragmentation Module (NEB) with 4 minutes fragmentation. The samples were purified using the RNA Clean and Concentrator TM-5 (Zymo).

Crosslinked RNA was bound to Hydrophilic Streptavidin Magnetic Beads (NEB) via the biotin tag on psoralen. 3′ cyclic phosphates were removed using 40 units of T4 Polynucleotide Kinase (NEB) (PNK) at 37°C for 4 hours. 5′ phosphates were added using PNK according to the manufacturer's instructions. Samples were then washed twice with PNK buffer (NEB) and proximity ligation performed with 30 units/µL of RNA Ligase 1 (NEB) in RNA ligase 1 buffer with 10 mM ATP and 2 µL of RNasin. This reaction was incubated for 16 hours at 16°C.

Beads were washed twice in 1 mL of wash buffer (2X saline-sodium citrate (Invitrogen) and 0.5% SDS) before being resuspended in 100 µL of PK buffer. The beads were incubated at 95°C for 10 minutes and then resuspended in 100 µL of Tri Reagent LS. 100 µL of chloroform was added and samples were centrifuged for 15 minutes at 14,000 x g at 4°C. RNA was extracted using the RNA Clean and Concentrator TM-5. Samples were eluted in 100 µL of H_2O and irradiated for five minutes in a UVC 500 Crosslinker (Hoefer) (254 nm) to reverse

crosslinking. RNA was precipitated by addition of 300 µL of 100% ethanol, 10 µL of 3M sodium acetate and 1 µL of 15 mg/mL GlycoBlue™ Coprecipitant (Invitrogen). After storage at -80°C for >1 hour the samples were centrifuged at 16,000 x g for 30 minutes at 4°C. The pellets were washed once in 70% ethanol and then resuspended in 8 µL of H_2O.

Libraries were prepared using the SMARTer smRNA-Seq Kit (Takara). Number of amplification cycles was determined by performing real time Polymerase Chain Reaction (qPCR). Samples were separated on 6% TBE gels and fragments of 300-450 bp selected. The gel was fragmented and incubated for 2 hours at 37°C in 450 µL of elution buffer (1 M sodium acetate, 1 mM EDTA). Samples were then centrifuged at 16,000 x g in a Costar Spin-x Column (Corning) and precipitated at -80°C in 1 mL of 100% ethanol with 0.5 µL of GlycoBlue. Library concentrations were assessed by Qubit and pooled into 2 nM libraries to sequence 1x150 bp on a NextSeq 500 (Illumina).

2.2.3 SPLASH data analysis

The adaptors were trimmed from sequencing reads using Skewer v0.2.2 (Jiang et al., 2014) and reads mapped back to the viral genomes using STAR v.2.5.3 (Dobin et al., 2013). Reads where <20 nucleotides mapped back to each of the viral segments were discarded. Remaining chimeric reads were deduplicated using CIGAR strings and alignment positions. Chimeric read coordinates were used to produce an interaction matrix in R using software developed by David Bauer (Francis Crick Institute). Interaction loci were selected manually and fitted with Gaussian curves to determine an interaction window. Interaction plots were generated using the Circlize package v.0.4.5 (Gu et al., 2014) in R v.3.5.1.

Intra-segment interactions were mapped and visualised using (currently unpublished) software made by Anob Chakrabarti (Francis Crick Institute). The software uses pblat (Wang and Kong, 2019) with step size of 5, tile size of 11, and minimum score of 15 to align reads to

the reference genome. To ensure reads with high confidence alignments are used in further analysis, reads with e-values ≤ 0.001 are then removed. Any solutions were both alignments overlap on the reference genome are removed (to filter out non-hybrid reads).

2.3 Investigating intra-segment RNA-RNA interactions in influenza

2.3.1 Chemical probing of influenza viruses

PR8, WSN, Wyoming, and reassortant viruses were grown and purified in the same manner as for SPLASH (section 2.2.1). The same viruses were used with the addition of A/WSN/33 (H1N1) (WSN). Two T175 flasks of cells were infected for each virus with the exception of Wyoming where 4 flasks were used. After purification viral pellets were resuspended in 160 µL of NTC. The virus was split into two 80 µL aliquots and 8.8 µL of 100 mM 1M7 (Tocris) or DMSO was added to each. Samples were incubated at 37°C for 90 seconds. Following this samples were subject to proteinase K digestion by adding 1.8 µL of proteinase K, 2.8 µL of 20% SDS, 2.2 µL of 0.5 M EDTA, and 10 µL of proteinase K buffer and incubating at 37°C for 30 minutes. Samples were then subject to Trizol LS (Ambion) extraction according to the manufacturer's instructions.

WSN was grown in media that contained 0.5% FCS, but not TPCK-trypsin or BSA. For WSN six T175 flasks were used to produce virus. Following purification viral pellet was resuspended in a volume of 400 µL. The resuspended virus was split into 4 aliquots and 25 µL of 1M bicine added to each. Each aliquot then had either 14 µL of 100 mM 1M7, 14 µL of DMSO, 1.67 µL of 10.67 M DMS, or 2 µL of 5.65 M EDC. DMS and EDC samples were incubated for 5 minutes at 37°C while 1M7 and DMSO samples were incubated at 37°C for 90 seconds. The DMS reaction then had 67.6 µL of 1 M bicine and 23.6 µL of 1 M Dithiothreitol (DTT) to quench the reaction. The 1M7 and DMSO samples had 79 µL of NTC added to them. The EDC reaction had 11.8 µL of 1 M DTT, was incubated for a further 5 minutes at 37°C, and then 79.2 µL of

NTC was added. All reactions were then subject to proteinase K digestion by adding 4.6 µL of 0.5 M EDTA, 2.3 µL of 19.5 mg/mL proteinase K, and 5.75 µL of 20% SDS. After incubating for 40 minutes at 37°C samples were subject to TRI Reagent LS® (Sigma) extraction according to the manufacturer's instructions.

2.3.2 Reverse transcription and library preparation

All samples were diluted such that 150 ng of RNA was present in 8.65 µL of water. A volume of 3.35 µL of a nonamer/dNTP mix (20 µM random ninemer primer and 30 mM dNTP mix) was added to each sample and they were then incubated at 65°C for 10 minutes and then 4°C for 2 minutes. Next 7 µL of betaMAP buffer (0.14 M Tris pH 8.0, 0.21 M KCl, 0.03 M DTT, 2.9 M betaine, 17 mM $MnCl_2$, and 0.03 M DTT) was added to each reaction and samples incubated at 23°C for 2 minutes. Then 1 µL of SuperScript II reverse transcriptase was added to the reaction mixture and incubated at 23°C for 2 minutes. Following this the samples were placed in a PCR machine and incubated for 10 minutes at 25°C and then 42°C for 90 minutes. This was followed by 10 cycles of 50°C for 2 minutes then 42°C for 2 minutes. Finally reactions were inactivated by incubating at 72°C for 2 minutes.

DNA was purified using 2.2 times the reverse transcription reaction volume (44 µL) of Agencourt RNAClean XP Beads according to the manufacturer's protocol. Samples were eluted in 34 µL of water. Second strand synthesis was performed by adding 4 µL of Second Strand Synthesis Buffer (NEB) and 2 µL of Second Strand synthesis Enzyme Mix (NEB) to each sample and incubating at 16°C for 2 hours. DNA was purified using PureLink™ PCR Micro Elution Columns according to manufacturer's instructions. Samples were eluted in 15 µL of TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) and RNA concentration determined by Qubit. Sequencing adaptors were added and libraries amplified using the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina according to the manufacture's guidelines. Between 4 and 7 cycles of amplification were used depending on the amount of input DNA. Library clean-up

was performed using Ampure XP beads (Beckman). A 0.9X volume (22.5 µL) of Ampure beads was added to each sample and incubated at room temperature for 5 minutes. Beads were captured with a magnetic stand and washed in 80% ethanol. After ethanol was removed sample was eluted from the beads by addition of 15 µL of buffer TE. The libraries were pooled to a final concentration of 1.8 pM and sequenced 1 X 150 on a NextSeq 500 System (Ilumina).

2.3.3 Data analysis and presentation

Data analysis was performed using RNA Framework software (Incarnato et al., 2018) with assistance from George Young (Francis Crick Institute). Reads were mapped to the genome using the rf-map module of RNA Framework. This module uses Cutadapt (Martin, 2011) to trim the sequencing adaptors from reads and bowtie2 (Langmead and Salzberg, 2012) to map the reads to the reference genome. The parameters used for Cutadapt were as default except: -m 25 (reads were only kept if there were at least 25 bases left after trimming) and -q 20,20 (trims bases below this quality from reads before adaptor trimming). The parameters used for bowtie2 were '--very-sensitive-local' with the addition of: --mp 3,1 (reduced the mapping quality penalty for individual bases not aligning to reference genome) and --rdg 5,1 (reduced the minimum, but not maximum, mapping quality penalty for gaps). The rf-count module was then used to calculate mutations per base. Default parameters were used except for: --min-quality 20 (minimum mapping quality was increased to 20), --eval-surrounding (also considers the quality of adjacent bases when assessing the quality of a mutation), --collapse-consecutive (collapses mutations to the 3' most mutation), --max-collapse-distance 6 (increased the maximum distance allowed between mutations that will be collapsed), and --right-deletion (considers only the 3' most base in a deletion as mutated).

The rf-norm module of RNA framework was used for normalisation of reactivity values. The Siegfried scoring method was use in which the raw reactivity rate of each base is calculated from its mutation rate in the 1M7 treated sample, relative to its mutation rate in the untreated

sample (Siegfried et al., 2014). The 2-8% normalisation method was used in which the raw reactivity data is normalised by taking the average of the top 90-98% of reactivity values and dividing all reactivity values by this number. The EDC and DMS samples used the '--norm-independent parameter' so that normalisation was performed individually for the 4 different nucleotides. The rf-fold module of RNA Framework was used for structure prediction and generation of graphics. Folding was performed using ViennaRNA (Lorenz et al., 2011). Parameters were default except: --maximum-distance 150 (maximum allowed pairing distance was 150), --pseudoknots (allows pseudoknot prediction), --pseudoknot-window 400 (the folding window used when searching for pseudoknots, --pseudoknot-offset 100 (how far the window moves in each iteration when searching for pseudoknots), and --no-lonely-pairs (doesn't allow formation of single base pairs). Additional graphics showing interactions were generated using Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013) and 2D RNA structure were generated using the online tool *forna*, which is part of the ViennaRNA online suite of tools (Gruber et al., 2015). Full data available online at https://figshare.com/s/6444d82a7bab5f8cbb74.

2.4 Investigating RNA structure in SARS-CoV-2

2.4.1 SARS-CoV-2 virus growth and purification

Vero Ccl-81 cells (ATCC) were grown in Dulbecco's modified Eagle medium (Merck) with 10% FCS, 10 units/mL penicillin (Gibco), 10 µg/mL streptomycin (Gibco), and 2 mM l-glutamine. Vero Cells were infected at 80% confluency with passage 3 SARS-CoV-2 England/02/2020 at an MOI of 0.03. For infection the same growth media was used, except that it contained only 1% FCS. After 72 hours, the media was removed and clarified by centrifugation at 4000 g for 10 minutes. Following this the virus was purified either by PEG precipitation or by centrifugation through a sucrose cushion.

The PEG Virus Precipitation Kit (Abcam) was used according to the manufacturer's instructions, with the virus left to precipitate overnight at 4°C. For sucrose cushion purification the clarified media was layered onto a cushion containing 10% sucrose in MSE buffer (10 mM MOPS, pH 6.8, 150 mM NaCl, and 1 mM EDTA) and subject to centrifugation at 10,000 x g for 4 hours at 4°C. Following this, the media was carefully removed, followed by the sucrose cushion. The virus pellet was then left to resuspend overnight at 4°C in MSE buffer.

2.4.2 Chemical probing of SARS-CoV-2

For each in *virio* repeat one T75 flask of Vero cells was infected and subject to purification by centrifugation through a sucrose cushion or PEG precipitation, as described in section 2.4. Virus was resuspended in 100 µL of MSE buffer and 25 µL of 1 M bicine, pH 8.0, was added. The sample was then added to a tube containing: 16.8 µL of 100 mM 1M7, 2 µL of DMS, or 16.8 µL of DMSO (for negative controls). The 1M7 and DMSO samples were incubated at 37°C for 90 seconds. DMS samples were incubated at 37°C for either 90 seconds or 5 minutes before addition of 81.1 µL of 1 M Bicine and 28.3 µL of 1M DTT. All samples were then subjected to Proteinase K digestion. EDTA was added to a final concentration of 1 mM, SDS to 0.5%, and proteinase K to 0.3 mg/mL. Samples were incubated at 37°C for 30 minutes before RNA was extracted using TRI Reagent LS® (Sigma) according to manufacturer's protocol. Reverse transcription and library preparation were performed as described in section 2.3.2. Analysis of data was performed as in section 2.4.3. The exception to this is that a windowed approach to folding was adopted when using the rf-fold module of RNA Framework. Max distance pairing distance was set to 500, folding window 3,000, fold offset 300, window trim 200, partition window 1,500, and partition offset 250. Full data available online at https://figshare.com/s/6444d82a7bab5f8cbb74.

2.4.3 SPLASH of SARS-CoV-2

For each in *virio* repeat two T75 flasks of Vero cells were infected and subject to purification by centrifugation through a sucrose cushion, as described in section 2.4.1. Virus was resuspended in 100 µL of MSE buffer and 25 µL of 1 M bicine was added. Resuspended virus (volume of 125 µL) was added to 14 µL aliquots of 1 mg/mL PTAz (Berry & Associates, Inc). The sample was then placed on a 4°C block in a Stratalinker 1800 UV Crosslinker (Stratagene) and subject to irradiation for 10 minutes at a wavelength of 365 nm. The sample was then subject to Proteinase K digestion, RNA extraction and size selection as described in section 2.4.3. The RNA was then fragmented using the NEBNext Magnesium RNA fragmentation module (NEB) as described in section 2.2.2. Biotin then needed to be added to the PTAz by click chemistry in order to allow subsequent enrichment of the crosslinked RNAs. Each 30 µL RNA sample had 7.6 µL of 10 mM Sulfo-Dibenzylcyclooctyne-Biotin, 2 µL of RNasin, and 0.4 µL of Tris pH 7.5 added to it. Samples were then incubated for 90 minutes at 37°C. To each 40 µL reaction, 200 µL of RNAClean XP beads and 288 µL of isopropanol were then added. Samples were incubated at room temperature for 8 minutes. Supernatant was removed and beads washed 3 times with 70% ethanol. RNA was recovered in 100 µL of water. Click reaction was performed by David Bauer (Francis Crick Institute). Enrichment of crosslinked RNA and sequencing library preparation was performed as described in section 2.2.2. Data analysis and presentation was performed as described in section 2.2.3 using software written by Anob Chakrabarti (Francis Crick Institute).

2.5 Structural studies on NP

2.5.1 Cloning of NP expression constructs

The A/Northern Territory/60/1968 (H3N2) (NT60) NP and NT60 R416A NP sequences were amplified from the corresponding pFL-TAP-NP vectors (Turrell, 2015), with sequence

optimised for expression in *Spodoptera frugiperda*. The amplification was performed with primers containing overhangs (Table. 2), adding BamHI and EcoRI restriction sites to the ends of the NP sequence. The PCR reaction contained 0.25 µL of Q5 DNA polymerase (NEB), 5 µL of 5X Q5 DNA polymerase buffer (NEB), 0.5 µL of 10 mM dNTPs, 1.25 µL of the 10 µM forward primer GGATCCATGGCTTCCCAGGGTAC, 1.25 µL of the 10 µM reverse primer GAATTCTTAGTTGTCGTATTCCTCAGC, 0.2 µL of the template, and 16.25 µL of water. The following PCR conditions were used: 98°C for 30 seconds, then 30 cycles of 98°C for 10 seconds, 55°C for 30 seconds, 72°C for 90 seconds, and then finally 72°C for 5 minutes. Fragments were run on a 1% agarose gel and the band of the expected size extracted and purified using the QIAquick gel extraction kit (Qiagen).

Source organism	A/Northern Territory/60/1968 (H3N2) influenza virus
DNA source	pFL-TAP-NP or pFL-TAP-NP R416A
Forward primer	[GGATCC][1]ATGGCTTCCCAGGGTAC
Reverse primer	[GAATTC][2]TTAGTTGTCGTATTCCTCAGC
Expression vector	pGEX-6P-1
Expression host	*E. coli* BL21 (DE3) cells
Complete amino acid sequence of the construct produced	[GPLGS][3]MASQGTKRSYEQMETDGERQNATEIRASVGKMIDGI GRFYIQMCTELKLSDYEGRLIQNSLTIERMVLSAFDERRNKYLE EHPSAGKDPKKTGGPIYKRVDGKWMRELVLYDKGEIRRIWRQ ANNGDDATAGLTHMMIWHSNLNDTTYQRTRALVRTGMDPRM CSLMQGSTLPRRSGAAGAAVKGVGTMVMELIRMIKRGINDRN FWRGENGRKTRSAYERMCNILKGKFQTAAQRAMMDQVRESR NPGNAEIEDLIFLARSALILRGSVAHKSCLPACVYGPAVASGYD FEKEGYSLVGIDPFKLLQNSQVYSLIRPNENPAHKSQLVWMAC NSAAFEDLRVLSFIRGTKVSPRGKLSTRGVQIASNENMDAMES STLELRSRYWAIRTRSGGNTNQQRASAGQISVQPAFSVQANL PFDKPTIMAAFTGNTEGRTSDMRAEIIRMMEGAKPEEMSFQG RGVFELSDEKAANPIVPSFDMSNEGSYFFGDNAEEYDN

[1] BamHI restriction site. [2] EcoRI restriction site. [3] Residues retained after cleavage that are not part of NP sequence.

Table 2: Cloning and expression of the NT60 NP.
Adapted from (Knight et al., 2021).

The extracted fragment was ligated into the pGEX-6P-1 expression vector. This vector allows expression of proteins with a cleavable Glutathione S-Transferase tag on the N-terminus. Both fragment and vector were digested with EcoRI and BamHI in SuRE/Cut Buffer B (Roche) for 2 hours at 37°C, followed by 15 minutes at 65°C to heat inactivate the enzymes. The fragment and vector were ligated together using the Quick-Stick ligase (Bioline) according to the manufacturer's instructions with a 3:1 insert to vector ratio. After incubating for 15 minutes at room temperature, the ligation mixture was added to competent DH5α cells and incubated at room temperature for 2 minutes, before plating onto ampicillin (100 µg/mL) agar plates and incubating overnight at 37°C. Colonies were picked and grown overnight at 37°C in Lysogeny Broth (LB) containing 100 µg/mL ampicillin. The QIAprep Spin Miniprep kit (Qiagen) was used to extract the plasmids. Constructs were confirmed by sequencing and transformed into *Escherichia coli* BL21 (DE3) cells. For the transformation, the cells were thawed on ice for 30 minutes and 1 µL of plasmid was then added to 50 µL of cells. After incubating for 2 minutes on ice the cells were incubated at 37°C for 45 seconds and then returned to ice for 5 minutes. The cells were then mixed with 250 µL of LB containing 100 µg/mL ampicillin and incubated at 37°C for 1 hour before being plated onto ampicillin agar plates and incubating overnight at 37°C. Colonies were picked and grown in LB ampicillin overnight at 37°C. A 50% glycerol was stock then made with this culture and stored at -80°C to expand for expression as required.

2.5.2 Expression and purification of NP

Starter cultures were grown by inoculating 10 mL of LB, containing 100 µg/mL ampicillin, with bacteria from the glycerol stock and incubating overnight at 37°C. The starter culture was used to inoculate 2 L of LB ampicillin media, at a 1:100 ratio. The culture was grown at 37°C until an OD_{600} of 0.6 was achieved. Isopropyl β-d-1-thiogalactopyranoside was added to a concentration of 1 mM to induce protein expression. The NT60 NP expressing bacteria were

incubated for 3 hours at 37°C and the NT60 R416A NP and NT60 E339A NP expressing bacteria overnight at 18°C.

The bacterial culture was then centrifuged at 4,000 g for 15 minutes at 4°C. Pellets were resuspended in 25 mL of Wash Buffer (50 mM HEPES-NaOH, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, and 0.05% (w/v) Octyl β-D-1-thioglucopyranoside (OTG)). The Wash buffer was supplemented with 50 µL of 1 M DTT, 2.5 mg of RNase A, one SIGMAFAST protease inhibitor tablet (Sigma), 10 µL of 250 units/ µL Benzonase Nuclease (Sigma), and 35 mg of lysozyme. The mixture was left to rotate at room temperature for 15 minutes prior to sonication and then centrifugation at 35,000 g for 45 minutes at 4°C. Per each original 2 L of culture, 1 mL of Glutathione Sepharose 4B beads (GE Healthcare) were added to the clarified supernatant and incubated at 4°C for 3 hours, with gentle rotation. The mixture was subject to centrifugation at 2,000 g at 4°C for 3 minutes. The supernatant was removed and 20 mL of High Salt Wash Buffer (50 mM HEPES-NaOH, pH 7.5, 1.5 M NaCl, 10% (v/v) glycerol, and 0.05% (w/v) OTG) was added to the beads. In the case where NP was left bound to endogenous nucleic acid from the expression host, these washes were instead performed with Wash Buffer (not high salt). The beads were incubated at 4°C for 10 minutes with gentle rotation and then centrifuged at 2,000 g at 4°C for 3 minutes again. This wash process was repeated 5 times before performing an additional wash with Wash Buffer (not high salt) containing 5 mM DTT. After the final wash, the beads were resuspended in 10 mL of Wash Buffer supplemented with 5 mM DTT, 0.2 mg of HRV 3C protease, and 5 µL of Benzonase Nuclease and incubated overnight at 4°C with gentle rotation.

Beads were pelleted at 2,000 g for 5 minutes at 4°C. The supernatant was removed and further clarified by centrifugation at 2,000 g for 5 minutes at 4°C. The resulting supernatant was then transferred to a 30 KDa Millipore Protein Concentrator (Merck) and centrifuged at 3,600 g at 4°C until a volume of ~ 0.5 mL was reached. The concentrated protein was then loaded onto a Superdex 200 Increase 10/300 GL column (GE Healthcare) equilibrated with a buffer containing 25 mM HEPES-NaOH, pH 7.5, and 150 mM NaCl. The flow through was collected

in 0.5 mL fractions. The fractions containing NP were pooled and concentrated in a 30 KDa Millipore Protein Concentrator. A fraction of the sample was subject to SDS-polyacrylamide gel electrophoresis with Coomassie blue staining to confirm its identity. The protein was either stored at 4°C for use the same day or frozen in liquid nitrogen and stored at -80°C.

2.5.3 Assessment of nucleic acid binding properties of the R416A NP

The R416A NP was incubated at room temperature for 10 minutes in a 1:1 molar ratio with a 5 (5′-AGUAG-3′) or 14 (5′-CCUCUGCUUCUGCU-3′) nucleotide long RNA (buffer 25 mM HEPES-NaOH, pH 7.5, and 150 mM NaCl). The mixture was then loaded onto a Superdex 200 Increase 10/300 GL column for Size Exclusion Chromatography (SEC), as described in section 1.2. The 260/280 ratio of the NP containing fraction was assessed to determine if the RNA had remained associated with NP. To assess DNA binding, the R416A NP was also mixed in a 4:1 molar ratio with a 100 nucleotide-DNA and subject to SEC.

A ThermoFluor assay (Walter et al., 2012) was performed on the R416A NP either alone, or with a G nucleotide, a 5′-AG-3′ dinucleotide, an 8-nucleotide RNA 5′-UAUGAGGC-3′, a 12-nucleotide RNA 5′-AAAAAAAAAAAA-3′, or a 14-nucleotide RNA 5′-GUAUAUGAGGCCCA-3′. The assay was performed with each condition in triplicate using an Mx3005P qPCR System (Agilent). An ‘expanding sawtooth’ method was used, with the sample heated gradually from 25°C to 95°C with 30 second stops at each new temperature before briefly returning to 25°C to measure fluorescence (temperature can affect fluorescence in a manner unrelated to protein unfolding). The excitation and emission filters were set to 492 and 585 nm respectively. The assay was performed in a 96 well PCR plate with a reaction volume of 40 µL in a buffer containing 25 mM HEPES-NaOH, pH 7.5, 150 mM NaCl, and a 1 in 100 dilution of SYPRO Orange (Invitrogen). The R416A NP was added at a concentration of 1.34 µM and RNA at a concentration of 20 µM. The data was analysed using the JTSA webserver (Bond, 2017). The

melt temperature was determined from the melt curve as the temperature at which 50% of maximum fluorescence was achieved (after subtracting base line fluorescence).

2.5.5 Crystallography

Crystallisation trials were set up using the vapour diffusion method, with the NT60 R416A NP at 10 mg/mL or the NT60 NP at 7 mg/mL. A wide range of conditions were screened including standard commercially available screens and fine screens around conditions that were found to form crystals. Crystallisation trials were performed both at room temperature and 4°C. Trials were performed with the NT60 NP alone or in the presence of: a G nucleotide, a 5′-AG-3′ dinucleotide, a 5-nucleotide RNA 5′-AGUAG-3′, an 8-nucleotide RNA 5′-UAUGAGGC-3′, a 12-nucleotide RNA 5′-AAAAAAAAAAAA-3′, a 14-nucleotide RNA 5′-GUAUAUGAGGCCCA-3′, or a 14-nucleotide DNA 5′-GTATATGAGGCCCA-3′ The details of the best diffracting crystals formed in the absence or presence of RNA are detailed in Table 3.

Crystal form	P1-2_1-1	C2-2-2_1
Method	Vapour diffusion	Vapour diffusion
Plate type	Swissci 3-drop	Swissci 3-drop
Temperature (K)	293	277
Protein concentration	10 mg/mL	10 mg/mL
Buffer composition of protein solution	25 mM HEPES-NaOH, pH 7.5, and 150 mM NaCl	25 mM HEPES-NA, pH 7.5, and 150 mM NaCl
Composition of reservoir solution	10% w/v PEG 8000, 20% v/v ethylene glycol, 0.02 M of each alcohol (1,6-hexanediol, 0.2 M 1-butanol, 0.2 M (RS)-1,2-propanediol, 0.2 M 2-propanol, 0.2 M 1,4-butanediol, 0.2 M 1,3-propanediol), 0.1 M MES/imidazole pH 6.5.	0.7 M tri-Sodium Citrate, 0.1 M bis-Tris Propane pH 7.0, 1.7 M excess of 14-nucleotide DNA (5′-GTATATGAGGCCCA-3′).
Volume and ratio of drop	200 nL (1:1)	200 nL (1:1)
Volume of reservoir	30 µL	30 µL

Table 3: Summary of the crystallisation conditions.
Adapted from (Knight et al., 2021).

Data sets were collected at the Diamond light source, Didcot, UK, from cryo-cooled crystals. The data was collected in 0.1° degree increments with a full 360° rotation, at a temperature of 100 K. The exact collection parameters for the best diffracting crystals in the absence or presence of nucleic acid are detailed in Table 4. Processing was performed using autoPROC (Vonrhein et al., 2011), with an anisotropic cut-off applied using STARANISO (Tickle et al., 2018).

Crystal form	P1-2_1-1	C2-2-2_1
Diffraction source	DLS-I24	DLS-I03
Wavelength (Å)	0.9686	0.9763
Temperature (K)	100	100
Detector	Dectris Pilatus 6M	Eiger2 XE 16M
Rotation range per image (°)	0.1	0.1
Total rotation range (°)	360	360
Exposure time per image (s)	0.005	0.003

Table 4: Crystallography data collection parameters.
Adapted from (Knight et al., 2021).

Initial assessment of data quality was performed using phenix xtriage (Zwart et al., 2005). Structures were solved by molecular replacement with the WSN R416A NP model (PDB ID 3ZDP) (Chenavas et al., 2013) using PHASER (McCoy et al., 2007). Refinement of the structure was then performed using phenix.refine (Afonine et al., 2012). Refinement strategies: 'XYZ (reciprocal space)', 'XYZ (real-space)', 'rigid body', 'individual B-factors', and 'occupancies' were selected with three cycles of refinement. The additional options 'optimise X-ray/stereochemistry weight' and 'optimise X-ray/ADP weight' were selected, but otherwise default parameters were used. Manual adjustments to the refined model were then made in *Coot* (Emsley et al., 2010) to correct for atom clashes, Ramachandran outliers, and missing atoms. In the C 2 2 2_1 space group twinning was detected with a significant twin fraction of 0.23, thus during refinement the twin law 1/2*h-1/2*k,-3/2*h-1/2*k,-l was used. The quality of the models produced was assessed using MolProbity (Williams et al., 2018a) to guide further iterative rounds of refinement using phenix.refine and *Coot*. Structural figures were made using chimera X (Goddard et al., 2018).

2.5.6 Cryo-EM

The purified NT60 R416A NP was mixed in a 3:1 molar ratio with a 47-nucleotide RNA (′5-AGUAGAAACAAGGGUAUUUUUCUUUACUAGUCUACCCUGCUUUUGCU-3′) or in a 3:1 or 4:1 molar ratio with a 100-nucleotide DNA. The NP was at a concentration of 0.5 mg/mL in a buffer containing 25 mM HEPES-NaOH, pH 7.5 and 150 mM NaCl. The sample was incubated on ice for 45 minutes before being applied to a SEC column and purified as described in section 2.5.2. The purified NT60 NP was mixed in 3:1 molar ratios with a 105-nucleotide DNA (′5-TCGAGATCTTCCACCACCTGGGCATCGTCGTCGTATTCCTCTTCGTCCTCATCTTCCTCTTCATCATCCAGTCCTTCCACATATCCTTCGGCGTCGCTGTCGGGG-3′) in a buffer containing 25 mM HEPES-NaOH, pH 7.5 and either 150 mM or 300 mM NaCl. For all samples a series of 3 serial 50% dilutions were made and all were used to make grids.

Cryo-EM was performed at the Oxford Particle Imaging Centre (University of Oxford) by either Jeremy Keown or Loïc Carrique (both University of Oxford). Samples were applied to glow discharged copper Quantifoil R2/1 200 mesh grids. A Vitrobot mark IV (FEI) was used to blot the grids for 3.5 seconds and then flash freeze them by plunging them into liquid ethane. Grids were screened on a Glacios (Thermofisher scientific) microscope operating at 200 kV and equipped with a Falcon 3 camera. EPU version 2.10 was used to control the microscope. For grids that were of a sufficiently high quality, a Titan-Krios operating at 300 kV equipped with a K2 Summit (Gatan) camera and a GIF Quantum energy filter was used for data collection.

Data were motion corrected and the Contrast Transfer Function (CTF) estimated using patch motion correction and patch CTF estimation, respectively. Particles were picked using the cryoSPARC (Punjani et al., 2017) blob picker to generate a set of particles for initial 2D classification. These 2D classes were then used for subsequent template picking again in cryoSPARC. From the template picking further 2D classification was used to remove bad particles. For 3D reconstructions particles from the desired classes were submitted to ab initio

model generation (to generate a starting model) followed by heterogeneous refinement (to sort particles into their most optimal 3D class). Where appropriate, non-uniform refinement was used to for refinement of promising 3D reconstructions from heterogeneous refinement.

3. Intersegment RNA-RNA interactions in the influenza genome

3.1 Chapter Summary

- Interactions between the eight influenza vRNAs are believed to facilitate their assembly for packaging.
- These interactions are captured *in virio* using SPLASH, revealing extensive redundant networks of inter-segment RNA-RNA interactions.
- Reassortant viruses are found to have interaction networks that are mostly derived from their parental strains.
- It is demonstrated that targeted mutations can generate new inter-segment interactions that can affect reassortment.
- This work has potential use in vaccine generation where reassortant viruses are produced containing the antigenic segments from the predicted seasonal strain.

3.2 Introduction

The influenza virus utilises a selective genome packaging mechanism that reduces the production of non-infectious virus particles (discussed in section 1.3). This is suggested to be mediated in part by inter-segment RNA-RNA interactions that encourage the formation of complexes containing the 8 different influenza genome segments (i.e. without duplicate or missing segments). These interactions were recently mapped on a genome wide scale for the first time by Dadonaite *et al* (Dadonaite et al., 2019), showing the presence of extensive networks of RNA-RNA interactions between the vRNAs.

SPLASH is a chemical crosslinking method used to identify RNA-RNA interactions (Fig. 7). Biotinylated psoralen is added to a sample (e.g. a virus) where it will intercalate into double stranded regions of RNA. Upon exposure to long wave UV radiation (365 nm), the psoralen

will crosslink the interacting RNAs together. The RNA is subsequently purified and fragmented. The psoralen crosslinked strands are then captured using streptavidin coated beads and proximity ligation is performed. This generates hybrid strands from the interacting segments of RNA. The psoralen is removed by exposure to short wave UV radiation (254 nm) which reverses the crosslinking. The hybrid RNA segment is then prepared for sequencing by addition of adaptors, reverse transcription, and amplification. The sequencing reads are mapped back to the reference genome. As the reads are hybrids, the read should map to two different regions (or segments) of the genome, indicating the presence of an interaction.

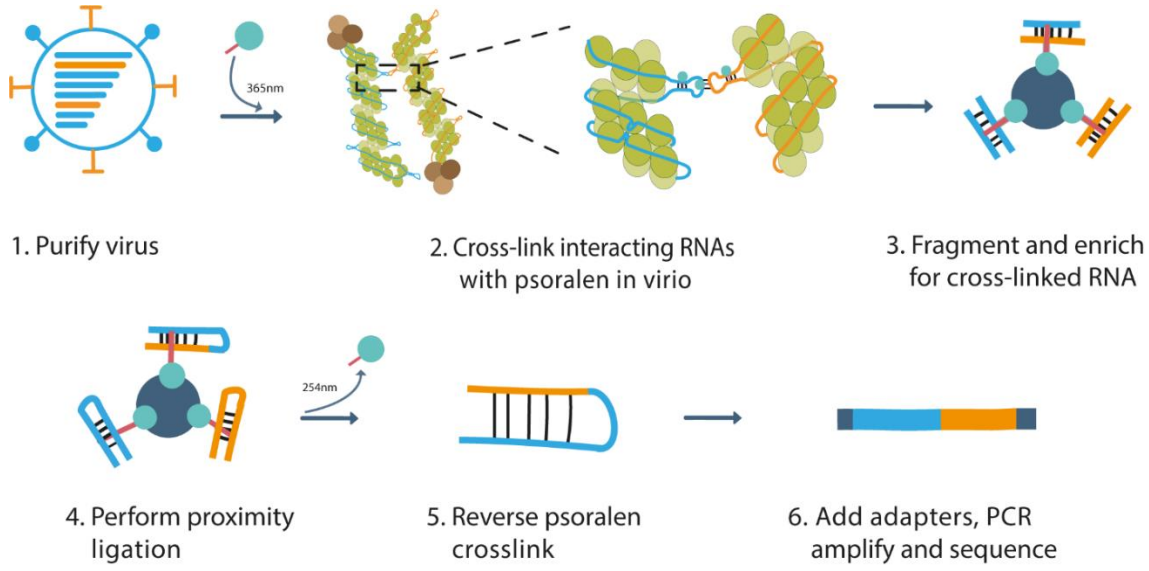


Figure 7: The process of performing SPLASH on an influenza virus.
1. Influenza virions are purified through a sucrose cushion and psoralen is added. 2. The psoralen intercalates into double stranded regions of RNA which can then be crosslinked together upon exposure to long wave UV radiation (365 nm). 3. The viral RNA is purified, fragmented, and then captured by making use of a biotin tag on the psoralen. 4. Proximity ligation is performed on the captured crosslinked RNAs to generate hybrid strands. 5. The RNA is exposed to short wave UV radiation (254 nm) to reverse the psoralen crosslinks. 6. The RNA is reverse transcribed and sequencing libraries are generated. Adapted from (Dadonaite et al., 2019).

Interactions between the influenza genome segments that mediate their bundling are also likely to be important to the process of reassortment. If two segments lack regions of sequence complementarity that facilitate the formation of RNA-RNA interactions between them, it may reduce the likelihood of them being packaged into the same viral particle. This may contribute to the large variance observed in reassortment rates, with more closely related strains reassorting more efficiently than those that are more distantly related (Villa and Lässig, 2017) (Gerber et al., 2014). Greater understanding of reassortment is required to aid in predicting the emergence of pandemic strains and because reassortment is required for the production of the seasonal influenza vaccine (Fulvini et al., 2011).

The aim of this chapter was to improve understanding of the structure and assembly of the influenza genome, both generally and in the context of reassortment. SPLASH was used to map the RNA-RNA interaction networks in reassortant viruses and their parent strains. In addition, targeted mutations were used to generate new interactions that influence reassortment.

3.3 Results

3.3.1 RNA-RNA interactions in the PR8 genome

To investigate the RNA-RNA interaction networks present in IAVs, SPLASH was first performed on the PR8 (H1N1) virus. Unique sequencing reads were mapped back to the PR8 genome, to produce interaction matrixes for each segment with each of its potential partners (i.e. the other 7 segments). The matrices were used to manually select interaction loci, which could be identified as regions of overlapping reads (Fig. 8A). A Gaussian curve was fitted to each locus based on read count, with its ‘full width at half maximum’ used to define an interaction window. The number of unique reads falling within the interaction window was used to define the ‘intensity’ of the interaction. The IntaRNA tool (Busch et al., 2008) was used to

predict the highest probability structure within the interaction window. A full table of interaction loci for all SPLASH datasets can be found online at https://figshare.com/s/6444d82a7bab5f8cbb74.

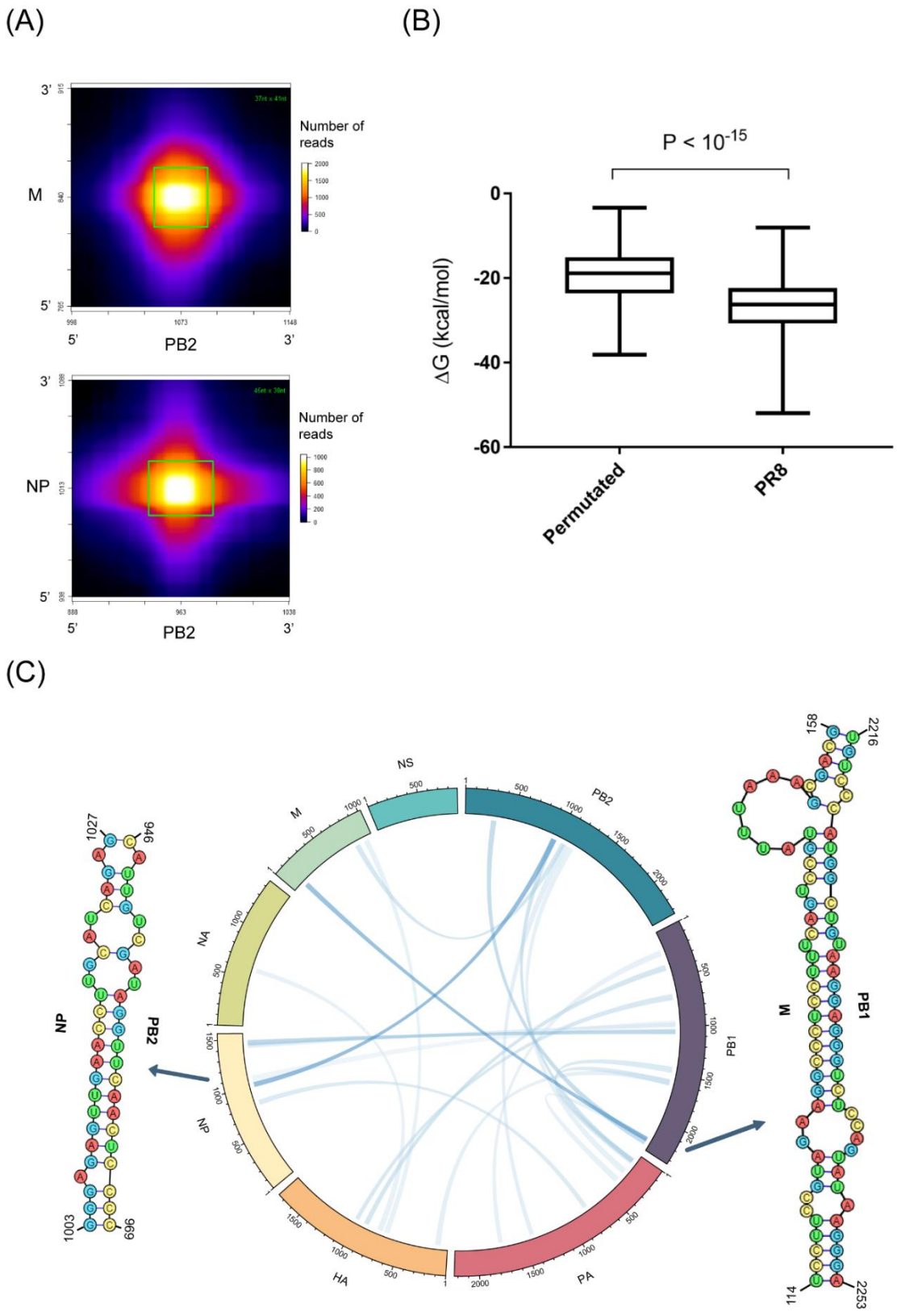


Figure 8: The inter-segment interactions of the PR8 virus

(A) The identification of interaction loci from SPLASH data. The unique sequencing reads are mapped to a grid in which the X and Y axis are the genome co-ordinates of two of the vRNA segments. Interactions can be identified manually as regions that display a large number of overlapping reads. The loci are then picked (region shown in green box) by fitting a Gaussian

distribution to the data. (B) A chart showing that the interactions identified by SPLASH have significantly lower free energy than the interactions formed when the same sequences are randomly permutated to have different partners. Centre line shows the median values, box edges the upper and lower quartiles, and the whiskers the range. P = $<10^{-15}$ Wilcoxon matched-pairs signed rank test. (C) Inter-segment RNA-RNA interaction map for the PR8 virus determined by SPLASH. The 8 influenza genome segments are arranged around the outside of the plot with connecting lines showing the location of interactions. The darker the blue line, the higher intensity the interaction. The 20 highest intensity interactions are displayed.

An extensive network of RNA-RNA interactions was identified connecting the PR8 genomic segments. A total of 600 discrete interaction loci were identified, 177 of which have more than 50 unique reads mapping to them. The interactions vary greatly in intensity, with the top 3% of interactions in the dataset accounting for 26% of the total mapped reads. When the ΔG of the observed interactions (mean -26.44 kcal/mol) is compared to a randomly shuffled data set of the identified interactions sites (mean -19.39 kcal/mol), the ΔG for the observed interactions is significantly lower than for the randomly permutated dataset (P $<10^{-15}$ Wilcoxon matched-pairs signed rank test), indicating that the interactions are not observed due to random chance (Fig. 8B). The predicted ΔG of an interaction does not correlate with the intensity order of the interactions (R = 0.114 with P value of 0.134, Spearman correlation), suggesting that the free energy of an interaction is not the determining factor of its intensity.

The interactions are dispersed across the length of the vRNAs and are not limited to non-coding or terminal regions (Fig. 8C). Some regions are seen to form interactions with multiple partners. This, along with the large number of possible interactions identified, suggests that there is redundancy in the interaction network and that there are multiple ways in which the influenza genome can be bundled, even for the same strain.

A separate isolate of PR8 (henceforth referred to as PR8B), with sequence identity >99%, had previously been subject to SPLASH (Dadonaite et al., 2019). Of the 20 highest intensity interactions in the PR8 dataset (i.e. the top 3.33%), 16 also fell in the top 20 interactions of the PR8B dataset, a further 3 were found within the top 6.67%, and one fell in the 40th percentile. Of the top 10% of PR8 interactions, 82% were present in the top 10% for PR8B and all were present at some level. Of the top 25% of PR8 interactions, 84.7% of interactions

were present in the top 25% of PR8B and 97.3% were present at some level. Some changes in interaction intensity are seen. This may indicate that the intensity order of the interactions from SPLASH is not a highly precise measure of their relative abundance in the population, and/or that the interactions vary in prevalence in different populations of the same virus.

The PR8 virus has 96% sequence identity to the WSN (H1N1) virus on which SPLASH has also been performed previously (Dadonaite et al., 2019). Despite the high sequence identity, the interactions networks differ substantially. Only 3 interactions fall in the 20 most intense interactions of both datasets and 6 of the interactions found in the top 20 of the PR8 dataset are completely absent from the WSN data. Of the top 25% of PR8 interactions, 32% are present in the top 25% for WSN and 65.3% were present at some level. This suggests that relatively small differences in sequence can have a large impact upon the inter-segment RNA-RNA interactions formed and their relative intensity.

Intra-segmental interactions can also be observed in the SPLASH data (Fig. 9). The vast majority of chimeric reads map to regions separated by very short distances (<100 nucleotides). This suggests that there is likely to be extensive short range RNA structure present. However, as these interactions are all very close together, it is not easy to identify discrete interaction loci (compare Fig. 8A to the short range interactions in Fig. 9). There is some evidence of a very small number of long range interactions loci (such as the interaction spanning from ~1,100 to ~2,000 in the PB2 segment (Fig.9)). However, the longer range interactions are of much lower intensity than the short range interactions observed. It should be noted that inter-segment interactions may contribute to these data due to cases where two copies of the same segment were packaged together.

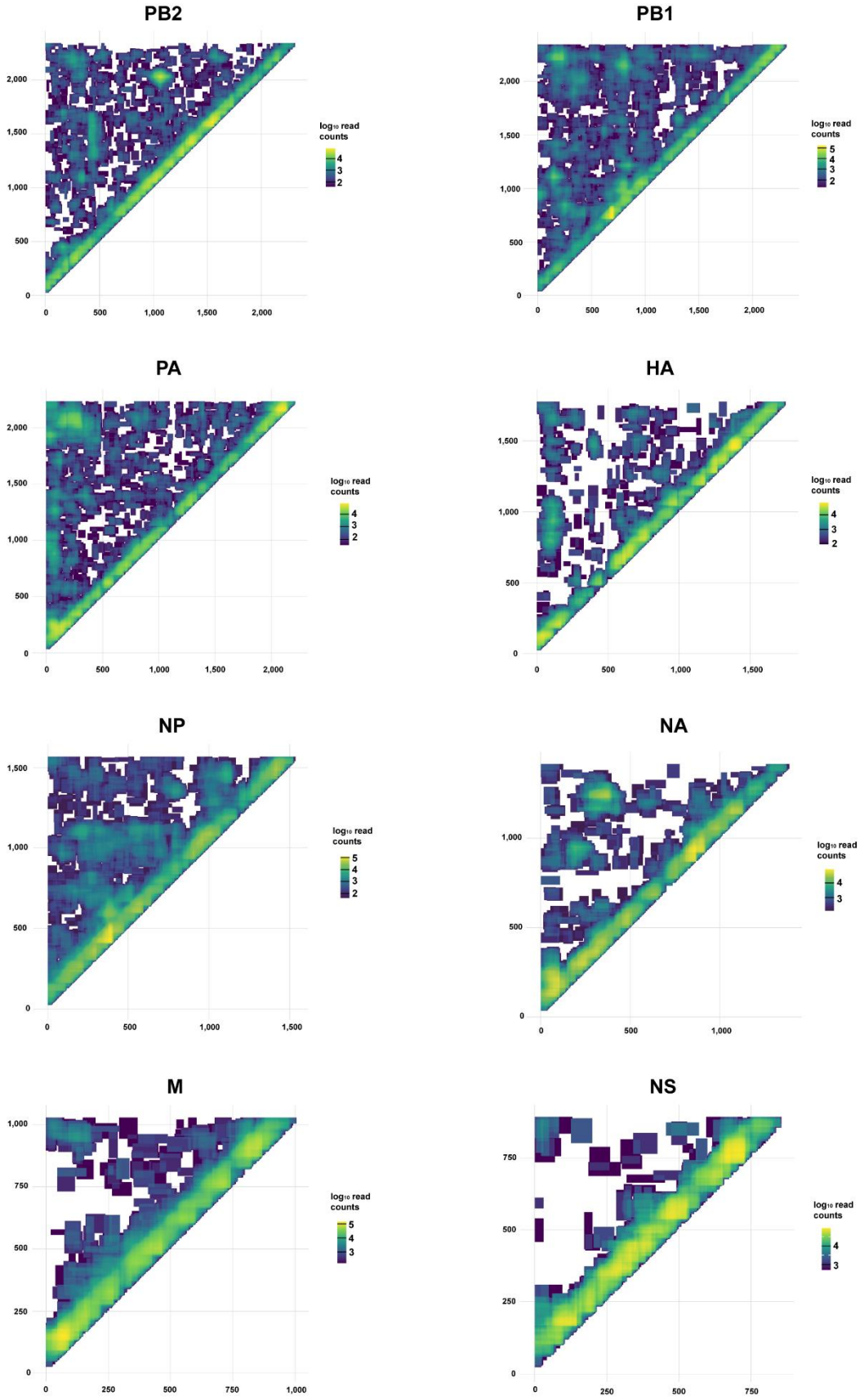


Figure 9: SPLASH-identified intra-segment interactions in PR8.
Contact maps from an in virio SPLASH experiment showing the intra-segment chimeric reads mapping to the PR8 genome. Reads have been log_{10} normalised.

3.3.2 Inter-segment interaction in reassortant viruses

In order to investigate the role of interactions in reassortment, SPLASH was next performed on the Udorn (H3N2) virus, and a reassortant with the PB2, PA, HA, NP, M, and NS segments from PR8 and the NA and PB1 segments from Udorn (denoted PR8::Udorn PB1 + NA) (Fig. 10 + 11). The Udorn and PR8::Udorn PB1 + NA viruses had 72 and 227 loci with 50 or more reads unique sequencing reads mapping to them respectively. The interaction networks differ greatly for the PR8 and Udorn viruses (which share 89.3% sequence identity), with few conserved interactions (~10% present in both data sets). This indicates that IAVs do not utilise a common set of inter-segment interactions.

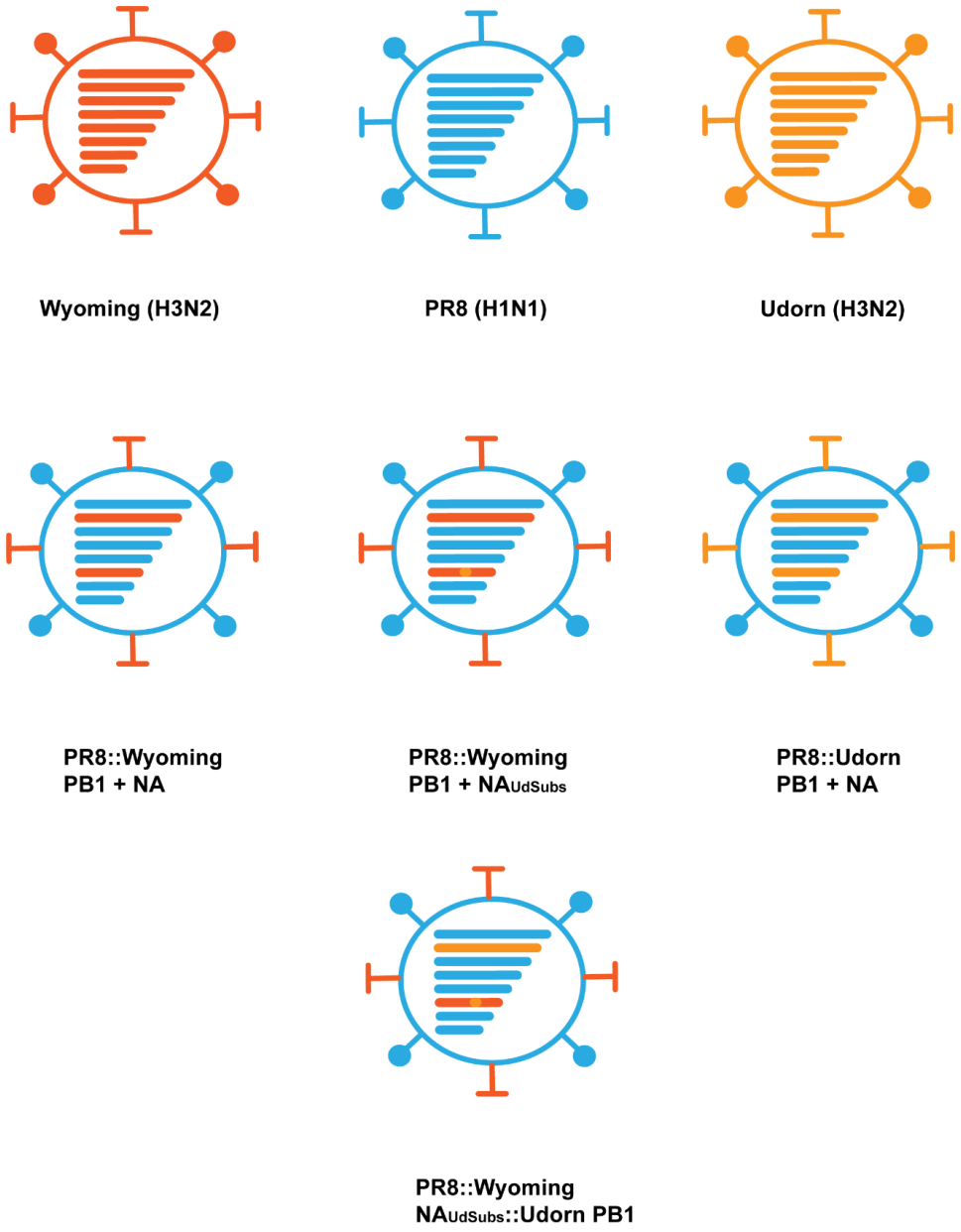


Figure 10: The viruses on which SPLASH was performed.

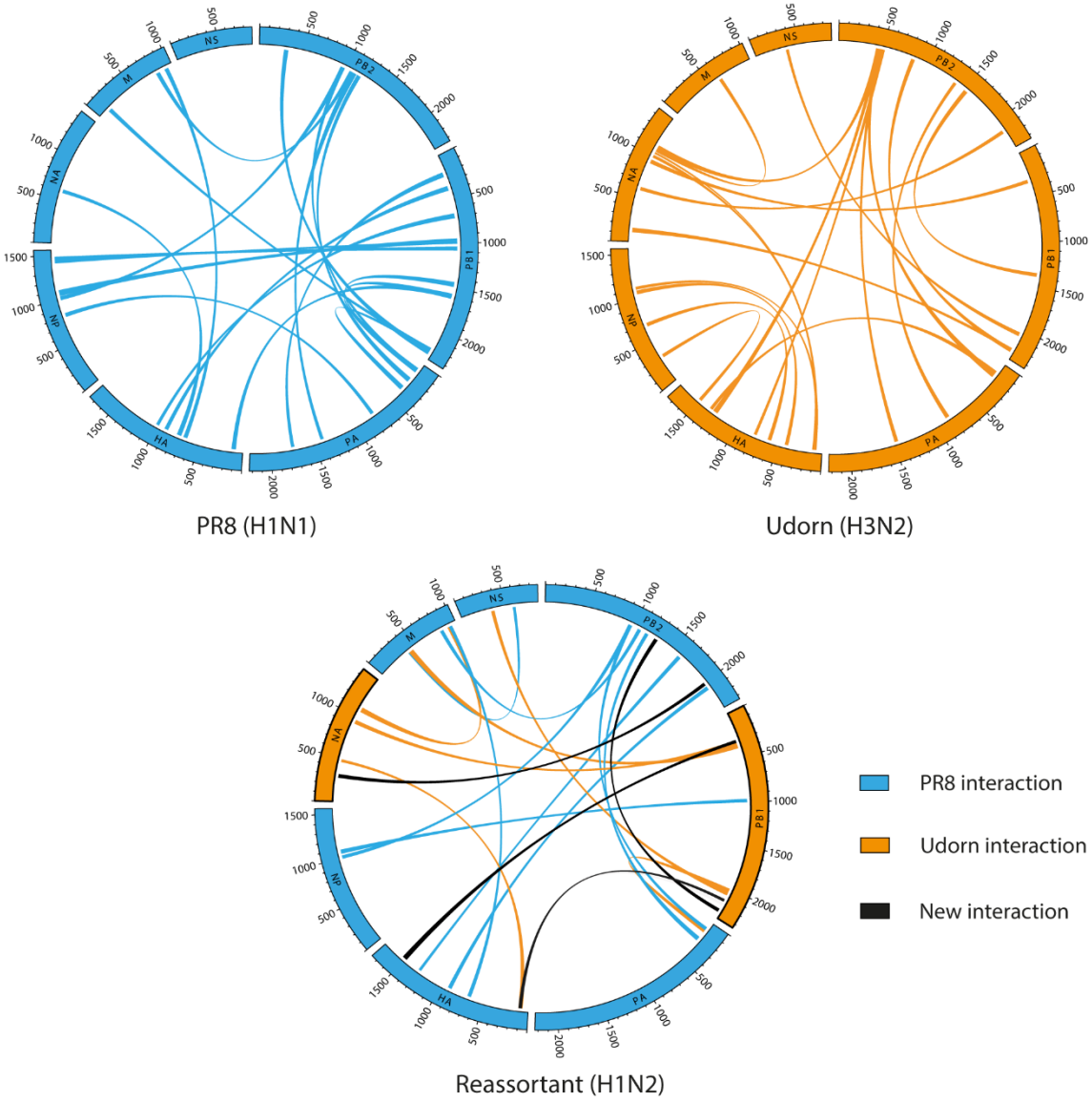


Figure 11: Comparison of inter-segment interactions in reassortant viruses.
RNA-RNA interaction maps generated for the PR8, Udorn and reassortant PR8::Udorn PB1 + NA viruses. The 20 highest intensity interactions are displayed for each virus.

The majority of interactions observed in the PR8::Udorn PB1 + NA reassortant virus are inherited from the two parental strains, with 81 of the 100 most intense interactions present in one of the parents. Of the 19 new interactions, 18 occur between Udorn and PR8 segments. This indicates that some new interactions are formed in order to facilitate bundling of the segments from the two different viruses. Overall, the interactions formed between segments from the same strain remain largely unchanged in the reassortant. However, the interactions do appear to change in prominence, with only 9 of the interactions in the top 20 of the

reassortant having fallen in the top 20 of one of the parents. In addition, of the top 20 interactions in the reassortant that were also present in PR8, 3 had risen by more than 40 places (~6%). This suggests that changes may also occur in the prominence of interactions to facilitate bundling of segments from different viruses.

3.3.3 Introducing new inter-segment interactions

Preferential co-segregation of the PB1 and NA segments from H3N2 viruses has been observed during vaccine seed strain production (Fulvini et al., 2011) (Cobbin et al., 2013). This is despite this conferring reduced replicative fitness on the virus when compared to reassortants containing only the NA and HA from the H3N2 virus (Cobbin et al., 2014). In addition, competitive reassortment experiments show viruses with the Udorn NA segment preferentially package the Udorn PB1 segment (Gilbertson et al., 2016) (Fig. 12A). The same study used chimeric constructs to narrow down the region responsible for this co-segregation to the 272-566 region of the PB1 segment. The SPLASH data indicates that a prominent interaction connects these two segments, falling at 305-338 on the PB1 segment, both in the Udorn (H3N2) virus (9^{th} highest intensity interaction) and the PR8::Udorn PB1 + NA reassortant (third highest intensity interaction) (Fig. 11).

Further competitive reassortment experiments (performed by Brad Gilbertson, University of Melbourne) (Fig. 12A) found that the preferential co-segregation of the NA and PB1 segments holds true for several H3N2 viruses (Udorn, A/Memphis/1/71, and A/Port Chalmers/1/73) (Dadonaite et al., 2019), but not for a seasonal strain isolated in Wyoming in 2003 (Fig. 12B). To investigate this observation SPLASH was performed on the Wyoming virus (Fig. 12C). The depth of the data is poor due to difficulties in growing the virus to high titres in tissue culture (i.e. achieving plaque forming units per mL (PFU/mL) of greater than 10^6). Only 16 interactions were observed, with the highest of these having only 24 unique reads mapping to it. However, it was interesting to note that the most prominent interaction observed was between the NA

and PB1 segments, but at a different site to that observed in Udorn (located in the 756-828 region of PB1 and 423-496 of NA). No interaction was observed at the Udorn NA-PB1 interaction site.

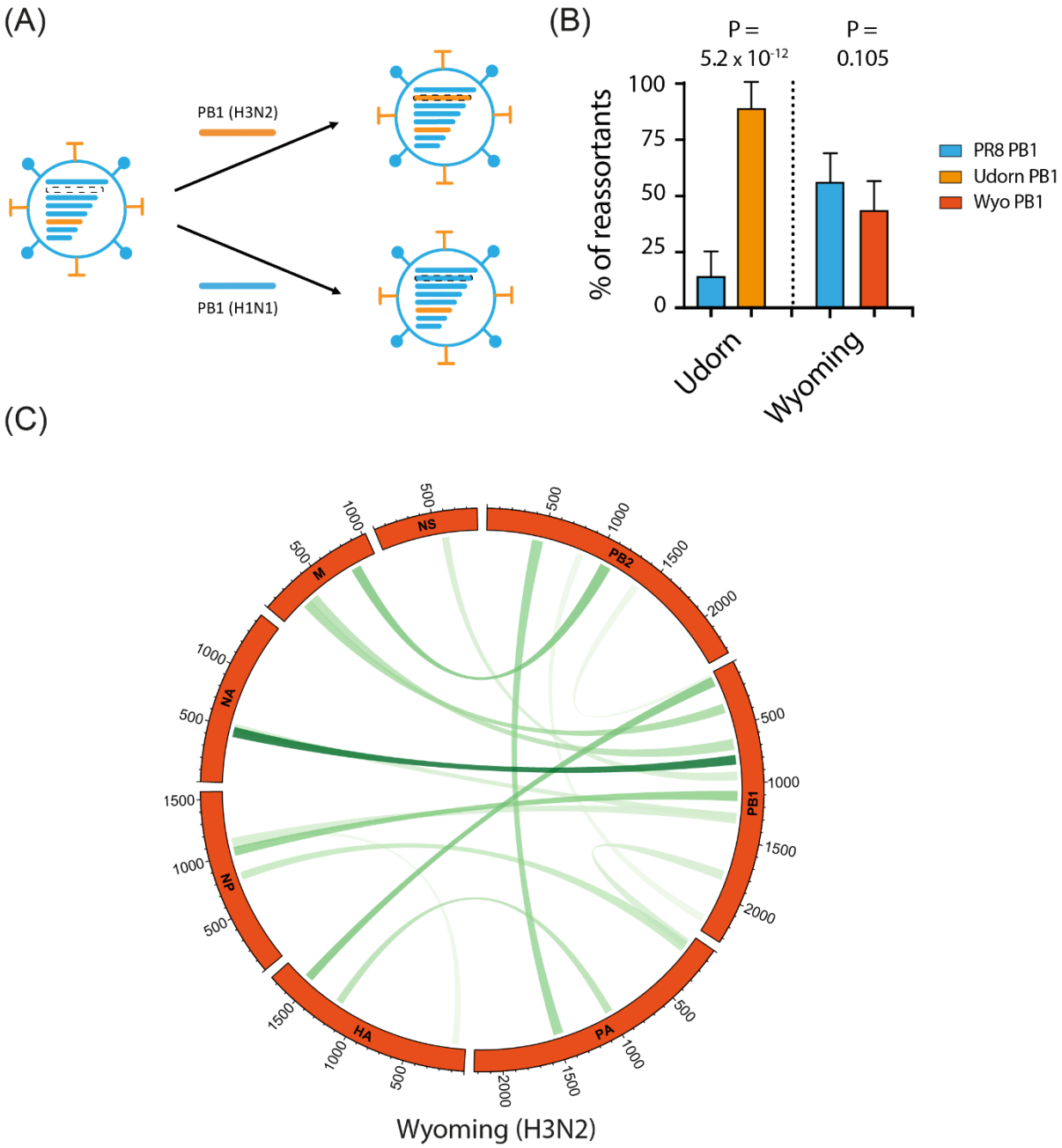


Figure 12: Reassortment in H3N2 viruses.
(A) The format of the competitive reassortment experiments performed by Brad Gilbertson (University of Melbourne). Cells are transfected with plasmids encoding the NA segment of the Udorn (H3N2) virus and the PB2, PA, HA, NP, M and NS segments of PR8 (H1N1). The cells are also transfected with both the Udorn and PR8 PB1 segments, giving the virus a

choice over which to incorporate. (B) The percentage of the progeny viruses that contained either the Udorn (orange) or PR8 (blue) PB1 segments are shown. The experiments were also performed for the Wyoming virus (red). Statistical test used was analysis of variance with Sidak correction for multiple testing. For Udorn n = 5 and P = 5.2 x 10^{-12}. For Wyoming n = 8 and P = 0.105. (C) The inter-segment RNA-RNA interaction plot for the Wyoming virus determined by SPLASH. The darker the green connecting line, the higher intensity the interaction. Very few chimeric reads were present, so data should be interpreted with caution.

The site of the NA-PB1 interaction in the Udorn virus contains four nucleotide differences in the Wyoming PB1 sequence (Fig. 13A). Only one of these differences changes the amino acid sequence, with A943G resulting in a lysine to arginine change. Based on the differences observed in the competitive reassortment experiments and the lack of an interaction observed at this site in the Wyoming SPLASH data, it was decided to further investigate the potential role of this interaction in reassortment. SPLASH was performed on a virus containing the Wyoming NA and PB1 segments in an otherwise PR8 genome (denoted PR8::Wyoming PB1 + NA) to determine if this interaction was present in a virus that did not show preferential co-segregation of these two segments (Fig. 13B). No interaction between the Wyoming NA and PB1 segments was identified at the NA-PB1 interaction site from the Udorn virus. SPLASH was then performed after four nucleotide substitutions were introduced to the Wyoming NA segment, making the region identical to that of the Udorn virus at the predicted interaction site. This mutant virus (denoted PR8::Wyoming PB1 + NA_{UdSubs}) was found to have an interaction at the NA-PB1 interaction site observed in Udorn (third most intense interaction in the dataset) (Fig. 13B). Competitive reassortment experiments (performed by Brad Gilbertson) showed that this mutant NA segment exhibits a significant preference for co-segregation with the Wyoming PB1 during reassortment (Fig. 13C). This suggests that RNA-RNA interactions between segments can influence the likelihood of them being packaged together during reassortment events.

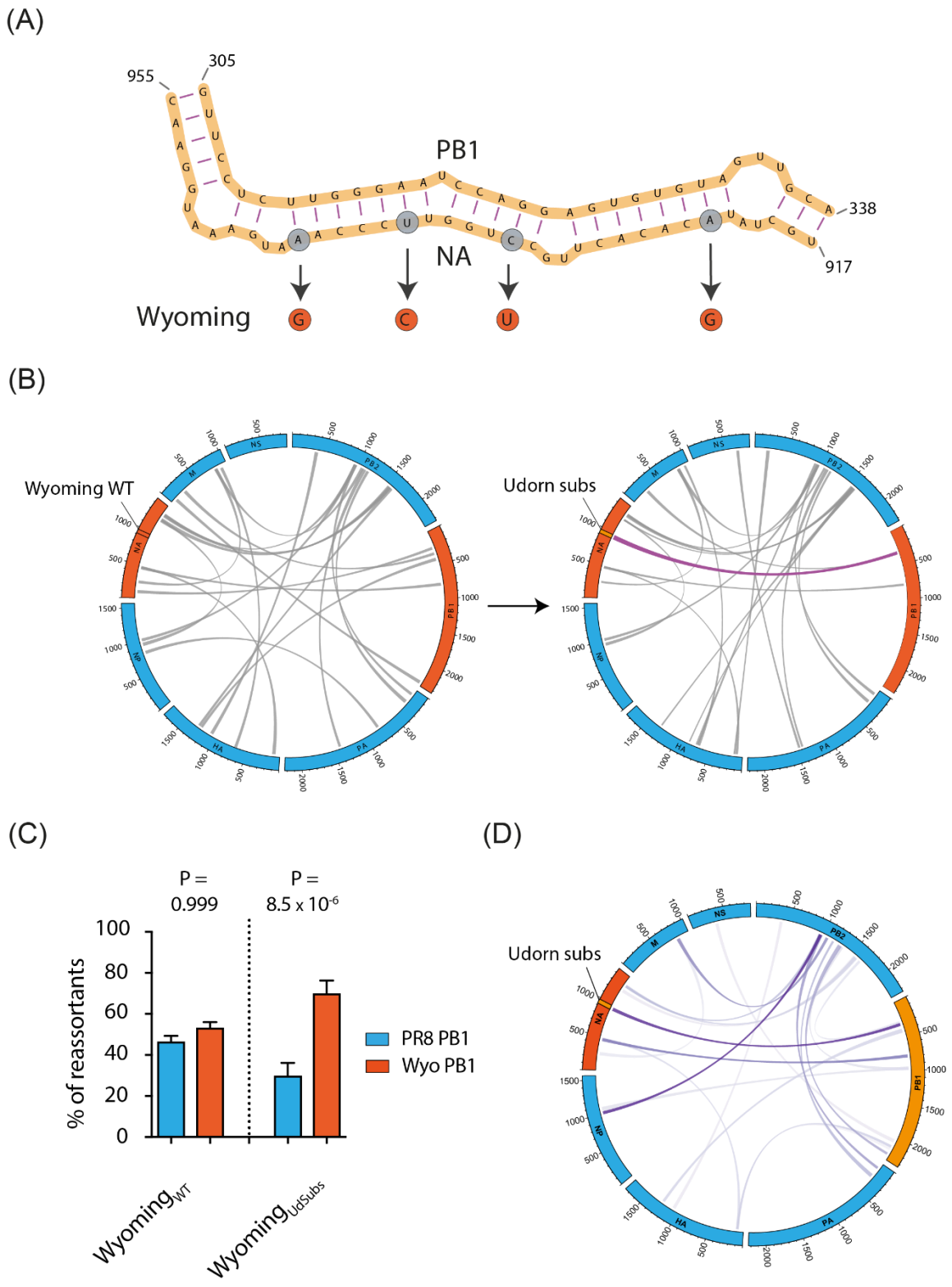


Figure 13: Inter-segment interaction influence reassortment.
(A) An interaction site identified by SPLASH between the NA and PB1 segments in the Udorn virus. This region of the Udorn virus NA segment differs from the Wyoming sequence by 4 nucleotides. (B) SPLASH was performed on the PR8::Wyoming NA + PB1 reassortant virus (left hand side). SPLASH was then performed on the same virus in which 4 nucleotide substitutions had been introduced in the region of NA identified as a NA-PB1 interaction site in the Udorn virus (PR8::Wyoming PB1 + NA_{UdSubs}). These mutations made this region

identical to that of the Udorn virus and led to the introduction of an NA-PB1 interaction at this site (shown in magenta). (C) Competitive reassortment experiments (performed by Brad Gilbertson, University of Melbourne). The Wyoming NA_{UdSubs} segment preferentially co-segregates with the Wyoming PB1 during reassortment. Statistical test used was analysis of variance with Sidak correction for multiple testing. For Wyoming n = 7 and P = 0.999. For $Wyoming_{UdSubs}$ n = 7 and P = 8.5 x 10^{-6} (D) The inter-segment RNA-RNA interaction plot for the PR8::Wyoming NA_{UdSubs}::Udorn PB1 reassortant virus. The 20 highest intensity interactions are displayed. The darker the purple connecting line, the higher the relative intensity of the interaction loci.

SPLASH was performed on a reassortant virus containing the Wyoming NA_{UdSubs} and Udorn PB1 segments (denoted PR8::Wyoming NA_{UdSubs}::Udorn PB1) (Fig. 13D). The Udorn like NA-PB1 interaction was present as the second most intense interaction. These segments also exhibited preferential co-segregation, which was not seen in the absence of the Udorn substitutions to the NA segment (Dadonaite et al., 2019). This further supports the idea that establishment of this NA-PB1 interaction influences reassortment.

Interestingly, all four viruses in which the Wyoming NA segment was present for which SPLASH was performed had a prominent interaction (at lowest the third most intense interaction) at a different site between the NA and PB1 segments (770-812 on PB1 and 443-473 on NA), not seen in the Udorn virus. Despite this, even in the reassortants where both NA-PB1 interactions were present, the NA-PB1 co-segregation was lower than for the two Udorn segments. This suggests that reassortment propensity of segments is complicated and will be influenced by the interactions networks as a whole and not just the interactions connecting two segments.

3.4 Discussion

This project sought to investigate the structure of the influenza virus genome and its role in reassortment. Extensive, redundant networks of interactions were found to exist between the influenza genome segments in virions. These interactions are likely involved in the bundling

of the segments prior to packaging. This is supported by the observation that in FISH experiments bundles of vRNPs in the cytoplasm rarely contain multiple copies of the same segment (Haralampiev et al., 2020) and the data in this study indicating that the interactions can influence the likelihood of segments co-segregating during reassortment.

When comparing a reassortant virus to its parent strains, ~80% of the interactions present in the offspring were present in one of the parental strains, though many had changed in prominence in the reassortant. This may reflect differences in the predominant way in which the virus is assembling its genome and provides the first illustration of how flexibility in inter-segment interaction networks can be utilised to accommodate reassortment. The appearance of interactions not present in either parent suggests that this flexibility extends beyond the interaction sites seen in the parents. Redundancy in the inter-segment interaction networks is likely to be advantageous to influenza in accommodating the genetic drift (from the error prone polymerase) and shifts (due to reassortment) that are necessary for influenza to evade populational immunity, whilst maintaining the ability to assemble its genome. During natural infection influenza exists as a quasispecies (Barbezange et al., 2018) and not relying upon a rigid, defined set of interactions, may also help to maintain this diversity. If small changes in sequence led to dramatic drops the ability of a segment to bundle, mutant segments could be easily outcompeted during bundling.

Studies using FISH indicate that bundles in the cytoplasm containing fewer than 8 segments can contain different combinations of segments, but that some compositions are favoured more than others (Lakdawala et al., 2014) (Haralampiev et al., 2020). This and the varied genome organisations observed in electron tomography reconstructions of virions (Noda et al., 2012) (Fournier et al., 2012), supports the flexible interaction networks observed in this study and suggests that bundling of the segments does not take place in a completely regimented order. It was noted that the PB2 and PB1 segments were more frequently observed in bundles containing fewer segments, suggesting that they may have a more central role in organising the bundling process (Haralampiev et al., 2020). This is supported by the

SPLASH data presented in this investigation showing that a disproportionately high proportion of reads map to the polymerase segments. For the PR8 virus 74.8% of chimeric reads map to the PB2 or PB1 segments. It is further supported by electron tomography data suggesting that the vRNPs are organised with seven of the segments surrounding one central segment, with the central segment determined to be one of the longer segments (Noda et al., 2012) (Fournier et al., 2012).

Le Sage and colleagues recently applied a sequencing approach, similar to SPLASH, called Dual Crosslinking, Immunoprecipitation, and Proximity Ligation (2CIMPL) to investigate inter-segment interactions (Le Sage et al., 2020). They performed 2CIMPL on WSN virus and identified extensive (300< loci), redundant interaction networks not limited to terminal regions of the segments, supporting the findings of this investigation. However, they reported only a 10% overlap with the interactions identified by SPLASH (Dadonaite et al., 2019). The 2CIMPL technique includes an extra UV exposure to crosslink the NP to the vRNA. Immunoprecipitation of NP is then used to capture the RNase digested strands before proximity ligation and library preparation. The authors argue that this allows proximity ligation to be performed under more native conditions than is the case for SPLASH. However, it could be argued that 2CIMPL biases itself towards the identification of interactions that are in proximity to regions of high NP binding (NP is not evenly distributed on the RNA (Williams et al., 2018b) (Lee et al., 2017). The authors report a significant correlation between regions with 2CIMPL identified inter-segment RNA-RNA interactions and regions of high NP binding, though this association also held true for the SPLASH WSN dataset.

The formation of complex redundant networks of interactions has also been supported by Ligation of Interacting RNA followed by high-throughput sequencing (LIGR-seq) (Takizawa et al., 2020). The LIGR-seq dataset showed 15% overlap with the combined SPLASH dataset and 27.5% of the identified interactions appeared in at least one replicate. Ultimately, all high-throughput sequencing techniques may have their own biases and the overlapping interactions identified by different techniques may provide the best indication of the ‘true’

interactions. Ideally techniques such as SPLASH, 2CIMPL, LIGR-seq (Sharma et al., 2016), PARIS (Lu et al., 2016), and COMRADES (Ziv et al., 2018) would all be performed for several strains to better understand the reproducibility between techniques. In addition, more work is needed to assess the functional implications of interactions identified, regardless of the technique used to identify them.

On two occasions in the last century (1957 and 1968) HA and N2 subtype NA segments of avian origin have been accompanied by avian PB1 segments in the generation of pandemic viruses (Taubenberger and Kash, 2010). The PB1 segment is also the most commonly found non-antigenic segment from seasonal strains in vaccine seed viruses, appearing in 45% of H3N2 and 63% of H1N1 seed viruses in one analysis (Cobbin et al., 2013). In addition, the NA segment has been shown to have a strong co-evolutionary relationship with PB1 (as well as PA, PB2, and NP) in an analysis of H3N2 viruses between 2005 and 2014 (Jones et al., 2021). It is not obvious why this relationship would be advantageous to the virus as the two proteins have seemingly unconnected functions. As such, it is possible that the relationship between the two segments in H3N2 viruses is down to maintenance of the RNA-RNA interactions between the segments, reported in this and other studies (Gilbertson et al., 2016).

From an evolutionary perspective influenza is interesting in that each individual genome segment is under selective pressure (rather than the genome as a whole). It could be speculated that the maintenance of an interaction between two segments could favour the continued existence of both segments during reassortment events. However, co-evolutionary dependencies have also been observed for NA and PB1 in H1N1 viruses (Jones et al., 2021). No high intensity NA-PB1 interactions were observed in the PR8 (H1N1) or WSN (H1N1) SPLASH datasets (highest intensity NA-PB1 interactions in datasets were 48th and 68th highest respectively). Whilst this may not be representative of more modern H1N1 viruses, it is possible that other factors may be responsible for the NA-PB1 relationship. For example, there is evidence that reassortant strains possessing the PB1 segment, as well as the NA and HA segments from a H3N2 virus have higher HA density on the surface of their viral particles

(Moulès et al., 2011) (Cobbin et al., 2013). Ultimately, attribution of sequence changes in seasonal strains to inter-segment interactions is extremely challenging as immune pressure, epistatic relationships between influenza proteins, drug resistance, intra-segment RNA structures, species adaptations, and random chance can all contribute.

The reassortant viruses used in this investigation had not been passaged prior to SPLASH. The rate of accumulation of sequence changes has been shown to be higher for viruses that have recently undergone reassortment (Neverov et al., 2014). It is possible that the reassortant viruses used in this investigation would accumulate mutations over time to better accommodate the new segments (with potentially the proportion and prominence of the inherited interactions decreasing). However, protein compatibility of the viral proteins, in particular between the Udorn PB1 and PR8 PB2 and PA, would also be likely to place a strong selection pressure on the virus (Li et al., 2008a) (Octaviani et al., 2011). This, along with adaptions to tissue culture (DuPai et al., 2019) and reports of NA and HA being under co-evolutionary pressure (Kryazhimskiy et al., 2011) (Jang and Bae, 2018), would make it hard to attribute sequence changes to importance at an RNA or protein level.

This investigation identified hundreds of putative interaction loci in influenza and demonstrated that one of these interactions can influence reassortment. However, it is not clear which of the interactions identified are of greatest importance. It is possible that lower intensity interactions may not be relevant to genome assembly. They may form as an accidental consequence of inbuilt redundancy in the networks or simply due to the close proximity of regions in the virion brought about by other interactions. Interactions were ranked by the number of reads mapping to them relative to the other interactions in that data set. However, this is not likely to be a precise measure of the importance of an interaction. Firstly, sequence bias when using psoralen as a crosslinking reagent could lead to over or under representation of an interaction (Cimino et al., 1985). Secondly, if a particularly strong interaction brings two segments into close proximity then other interactions in the vicinity may be stabilised, increasing their read count in the sequencing data, regardless of their importance to the assembly process. Finally,

some interactions may be important in the process of assembly, but become less predominant in the final assembled genome (or vice versa). This could potentially be investigated by determining the RNA-RNA interaction networks of vRNPs in the cytoplasmic or nuclear compartments of infected cells. However, due to the much higher prevalence of higher order bundles (i.e. bundles with eight or close to eight segments) (Lakdawala et al., 2014) (Haralampiev et al., 2020) seen in infected cells, this may not yield much information unless bundles of different order could be separated from one another. This could be performed by adding a glycerol gradient purification step post crosslinking, but before RNA extraction. It may also be interesting to investigate the interactions in cells at different time points. Defective particle formation is seen to increase at later stages of infection (Frensing et al., 2016), and this may be reflected in the interactions formed and their prominence.

Assisting with the prediction of pandemic emergence has been touted as a potential application of studying inter-segment interactions (Dadonaite et al., 2019) (Jones et al., 2021). The high complexity of the networks and inbuilt flexibility is likely to make this challenging. It is difficult from SPLASH data alone to establish which interactions are going to be most influential in reassortment. More functional studies and investigations into the role of different interactions in reassortment events will be required. It is also possible, given the apparent high level of redundancy seen in the interaction networks in this and other studies (Le Sage et al., 2020), that inter-segment interactions play a secondary role in the likelihood of pandemic emergence to factors such as protein compatibility, populational immunity, and host adaptations.

A more short term potential application of understanding RNA-RNA interaction networks is in vaccine production. The majority of influenza vaccine is produced in chicken eggs and this requires the production of reassortant viruses containing the HA and NA segments from seasonal strains with segments from an egg adapted strain. As previously noted, this frequently results in the production of vaccine seed candidate viruses containing additional non-antigenic segments from the seasonal strain (Bergeron et al., 2010) (Fulvini et al., 2011)

(Cobbin et al., 2013), even in instances where this results in reduced replicative efficiency (Cobbin et al., 2014). By analysing the inter-segment interactions it may be possible to introduce silent mutations that will introduce new interactions between the egg adapted segments and those from the seasonal strain, in order to improve the efficiency of vaccine seed strain generation. This may even allow for increased yields if the egg adapted segments and seasonal strain segments are poorly matched and do not bundle efficiently. This may also be relevant for live attenuated vaccines. In the 2013-14 and 2015-16 seasons the FluMist vaccine (a live attenuated vaccine) gave low protection against H1N1 viruses. This was attributed to low replicative efficiency (Hawksworth et al., 2020). The FluMist H1N1 strain was also shown to produce a relative high number of non-infectious particles (Gould et al., 2017) and to be outcompeted by the other strains present in tri or quadrivalent formulations (Dibben et al., 2021). It is possible that inter-segment interactions may have contributed in this instance and that analysing their impact may be of benefit in future vaccines.

4. Intra-segment RNA-RNA interactions

4.1 Chapter summary

- A chemical probing technique called SHAPE is used to aid in the prediction of intra-segment RNA structures in influenza virions.
- Analysis of H1N1 and H3N2 viruses reveals that few RNA structures are conserved between the two sub-types.
- SHAPE analysis is extended to reassortant viruses and reveals that the intra-segment RNA structures formed by a segment are largely independent of the other segments present in the virion.
- Different chemical probes (DMS and EDC) were investigated to try to validate the existence of the structures predicted using SHAPE reagents.
- The results suggest EDC, which had not previously been established for use in mutational mapping experiments, can provide information about the pairing probability of G and U nucleobases. This has the potential to complement the well-established probing of A and C nucleobases by DMS.
- The intra-segment RNA structures presented provide targets to investigate for potential functional roles.

4.2 Introduction

In chapter 3 SPLASH was employed to investigate inter-segment RNA structure in influenza viruses. However, this technique is not effective for identifying the short range intra-segment interactions which have been predicted to occur in the influenza RNAs (Gultyaev et al., 2014) (Kobayashi et al., 2016) (Simon et al., 2019). The use of chemical probing techniques to improve the accuracy of RNA structure prediction is well-established and these techniques do not suffer from the same minimum distance limitations as SPLASH (Siegfried et al., 2014). Recently the chemical probing technique SHAPE was applied to influenza virions for the first time, guiding modelling of intra-segment RNA structure across the entire influenza genome

(Dadonaite et al., 2019). The study found a large number of structures across all of the genomic segments.

SHAPE reagents can be added to influenza virions and will preferentially acylate single-stranded RNA (Fig. 14). Upon reverse transcription acylated nucleobases in the RNA are likely to result in errors being introduced into the resulting cDNA. The reads are sequenced and the location of these reverse transcription errors is recorded. The mutation rate of each nucleobase can then be determined, with this corresponding with its likelihood of being unpaired (higher mutation rate = higher likelihood of being unpaired). The SHAPE reactivity data can then be used to compliment RNA structure prediction algorithms (for more information see section 1.6).

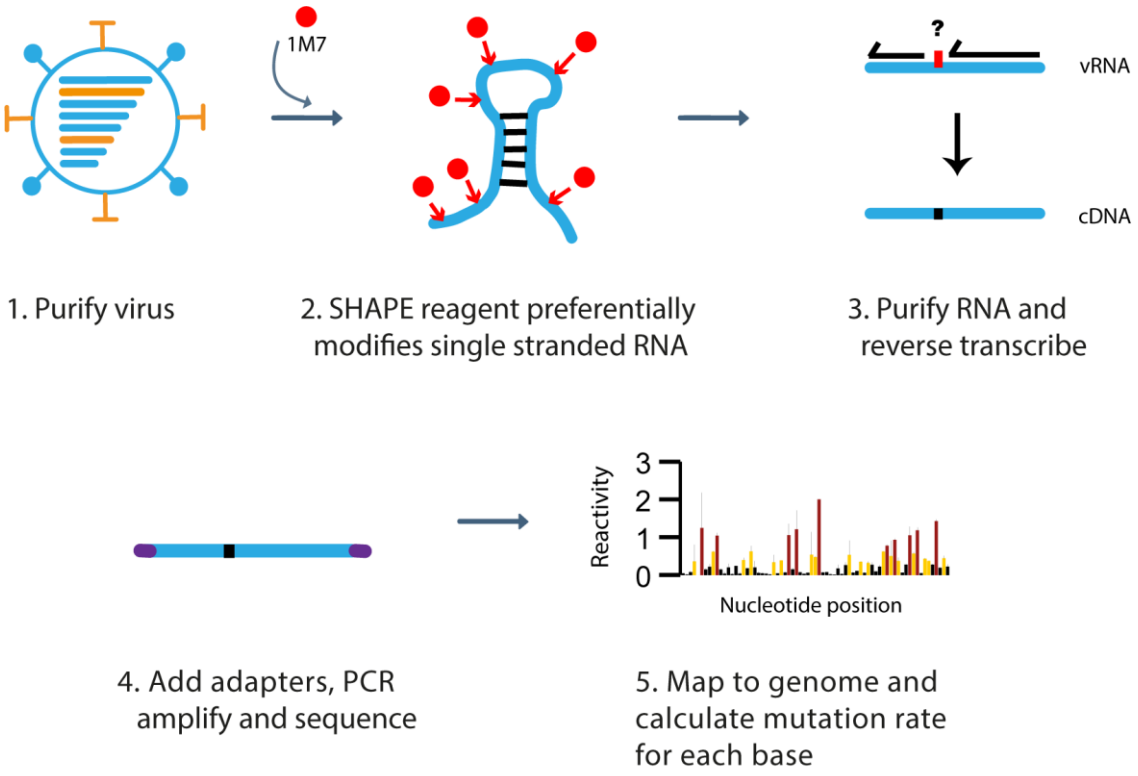


Figure 14: The process of performing SHAPE in influenza virions
The SHAPE reagent 1M7 is added to purified virions. It will preferentially modify single stranded RNA. RNA is then extracted and reverse transcribed. Errors are introduced when the reverse transcriptase hits modified bases. A sequencing library is prepared and reads are mapped back to the reference genome. The site of reverse transcription mutations are

recorded and counted for each base relative to a non-modified control sample. This is used to calculate the relative reactivity rate of each base (higher reactivity rate = higher likelihood of being unpaired).

The aim of this chapter was to identify the intra-segment RNA structures present in a range of IAVs to investigate RNA structure conservation. This included probing reassortant viruses to investigate whether intra-segment RNA structure is affected by the other segments present in a virion. Additional chemical probing techniques were also explored in order to try to validate the RNA structures predicted by SHAPE.

4.3 Results

4.3.1 SHAPE of PR8

SHAPE was first performed on the PR8 virus. Virions were purified through a sucrose cushion. Samples were split in two, with half treated with the chemical probing reagent 1M7 and half with DMSO (referred to henceforth as untreated). Libraries were then generated and sequenced. Median read depths were in excess of 37,000 for each of the two replicates for both the treated and untreated samples (Fig. 15A, Table. S1). The median mutation rates of the 1M7 treated samples were ~3 times higher than in the untreated samples (Table. S1). The reactivity rates for each base were calculated using the Siegfried method, in which the raw reactivity rate of each base is calculated from its mutation rate in the 1M7 treated sample, relative to its mutation rate in the untreated sample (Siegfried et al., 2014). The raw reactivity data was then normalised by taking the average of the top 90-98% of reactivity values and dividing all reactivity values by this number. The finalised SHAPE reactivity rates showed good reproducibility with a Spearman correlation R value of 0.91 for the two replicates. The data from the two replicates were combined to produce reactivity profiles for each segment (Fig. 15B, Fig. 16). Regions of low reactivity can be seen across all segments indicating the

presence of paired bases. Tables containing full information on SHAPE reactivity values and pairing probabilities for all viruses probed in this investigation can be found online at https://figshare.com/s/6444d82a7bab5f8cbb74.

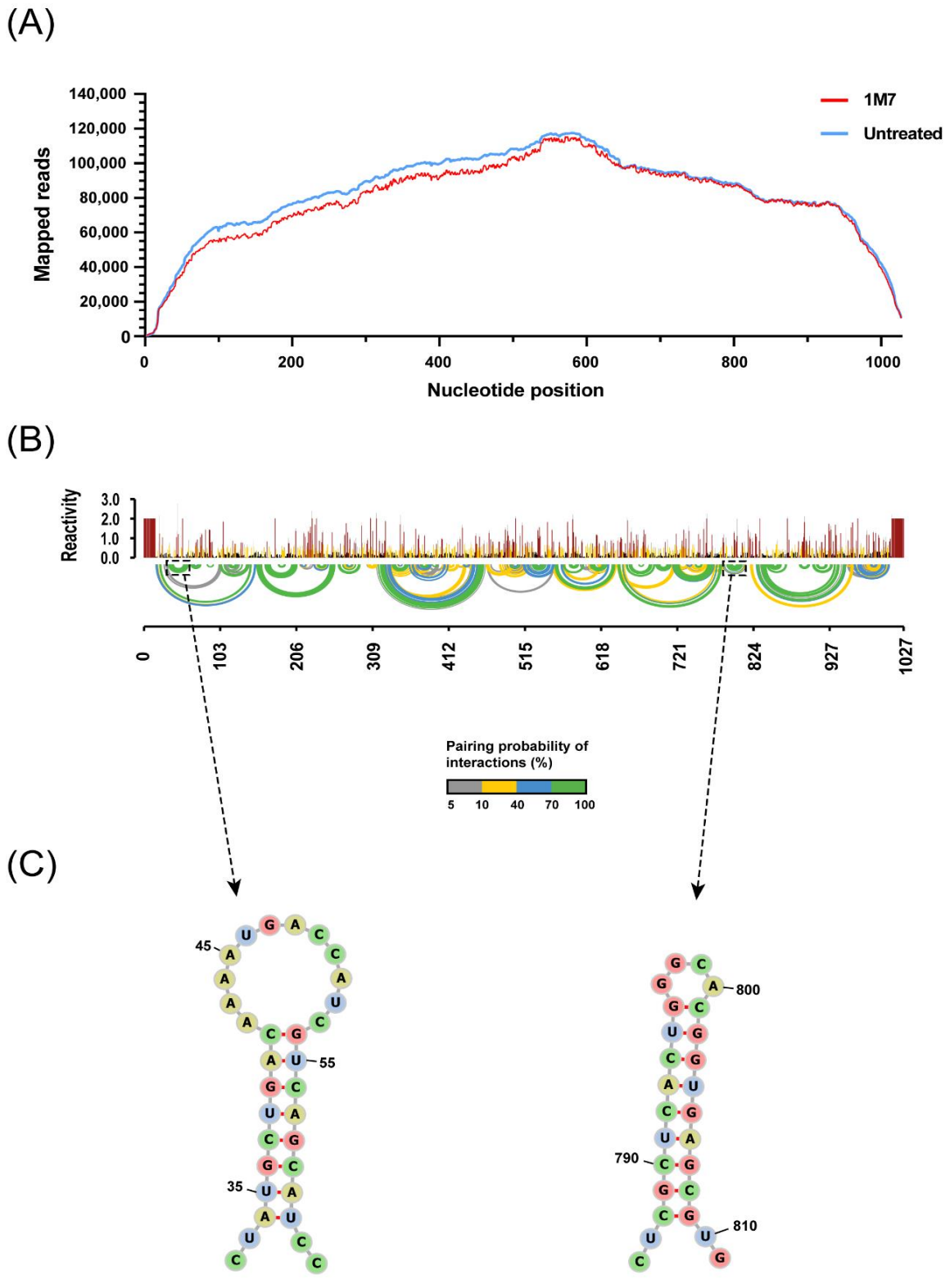


Figure 15: The intra-segment RNA structure of the PR8 M segment
(A) The sequencing depth across the M segment for 1M7 and untreated PR8 SHAPE samples. Data are for two replicates combined. (B) The SHAPE reactivity profile for the M segment of PR8. The arcs indicate the predicted interactions. (C) The structures of two hairpins in the M segment the existence of which had previously been supported by evolutionary analysis (Kobayashi et al., 2016).

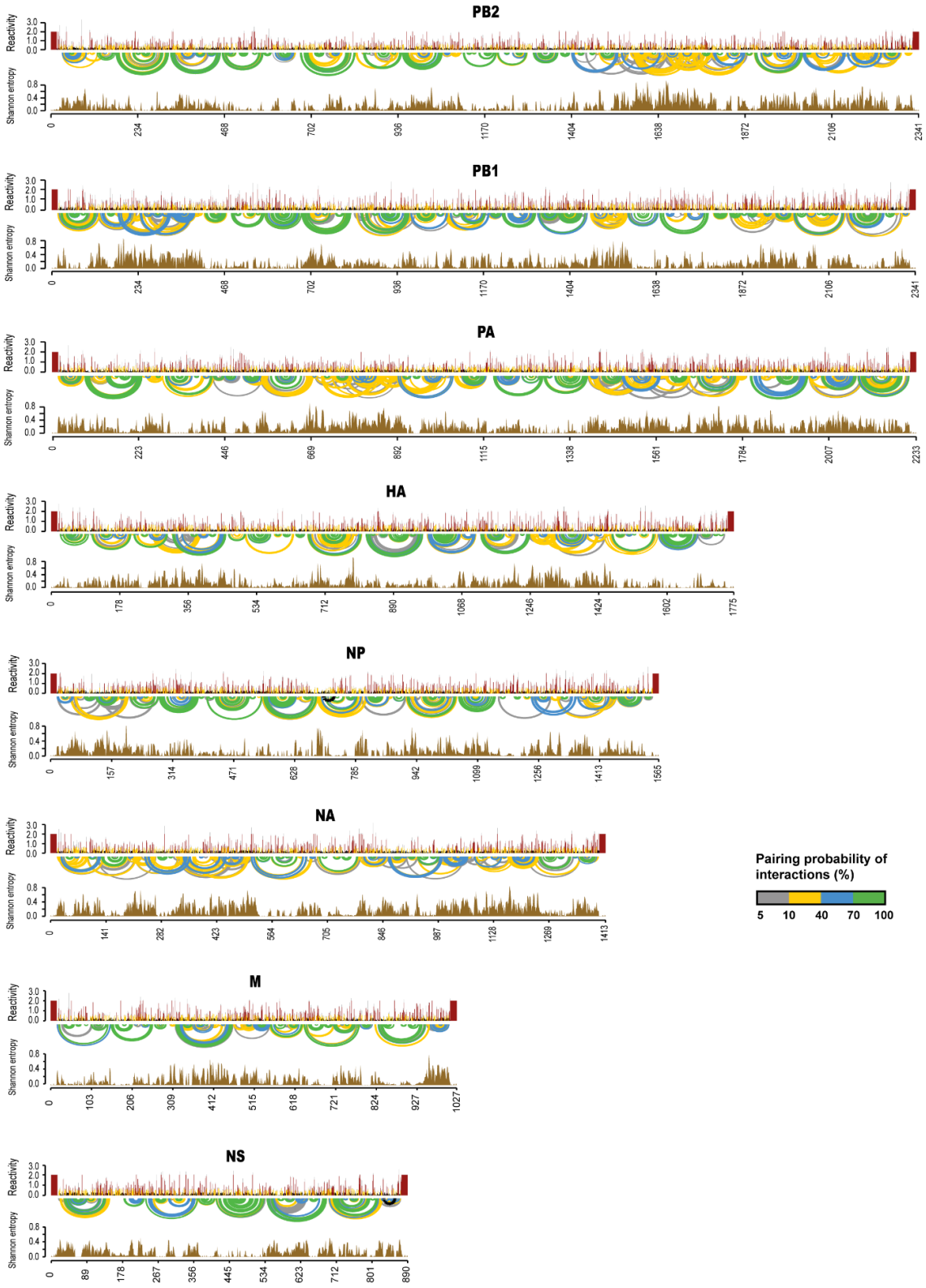


Figure 16: The intra-segment RNA structure of the PR8 virus.
The SHAPE reactivity profiles and predicted interactions for the PR8 vRNAs. Pseudoknots are shown with black lines. Shannon entropy values are also displayed (high values indicating regions that are likely to form one or more alternative structures).

SHAPE reactivity values were used as soft constraints to guide RNA structure prediction. The first and last 16 bases for each segment were excluded from pairing. These regions of the influenza genome are partially complimentary and are known pair to one another, making them unlikely to be involved in other RNA structures. In addition, the 17th nucleobase was the first that had above 5,000 reads mapping to it in both replicates (5,000 reads is seen as benchmark for achieving accurate predictions (Busan et al., 2019)). A maximum pairing distance of 150 nucleotides was imposed.

A large number of intra-segment interactions are predicted to occur spanning the length of all 8 vRNAs (Fig. 15B, Fig.16). This includes two hairpins in the M segment (34-61 and 788-809) (Fig. 15A) the existence of which had previously been supported by evolutionary and mutational analysis (Kobayashi et al., 2016). All of the segments in this study contain multiple regions of high Shannon entropy (Fig. 16) (see section 1.6 for explanation of Shannon entropy). This indicates that these regions are likely to form alternative structures, suggesting that there are many regions of the influenza genome that do not form single stable secondary structures.

A previously predicted pseudoknot in the NP segment (Gultyaev et al., 2014) was not predicted to form in the analysis performed in this investigation. However, a different pseudoknot was predicted to form NP (689-823). A further pseudoknot was predicted in the NS segment (816-873) (Fig. 16, Fig. S1). The predicted NP pseudoknot has low pairing probabilities (10-40%) in the lower part of its stem. Likewise the NS pseudoknot occurs in a region of considerable paring uncertainty, with no base pairs predicted with greater than 70% pairing probability. Neither structure was recapitulated when max pairing distances of 250 or 300 nucleotides were used (data not shown). Overall, the likelihood of these pseudoknots existing seems low.

A previous study also performed structure probing on PR8, though data was processed using different software and with slightly different folding parameters (Dadonaite et al., 2019). Of the 282 interactions with greater than 80% pairing probability predicted in that study, 58.2% were

predicted to have above 80% pairing probability in this study, 14.5% were predicted to have 30-80% pairing probability, and 27.3% had paring probability below 30% (Fig. S2).

A previous study has investigated the distribution of NP across the PR8 genome using Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (Williams et al., 2018b). There was no significant difference in SHAPE reactivity between high NP binding regions, low NP binding regions, or all bases (Kruskal-Wallis test P = 0.972) (Fig. 17). This suggests that NP binding (or lack thereof) does not affect the likelihood of a region being involved in intra-segment RNA structures.

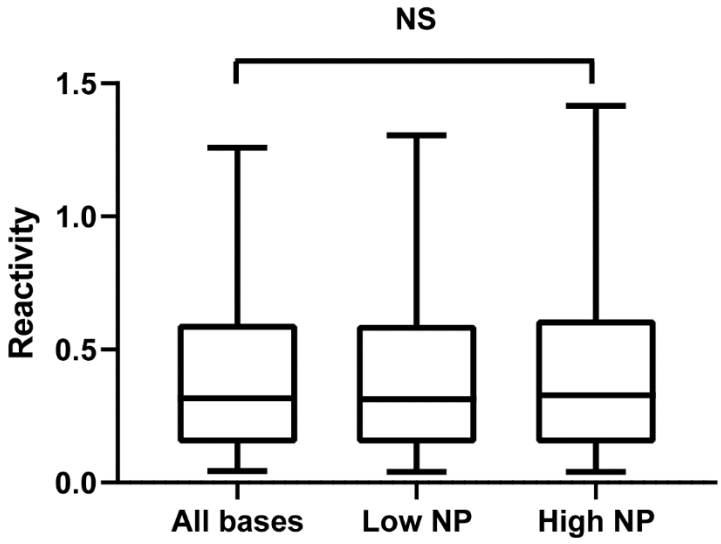


Figure 17: SHAPE reactivity does not correlate with NP distribution.
Comparing the SHAPE reactivity values of regions found to exhibit low or high NP binding in PAR-CLIP experiments (Williams et al., 2018b). In each plot the middle line is the median, the edges of the box show the lower and upper quartiles, and the whiskers show the 5th and 95th percentiles. Kruskal-Wallis test showed no significant difference between the three groups P=0.972. High NP = 309 bases, Low NP = 1,172 bases, All bases = 13,275.

4.3.2 RNA structure in reassortant viruses

SHAPE was then performed on the WSN (H1N1), Wyoming (H3N2) and Udorn (H3N2) viruses. An additional, 3 reassortant viruses were also included: PR8::Udorn PB1 + NA, PR8::Wyoming PB1 + NA, and PR8::Wyoming PB1+ NA_{UdSubs} (Fig. 18-25) (the same viruses

were subject to SPLASH in chapter 3). The similarity of the SHAPE reactivity profiles for segments from different viruses was measured by Spearman R correlation (Fig. 18A-25A). Similarity in SHAPE reactivity profiles between segments correlated strongly with their sequence identity (Spearman R value of 0.97 ($P<0.0001$)). For example, the Spearman R correlation values between the H1N1 PR8 and WSN segments ranged from 0.65 (for HA) to 0.76 (for PB2). Whilst between PR8 and the more distantly related Udorn (H3N2), values ranged between 0.14 (for HA) to 0.75 (for M).

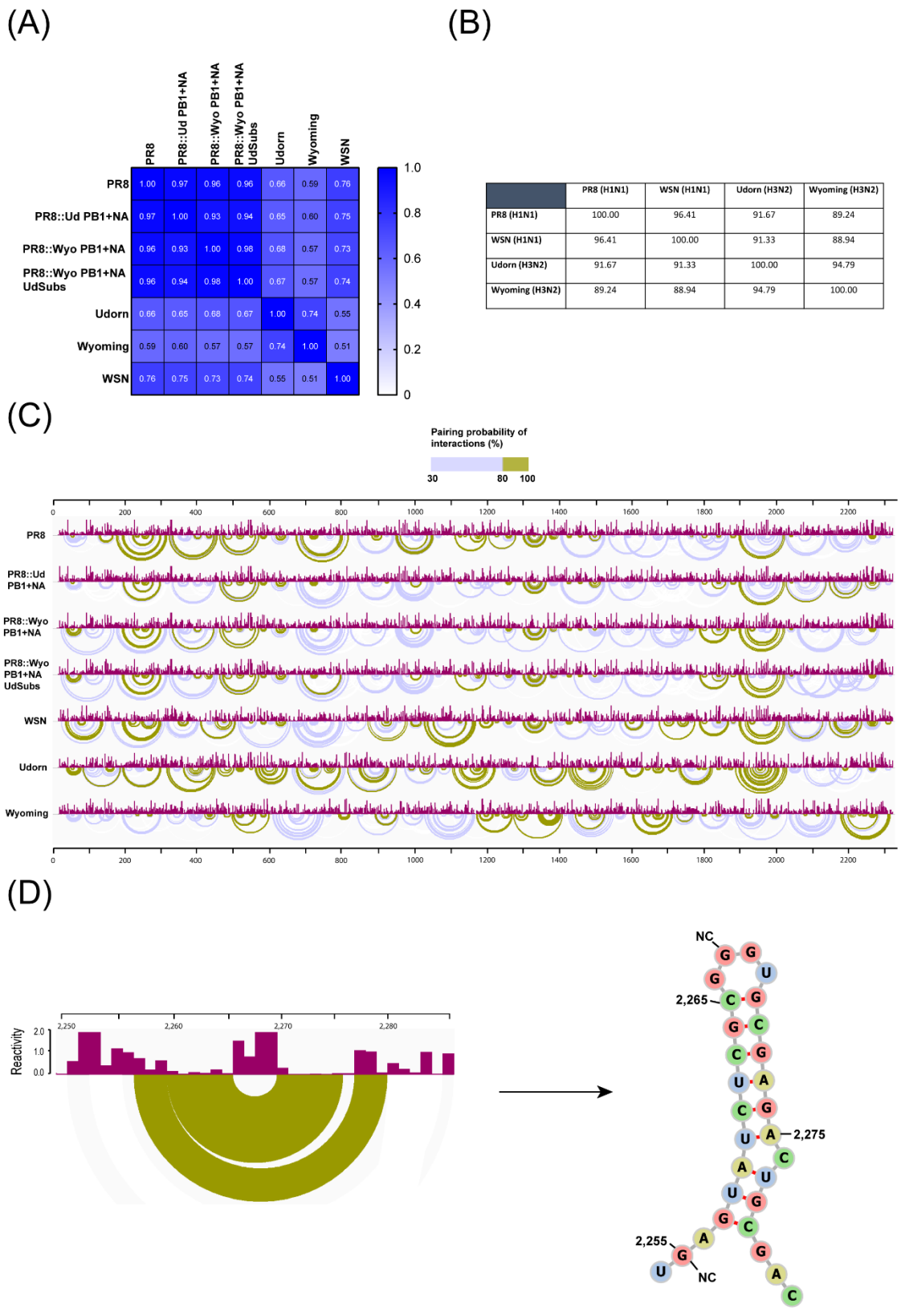


Figure 18: Comparison of PB2 intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the PB2 segments. (B) The sequence identities between the PB2 vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the PB2 segments. (D) Reactivity profile and structure of a hairpin predicted to be conserved amongst the viruses tested. NC indicates that this residue is not conserved amongst the 4 virus sequences investigated.

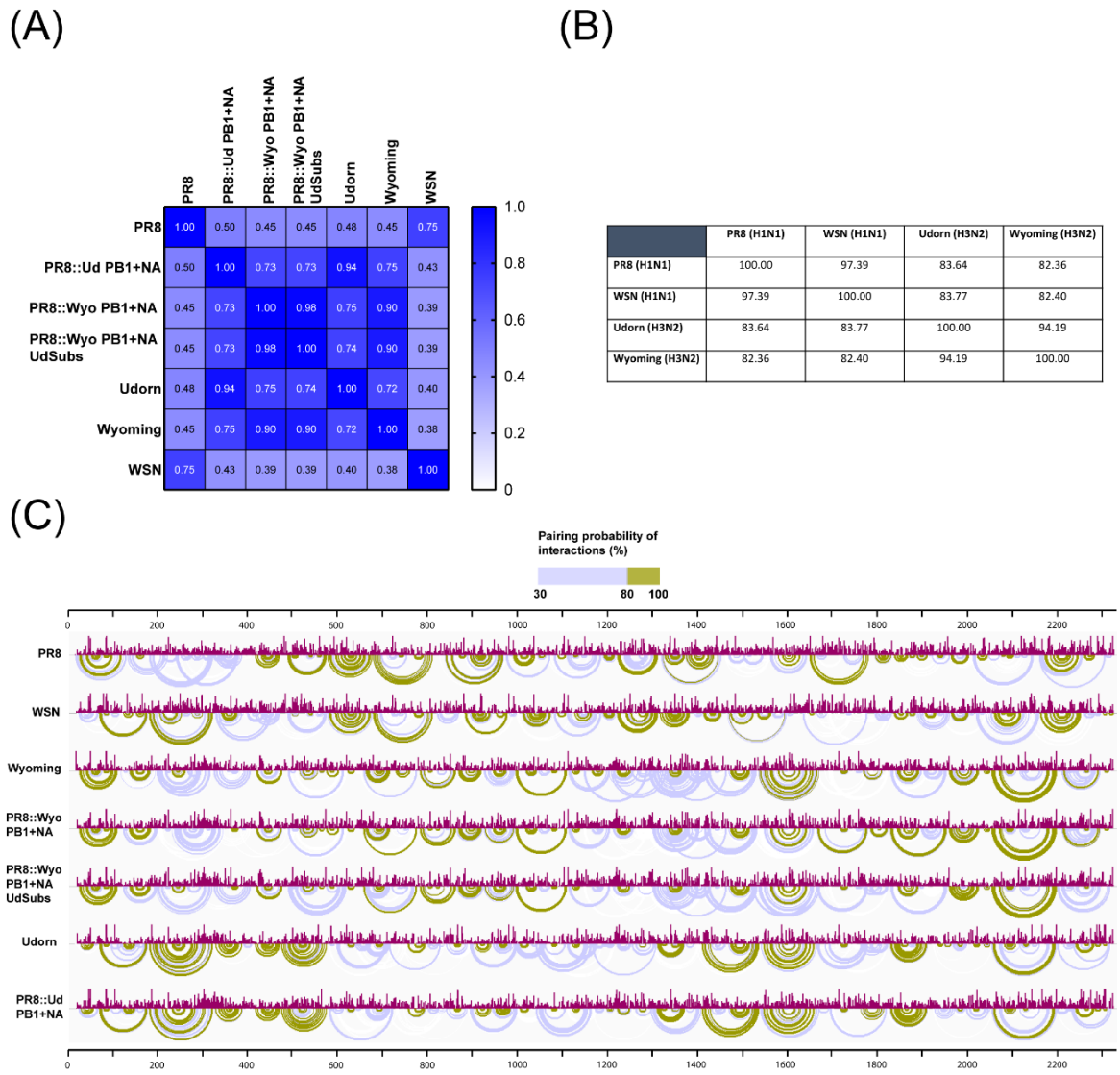
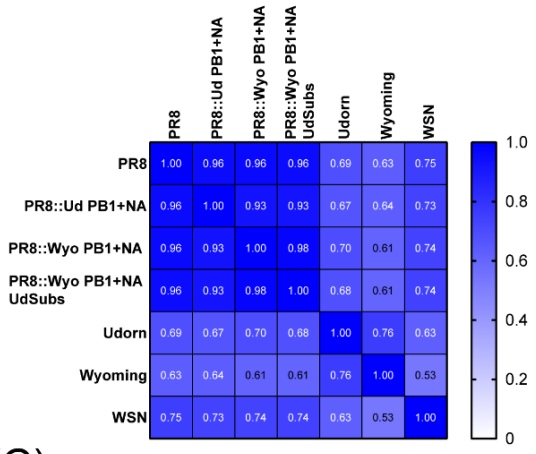


Figure 19: Comparison of PB1 intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the PB1 segments. (B) The sequence identities between the PB1 vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the PB1 segments.

(A)



(B)

	PR8 (H1N1)	WSN (H1N1)	Udorn (H3N2)	Wyoming (H3N2)
PR8 (H1N1)	100.00	97.13	92.66	89.79
WSN (H1N1)	97.13	100.00	92.92	89.70
Udorn (H3N2)	92.66	92.92	100.00	94.89
Wyoming (H3N2)	89.79	89.70	94.89	100.00

(C)

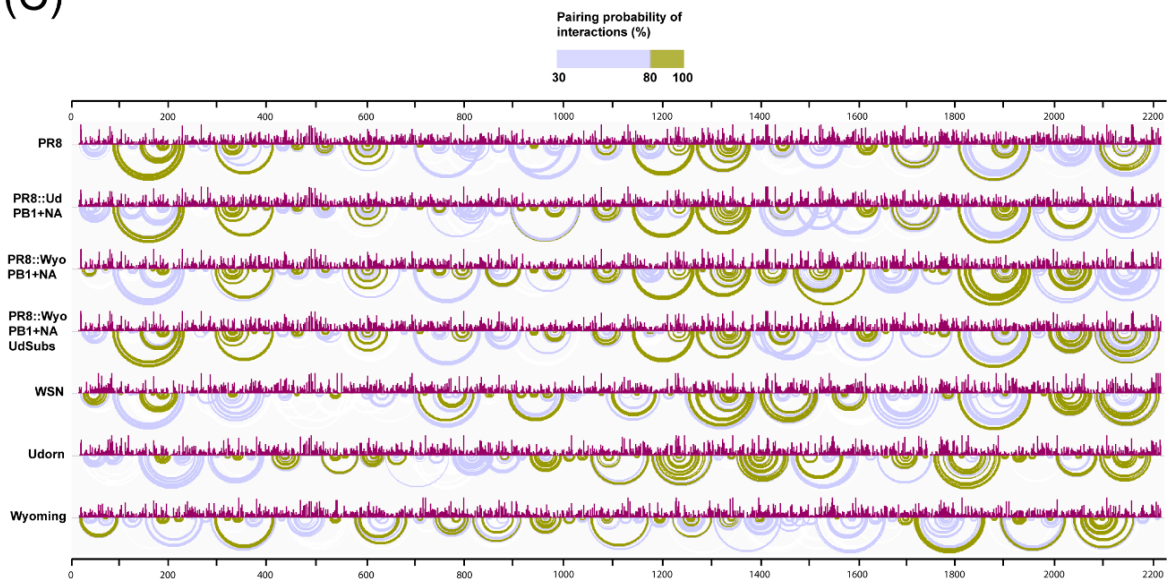


Figure 20: Comparison of PA intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the PA segments. (B) The sequence identities between the PA vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the PA segments.

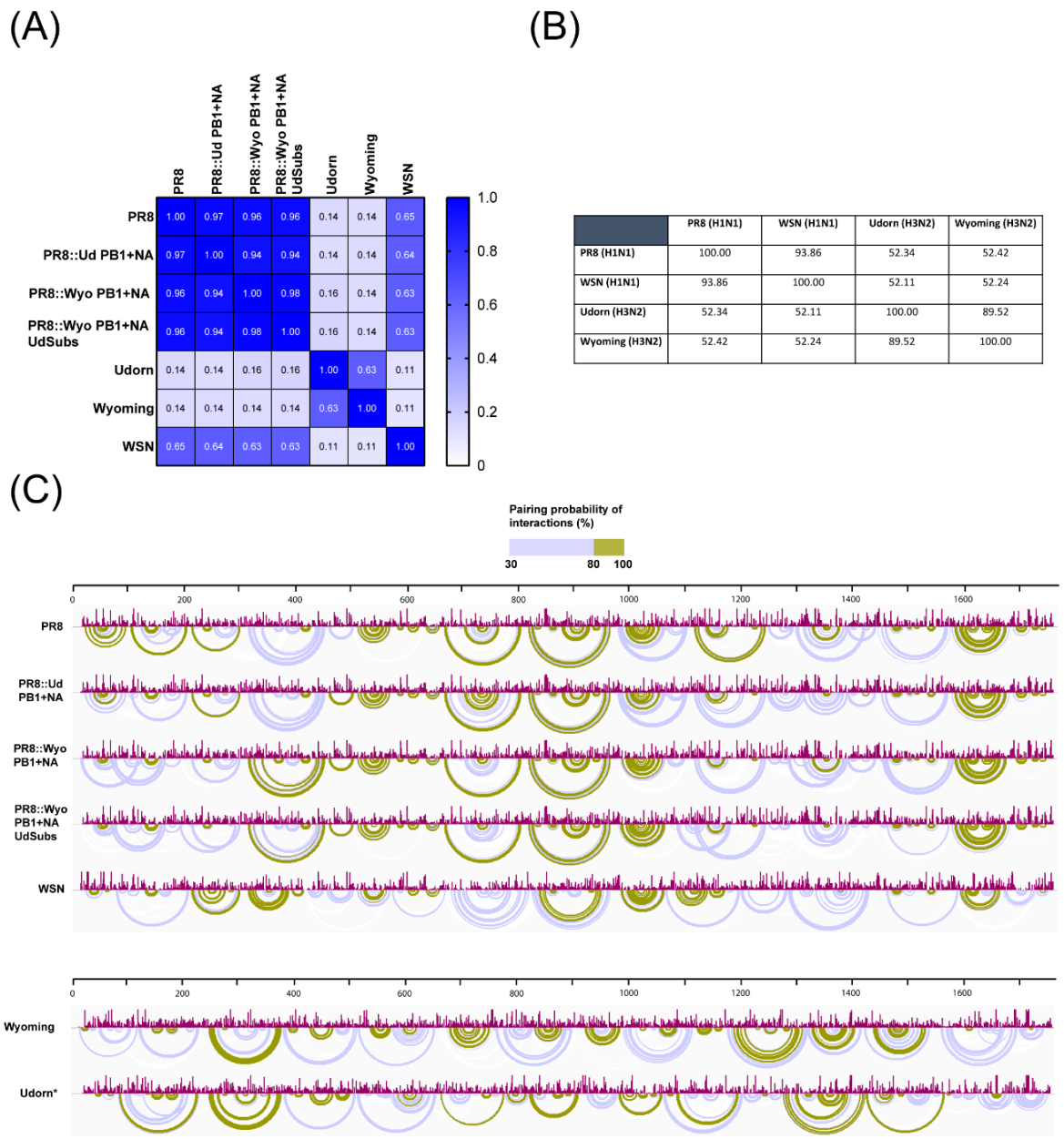
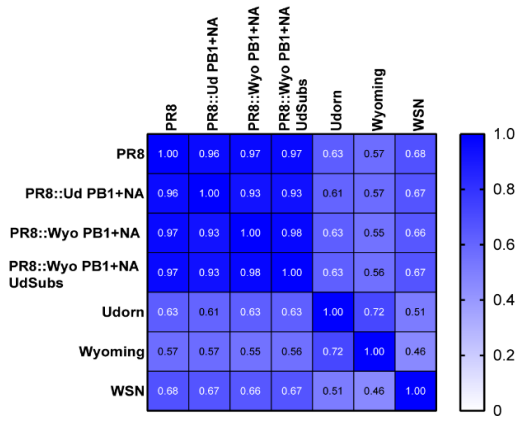


Figure 21: Comparison of HA intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the HA segments. (B) The sequence identities between the HA vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the HA segments. ***Udorn** reactivity profile is shifted by 5 nucleotides to bring it into sequence alignment with Wyoming (Wyoming has 5 extra nucleotides that do not align at positions 21-26).*

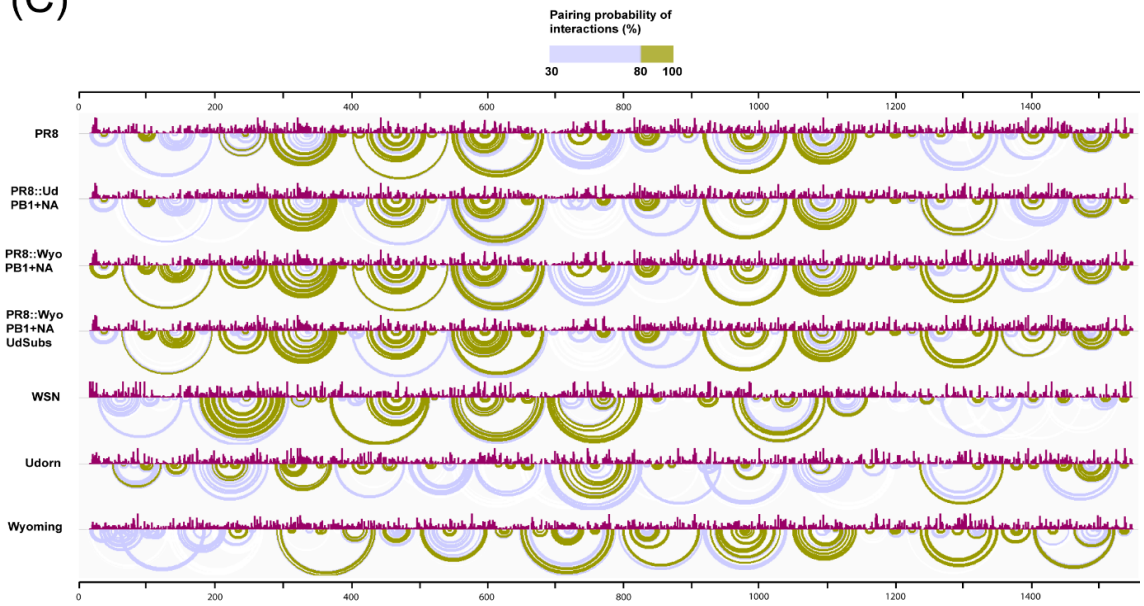
(A)



(B)

	PR8 (H1N1)	WSN (H1N1)	Udorn (H3N2)	Wyoming (H3N2)
PR8 (H1N1)	100.00	96.04	91.18	88.43
WSN (H1N1)	96.04	100.00	91.31	88.56
Udorn (H3N2)	91.18	91.31	100.00	94.70
Wyoming (H3N2)	88.43	88.56	94.70	100.00

(C)



(D)

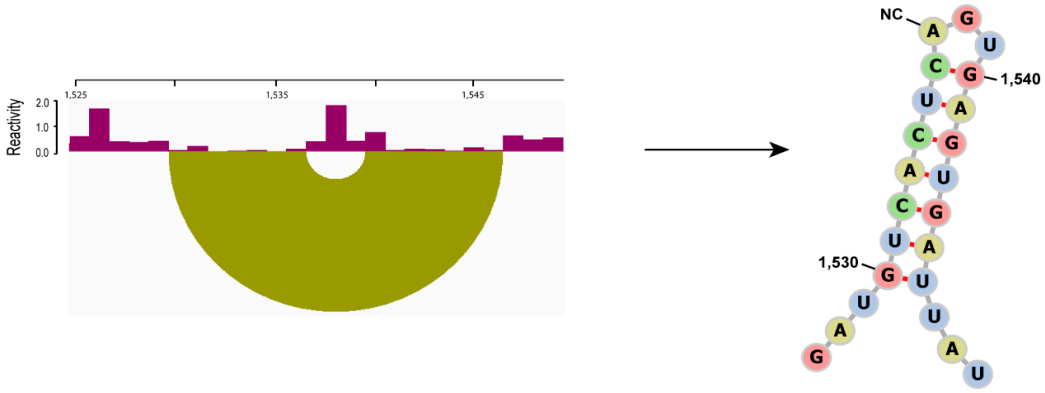


Figure 22: Comparison of NP intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the NP segments. (B) The sequence identities between the NP vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the NP segments. (D) Reactivity profile and structure of a hairpin predicted to be conserved amongst the viruses tested. Its existence had previously been supported by evolutionary analysis (Gultyaev et al., 2014). NC indicates that this residue is not conserved amongst the 4 virus sequences investigated.

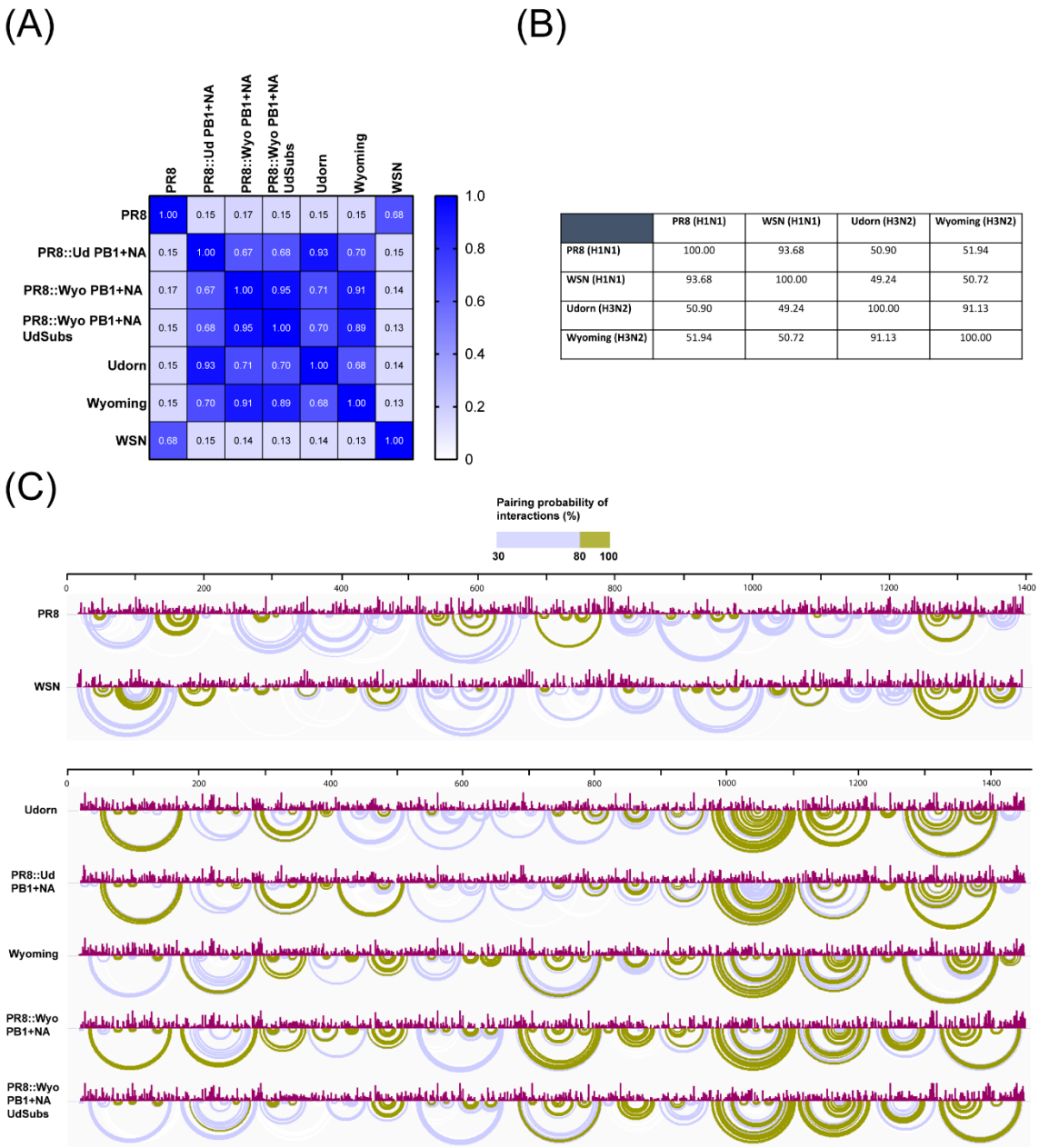


Figure 23: Comparison of NA intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the NA segments. (B) The sequence identities between the NA vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the NA segments.

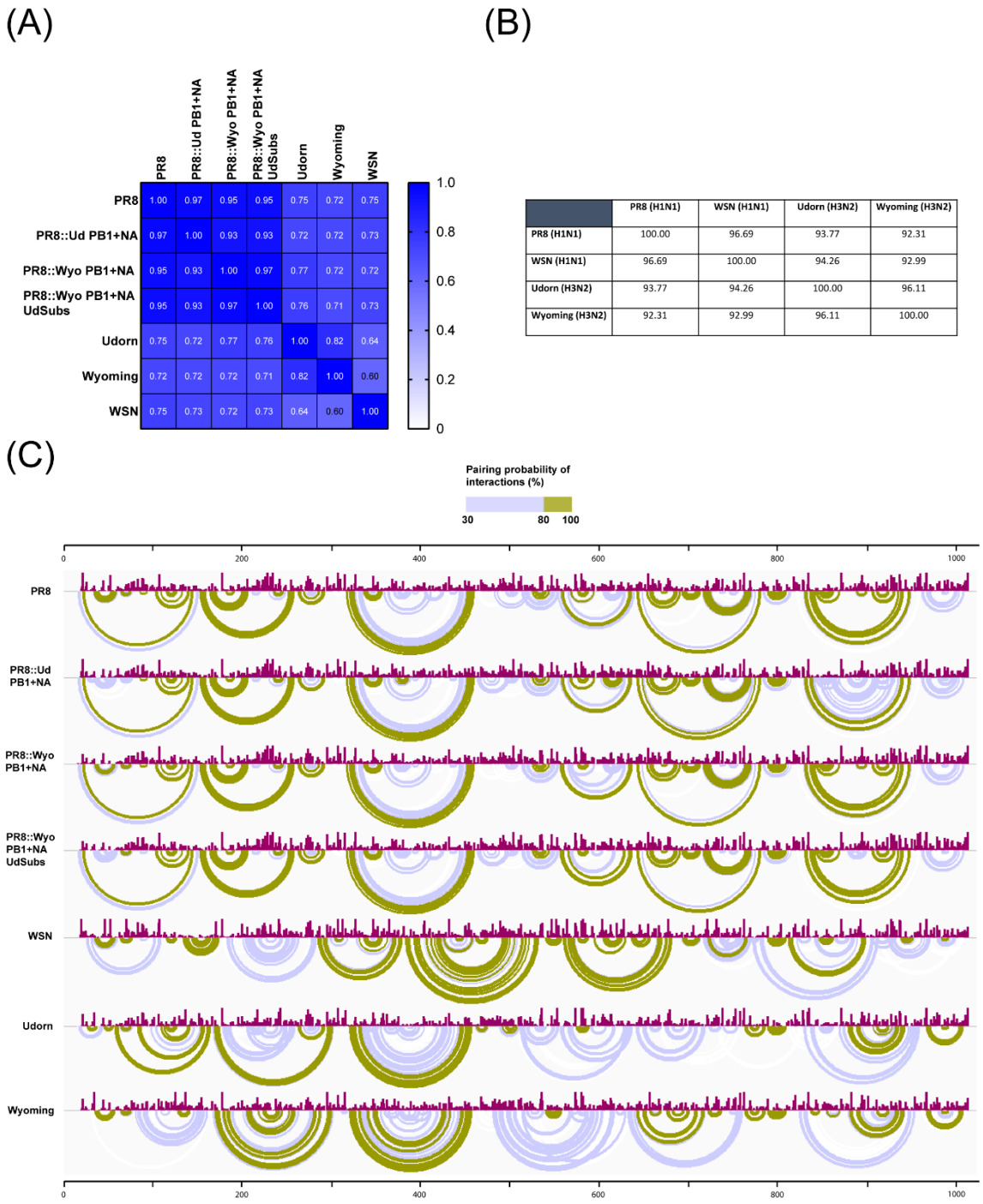
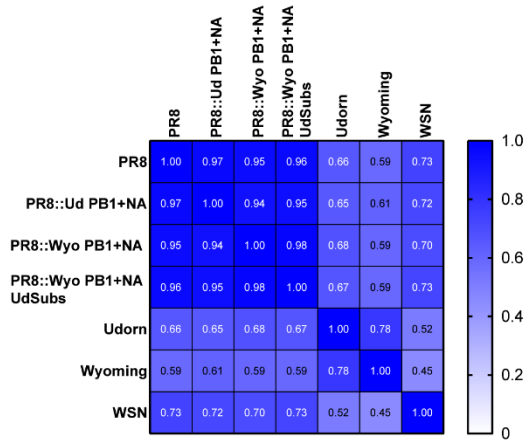


Figure 24: Comparison of M intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the M segments. (B) The sequence identities between the M vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the M segments.

(A)



(B)

	PR8 (H1N1)	WSN (H1N1)	Udorn (H3N2)	Wyoming (H3N2)
PR8 (H1N1)	100.00	96.29	91.12	89.21
WSN (H1N1)	96.29	100.00	92.13	89.78
Udorn (H3N2)	91.12	92.13	100.00	95.39
Wyoming (H3N2)	89.21	89.78	95.39	100.00

(C)

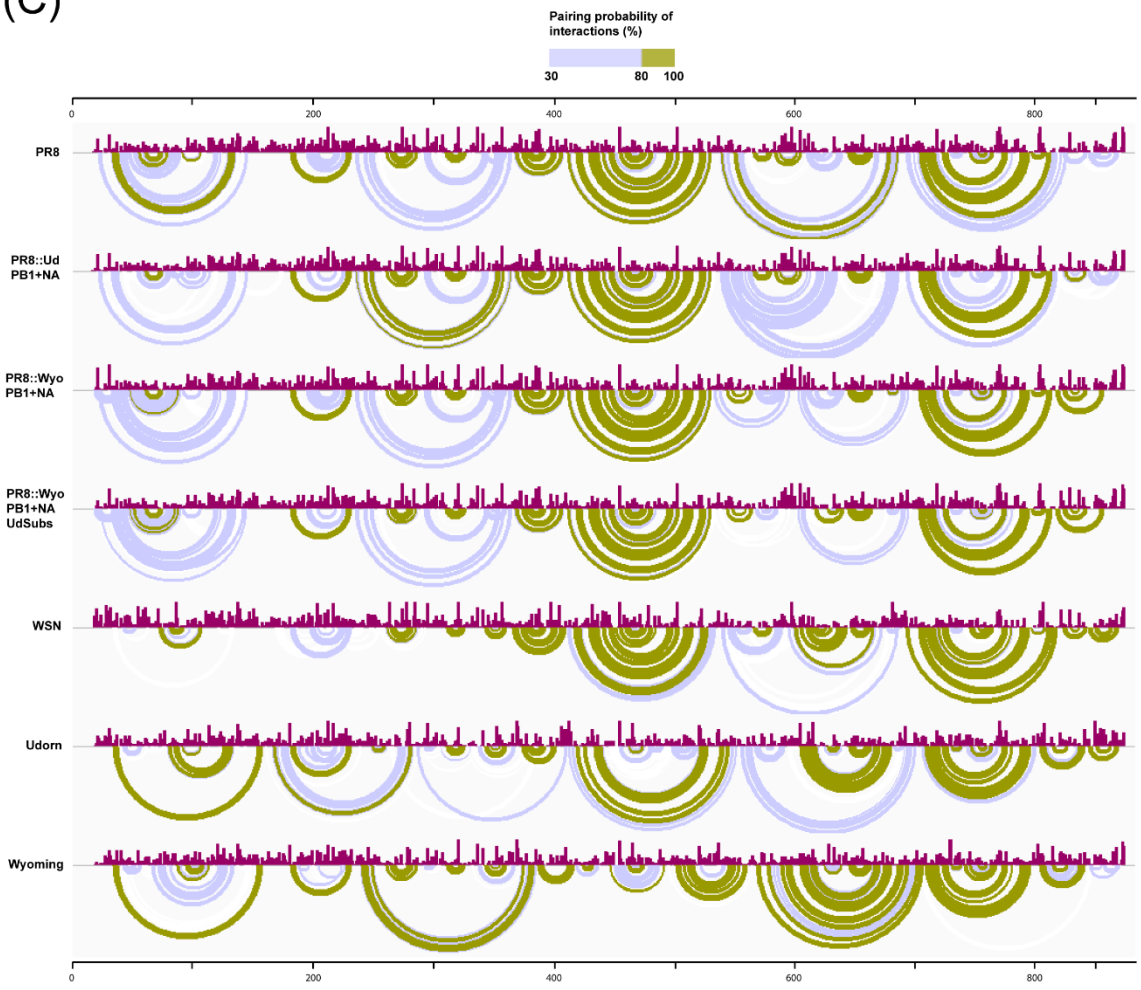


Figure 25: Comparison of NS intra-segment RNA structure in different IAVs.
(A) The Spearman R correlation values of the SHAPE reactivity profiles between the M segments. (B) The sequence identities between the NS vRNAs. (C) SHAPE reactivity profiles and predicted interactions of the NS segments.

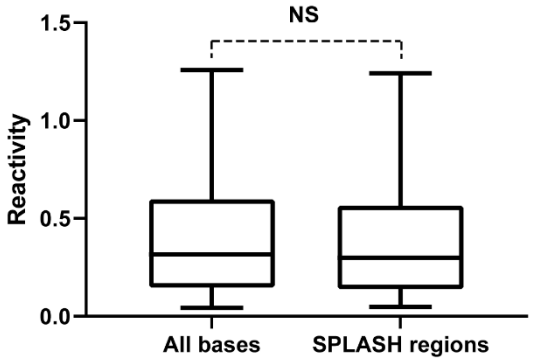
The SHAPE reactivity profiles of segments appear to be largely unaffected by the other segments present in the virus (i.e. SHAPE reactivity profiles of segments in a reassortant virus correlate highly with the same segment in the parental strain). For example, the PR8 PB2 segment is present in the PR8 virus and all 3 of the reassortant viruses used in this study. The Spearman correlation R values for SHAPE reactivity between the 4 viruses vary from 0.93-0.98 (Fig. 18A). In addition, the interactions predicted for the segments are almost identical (Fig. 18C). This suggests that intra-segment RNA structure *in virio* is largely unaffected by changes in inter-segment interactions and/or the presence of segments and proteins from a different virus.

There are predicted to be relatively few RNA structures conserved across all of the viruses used in this study. The more distantly related segments tended to see fewer conserved interactions. For example, no conserved interactions were observed between the H1N1 and H3N2 virus HA segments (Fig. 21C). In contrast, many structures in HA are conserved within the sub-types, especially between WSN and PR8. Amongst the more conserved segments cross sub-type structures are observed, though are still few in number. Examples of conserved structures include hairpins in the PB2 (2,257-2,279) (Fig. 18D) and NP segments (1,530-1,546) (Fig. 22D), the latter of which is located in a UTR.

SHAPE was performed on the same viruses on which SPLASH had been performed (with the exception of WSN). Analysis of the sequence regions encompassed by SPLASH interactions (the 20 most prevalent interactions) showed that the reactivity of these regions did not differ significantly from the reactivity values across the whole sequence (Mann-Whitney test P=0.09) (Fig. 26A). Chapter 3 (Section 3.3.3) detailed the introduction of an inter-segment interaction between the NA and PB1 segments of the PR8::Wyoming PB1 + NA virus by the introduction of 4 mutations to the NA sequence. In the mutated virus this interaction was one of the most prominent interactions in the SPLASH dataset, its existence has been supported in other studies (Gilbertson et al., 2016), and its introduction appears to influence reassortment (Chapter 3) (Dadonaite et al., 2019) . Despite this, the introduction of this interaction did not

greatly change the SHAPE reactivity in the region (Fig. 26B). The same intra-segment structures are predicted to form in both the mutant and non-mutant segment (Fig. 23C). These data suggest that individual inter-segment interactions are not prevalent enough in the population *in virio* to greatly effect SHAPE reactivity data. Thus it does not appear that inter-segment interactions need to be accounted for when making intra-segment structural predictions based on SHAPE reactivity.

(A)



(B)

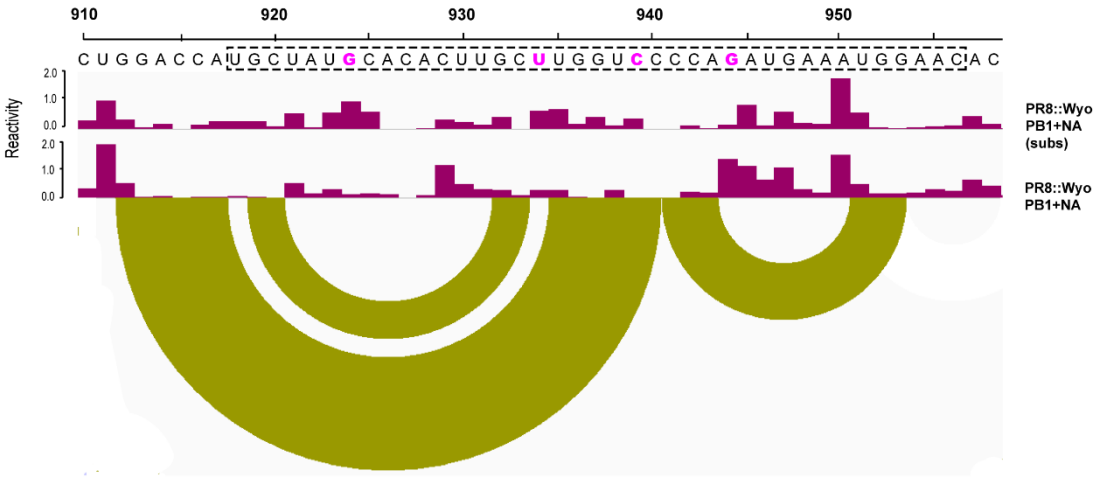


Figure 26: SPLASH interactions regions do not correlate with SHAPE reactivity.
(A) Comparing the SHAPE reactivity's between regions involved in top 20 SPLASH interactions and the dataset as a whole for the PR8 virus. In each plot the middle line is the median, the edges of the box show the lower and upper quartiles, and the whiskers show the

5th and 95th percentiles. Mann-Whitney test shows that there is not a significant difference between the groups P = 0.09. N = 1,298 for SPLASH interactions and 13,241 for All bases. (B) Residues in pink were mutated in the PR8::Wyo PB1 + NA_{UdSubs} virus. Interactions shown are for the PR8::Wyo PB1 + NA virus. The interactions seen for the mutant virus were the same but some were of lower confidence (Fig. 23). The dotted box indicates the region involved in the NA-PB1 inter-segment interaction in the PR8::Wyo PB1 + NA_{UdSubs} virus but not in the PR8::Wyo PB1 + NA virus.

4.3.3 DMS and EDC probing of WSN

Chemical probes tend to each have their own biases. For example, there is the possibility that SHAPE reagents may have reduced accessibility to NP-bound RNA, due to NP binding to the RNA backbone (Ng et al., 2008). To try to account for this factor, two additional chemical probes were used (both of which react at the base pairing interface). Probing with DMS for mutational mapping has been validated for use as soft constraints in RNA structure prediction (Cordero et al., 2012) and has previously been applied to study viral RNA structure (Watts et al., 2009) (Mauger et al., 2015) (Simon et al., 2019). However, DMS reacts preferentially with A and C bases. EDC allows probing of G and U bases (Mitchell et al., 2019) (Wang et al., 2019), but previous studies have relied upon analysing reverse transcription stops rather than mutations. This is due to the difficulty in reverse transcriptase's reading through the larger adduct produced by EDC probing. Here EDC-MaP was attempted on WSN by performing reverse transcription with Superscript II in the presence of manganese (II) ions. DMS-MaP was also performed (separately) on the WSN virus.

For EDC median read depths were in excess of 17,000 for both replicates with median mutation rates of 0.12%, 0.16%, 0.41%, and 0.43% for A, C, G and U respectively. By comparison, the median reactivity was 0.18% in the untreated and 0.31% in the 1M7 treated sample. This suggests that EDC reverse transcription mutations were successfully introduced at modified G and U bases. The DMS sample had very low median read depths of ~2,000 for both replicates. The mutation rates were considerably higher, with median values of 3.97%, 4.99%, 1.09%, and, 1.01% for A, C, G, and U respectively.

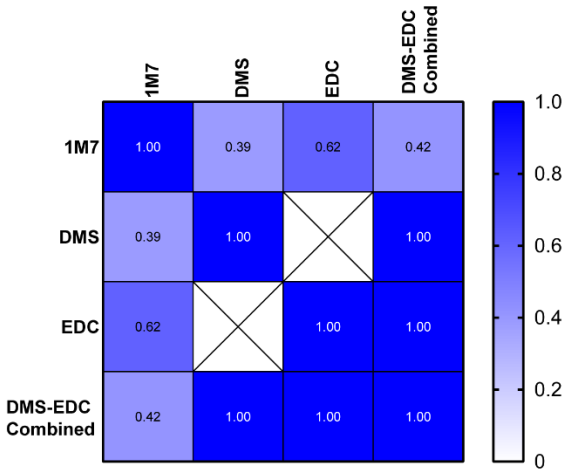
The SHAPE reactivity values for the EDC and DMS samples were calculated the same way as for the SHAPE experiments with the exception that normalisation was done separately for each base. The reactivity of A and C to DMS is not equal, nor is EDC's for G and U (at least when using reverse transcription stops) (Mitchell et al., 2019). As such, especially when attempting to combine data from the two probes (with DMS having much higher reactivity) per base normalisation is necessary. This resulted in median reactivity values of 0.53, 0.57, 0.40, and 0.35 for A, C, G, and U respectively. The reactivity rates were on average higher in the EDC and DMS samples as compared to in the 1M7 samples (Table. 5).

	1M7	DMS	EDC	DMS-EDC Combined
Minimum	0	0	0	0
25% Percentile	0.124	0.370	0.183	0.259
Median	0.287	0.545	0.372	0.468
75% Percentile	0.584	0.718	0.616	0.679
Maximum	21.24	2.519	5.468	5.468
Mean	0.431	0.562	0.462	0.510
Std. Deviation	0.530	0.286	0.41	0.359

Table 5: Comparison of reactivity profiles of different chemical probes.

The Spearman correlation R value of the SHAPE reactivity values for the 1M7 and EDC samples was 0.62 (Fig. 27A). This is higher than the correlation between the 1M7 and DMS samples (0.39). The correlation between the 1M7 and EDC sample supports the idea that the mutation rate of a base is at least partially linked to its likelihood of being unpaired (assuming this is true for 1M7 probing). However, the fact that the correlation between the three different probes is not higher suggests that the probe used has a large impact on the reactivity rate of a base.

(A)



(B)

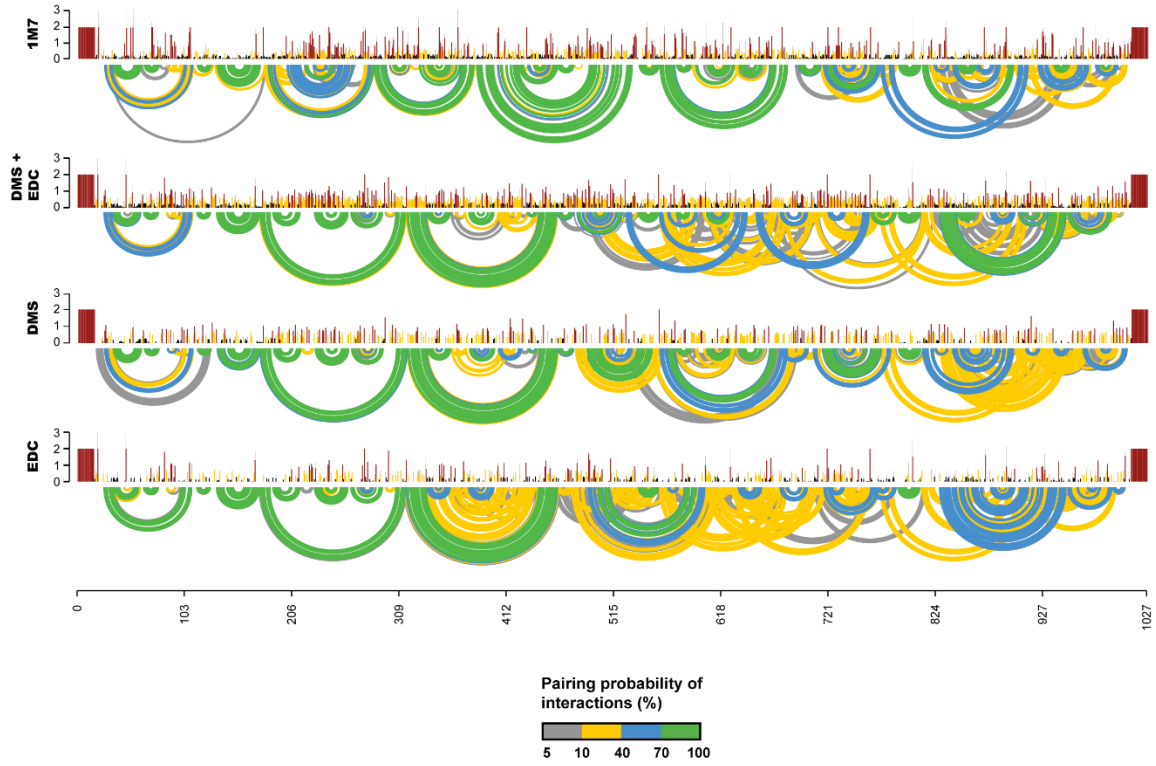


Figure 27: Comparing RNA structure predictions using different chemical probes.
(A) The Spearman R correlation values of the 1M7, EDC, and DMS reactivity profiles (all segments combined). Correlation for EDC only includes G and U base reactivity and DMS only includes A and C. (B) The reactivity plots and predicted interactions for the WSN M segment based on 1M7, EDC, DMS, or EDC and DMS probing combined.

Structure predictions were performed using the EDC and DMS reactivity values separately and in combination (Fig. 27B, Fig. 28-30). The interactions predicted by the combined EDC and DMS probing do not entirely support the predictions made by the 1M7 probing (Fig. 28-30). In the example of the M segment (Fig. 14B) the data are in fairly good agreement at the 5′ end, but the 3′ half contains considerable uncertainty in the EDC-DMS probed datasets. Of the 26 interactions in the M segment with greater than 80% pairing probability in the 1M7 generated structure, 46.1% also have greater than 80% pairing probability in the EDC-DMS predicted structure, 19.2% have 10-70% pairing probability, and 34.6% have less than 10% pairing probability.

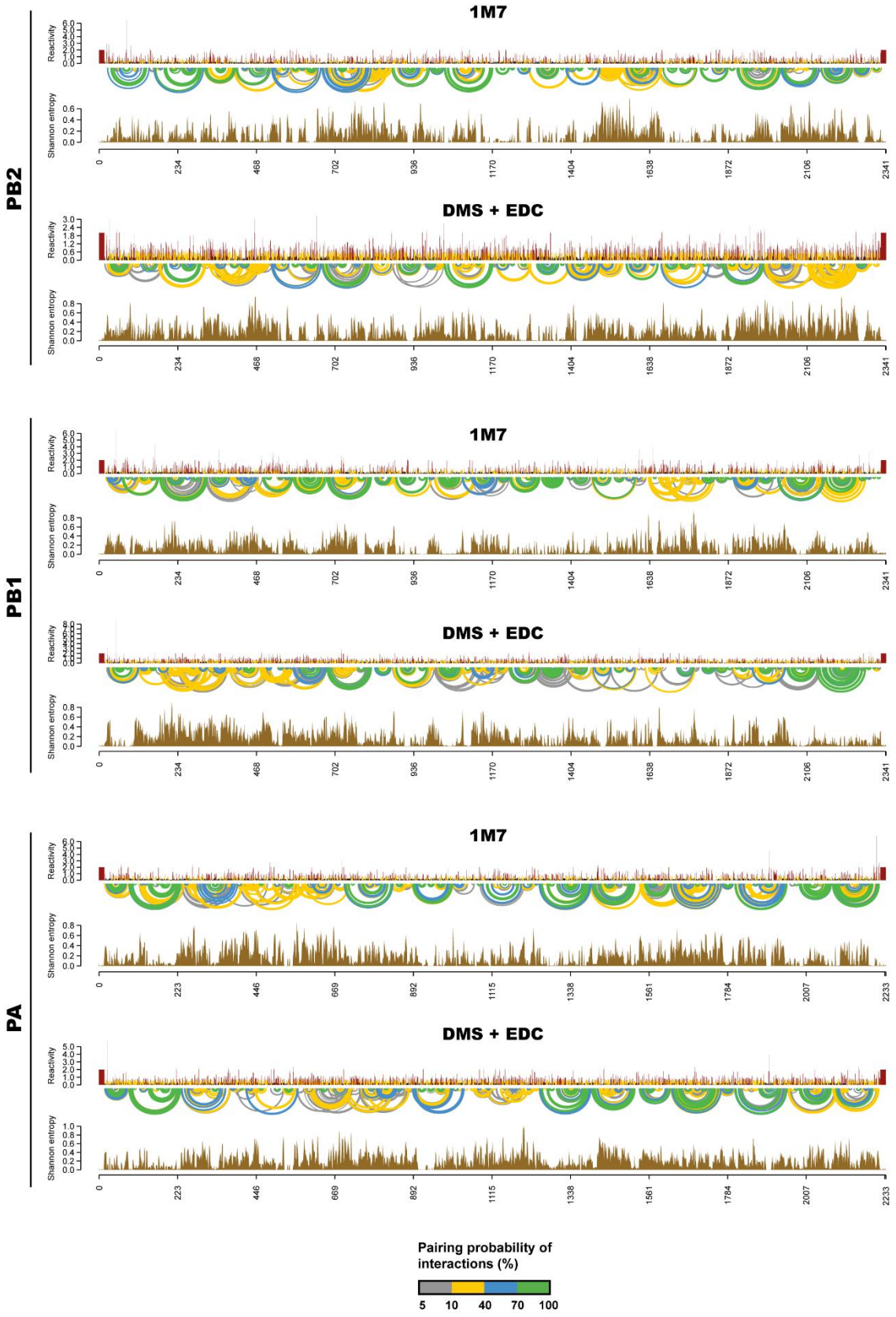


Figure 28: Comparing 1M7 and DMS + EDC predictions for PB2, PB1, and PA.
The reactivity plots, Shannon entropy's, and predicted interactions for the WSN polymerase segments based on 1M7 or combined EDC and DMS probing.

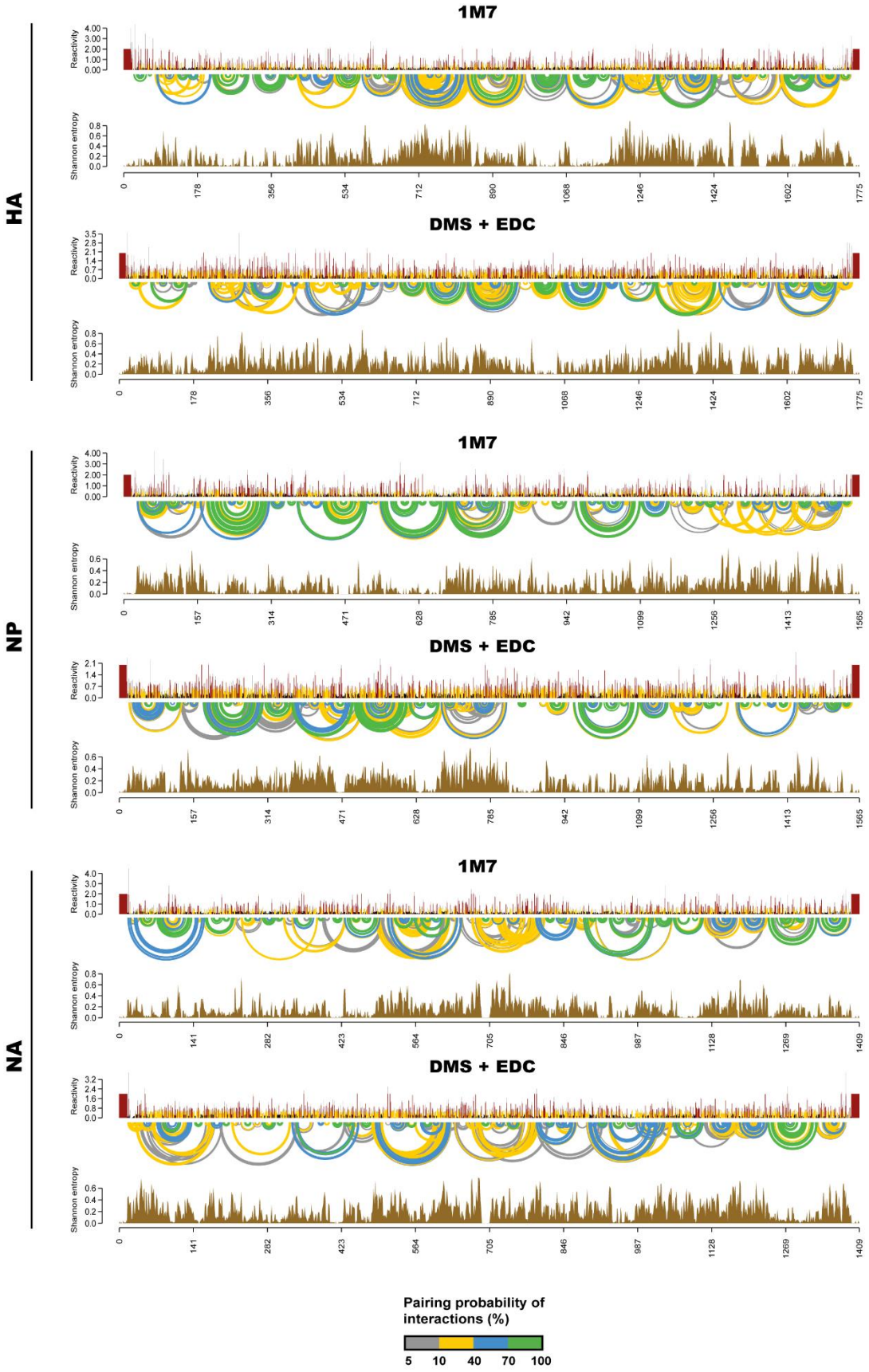


Figure 29: Comparing 1M7 and DMS + EDC predictions for HA, NP, and NA.
The reactivity plots, Shannon entropy's, and predicted interactions for the WSN HA, NP, and NA segments based on 1M7 or combined EDC and DMS probing.

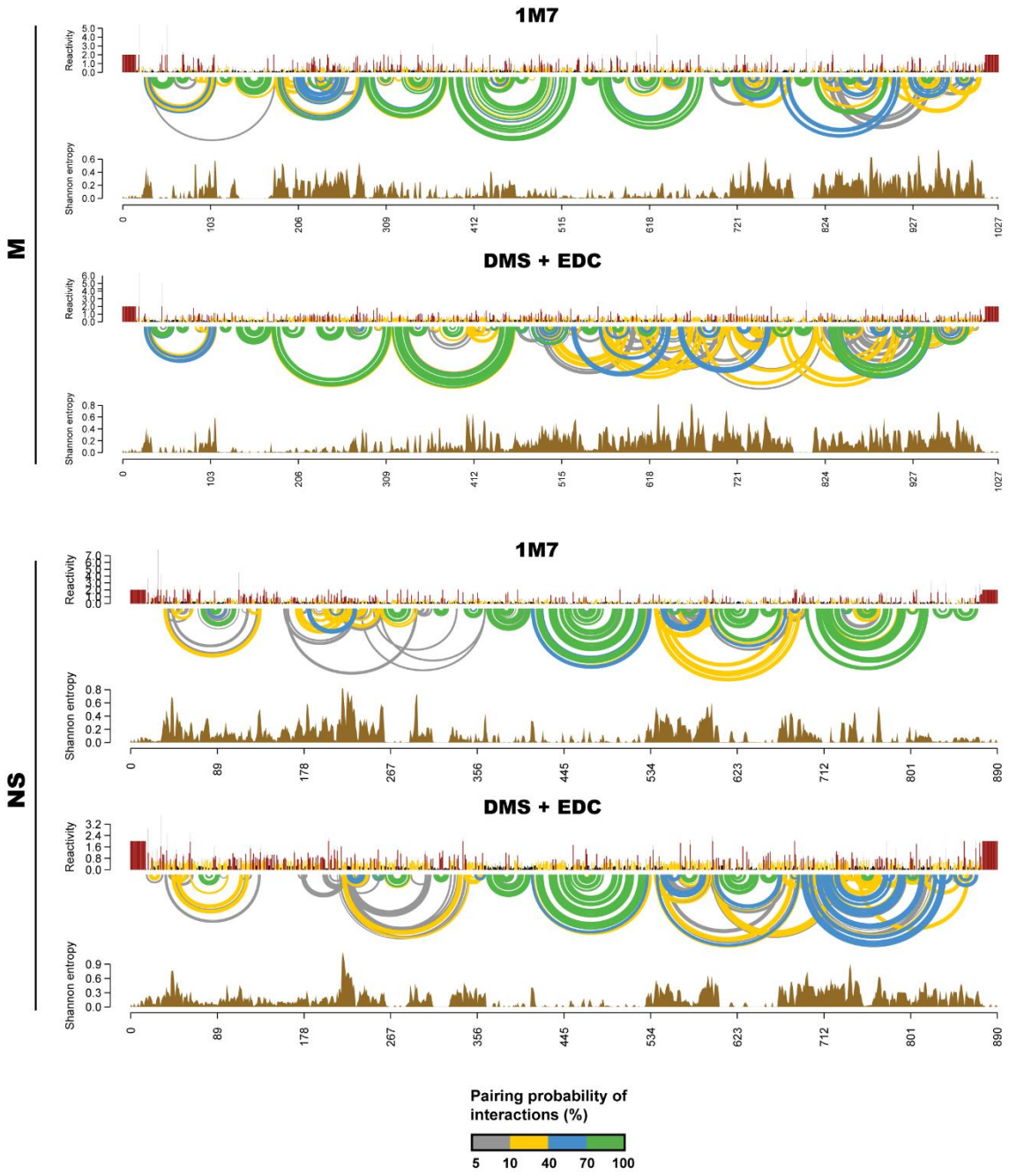


Figure 30: Comparing 1M7 and DMS + EDC predictions for HA, NP, and NA.
The reactivity plots, Shannon entropy's, and predicted interactions for the WSN M and NS segments based on 1M7 or combined EDC and DMS probing.

4.4 Discussion

This investigation has presented the predicted intra-segment interactions on a genome wide scale of 4 IAVs and 3 of their reassortants. The genomes appear to contain a large number of RNA structures across all segments. There appears to be relatively little conservation between viruses from different sub-types, though several conserved structures are observed. On the more distantly related HA and NA segments, there is no discernible conservation of structure between segments from the two subtypes (Fig. 21, Fig. 23). Conserved structures were found to be more prevalent on segments sharing higher sequence conservation, such as PB2, NP, and M (Fig. 18, Fig. 22, Fig. 24).

Two of the stem-loops on M (34-61 and 788-804) (Fig. 15C) had been predicted to occur in a computational study (Kobayashi et al., 2016). The 32-61 and 788-804 stem-loops were predicted to occur respectively in 98% and 99% of the genomes of 1,865 IAVs from 88 sub-types. Both structures are located in regions determined to be important to packaging of the M segment (Hutchinson et al., 2008) (Ozawa et al., 2009). Both also occur in regions where synonymous mutations appear to be relatively infrequent when compared to other regions of the M segment (Moss et al., 2011) (Kobayashi et al., 2016). Disruption of these stem-loops by synonymous mutation resulted in reduced viral growth rates that, in the case of the 34-61 stem-loop, could be recovered with compensatory mutations that restored base pairing (Spronken et al., 2017). Disruption of the 788-804 structure was also found to produce an increase in production of non-infective particles (Kobayashi et al., 2016). Both structures are also predicted to form in the M mRNA (Simon et al., 2019), so it is hard to assess whether it is important functionally at a vRNA and/or mRNA level. This challenge applies across all segments. One study has performed DMS probing of PR8 mRNAs (Simon et al., 2019). Of the structures formed by the NS mRNA 56% were identical to those in the vRNA predicted by the SHAPE data in this investigation, 11% were partially the same and 33% were not present in the vRNA (Simon et al., 2019). Given that some of the differences are likely attributable to

the different probe and folding parameters used (e.g. max pairing distance was 500), this suggests many of the RNA structures in the influenza genome form in both senses.

One study used covariation analysis to predict the presence of conserved structures in the NP segment, highlighting 6 structures believed to contain co-varying base pairs (Gultyaev et al., 2016). The 1,526-1,546 stem-loop was predicted for every virus in this study, including in the DMS-EDC probing (Fig. 22, Fig. 27B) (Gultyaev et al., 2014). This occurs in a region known to be a packaging signal for NP (Ozawa et al., 2007) and that was found in PAR-CLIP experiments to have low NP binding (Williams et al., 2018b). Disruption of the loop by mutation leads to defects in replication and an increase in defective particle formation (Williams et al., 2018b). The structure has previously been suggested to be important to regulating translation of the NP mRNA (it is also predicted to form in the positive sense) (Park et al., 1999). The other 5 structures predicted to contain co-varying residues were not replicated in this study. This includes a predicted pseudoknot in the 90-129 region. Whilst this structure was not recapitulated in this analysis, the region contains very low SHAPE reactivity values (mean = 0.18). It is possible that the formation of this structure somehow inhibits probe access to the unpaired bases, whether through binding of a protein or lack of solvent access.

The lack of cross subtype structural conservation of HA observed in this investigation is supported by structural covariance analysis, which found no evidence of structures maintained across viruses from different subtypes (Gultyaev et al., 2016). However, the study did show that, within some subtypes, covariation indicative of RNA structure maintenance was present. No H1N1 structures were found to be maintained, but two predicted H3N2 structures were each shown to have two base pairs exhibiting some degree of covariation. Of the predicted structures, only one was supported by the SHAPE data and only in Wyoming (685-757) (Fig. 21). A study using DMS probing of influenza mRNA found similar issues in reproducing computationally predicted structures, with only ~50% of *in silico* predicted structures supported by their probing data (Simon et al., 2019). It is typically accepted that chemical probing increases the accuracy of structure prediction (Hajdin et al., 2013). As such, the structures

presented in this and other chemical probing experiments may form a basis from which to improve future covariance analysis.

The structures presented here are highly unlikely to represent a completely accurate picture of the intra-segment structure of the influenza virus genome. Base pairing accuracy can be in the region of 90% for SHAPE informed structure prediction, although this figure tends to decline for longer RNAs (Deigan et al., 2009) (Weeks, 2021). In order to try to validate the existence of SHAPE predicted structures, this study used DMS and EDC probing. Different chemical probes have different biases. For example EDC and DMS react with the base pairing interface, rather than the backbone of the RNA which may mean they are less affected by NP binding (although it is not obvious that 1M7 probing is affected by this phenomenon (Fig. 17)) and base stacking. Structures that can be supported by different probing techniques may provide more reliable targets for functional studies.

The results here suggest that EDC is an effective chemical probe for mutational mapping of G and U bases. However, EDC probing with mutational mapping should be performed on RNAs of known structure. This will allow validation of its accuracy and can act as a 'ground truth' to allow establishment of better parameters to use as the gradient and intercept when applying free energy terms to bases during structure prediction (a process termed jack-knifing) (Incarnato et al., 2018). The best parameters to use vary by probe and optimising them can result in more accurate structure predictions (Hajdin et al., 2013) (Marinus et al., 2021). It is clear the reactivity values are distributed quite differently for all 3 probes used in this investigation (Table. 5), with different median reactivities, interquartile ranges, and standard deviations. As such, it is highly unlikely that base pairing probability correlates the same relative to reactivity for all 3 probes. There is high potential to combine EDC mutational mapping of G and U bases with the A and C probing of DMS. However, combining the data presents challenges that need to be resolved. For example, both probes are likely to have different optimal slope and intercept values for free energy prediction and there is currently no way of setting these values to be different for different bases with the available software. As

such, the standard parameters for the 1M7 reagent were used in this investigation for the combined folding (Hajdin et al., 2013). Regardless of the use of EDC, allowing different parameters for each base may be beneficial for the accuracy of DMS informed predictions given its different reactivity for A and C nucleobases.

The DMS probing in this investigation was performed at too high a concentration. The mutation rates of the A and C bases were almost 5% meaning there would have been ~3 modifications per read on average. This may explain the low sequence coverage, as the high modification rate may have reduced the efficiency of reverse transcription (there is always a chance of a reverse transcription stop at a methylation site rather than a read-through) or led to instances where the large number of mutations in a read prevented it from mapping back to the genome. This is supported by the fact that only ~10% of DMS reads mapped to the genome compared to ~65% for 1M7 and EDC (despite the probing being performed on the same samples). There is the additional concern that despite the rapid reactivity of DMS, over-modification could begin to alter the RNA structure. Future attempts at DMS probing of influenza virions would likely benefit from reducing the DMS concentration to seek to lower this modification rate. Increased sequence coverage would also allow further validation of interactions by correlated mutation analysis (Mustoe et al., 2019).

Another limiting factor in the accuracy of structure prediction is the parameters used for folding. The maximum pairing distance allowed for a interactions in this investigation was 150 bases, which is considerably shorter than the 500-600 that is often used (Simon et al., 2019) (Manfredonia et al., 2020). Changing this distance can result in substantial differences in the interactions predicted. The logic for using this smaller distance relates to cryo-EM studies of influenza vRNPs. These studies have shown that filaments vary substantially in length from ~50-150 nm, whilst their width is more consistent (~15 nm) (Arranz et al., 2012) (Gallagher et al., 2017). Both vRNA termini are bound by the polymerase. The shortest segment is ~890 bases in length (NS). This means the segment must fold back on itself at approximately nucleotide 445. If an intra-segment interaction were to connect base 20 to 320 then this could

not easily be resolved into a rod shaped structure (it would instead form more of a blob). A pairing distance of 150 was also found to provide the highest accuracy when probing RNA's of known structure, as longer distances increased the false positive rate beyond the benefits gained from potentially identifying interactions exceeding 150 bases (Lange et al., 2012).

The maximum pairing distance was not increased for longer segments. The length differential between the shortest and longest vRNPs observed by Cryo-EM (~3 times) (Arranz et al., 2012) (Gallagher et al., 2017) approximately matches that of the difference in size between the shortest and largest vRNA segments (~2.6 times). This, combined with the similar widths observed for vRNPs, suggests that the longer segments do not tend to contain longer range interactions. However, there is a flaw to this logic in that vRNPs fold back on themselves, potentially creating circumstances in which interactions spanning 2,000 bases are more likely than those spanning 500 (in the case of the polymerase segments which exceed 2,000 bases in length). This could potentially be accounted for by giving each base two windows in which it can find a partner. One window 150 bp either side of it and then a similar sized window on the 'return' part of the segment (e.g. base 100 of the PB2 segment would be allowed to pair with bases 1-350 and 2,091-2,341). However, this would rely on a number of assumptions, including that the tip of the segment furthest from the polymerase occurs exactly halfway through the sequence and that such long range interactions even occur (which is largely not supported by SPLASH (Chapter 3)). One way to investigate some of these assumptions would be to use FISH probes with fluorescence resonance energy transfer dye pairs against different regions of the genome to see where they fall in the vRNP (for example determining if the halfway point on the segment is located at the tip of the vRNP).

An additional problem when predicting RNA structure is the potential for regions of the RNA to form multiple structures. All of the segments were shown to contain regions of high Shannon entropy suggesting that there are one or more alternative structures with similar free energy. A challenge of SHAPE reactivity values is that they present an average and a base may have intermediate reactivity when in fact it may exist in two states, one with high and one with low

reactivity. Recent advances have begun to provide ways of deconvoluting these structures. This includes the algorithm DRACO. DRACO exploits the high modification rate of DMS, which results in multiple modifications per transcript, to try to establish which bases are likely to be modified together (or not) (Morandi et al., 2021). This can allow the reactivity profile to be decomposed into its constituent structures, reportedly allowing accurate prediction of the relative abundances of these structures. It was attempted to apply DRACO to the DMS data from this study but sequencing coverage and read lengths were insufficient. Repeating these experiments with reduced DMS concentration could thus be of great use in improving structure predictions in regions that may fold into multiple structures.

Studies of RNA structure may help in the design of anti-sense oligonucleotides that could act as influenza drugs. One anti-sense oligo (Radavirsen) targeting IAVs is in clinical trials (Beigel et al., 2018). It seeks to inhibit the translation and splicing of the M mRNA. The M segment may present the best target for such therapies as this and other studies have found it to contain a number of conserved structures in regions that seem to have low tolerance for mutation whether synonymous or otherwise (Kobayashi et al., 2016) (Spronken et al., 2017). Targeting other segments may be more sub-type specific which may suggest they will be more prone to escape mutations. The effect of anti-sense oligonucleotides can be hard to predict, but generally targeting loops is considered more successful than paired regions (Szabat et al., 2020). Targeting regions of the NP (Michalak et al., 2019) and M (Lenartowicz et al., 2016) segments with Anti Sense Oligonucleotides (ASOs) based on regions predicted to be single stranded produced mixed results in inhibition of viral growth in tissue culture, with some probes ineffective and some producing up to 10 fold drops in titre. There was no clear pattern as to what would be most effective, as some of the regions targeted were single stranded and some paired (in both their structure predictions and those made in this investigation). As such, it is not been clearly demonstrated that RNA structure predictions will be of use in antisense oligo design, at least without more functional information (e.g. whether proteins bind to some of the loops).

Overall this work has uncovered the RNA structure in a number of IAVs. The lack of conserved structural features in the terminal ‘packaging’ regions suggest that there is not a particular RNA structure responsible for packaging. It seems more likely that packaging is governed by protein-protein interactions involving 3P, NP, and M. A number of conserved structures have been identified. This includes 2,257-2,279 (PR8 numbering) in PB2, 1,530-1,546 (PR8 numbering) of NP and several structures across the M and NS segments. It would be informative to perform mutational analysis on these interactions. It would also be interesting to investigate RNA structure in other *genera* of *Orthomyxoviridae* as conservation across *genera* would likely be highly indicative of functional importance.

5. Structure of the influenza virus nucleoprotein

5.1 Chapter Summary

- NP has potential as a drug target, however the structure of a NP from a H3N2 influenza virus had not been determined, nor had the structure of NP in complex with RNA.
- The structure of the NT60 (H3N2) R416A NP was determined using X-ray crystallography at 2.2 Å resolution.
- The structure is highly similar to the other available IAV NP structures, but the conformation of an RNA binding loop (residues 73-90), is seen to differ.
- Cryo-EM analysis shows that NP forms oligomeric structures that are highly heterogeneous. These complexes could not be resolved to high enough resolution to elicit information on RNA binding.

5.2 Introduction

Structural proteins that bind to and encapsidate the viral genome are common features of RNA viruses. This includes influenza, where this role is carried out by the 56 KDa NP. NP is a multifunctional protein that influences the structure of the vRNA and is essential for its replication (see section 1.2). In order to support infection NP is required to interact with a number of viral and host proteins, as well as to bind in a non-sequence-dependent manner to RNA (Eisfeld et al., 2015). These factors are likely to contribute to the extremely high amino acid sequence conservation seen amongst NPs. One study of 40 cross-species IAV NP sequences from various sub-types, found sequence identity varied by a maximum of ~11% (Cianci et al., 2013). By contrast, only 43.7% sequence identity is shared between the WSN (H1N1) and NT60 (H3N2) HA proteins.

There are currently two licensed classes of drugs targeting influenza viruses, neuraminidase inhibitors (e.g. oseltamivir) and M2 ion channel blockers (e.g. amantadine). Mutations conferring resistance to both classes of drugs are becoming increasingly prevalent amongst

circulating influenza viruses (Ison, 2011) (Hu et al., 2017). There is a need for new influenza anti-virals to replace, or use in combination with, these existing drugs. The high conservation amongst IAV NPs makes them an attractive target for influenza interventions, as drugs targeting them are likely to have cross-strain efficacy. A number of NP-targeting drugs are in various stages of research or clinical testing (Hu et al., 2017). In addition, putative vaccine candidates are attempting to target regions of NP in the effort to generate universal influenza vaccines, effective against multiple strains of influenza (McMahon et al., 2019) (Sun et al., 2020) (Pleguezuelos et al., 2020).

The structures have been determined for NPs from 3 of the 4 influenza *genera*, with structures available for influenza A, B (Ng et al., 2012) and D (Donchet et al., 2019) virus NPs. For IAVs, structures have been determined for the NPs of the HK97 (H5N1) (Ng et al., 2008) and WSN (H1N1) (Ye et al., 2006) viruses. NP forms homo-oligomers, but monomeric mutants can be produced by disrupting an interaction that occurs between the R416 residue, located in the oligomerisation loop, and the E339 residue in the body domain of a neighbouring NP. The structure of one such mutant, the R416A WSN NP has also been determined (Chenavas et al., 2013).

Despite being one of only two IAV subtypes circulating in humans, the structure of a H3N2 virus NP had not been determined, nor had the structure of NP in complex with RNA. The objectives of this chapter were to determine the structure of the NP from a H3N2 virus and to solve the structure of NP in association with RNA. Understanding the structural conservation amongst NPs from different sub-types is important for the development of universal interventions against influenza. Determining an NP-RNA structure could improve our understanding of vRNP organisation and aid in the design of new anti-viral drugs.

5.3 Results

5.3.1 Nucleic acid binding experiments

NP has potential for high sample heterogeneity, with the possibility of different oligomeric states and flexibility between NPs within an oligomer. It was thought that monomeric NP bound to short fragments of RNA may provide greater sample homogeneity and prove more amenable to crystallisation. In order to pursue this possibility the RNA binding properties of the monomeric mutant were investigated.

The NT60 (H3N2) R416A NP was expressed in *E. coli* and purified by SEC. Early attempts at purification yielded NP with high 260/280 ratios (>1.3), indicative of NP being bound to endogenous nucleic acid. Subsequently, a mixture of nuclease treatments and washes with buffer containing 1.5 M NaCl were introduced during purification. This produced a SEC absorbance profile with a single elution peak (Fig. 31A) at a position consistent with the mass of monomeric NP. The 260/280 ratio of the purified NP was 0.49, indicating that the sample had successfully been purified of nucleic acids from the expression host.

NP can also bind to DNA (Newcomb et al., 2009) which has the advantage of being much less expensive to have synthesized, particularly for longer oligonucleotides. The ability of the R416A NP to bind to DNA was assessed by mixing the purified NP in a 4:1 molar ratio with a 100-nucleotide DNA (Fig. 31A). This produced a second, earlier eluting absorbance peak during SEC, likely formed from multiple NP's binding to the same piece of DNA. This peak indicates that the R416A NP retains the ability to bind to DNA observed for the non-monomeric protein (Newcomb et al., 2009). It is not possible to give a precise measure of the number of NPs bound to the DNA, as elution volume is not entirely consistent with mass (shape is an important factor). However, the complex has a mass in the region of 200 KDa, so it is likely that 3-4 NPs were bound to the 100 nucleotide-DNA (DNA mass ~30.7 KDa). This is fairly consistent with the pattern of NP binding to vRNA, where it is estimated to bind on average every ~25 nucleobases (Ortega et al., 2000) (Hutchinson et al., 2014).

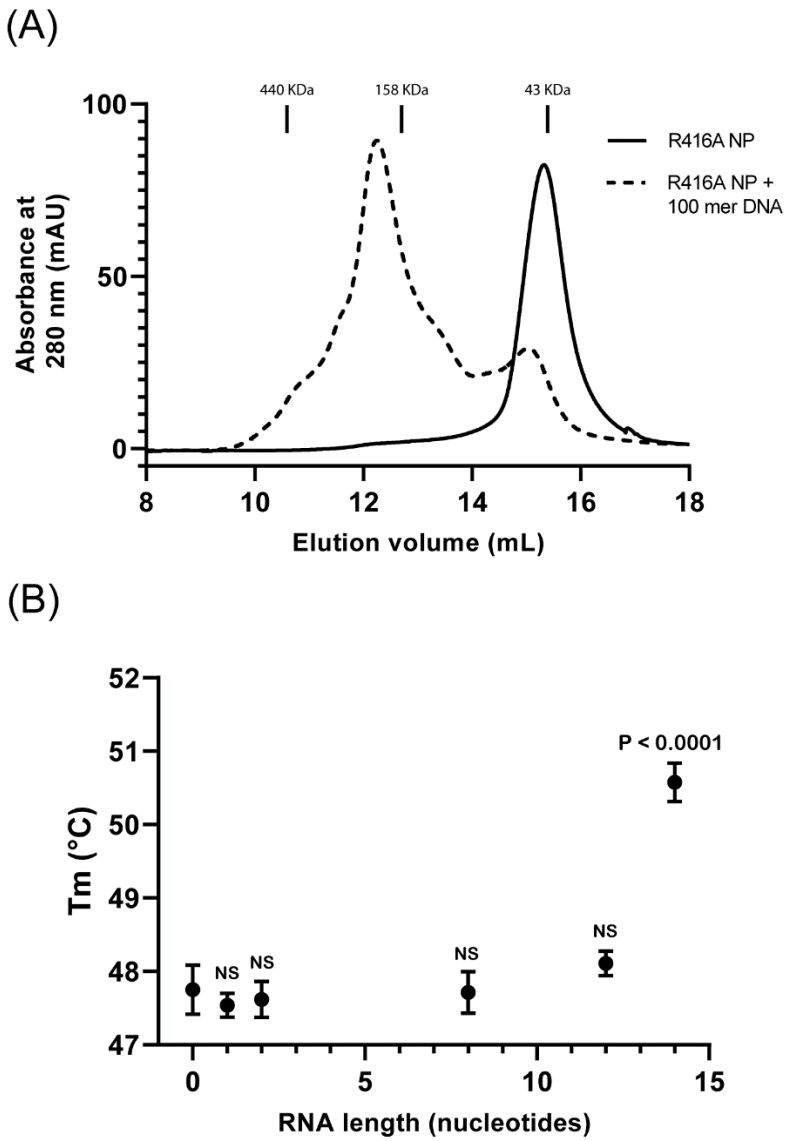


Figure 31: Nucleic acid binding of R416A NP
(A) The absorbance profile from SEC on the NT60 R416 NP in the presence or absence of a 100-nucleotide DNA. (B) The melt temperature of the NT60 R416A NP in the presence of different length RNAs, determined by ThermoFluor assay. A one-way ANOVA was performed comparing the NP samples with different length oligonucleotides to NP in the absence of any nucleotide. Figure adapted from (Knight et al., 2021).

A ThermoFluor assay was employed to investigate RNA binding by the R416A NP. The ThermoFluor assay measures the thermal stability of a protein using a fluorogenic dye that will fluoresce when bound to hydrophobic surfaces. As a protein becomes denatured, the

hydrophobic regions that are usually buried become accessible to the dye. Fluorescence is measured as the temperature of the protein is gradually increased, producing a melt curve. The melting temperature (Tm) can be calculated as the temperature at which 50% of the maximum fluorescent signal is achieved, relative to baseline fluorescence. The binding of ligands to a protein often leads to an increase in thermal stability (Lo et al., 2004) (Kranz and Schalk-Hihi, 2011). Thus, ligand binding can be assessed by comparing the Tm of the protein in the absence or presence of different ligands. Higher thermal stability of a protein is associated with higher chance of successful crystallisation (Dupeux et al., 2011).

The thermal stability of the NT60 R416A NP was determined in the absence or presence of oligoribonucleotides of different lengths (Fig. 31B). The Tm of NP was increased by 2.8°C in the presence of a 14-nucleotide RNA ($P < 0.0001$, one-way ANOVA) whilst the other oligoribonucleotides tested (of length 12-nucleotides and below) did not significantly affect the Tm. This suggests that the shorter RNAs tested may not be binding to the R416A NP. The WSN NP was shown to bind to an 8-nucleotide RNA with a K_d of 70 nM (at 300 mM NaCl) (Tarus et al., 2012). However, reduced affinity for RNA has been reported for the monomeric WSN R416A NP (Chan et al., 2010) (Elton et al., 1999) and it is possible that binding to shorter oligoribonucleotides is abolished in this monomeric mutant. It is also possible that the shorter RNAs used in this assay bind, but do not increase, the thermal stability of the NP.

To further assess RNA binding, R416A NP was mixed in a 1:1 molar ratio with different length RNAs and subject to SEC. The 260/280 of the eluted NP post-SEC was measured at 0.53 and 1.05 where it was mixed with a 5 or 14-nucleotide RNA respectively (R416A NP 260/280 of 0.49 without RNA). This suggests the 14-nucleotide RNA binds and is able to remain bound to R416A NP through SEC, whilst the 5-nucleotide RNA is not.

5.3.2 Crystallography

A range of crystallisation trials were set up with purified NT60 R416A NP at 10 mg/mL. The best diffracting crystal gave a maximum resolution of 2.2 Å (Tables 6 and 7). The protein was present in the P1-2_1-1 space group with two NP's in each asymmetric unit (referred to as chains A and B). The structure could be resolved for residues 21-389. Density was then missing for much of the oligomerisation loop, which is comprised of residues 402-428 (missing residues 390-417 in chain A and 390-437 in chain B). In both chains, residues 452-461 and 497-498 could not be resolved. NP forms a crescent-shaped structure (Fig. 32), with a body domain composed of residues 21-147, 278-396, and 463-489, and a head domain composed of residues 152-265 and 438-450. These two domains are formed predominantly from α-helices and the predicted RNA binding site is located in a groove formed at their interface.

Crystal form	P1-2_1-1	C2-2-2_1
***a*, *b*, *c* (Å)**	87.775, 63.376, 105.95	165.3, 286.3, 118.4
α, β, γ (°)	90.0, 98.31, 90.0	90.0 90.0 90.0
Resolution range (Å)	86.85 - 2.22 (2.3 - 2.22)	82.65 - 2.29 (2.38-2.29)
Total No. of reflections	250461 (11323)	622781 (31597)
No. of unique reflections	37998 (1900)	45852 (2293)
Completeness (%) (ellipsoidal)	90.9 (55.8)	90.7 (67.5)
Redundancy	6.6 (6)	13.6 (13.8)
⟨ *I*/σ(*I*)⟩	8.1 (1.6)	13.2 (1.8)
$R_{r.i.m.}$	0.07 (0.56)	0.05 (0.45)
Overall *B* factor from Wilson plot (Å²)	46.1	49.0

Table 6: X-ray crystallography merging statistics for the NT60 R416A NP.
Values for the outer shell are given in parentheses. Table adapted from (Knight et al., 2021).

Crystal form	P1-2_1-1	C2-2-2_1
Resolution range (Å)	72.21 - 2.22 (2.30 - 2.22)	82.65 - 2.29 (2.38 - 2.29)
Completeness (%) (spherical)	66.6 (3.4)	36.59 (2.06)
No. of reflections, working set	38052 (203)	45840 (254)
No. of reflections, test set	1830 (11)	2324 (16)
Final R_{cryst}	0.21 (0.26)	0.22 (0.38)
Final R_{free}	0.26 (0.38)	0.24 (0.35)
No. of non-H atoms	6813	10186
Protein	6762 (858)	10162 (1287)
Water	51	24
R.M.S. deviations		
Bonds (Å)	0.003	0.006
Angles (°)	0.55	0.81
Average *B* factors (Å^2)		
Protein	60.0	60.6
Water	46.5	53.8
Ramachandran plot		
Most favoured (%)	97.16	97.69
Allowed (%)	2.73	2.23
Molprobity score	1.28	1.88

Table 7: Structural parameters and refinement scores for the NT60 R416A NP.
Values for the outer shell are given in parentheses. Table adapted from (Knight et al., 2021).

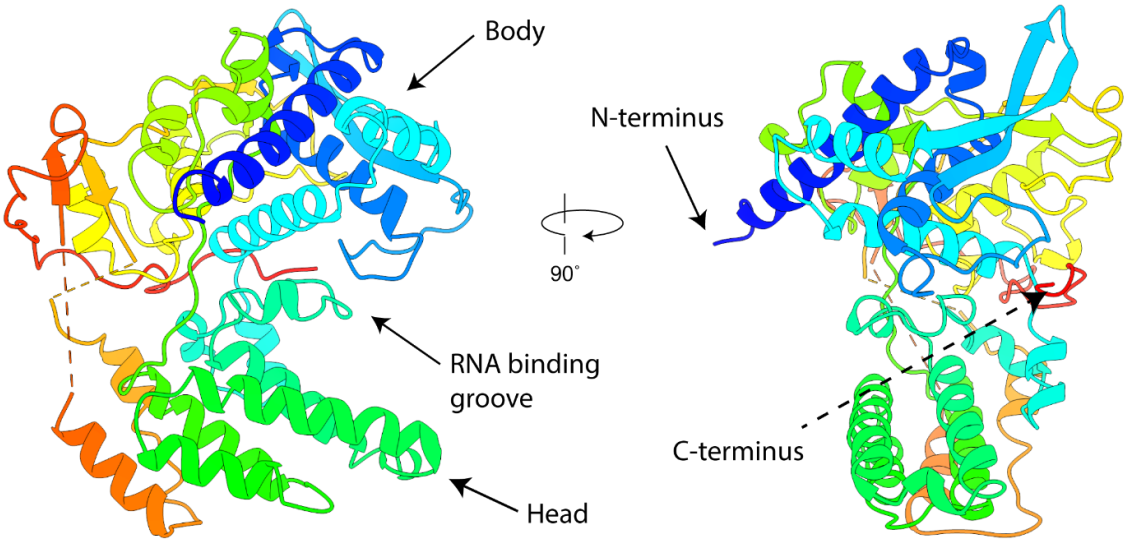


Figure 32: The structure of the NT60 R416A NP.
Structure is rainbow coloured progressing from violet at the N-terminus to red at the C-terminus. Adapted from (Knight et al., 2021).

Structures have been solved for the WSN (H1N1) NP (Ye et al., 2006), with which the NT60 NP shares 93.6% amino acid sequence identity and the HK97 (H5N1) NP (Ng et al., 2008), with which the NT60 NP shares 91.4% sequence identity. The residues not conserved amongst these 3 structures are not clustered, being widely dispersed both in the sequence (Fig. 33A) and spatially in the structure (Fig. 33B). The non-conserved residues are mainly located away from the basic groove (Fig. 33C) that is predicted to be the site of RNA binding. An exception to this is K77, which is replaced by an arginine in the HK97 sequence. This change maintains the basic charge of the groove, though lysine residues have been suggested to contribute less to NP RNA binding than arginine residues (Elton et al., 1999).

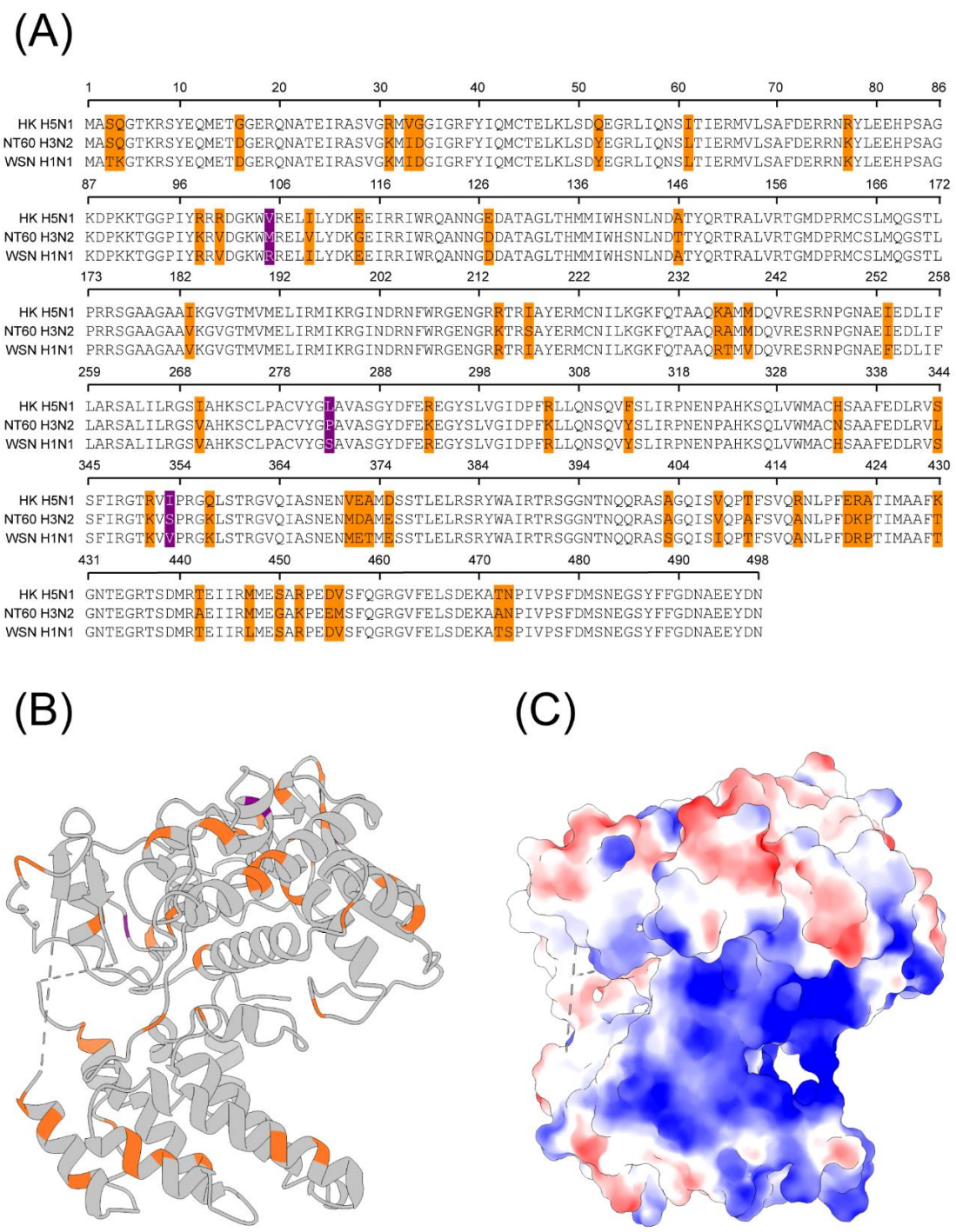


Figure 33: Conservation of NP.
(A) A sequence alignment of the NT60, WSN and HK97 NP sequences. Residues where one sequence differs are highlighted in orange. Residues where all three sequences differ are highlighted in purple. (B) The structure of the NT60 R416A NP coloured by amino acid sequence conservation relative to the WSN and HK97 NPs. Colouring matches that in the previous panel. (C) The structure of the NT60 R416A NP with surface charge (Coulombic electrostatic potentials) displayed. Blue = basic and red = acidic. Adapted from (Knight et al., 2021).

The structure of NP also appears to be highly conserved. The NT60 R416A NP structure has a Root-Mean-Square Deviation (RMSD) of 4.0 Å (across 393 pairs) compared to WSN NP, and of 5.5 Å compared to the HK97 NP (across 429 pairs). Much of the difference can be attributed to the folding of the C-terminal tail. In the NT60 R416A NP this tail folds into the predicted RNA binding groove. The same folding of the C-terminal tail is observed in the R416A WSN structure and its structure has an RMSD of 1.2 Å when compared to the NT60 R416A NP structure. The folding of this tail observed in the monomeric mutants is likely to reduce the charge of the predicted RNA binding groove and has been suggested as a reason for the lower RNA binding affinity observed for the monomeric mutant (Chenavas et al., 2013).

When compared to the other IAV NP structures, the biggest difference in conformation is seen in the 73-90 loop. This loop protrudes from the body domain into the basic groove (Fig. 34A). Another loop, located opposite in the structure, (comprised of residues 169-175) also protrudes into this groove from the head domain. These two loops contain a number of residues that together have been shown to be critical to RNA binding (Ng et al., 2008). This region could not be resolved in the WSN NP structure (PDB: 2IQH), but was partially resolved in the HK97 structure (PDB: 2Q06) (missing density for residues 79-86) and fully resolved in the WSN R416A NP structure (PDB: 3ZDP) (Fig. 34B). In the NT60 R416A NP structure residues 82-86 are observed to extend further away from the body domain of NP than in the WSN R416A NP. Residues 88-90 adopt a similar confirmation in the NT60 R416A NP and WSN R416A NP structures, but differ in the HK97 structure, with Pro89 flipped away from the body domain. The 73-81 region of the loop appears more structurally conserved, with all three structures displaying a very similar fold.

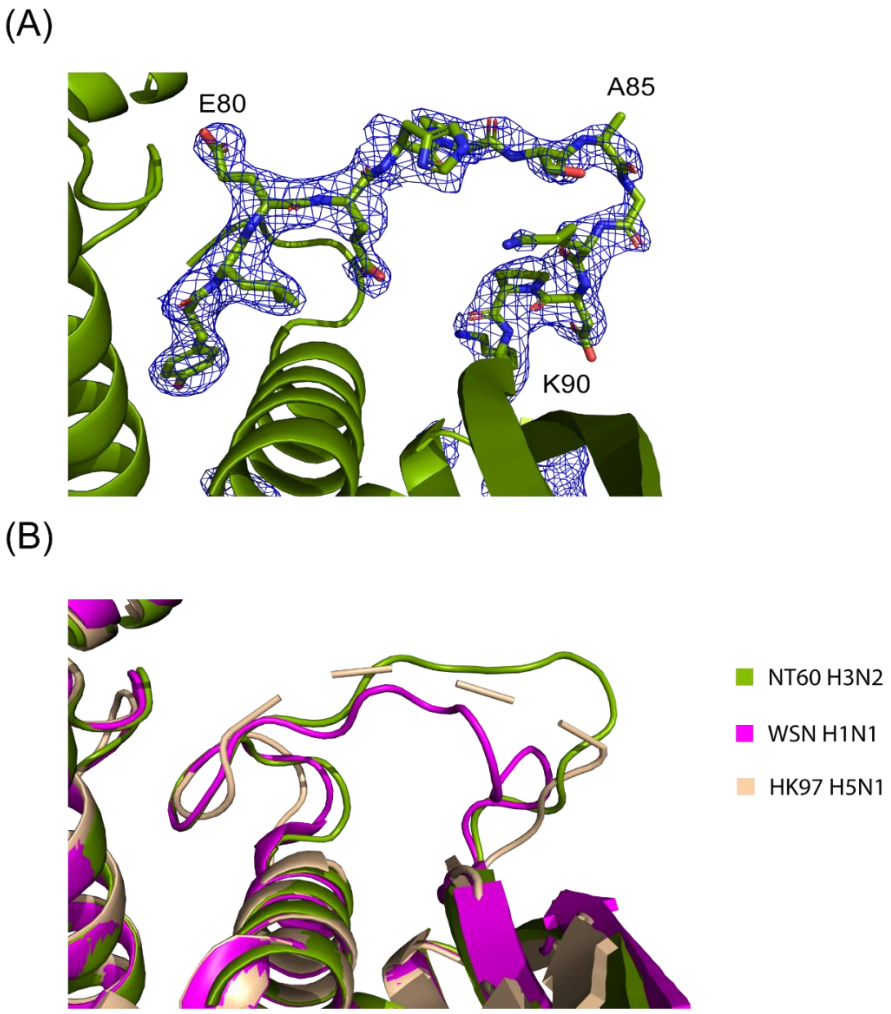


Figure 34: The NP 73-90 loop.
(A) The electron density (shown at level 1.0) of residues 78-90. (B) A comparison of the structure of the 73-90 loop in the NT60 R416A, WSN R416A (PDB: 3ZDP) and HK97 (PDB: entry 2Q06) NPs. Adapted from (Knight et al., 2021).

Crystallisation trials were also performed for NP in the presence of nucleic acids to determine the mode of RNA binding. Trials were attempted with a 14-nucleotide RNA with sequence 5′-GUAUAUGAGGCCCA-3′. The sequence of this RNA was designed by selecting a region of the vRNA (nucleotides 607-620 of the PR8 M segment) that had been shown to have high levels of NP binding in PAR-CLIP experiments (Williams et al., 2018b). This RNA had also been shown to increase the Tm of NP (Fig. 31B), which has been shown to increase the likelihood of crystallisation (Dupeux et al., 2011), and 14-nucleotide RNA had been shown to remain associated with NP through SEC (Section 5.2.1). Trials were also performed using a

14-nucleotide DNA (same sequence except U's replaced with T's), a truncated 8-nucleotide version of the RNA, and a 12-nucleotide poly-A RNA.

Crystals diffracting to high resolution formed, however in no instance could nucleic acid be resolved in the NP structures produced. The crystals produced gave a different space group ($C2\text{-}2\text{-}2_1$) to the NP crystals produced in the absence of nucleic acid (Fig. 35). The best diffracting crystal was produced in a 1.7 molar excess of 14-nucleotide DNA (Tables 6 and 7). These crystals diffracted to a resolution of 2.3 Å with 3 NPs present in each asymmetric unit. No regions of NP were resolved that were not already determined in the $P1\text{-}2_1\text{-}1$ structure (all chains were missing residues: 1-20, 82-87, 389-417, 432-435, 452-461, and 498). The regions encompassed by these missing residues have an average IUPred score (Dosztányi, 2018) of 0.59, indicating that these regions are likely to be disordered. Despite the different crystal packing, the structures produced are extremely similar to that of the NP structure observed in the $P1\text{-}2_1\text{-}1$ space group, with a comparative RMSD value of 0.8 Å (across 429 pairs).

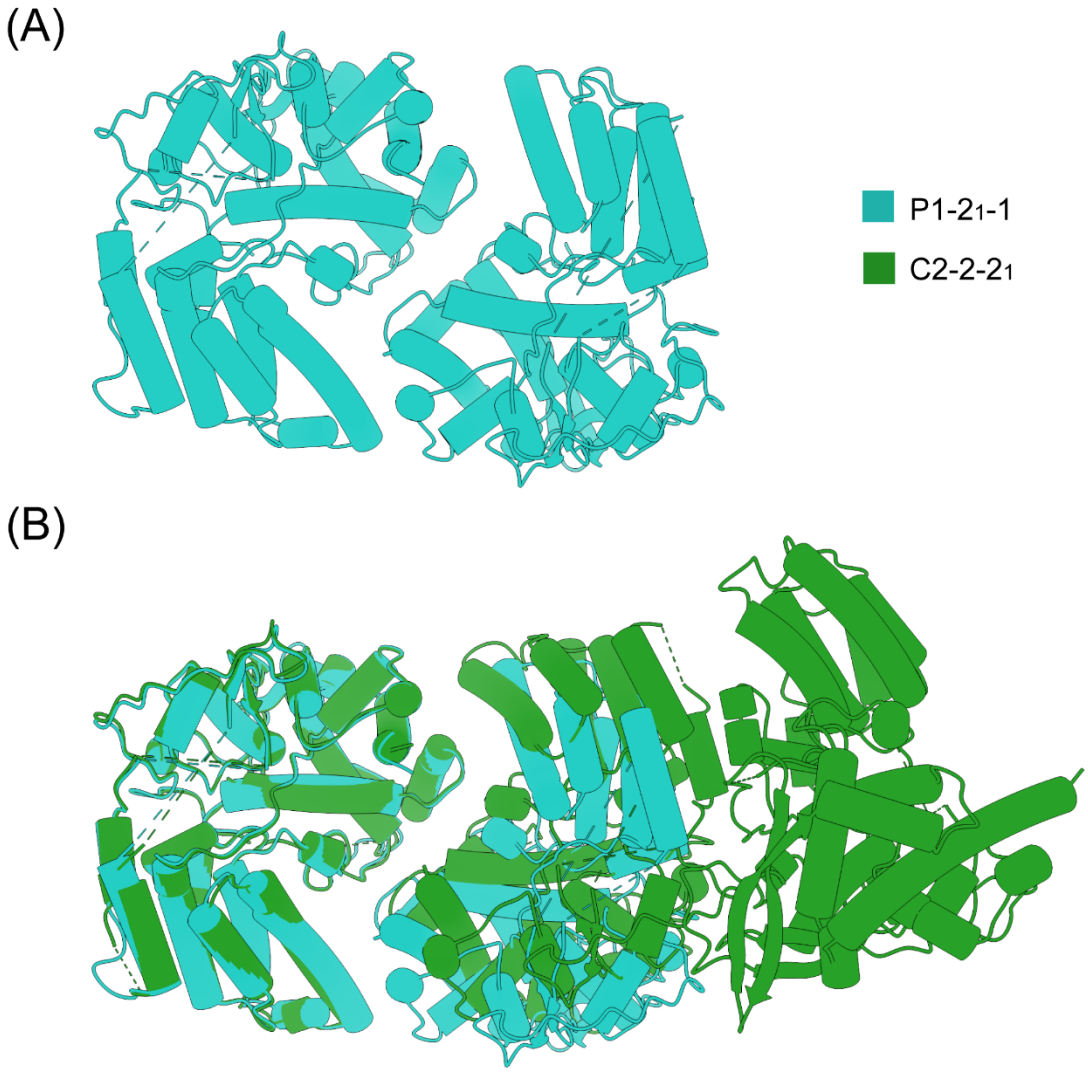


Figure 35: Comparison of R416A NP space groups.
(A) Cartoon representation of the unit cell of the NT60 R416A NP in the P1-2_1-1 space group (B) Comparison of the unit cells for the NT60 R416A NP in the P1-2_1-1 and C2-2-2_1 space groups.

Many further attempts were made at crystallisation trials with nucleic acids. This included: performing fine screens around conditions that had successfully produced crystals, soaking RNA into already formed crystals, crystallisation at 4°C, and performing an additional round of SEC post-mixing with nucleic acid (confirming nucleic acid bound by 260/280). Attempts were also made using other nucleic acids including: individual nucleotides, di-nucleotides, and 5-nucleotide RNAs in high molar excesses. Whilst some of these conditions produced crystals, none gave structures where nucleic acid could be resolved. Attempts were also made with the

non-monomeric version of the NT60 NP. No crystals were produced, although trials were more limited and had to be performed at lower protein concentrations, due to poor protein expression. Following these attempts, it was decided to pursue non-crystallographic methods of structure determination.

5.3.3 Cryo-electron microscopy

After being unable to resolve the structure of NP in complex with nucleic acid by crystallography, it was decided to try Cryo-EM approaches. Cryo-EM typically requires complexes >200 KDa. NP is 56 KDa, so an oligomer of NP, or multiple monomeric NPs bound to the same nucleic acid, would likely be required to achieve a high resolution structure. The *E. coli* expressed NT60 R416A monomeric NP was mixed with 47-nucleotide RNA or 105-nucleotide DNA. A 105-nucleotide DNA was used as this had been shown to produce a complex with NP in excess of 200 KDa (Fig. 31). The R416A NP with 105-nucleotide DNA formed visible complexes (Fig. 36A) on Cryo-EM grids (Cryo-EM performed by Jeremy Keown and Loic Carrique, University of Oxford). However, the complexes were extremely heterogeneous in shape, forming string-like structure of varying length and morphology. This heterogeneity made classifying and resolving the complexes impossible. `

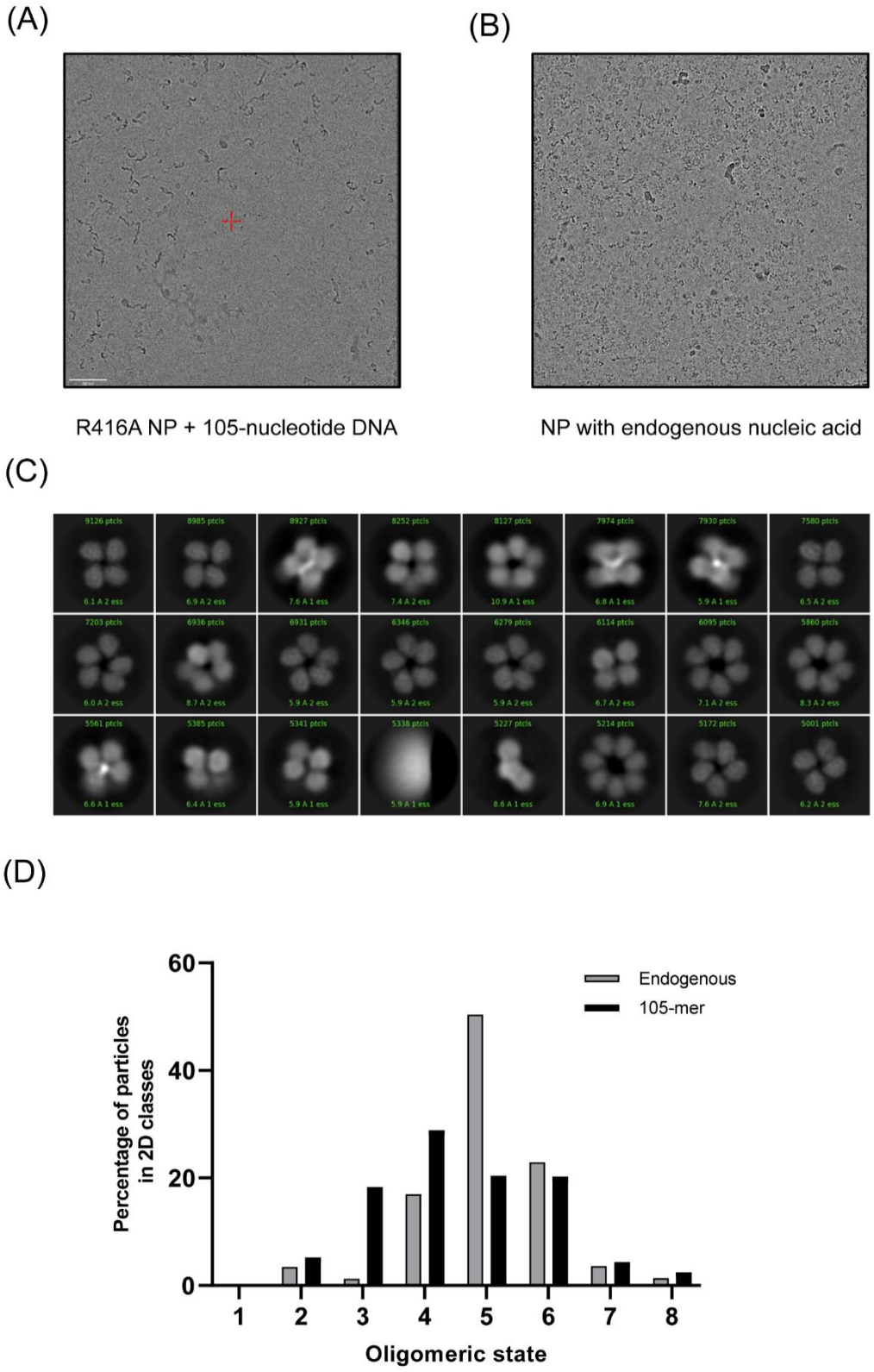


Figure 36: Cryo-EM of the NT60 NP.
(A) Cryo-EM analysis of NT60 R416A NP mixed with a 105-nucleotide DNA. (B) Cryo-EM analysis of NT60 NP (i.e. non-monomeric) with endogenous nucleic acid. Cryo-EM was performed by Jeremy Keown and Loic Carrique (C) A selection of 2D classes from analysis of NT60 NP with nucleic acid from expression host (presumably of mixed length). (D) A comparison of the oligomeric states of NP formed either with endogenous nucleic acid from the expression host or a 105-nucleotide DNA. Stacked rings were excluded from the analysis.

It was subsequently decided to use the NT60 NP (i.e. not the monomeric mutant), to try to form more homogenous complexes. Cryo-EM grids were made for NP that was not stripped of endogenous nucleic acids from the expression host (260/280 of the sample was 1.30). The NP on these grids formed ring-like oligomeric structures (Fig. 36B). A 2D classification of these complexes (Fig. 36C) showed that multiple oligomeric states of NP were present ranging from dimers through to septamers. There also appeared to be classes in which two pentameric NP rings were stacked, one on top of the other.

A wide range of 2D classes makes achieving high resolution difficult, as there is not a large amount of data for any given 2D class. It was decided to try to bias the oligomerisation of NP to a particular class by stripping it of endogenous nucleic acid (260/280 of 0.75) and adding a 105-nucleotide DNA (in a 3:1 molar ratio of NP to DNA). This was also performed at two different NaCl concentrations (150 mM or 300 mM), as there is evidence that salt concentration may alter the oligomeric state of NP (Labaronne et al., 2016). Screening of the grids indicated that salt concentration did not greatly affect oligomeric state and so data was only collected on the 150 mM NaCl sample. A wide range of oligomeric states were observed in the 2D classification (Fig. 37). The distribution of oligomeric states did differ slightly from the NP with endogenous nucleic acid, in particular there was an increase in the abundance of trimers observed (Fig. 36D). Most notably, none of the stacked ring structures were observed in the 105-nucleotide dataset. Without multiple repeats it is hard to attribute differences in oligomeric state to the nucleic acid present. There may have been slight differences in sample preparation that could also have affected the oligomeric state.

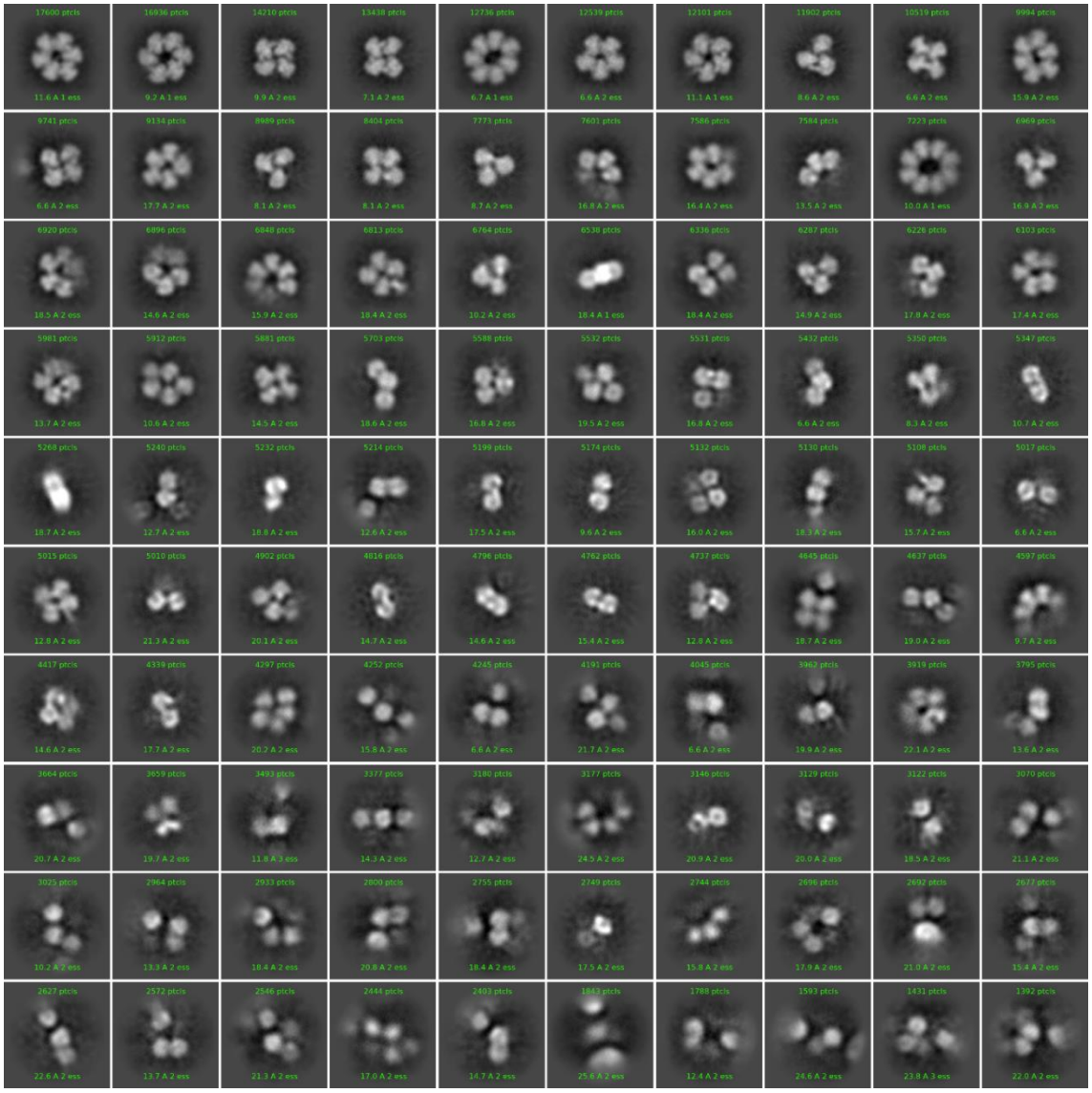


Figure 37: 2D classes from cryo-EM analysis of NT60 NP with 105-nucleotide DNA.

It was attempted to fit the NT60 R416A NP crystal structure to the density of one of the tetrameric classes from the Cryo-EM dataset of NT60 NP with 105-nucleotide DNA. The crystal structure could not be unambiguously fitted to the model as the resolution is too low and there is no secondary structure present in the map to aid fitting (it was also not possible to resolve nucleic acids). The oligomeric structures formed have a high degree of flexibility in the positioning of the NPs relative to one another meaning that high resolution information cannot be averaged out from the data. This flexibility meant applying internal symmetry to the ring structures during processing was ineffective at increasing resolution.

5.4 Discussion

The structure of the NT60 R416A NP has been resolved at 2.2 Å resolution, the first structure of a NP from a H3N2 virus. However, the structure of NP in association with RNA could not be resolved. Since the conclusion of this work crystal structures were obtained for a monomeric HK97 NP (H5N1) (with C-terminal truncations) in complex with 8 and 9 nucleotide 2'-O-methylated RNAs (Tang et al., 2021). The authors were able to resolve 3 nucleotides and found that they mainly formed interactions with residues R65, R74, R75, and K87. The R74 and R75 residues are part of a loop (the 73-90 loop) that was seen to be the main difference in structure between the NT60 R416A NP presented in this investigation and the other NP structures. Deletion of this loop in HK97 has been shown to produce a 6.4-fold drop in RNA binding affinity (Ng et al., 2008) and deletion of the equivalent loop in influenza B virus (residues 125-149) produced a 14-fold drop (Ng et al., 2012).

The 73-81 region of the loop shows high conservation amongst the non-RNA bound structures, adopting a very similar fold in R416A NT60, R416A WSN, and HK97 (Ng et al., 2008) (Tang et al., 2021). In contrast, in the RNA bound structure the 73-81 region is found to be orientated downwards, away from the opposite loop containing residues R174 and R175 (Tang et al., 2021). This opens up the RNA binding groove and is suggested to be required in order to physically provide the space required for the RNA to fully enter the groove. In the HK97 NP the mutation of the R74 and R75 residues, in concert with the mutation of the R174, R175, and R221 residues located on the opposite side of the groove, led to almost complete abolition of RNA binding (Ng et al., 2008), further demonstrating the importance of the 73-81 region in RNA binding. The 82-86 region is observed to project further away from the body domain of NP in the NT60 R416A NP structure presented here, compared to the WSN R416A NP structure (not resolved in HK97). This region could not be resolved in any of the other structures.

An examination of over 4,000 cross-species IAV NP sequences found residues Glu73, Arg74, Arg75, Asn76, Tyr78, Glu80, Pro83, Lys87, Asp88, Pro89, and Lys90 of the 73-90 loop to be highly conserved (Kukol and Hughes, 2014). Of these residues Asn76, Tyr78, Asp 88 and Pro89 were found to be invariant. It should be noted that 59% of residues across the entire structure were classified as highly conserved, compared to 61% in this loop, so overall it is not strikingly more conserved than other parts of the NP at a sequence level. However, many of the non-conserved residues are clustered in the 82-86 region loop where the structure is seen to vary between the NT60 and WSN R416A NP structures. Only one highly conserved residue is located in this region (Pro83). Two of the residues, Ser84 and Ala85, were found to be highly variable in sequence. So far, the 82-86 residues have only been resolved in monomeric structures. The C-terminal tail of the NP in the R416A structures inserts into the basic groove, close to this loop, and may stabilise its structure. As such, it is possible that this region of the loop is more flexible in the oligomeric protein and will not strictly adopt the confirmations seen in the monomeric structures.

The 73-81 region of NP may present a target for universal interventions against influenza. This region contains residues essential for RNA binding and appears to be conserved, both at a sequence and structural level. However, so far drugs blocking RNA binding have focused on the 169-175 residues (Hu et al., 2017). These residues are located in a loop situated opposite the 73-90 loop in the RNA binding groove. This loop is highly conserved in the structure presented here and all of the available IAV NP structures (including RNA bound). In addition, all but one of these residues (R174 - which is moderately conserved) are highly conserved (Kukol and Hughes, 2014). Preliminary experiments have shown drugs targeting this site can inhibit influenza virus replication (Kakisaka et al., 2015) (Hu et al., 2017).

A putative universal vaccine was designed by targeting regions of the M1, M2, NP and PB1 proteins that were found to be both highly conserved and to contain multiple reactive T-cell epitopes (Stoloff and Caparros-Wanderley, 2007). This vaccine has been shown to provide cross-protective immunity in ferrets (McMahon et al., 2019). One of the epitopes used in this

vaccine is the 255-275 region of IAV NP. This region encompasses part of the head domain and a linker connecting to the body domain. The structure presented here and the other available IAV structures show very high structural conservation in this region, supporting its use as a universal intervention target. Overall the data presented here, showing a high degree of structural conservation, along with the high sequence conservation and the large number of interactions it forms (including with host proteins) (Gabriel et al., 2008) (Eisfeld et al., 2015), supports the case for using NP as a target for influenza interventions.

In this investigation crystallographic approaches failed to resolve RNA in complex with NP, despite NP crystals being formed in the presence of RNA. It seems likely that NP crystals were only achieved in conditions in which the NP-nucleic acid association was disfavoured. The resolution of 3 residues was achieved using a monomeric mutant with a C-terminal truncation (Tang et al., 2021). This may suggest that achieving an overall more stable complex is required to resolve RNA. The authors suggest that the inability to resolve more than 3 residues may be down to flexibility in their positioning. Future attempts at resolving further residues using crystallisation could seek to utilise RNAs containing regions of secondary structure with short single stranded regions. This may promote NP binding to RNA in a single non-flexible mode, although NP is known to melt RNA structures (Dadonaite et al., 2019). Crosslinking an RNA to NP could also be attempted, such as by using an RNA containing 4-thiouridine. However, crosslinking efficiencies are low and it is still possible that the RNA could be crosslinked into the binding site in multiple different positions, making the sample too heterogeneous to crystallise.

Cryo-EM approaches to solving the NP-RNA structure also proved unsuccessful. Such approaches require larger complexes than crystallography. Binding multiple NPs to the same RNA formed oligomers with no defined shape at a high resolution. The non-monomeric protein formed ring like structures, similar to those observed in a previous study (Gallagher et al., 2017). The 2D classifications also contained stacked ring structures, which have not previously been reported. The biological significance, if any, of these ring structures is not

known. NP in vRNPs is thought to form a helical structure with ~10-12 NPs per turn (Arranz et al., 2012) (Coloma et al., 2020), although electron tomography reconstructions suggest that regions containing ring structures may also be present with the vRNPs (Gallagher et al., 2017). It is possible that a mixture of ring and helical NP structures allow for more overall flexibility in the vRNP structure.

The large range of 2D classes for the oligomeric NP highlights the high degree of flexibility NP is afforded in oligomerisation. This is supported by the range of rotation angles observed between NP monomers in the Cryo-EM reconstruction of vRNPs (Arranz et al., 2012) (Coloma et al., 2020) and the differences in positioning of the oligomerisation loop relative to the main body of NP observed between the WSN and HK97 oligomeric crystal structures. In addition, NP-binding on vRNAs has been shown to be non-uniform in nature (Williams et al., 2018b) (Lee et al., 2017) and electron tomography suggests that vRNPs have high structural variability, with kinks and regions of local unwinding (Gallagher et al., 2017) (Coloma et al., 2020). A high degree of flexibility in NP oligomerisation may be advantageous to influenza, perhaps in accommodating variable RNA structures.

None of the oligomeric ring structures could be solved to high resolution, with flexibility between the NPs within NP oligomers hindering averaging and application of internal symmetry during processing. Application of symmetry has been a key in solving the nucleocapsid structures of other viruses by Cryo-EM, such as for the Nipah (Ker et al., 2020) and measles (Desfosses et al., 2019) viruses. To resolve RNA binding, use of monomeric NP may be required to reduce heterogeneity. Future attempts could overcome this lack of size by binding NP to a megabody. Megabodies are nanobodies that have been grafted onto protein scaffolds. They are used to add mass to proteins that are too small to be resolved by Cryo-EM alone. They can also help to overcome preferential orientation of proteins on Cryo-EM grids (Uchański et al., 2021), as was the case with the NP oligomers (i.e. the vast majority of views are looking down on rings). In addition, they can help to stabilise proteins in a particular conformation, potentially further reducing heterogeneity (Uchański et al., 2021). This could

also be performed on NP with C-terminal truncations as proved successful in crystallography (Tang et al., 2021). Other possibilities include trying larger RNP complexes containing other proteins, such as the trimeric influenza polymerase, in addition to RNA. This would add mass and may lead to the formation of more homogeneous structures.

6. The structure of the SARS-CoV-2 genome

6.1 Chapter summary

- The genomes of viruses belonging to the *Coronaviridae* family are believed to contain a number of conserved functional RNA structures.
- Here the SARS-CoV-2 genome is subjected to chemical probing *in virio* to guide prediction of RNA structure.
- Extensive RNA structure is found to be present, spanning the entire length of the genome.
- Features conserved amongst coronaviruses are observed, including several stem-loops in the 5′ UTR.
- SPLASH experiments were also performed and no stable, specific, long range interactions were identified *in virio*.
- The RNA structures presented provide targets for future functional studies.

6.2 Introduction

During the course of this project a novel coronavirus, subsequently named Severe Acute Respiratory Syndrome-Coronavirus-2 (SARS-CoV-2), emerged in China. The disease caused by SARS-CoV-2 infection is called COVID-19 and can produce a range of symptoms, but most commonly fever, cough and loss/change of sense of taste/smell. Within months the virus had spread around the world and COVID-19 was declared a pandemic by the World Health Organisation. At the time of writing COVID-19 has been the cause of more than 2 million deaths and has produced massive economic disruption.

Coronaviruses infect a range of mammals and birds and are divided into four genera, the *alphacoronaviruses*, *betacoronaviruses*, *gammacoronaviruses*, and *deltacoronaviruses*. A number of viruses from the a*lpha* and *betacoronavirus* genera circulate in humans, generally producing mild symptoms and accounting for ~15% of cases of the common cold

(Chathappady House et al., 2021). The last 20 years have seen the emergence of three highly pathogenic *betacoronaviruses*, Severe Acute Respiratory Syndrome-Coronavirus-1 (SARS-CoV-1), Middle East Respiratory Syndrome-Coronavirus (MERS-CoV), and SARS-CoV-2. SARS-CoV-1 and MERS-CoV have largely been contained, producing only localised outbreaks (Haagmans et al., 2014). SARS-CoV-2 however has rapidly achieved sustained global transmission. Spread of the virus is thought to occur predominantly through airborne transmission (Greenhalgh et al., 2021) and contact with surfaces contaminated with respiratory droplets (Van Doremalen et al., 2020), though other routes of transmission may contribute (Xu et al., 2020).

SARS-CoV-2 is an enveloped virus with a ~30 kb capped and polyadenylated, positive sense RNA genome, which is encapsidated by the viral nucleoprotein (N). The viral Spike (S), Membrane (M) and Envelope (E) glycoproteins are all embedded in the viral membrane. The S protein consists of S1 and S2 domains. A polybasic cleavage site is located between the S1 and S2 domains and is cleaved by host cell furin proteases, shortly after the spike protein is produced (Hoffmann et al., 2020). The receptor binding domain of S1 binds to Angiotensin-Converting Enzyme 2 (ACE2), which is expressed on the surface of a wide range of cells, including in the lungs, blood vessels, gastrointestinal tract, and liver (Hamming et al., 2004) (Zou et al., 2020). Upon binding to ACE2 the S can then undergo cleavage at the S2′ site (note this is different to the S1/S2 cleavage) by another membrane protein TMPRSS2, which exposes the fusion peptide of the S2 domain, facilitating virus entry. Alternatively, the virus can be endocytosed upon ACE2 binding, with the S2′ cleavage then performed by cathepsin L, allowing the virus to escape the endosome and enter the cell (Shang et al., 2020).

The SARS-CoV-2 genome contains multiple Open Reading Frames (ORF). The first two of these open reading frames, ORFs 1a and 1b, overlap and account for ~2/3 of the genome, encoding 16 Non-Structural Proteins (NSPs). The structural proteins S, N, M and E, as well as a number of accessory proteins, are encoded in the 3′ third of the genome. Upon infection, ORF1a and b can be translated directly from the positive sense genome. ORF1b is produced

by a -1 ribosomal frameshifting event, allowing differential expression of the two polyproteins, pp1a (~450 KDa) and pp1ab (750 KDa) (Finkel et al., 2021). The polyproteins are cleaved by protease domains in NSP3 and NSP5. The viral RNA-dependent RNA polymerase is comprised of NSP12 and its two co-factors NSP7 and (two copies of) NSP8. This complex is responsible for both replication and transcription of the viral RNAs (Hillen et al., 2020) (Gao et al., 2020b). Infection by SARS-CoV-2 results in the formation of double membrane vesicles, likely mediated by the NSP3, NSP4, and NSP6 proteins (Angelini et al., 2013), which act as the site of viral replication and transcription (Zhang et al., 2021) (Snijder et al., 2020). Replication of the viral genome occurs via the production of a full-length negative sense intermediate, which can act as a template to produce more full-length genomic RNA.

Expression of the structural viral proteins requires the production of subgenomic RNAs (sgRNAs) containing the same 75-nucleotide 5′ leader sequence as the full-length viral genome. This first requires the production of negative sense versions of these sgRNAs via a discontinuous transcription mechanism that is mediated by Transcription Regulatory Sequences (TRS). The SARS-CoV-2 TRSs have the consensus sequence 5′-ACGAAC-3′, with the majority containing the extended consensus 5′-CUAAACGAAC-3′. Several TRS-Body (TRS-B) sequences are present in the 3′ third of the genome, located immediately upstream of their corresponding ORFs. During transcription the polymerase complex can undergo template switching and skip from the TRS-B sequence to the TRS-Leader (TRS-L) sequence which is located at the end of the 5’ leader sequence (V’kovski et al., 2021). This is thought to be mediated by binding between the complimentary sequences of the nascent negative sense TRS-B sequence and the positive sense genomic RNA TRS-L sequence. This results in the production of a set of at least eight negative sense genomic RNAs with a common 3′ sequence that can be transcribed into their positive sense counterparts by the viral transcription/replication complex (Kim et al., 2020). The positive sense RNAs produced are referred to as nested, in that they all contain the same 5′ leader sequence and 3′ UTR.

The mechanisms of viral budding are not well understood, but spike protein is thought to accumulate at the endoplasmic reticulum/Golgi network. The mature spike complexes are then thought to be transported in small vesicles to larger single membrane vesicles which contain the encapsidated genomic RNA. The virus may then exit the cell via exocytosis and/or use of the endosomal pathway (Ghosh et al., 2020) (Mendonça et al., 2021). Formation of virions is dependent on the presence of the S, M, and E proteins (Swann et al., 2020).

Several studies have investigated the minimal sequence requirements for replication in coronaviruses using DI RNAs. The minimal 5′ sequence was found to be 467 nucleotides for Mouse Hepatitis Virus (MHV) (Luytjes et al., 1996) and 498 nucleotides for Bovine Coronavirus (BCoV) (Chang et al., 1994), both encompassing part of ORF1a. The minimal 3′ sequence for DI replication only required the UTR (De Haan et al., 2002) (Goebel et al., 2004). Despite fairly low sequence conservation, there is a high degree of structural conservation present in the terminal regions across the *alphacoronavirus* and *betacoronavirus* genera (Madhugiri et al., 2014) (Chen and Olsthoorn, 2010) (Yang and Leibowitz, 2015). This structural conservation suggests that the secondary structures formed in these regions are important to the viral replication cycle.

The objective of this chapter was to identify RNA structures present in the SARS-CoV-2 genomic RNA. RNA structures in viruses can have important functions and show potential as drug targets. SHAPE and DMS-MaP were used to guide prediction of short range interactions whilst SPLASH was employed to look for longer range interactions.

6.3 Results

6.3.1 Chemical probing of the SARS-CoV-2 genome

A bio-safety level 3 facility must be used when working with SARS-CoV-2. This meant that ultracentrifuges were not available for purification and concentration of virions. Two alternative

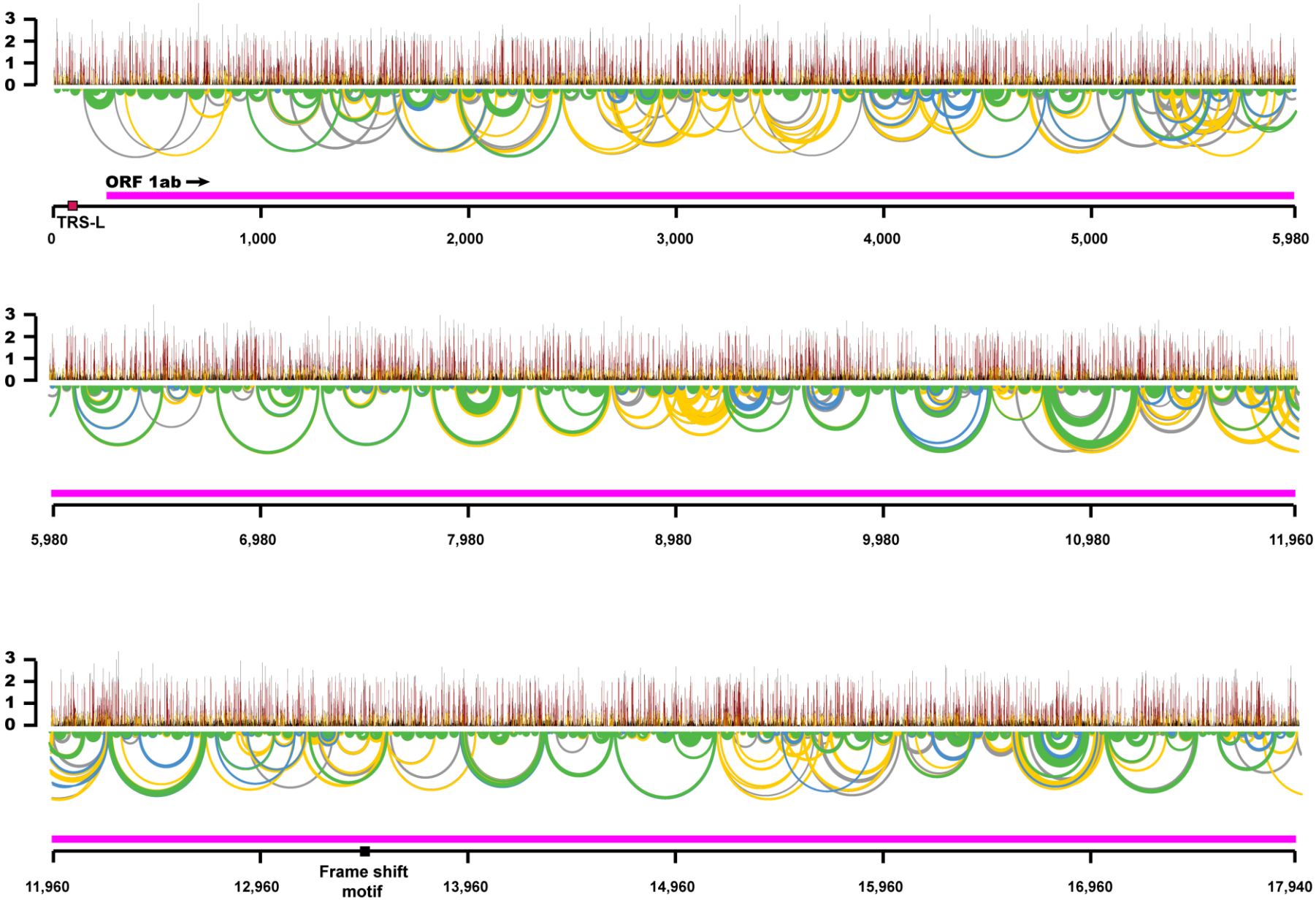
virion purification strategies were attempted. The first of these was PEG precipitation, a technique that has been successfully employed for a large number of viruses (Lewis and Metcalf, 1988). SARS-CoV-2 was amplified in Vero cells and the growth media then mixed with PEG overnight. The precipitated virus could then be pelleted by centrifugation and resuspended in a smaller volume. The precipitated virus was resuspended in one hundredth of the volume it was collected in and had a titre ~25 times greater after precipitation (average titre increased from $7.5x10^6$ to $1.85x10^8$ PFU)/mL.

Separately, purification of SARS-CoV-2 virions by centrifugation through a sucrose cushion was also attempted. Due to the lack of an ultracentrifuge, a lower sucrose concentration was used compared to when purifying influenza virions (10% vs 30%). SHAPE was performed on both sucrose cushion and PEG purified samples. In both cases median read depths were in excess of 20,000 for both 1M7 treated and untreated samples (Table. 8). A greater proportion of reads mapped to the SARS-CoV-2 genomic RNA in the sucrose purified samples (61%) compared to in the PEG purified samples (45%). Reactivity rates in the 1M7 treated samples were higher relative to the untreated samples in the sucrose purified virions (median reactivity rates 1.7 times higher in sucrose compared to 1.2 times in the PEG purified). The low reactivity rate of the 1M7 treated PEG purified virions meant that there was low signal to noise ratio resulting in poor correlation between replicates (Spearman R correlation of 0.60). The sucrose treated sample showed greater reproducibility with a Spearman R correlation of 0.88 between replicates. Given the higher signal to noise ratio, better reproducibility, and potentially higher sample purity (PEG may precipitate nucleic acids not contained in virions), the sucrose purified samples were used for RNA structure prediction. Tables containing full information on SHAPE reactivity values and pairing probabilities can be found online at https://figshare.com/s/6444d82a7bab5f8cbb74.

Sample	Median read depth	Median mapping quality (phred scaled)	Median base call quality (phred scaled)	Median read length	Median base mutation rate (%)
1M7-PEG-1	21,676	40.4	43.2	144.7	0.29
1M7-PEG-2	21,435	40.9	43.0	144.7	0.26
DMSO-PEG-1	19,400	40.8	46.2	142.4	0.22
DMSO-PEG-2	20,042	40.4	43.4	144.4	0.23
DMS-PEG-1	25,514	40.1	43.6	144.5	0.44
DMS-PEG-2	26,125	39.9	43.9	144.4	0.50
1M7-Sucrose-3	25,264	40.8	42.9	144.6	0.34
1M7-Sucrose-4	23,970	39.8	43.9	144.4	0.45
DMSO-Sucrose-3	28,716	40.7	42.7	144.9	0.19
DMSO-Sucrose-4	27,259	40.2	42.7	145	0.27
DMS-Sucrose-3	24,695	39.4	46.6	142.5	0.53
DMS-Sucrose-4	23,703	39.7	44.0	144.3	0.49
DMS-PEG-5	20,915	42.2	42.8	74.2	0.66
DMS-PEG-6	19,421	42.1	43.1	74.1	0.72
DMSO-PEG-5	17,958	42.5	42.4	74.2	0.30
DMSO-PEG-6	20,080	42.3	41.8	74.1	0.28

Table 8: Sequencing statistics from chemical probing experiments.
A table showing statistics on the SHAPE and DMS-Seq experiments on SARS-CoV-2 virions purified by PEG precipitation or centrifugation through a sucrose cushion. Samples use the naming system X-Y-Z. X = the reagent used for probing (DMSO = untreated). Y = the method of purification. Z = the biological replicate (i.e. all samples marked with a 1 are from the same batch of purified virions).

The SHAPE reactivity data was used as soft constraints to guide RNA structure prediction. The SARS-CoV-2 genome is too large to fold as a single entity, so a windowed approach to folding was adopted (Siegfried et al., 2014). Different folding windows (2,000, 3,000, and 4,000 nucleotides) were tested and this had little effect on the structures predicted, with less than 5% of bases adopting a different structure. Ultimately, a window of 3,000 was chosen as this gave the lowest mean reactivity for paired bases (0.132 vs 0.135 and 0.132 for the 2,000 and 4,000 windows respectively). It also gave the lowest percentage of paired bases with SHAPE reactivity values above 0.8 (2.36% vs 2.76% and 2.52% for the 2,000 and 4,000 windows respectively). The SARS-CoV-2 genome is highly structured, with regions of low reactivity spanning the entire length of the genome (Fig. 38).


3
2
1
0
ORF 1ab
TRS-L
0
1,000
2,000
3,000
4,000
5,000
5,980
3
2
1
0
5,980
6,980
7,980
8,980
9,980
10,980
11,960
3
2
1
0
11,960
12,960
Frame shift
motif
13,960
14,960
15,960
16,960
17,940

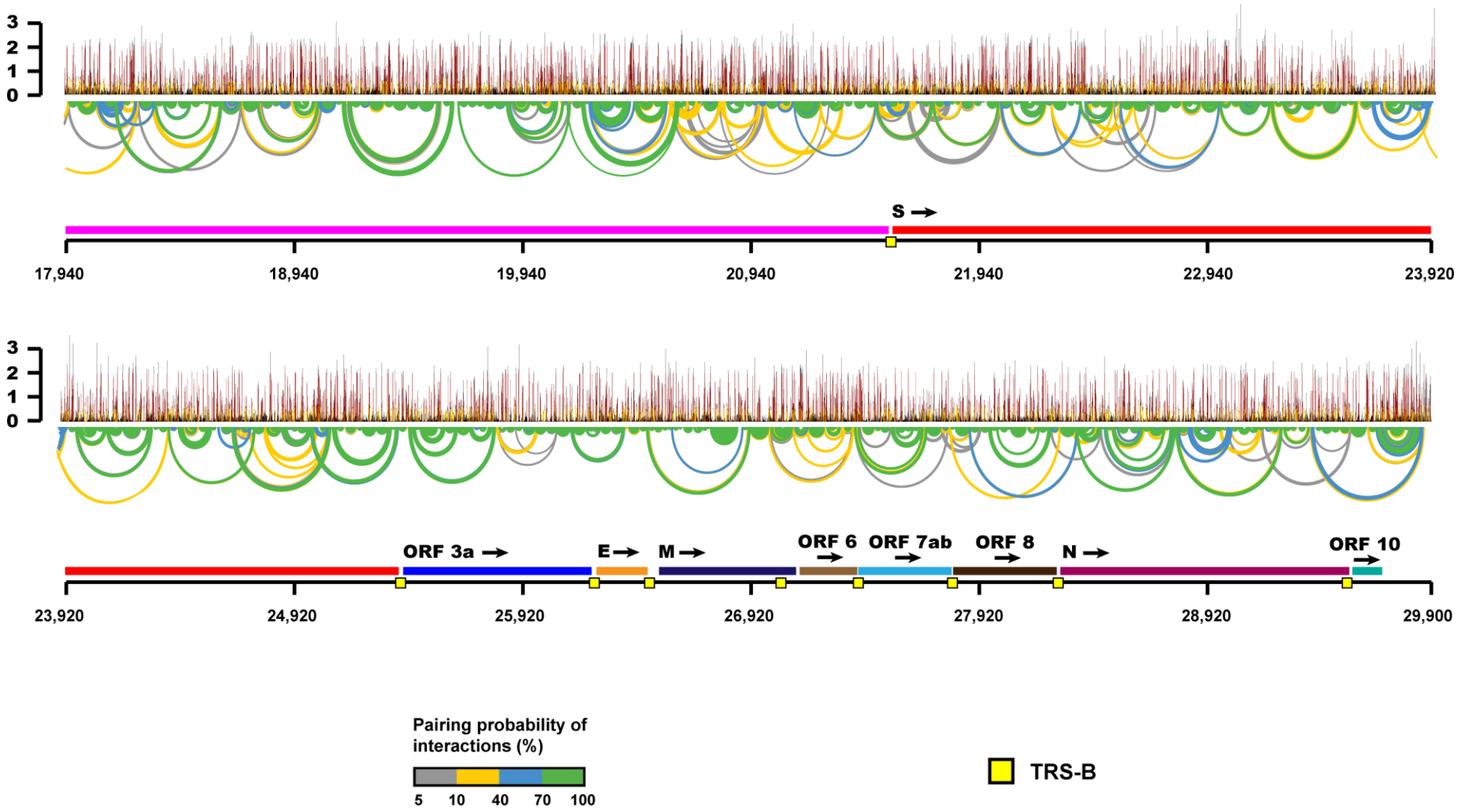


Figure 38: The reactivity profile from SHAPE performed on SARS-CoV-2 virions.
Predicted RMA interactions are shown as arcs with colour indicating pairing probability. The positions of protein coding regions are indicated along with the positions of the TRS sequences.

SHAPE guided structure prediction indicates that the 5′ and 3′ regions of the SARS-CoV-2 genome contain a number of stem-loop structures homologous to those identified in other coronaviruses (Fig. 39) (Chen and Olsthoorn, 2010) (Madhugiri et al., 2016). This includes SL1, SL2, SL4 and, SL5 which are widely conserved amongst *alpha* and *betacoronaviruses* (Fig. 40) (Madhugiri et al., 2016). SL1 begins just seven nucleotides into the SARS-CoV-2 genomic RNA and consists of upper and lower stem sections separated by a small bulge. SL2 is a very small structure with a 5 base pair stem enclosing a 5-nucleotide loop. This structure is the most conserved of the *Coronavirinae* subfamily, with the loop always having the consensus sequence CUUG(U/C) (Liu et al., 2007). The loop was shown to form a CUYG tetraloop fold in the nuclear magnetic resonance determined structure of this region for SARS-CoV-1 (Lee et al., 2011). The 5^{th} nucleotide of the loop was found to be flipped in the opposite orientation to the other four bases.

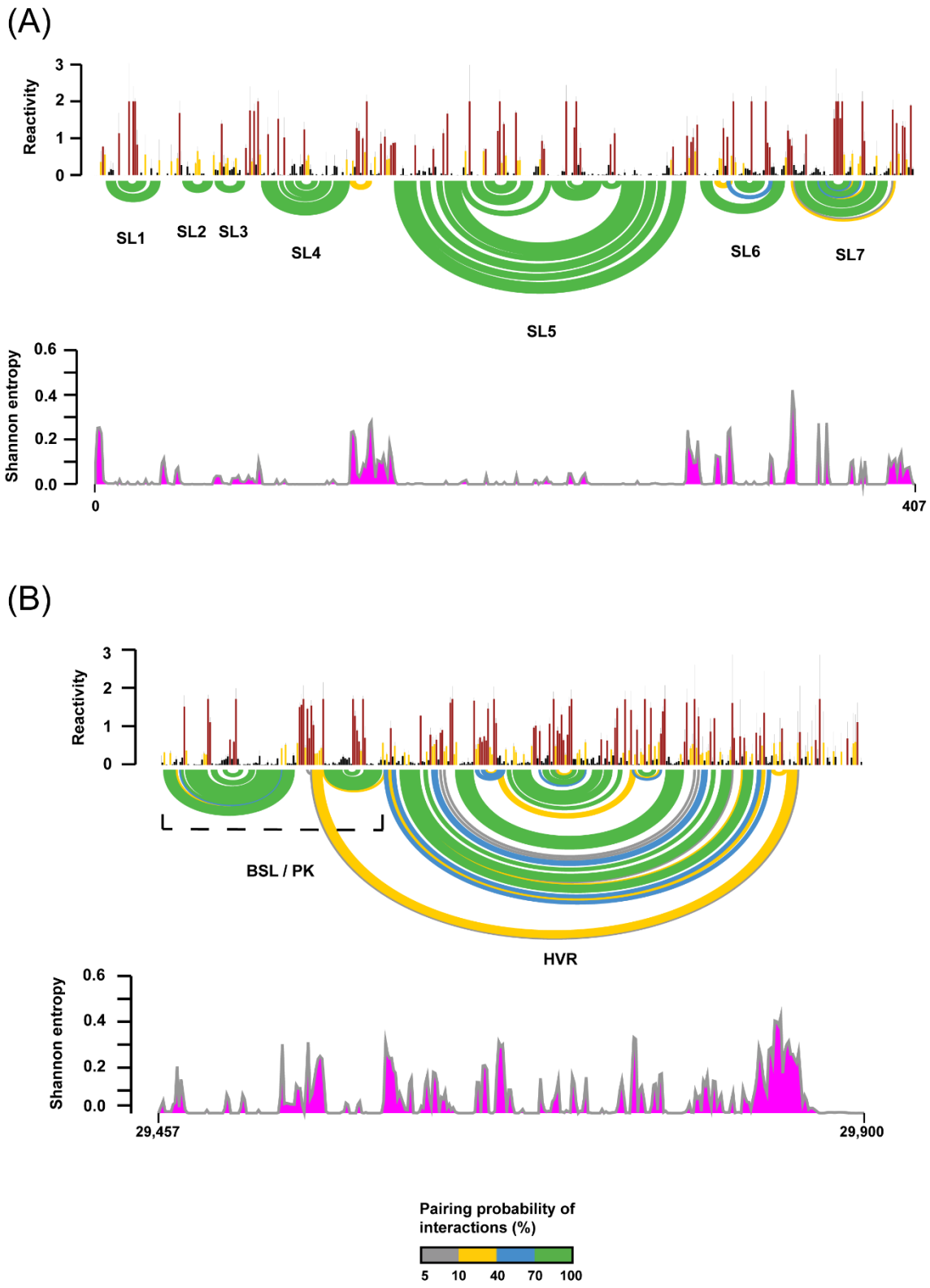


Figure 39: The 5′ and 3′ regions of the SARS-CoV-2 genome.
SHAPE reactivity profiles for the 5′ (A) and 3′ (B) regions of the SARS-CoV-2 genome. Arcs indicate interacting base pairs. Shannon entropy is also displayed, giving an indication of the likelihood of alternative structures forming in a region.

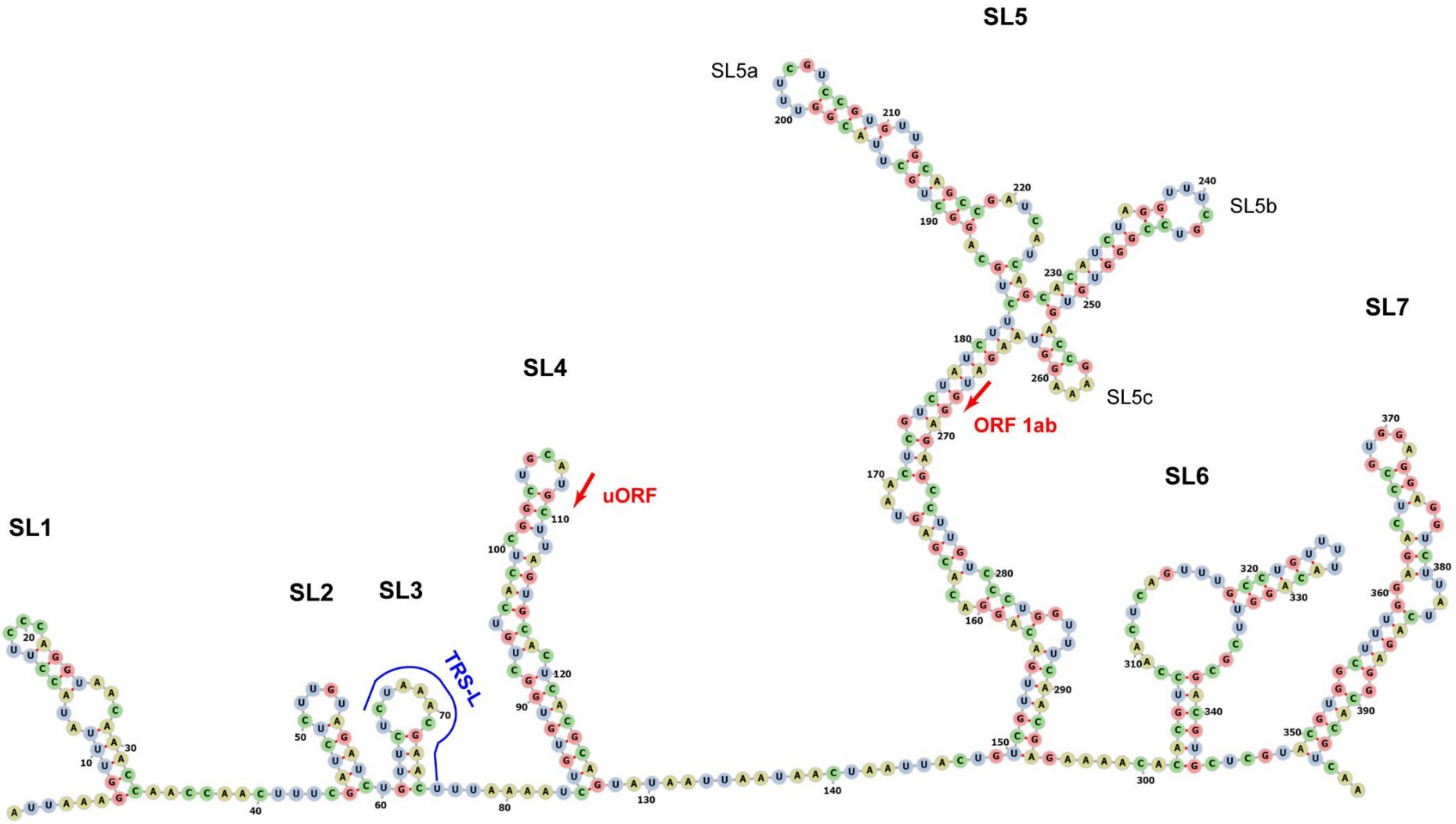


Figure 40: The SHAPE predicted structure for the 5′ region of the SARS-CoV-2 genome.

SL4 is a ~40-nucleotide stem-loop with two small bulges (Fig. 40). It contains the start codon for a short upstream open reading frame (uORF) encoding an 8 amino acid polypeptide. The start codon for ORF-1ab is located in SL5, a large multi-partite stem-loop structure spanning almost 150 nucleotides. This structure is conserved across many coronaviruses and is often divided into SL5a, SL5b, and SL5c, based on its three loops (Fig. 40) (Madhugiri et al., 2016). SL5a and SL5b both have loops with the sequence 5′-UUUCGU-3′, whilst SL5c forms a GNRA tetraloop. Downstream of this are SL6 and SL7 which are conserved amongst some *betacoronaviruses* (Madhugiri et al., 2016).

A short stem-loop (SL3) is predicted to encompass the TRS-L sequence (Fig. 40). This marks the end of the leader sequence which is present at the 5′ of the nested sgRNAs. This region is predicted to be single stranded in a number of coronaviruses, including MHV and MERS-CoV (Chen and Olsthoorn, 2010) (Madhugiri et al., 2014). However, the SL3 structure is predicted to form in SARS-CoV-1 and BCoV (Yang and Leibowitz, 2015). The TRS-L and TRS-B regions exhibit sequence complementarity to facilitate discontinuous replication (Sola et al., 2005). There are 9 TRS-B sequences located in the SARS-CoV-2 genome (Fig. 41). Six of these are predicted to occur on the 5′ side of stem-loop structures. The ORF8 TRS-B is predicted to be in a single stranded region and the N TRS-B on the 3′ side of a stem-loop. There is a partial TRS-B sequence prior to ORF10 (5′-CUAAAC-3′) located predominantly within the bulge of a bulged stem-loop structure. However, it is not clear if ORF10 is actually expressed (Pancer et al., 2020) (Kim et al., 2020).

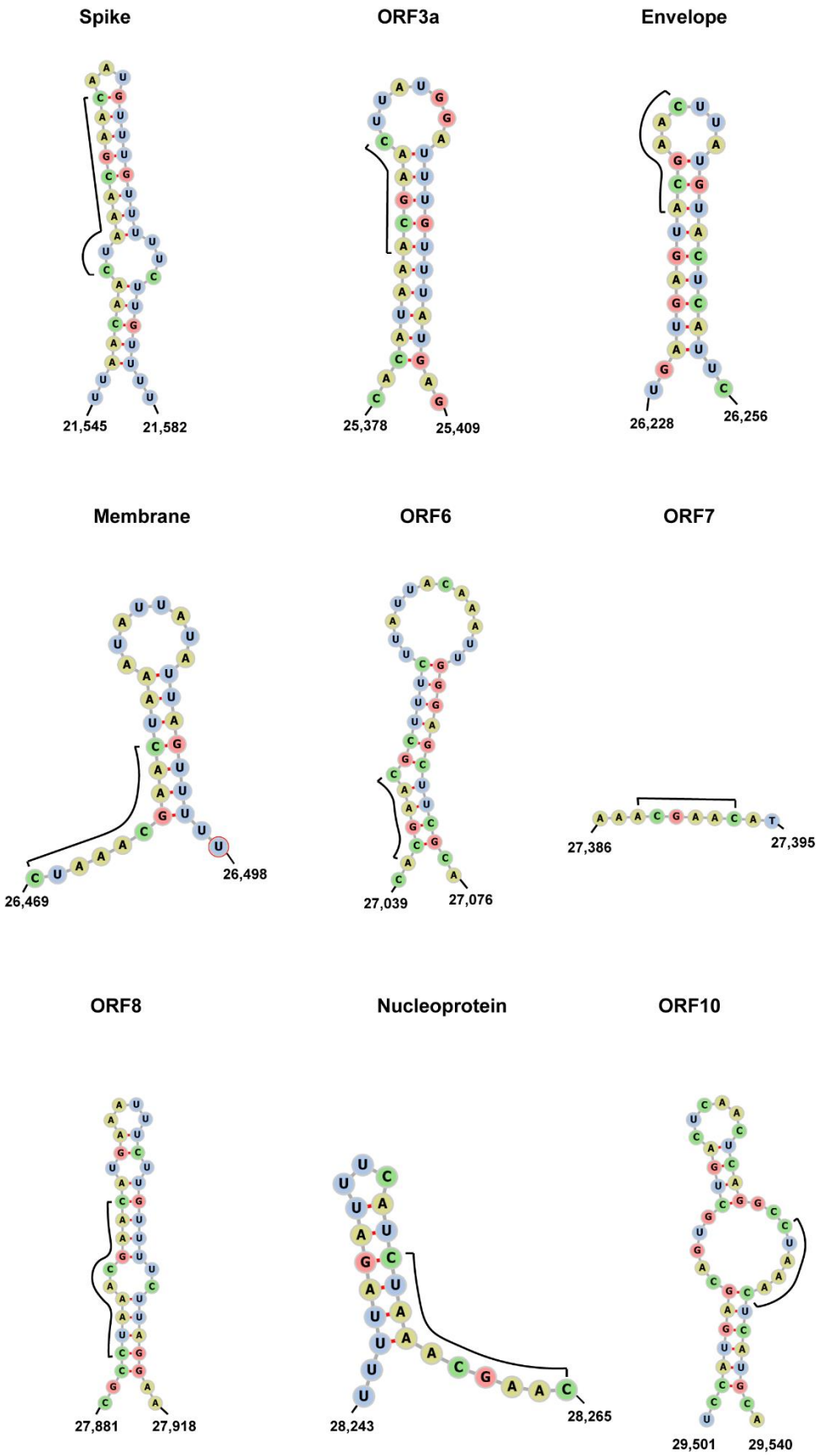
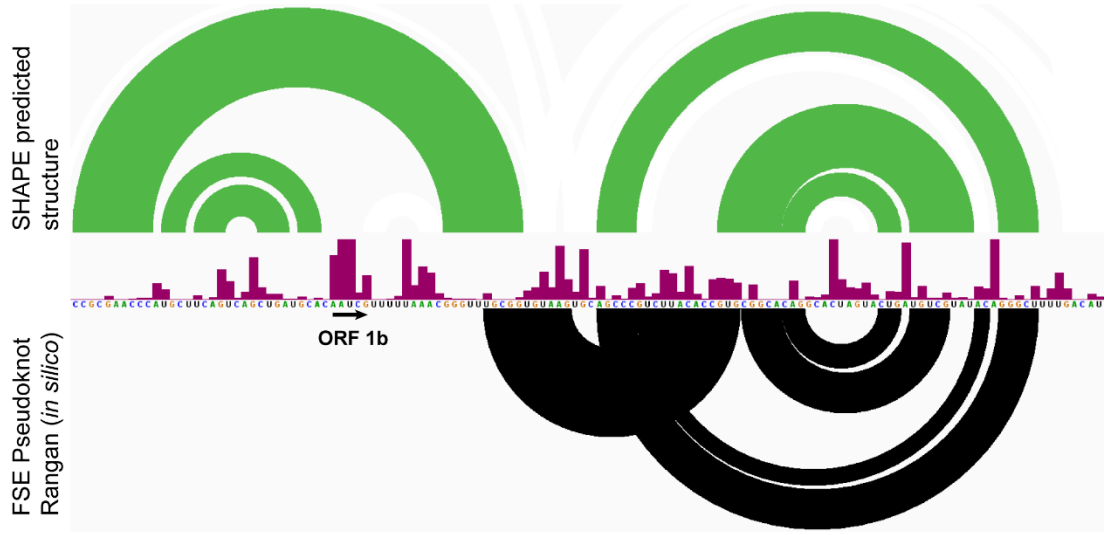


Figure 41: The structures encompassing the SARS-CoV-2 TRSs.
The SHAPE predicted structures for the regions encompassing the TRS-B sequences of the SARS-CoV-2 genomic RNA. Black lines indicate the TRS-B sequences.

A ribosomal frameshift is required for translation of ORF1b and occurs in a region known as the Frame Shift Element (FSE). This region contains a 5′-UUUAAAC-3′ consensus sequence that is proposed to be the site of ribosomal slippage in coronaviruses (Baranov et al., 2005). This ‘slippery’ sequence is proposed to be followed by a pseudoknot structure that mutational analysis in Infectious bronchitis virus (IBV) has indicated is essential for ORF1b expression (Brierley et al., 1989). Two different pseudoknots have been postulated to form in this region of SARS-CoV-2 (Rangan et al., 2020) (Huston et al., 2021). However, neither pseudoknot was supported by the SHAPE-informed *in virio* structure prediction (Fig. 42). Instead a bi-partite bulged stem loop is predicted to form (Fig. 43). The slippery sequence is predicted to be single stranded, located in the bulge separating the upper and lower stems. The upper stem and loop region is known as the attenuator stem and has been shown to be important to frameshifting in SARS-CoV-1 (Cho et al., 2013).

(A)



(B)

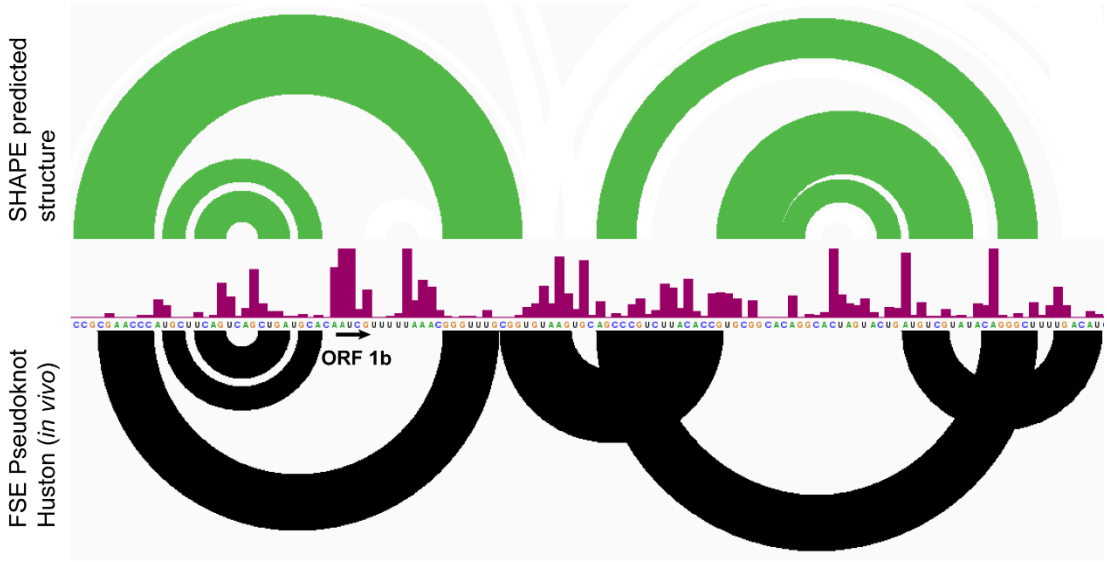
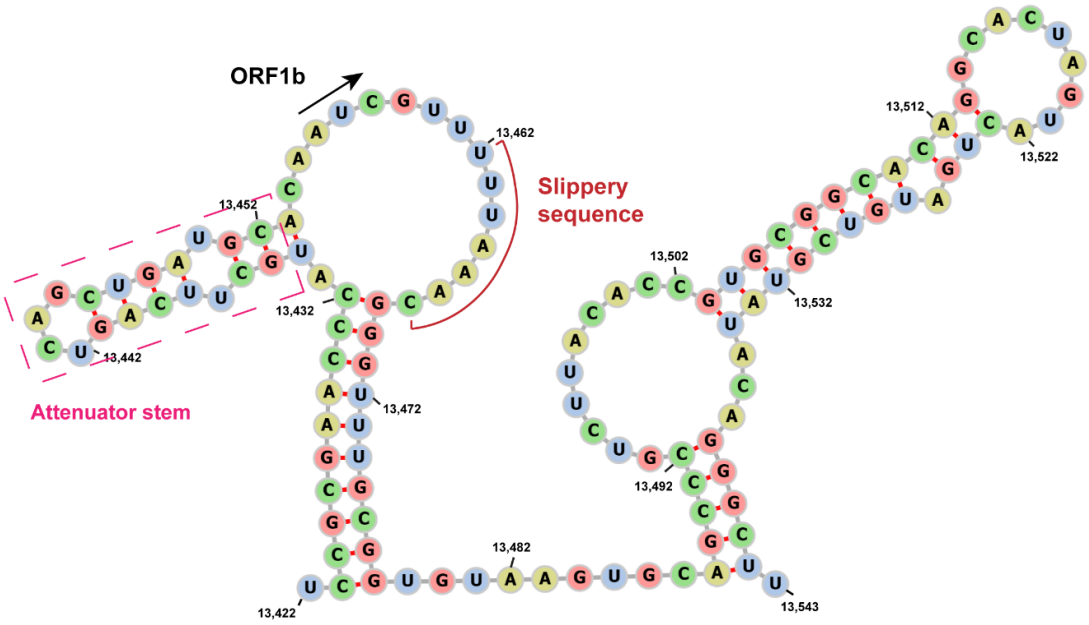


Figure 42: Comparison of predicted SARS-CoV-2 FSE structures.
The SHAPE reactivity values for the FSE region are shown. Green arcs on the top show the in virio *SHAPE-informed structure prediction for the region. The arcs in black beneath the reactivity profile show a pseudoknot predicted (A) computationally (Rangan et al., 2020) or (B) based on in cell SHAPE probing (Huston et al., 2021).*

(A)



(B)

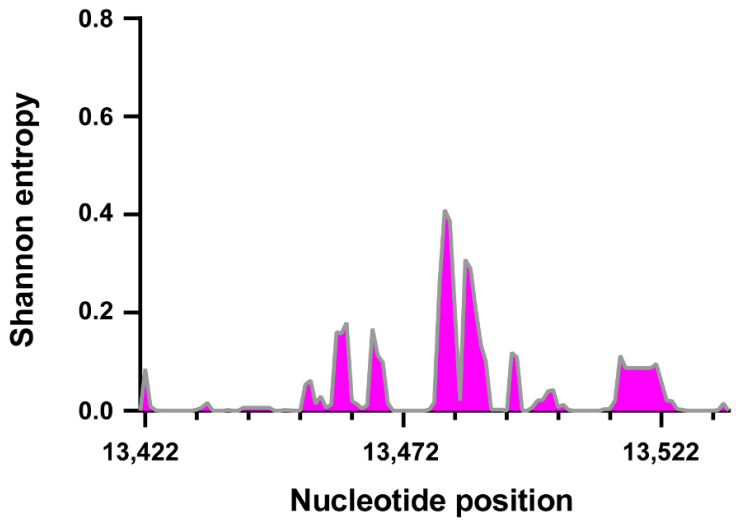
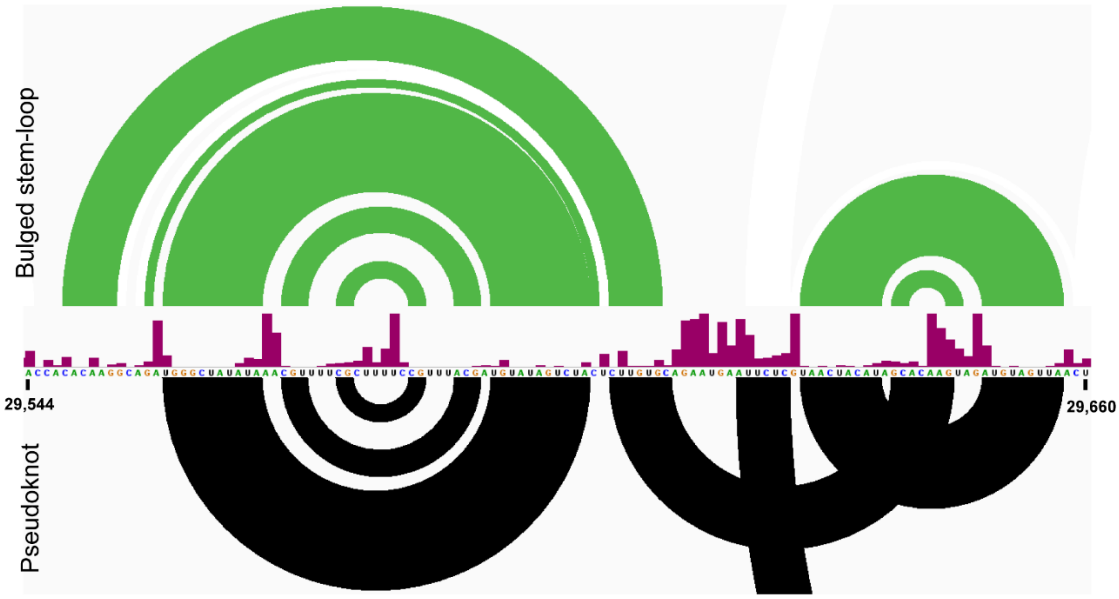


Figure 43: The structure of the SARS-CoV-2 FSE.
(A) In virio *SHAPE-informed structure prediction for the FSE of SARS-CoV-2. (B) The Shannon entropy for the SARS-CoV-2 FSE region, giving an indication of the likelihood of alternative structures forming in a region.*

The last ~350 nucleotides at the 3′ end of the SARS-CoV-2 genome are often described as the UTR (even though this region contains the putative ORF10). The 3′ most end of this region is predicted in *betacoronaviruses* to form a structure known as the Bulged Stem-Loop/Pseudoknot (BSL/PK). These two structures are mutually exclusive and cannot exist concurrently, with the bottom stem of the BSL required to unpair in order to form the PK. The requirement for both structures to be able to form has been supported by mutagenesis studies in MHV (Goebel et al., 2004) and by covariance analysis (Madhugiri et al., 2014). The SHAPE guided structure prediction indicates that the region exists, at least predominantly, in the BSL form in virions (Fig. 44). It has been proposed that the two conformations may both be adopted and that the region may act as a molecular switch (see discussion) (Goebel et al., 2004).

(A)



(B)

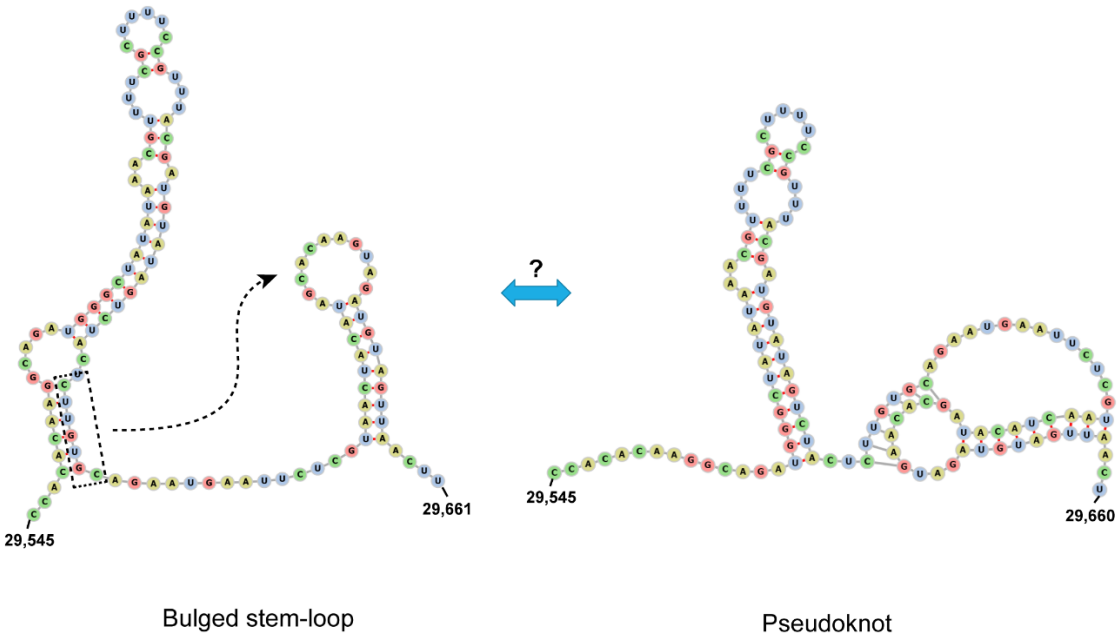


Figure 44: The structure of the BSL/PK.
(A) The SHAPE reactivity values for the BSL-PK region are shown. Green arcs on the top show the in virio *SHAPE informed structure prediction for the region. The arcs in black beneath the reactivity profile show a pseudoknot predicted computationally (Rangan et al., 2020). (B) The BSL (left) and PK (right) RNA structures which have been proposed to act as a molecular switch.*

The 3′ most structure in the SARS-CoV-2 genome is known as the Hypervariable Region (HVR). SHAPE-directed structure prediction shows this region forms a multipartite stem loop that spans almost 120 bases (Fig. 45). The top part of the HVR structure contains a Stem–Loop II-like Motif (S2M) seen in astroviruses and equine rhinovirus (Jonassen et al., 1998). The nucleotide sequence of the S2M is highly conserved amongst coronaviruses with 75% of the residues invariant. The highly conserved octanucleotide sequence (5'-GGAAGAGC-3') is located towards the 3′ end of the HVR. This sequence spans a single stranded bulge (5 of the bases) and the edges of two stems enclosing the bulge.

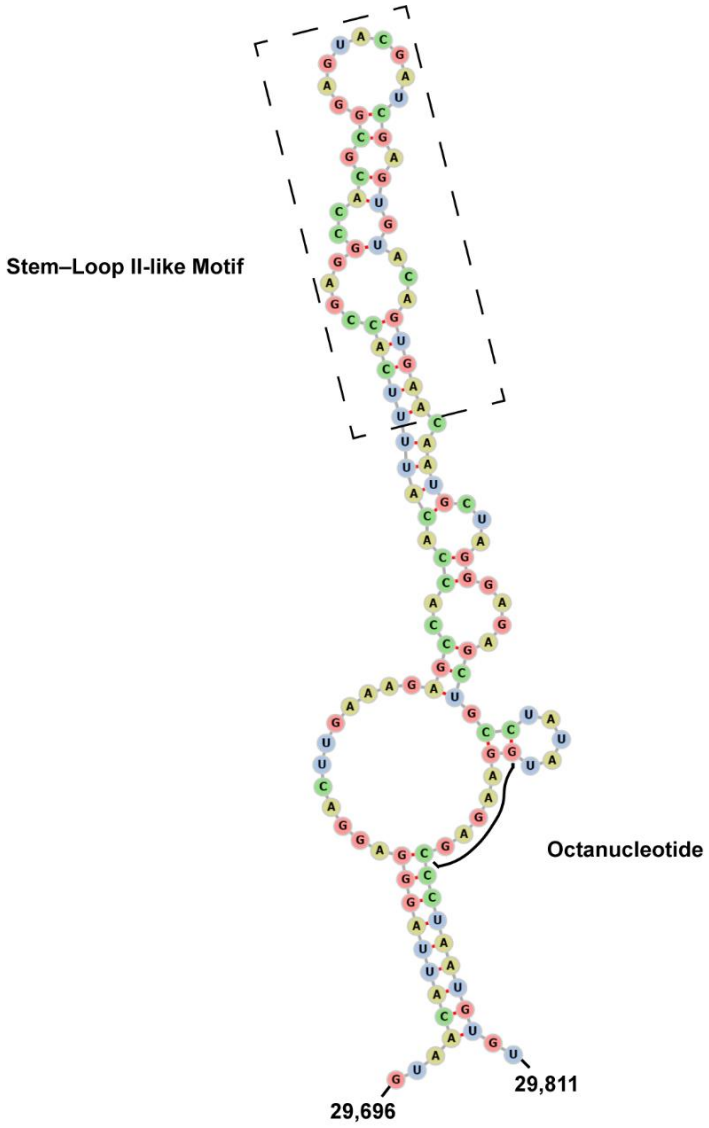


Figure 45: The structure of the SARS-CoV-2 HVR.

DMS probing of SARS-CoV-2 was also performed in order to support the SHAPE data. PEG- and sucrose -purified samples showed good correlation (Spearman R correlation values between 0.93-0.98). Mutation rates in treated samples showed median mutation rates on average 2.2 times higher than the untreated controls (Table. 8). A further two replicates were performed on PEG-purified sample in which the DMS concentration was doubled, with these resulting in mutation rates 2.4 times higher than the untreated samples. The DMS reactivity rates were used as soft constraints (for A and C nucleobases only) for RNA structure prediction. The resulting structures provided strong support for those predicted using the SHAPE reactivity data (Fig. 46). The only changes seen were for a number of the longer-range interactions predicted and the 3′, 5′, and FSE region structures were supported.

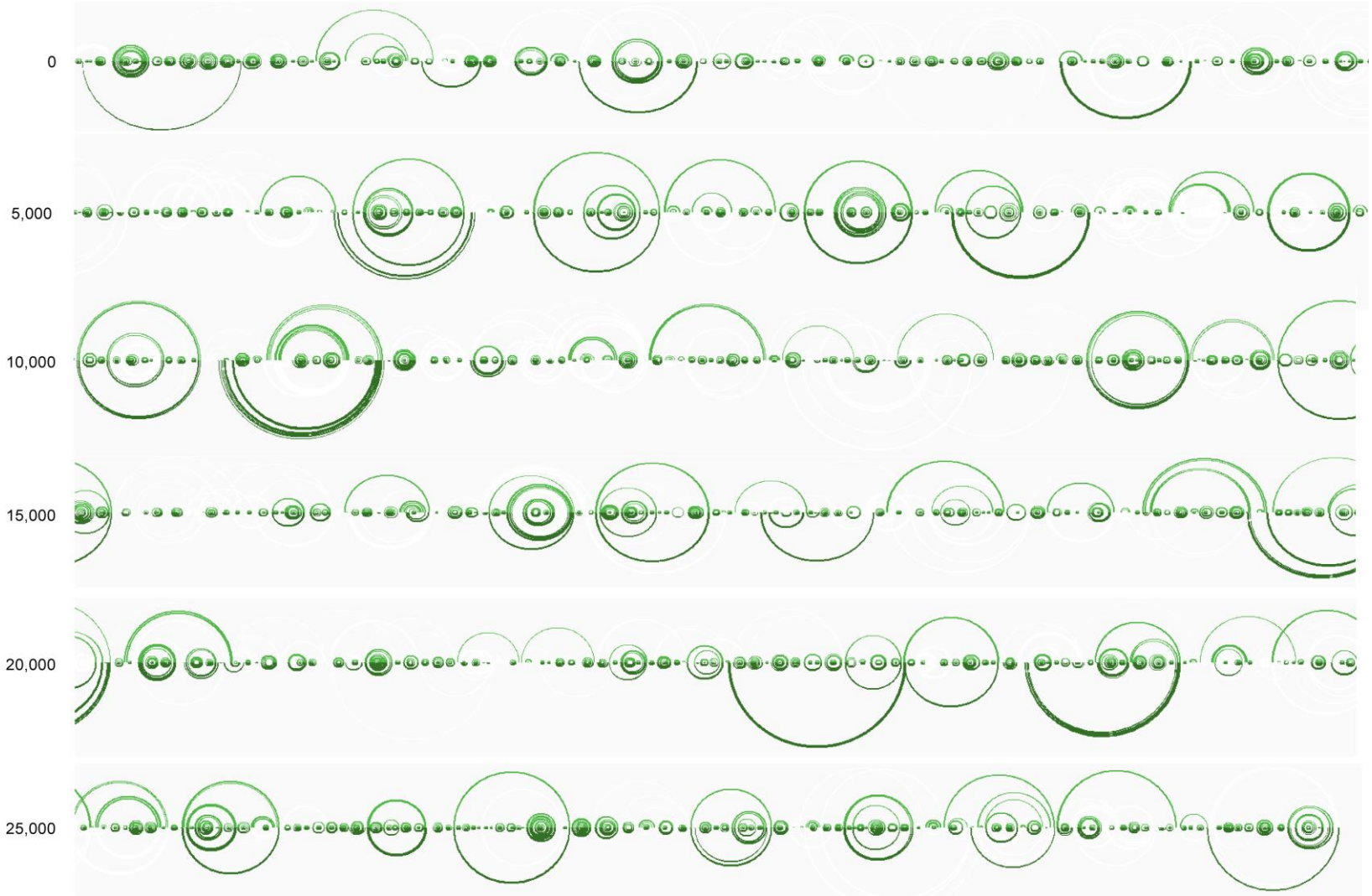


Figure 46: Comparison of SHAPE and DMS-informed structure prediction.
Figure 9: *A comparison between the RNA structures predicted for the SARS-CoV-2 genome when SHAPE (the top arcs in light green) or DMS-reactivity data (the bottom arcs in dark green) were applied as soft constraints. Numbers indicate the position in the genome.*

6.3.2 Long range RNA-RNA interactions in the SARS-CoV-2 genome

Chemical probing is limited in the distance of interactions it can accurately identify due to the increased false positive rate as the maximum pairing distance increases. This can make structure determination of longer RNAs very challenging. In order to investigate the existence of long range RNA-RNA interactions in the SARS-CoV-2 genome SPLASH was performed on SARS-CoV-2 virions that had been purified through a 10% sucrose cushion (Fig. 47A). Surprisingly, no distinct interaction loci were observed. This suggests that stable long range interactions are not present in *virio* as the dominant conformation. Certain regions do appear to have a higher propensity to be involved in long range interactions, as seen by the disproportionately high number of chimeric reads mapping to them (Fig. 47B). Examples of this include the 2,400-2,500, 6,400-6,500, 21,000-21,100, and 25,500-25,600 regions. These regions may be free to transiently form interactions with more distant partners due to lower local structural stability (although this does not appear to be supported by the SHAPE data). It could also be that these regions are arranged in a way inside virions that gives them a greater likelihood of interaction with other regions (e.g. maybe they are central in an arrangement or are free from N binding).

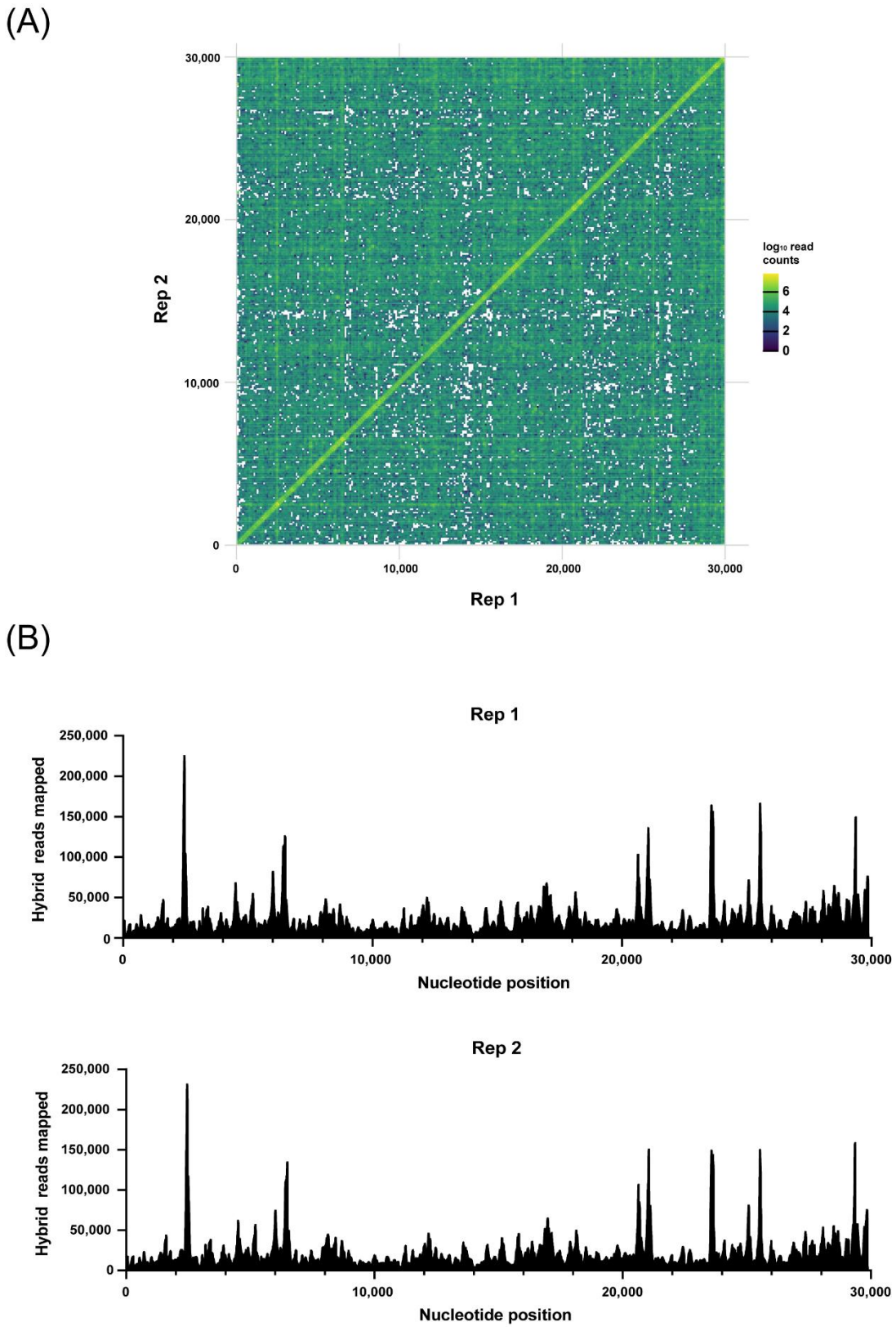


Figure 47: Long Range RNA-RNA interactions in SARS-CoV-2
(A) The interaction matrix for the two replicates of SPLASH performed on SARS-CoV-2 in virio*. Interactions are displayed on a log_{10} scale. (B) Graphs showing the number of hybrid reads from the SPLASH experiment mapping to different regions of the SARS-CoV-2 genome. This illustrates that certain regions of the genome appear to have a higher propensity to be involved in long range interactions.*

6.4 Discussion

Chemical probing was used to guide RNA structure prediction showing that there is extensive RNA structure across the SARS-CoV-2 genome in *virio*. Many of the structures observed in the terminal regions are highly conserved across *betacoronaviruses* and in some cases, the entire *Coronavirinae* subfamily (Madhugiri et al., 2016). The structure presented here for the 5′ region of SARS-CoV-2 *in virio* is more or less identical to those predicted by *in silico* prediction (Andrews et al., 2020) (Rangan et al., 2020), *ex virio* SHAPE experiments (Sanders et al., 2020), in cell SHAPE experiments (Sun et al., 2021) (Manfredonia et al., 2020) (Huston et al., 2021), and in cell DMS probing (Lan et al., 2021).

SL1 is the 5′ most RNA structure in the SARS-CoV-2 genome and is a bi-partite stem-loop with the two stems separated by a bulge. In MHV mutations disrupting the upper stem have been shown to be lethal or highly detrimental to virus replication, whilst mutations in the lower part are more well tolerated (Li et al., 2008b). Lethal mutations were shown to fail to produce negative sense sgRNA, suggesting a role for this structure in discontinuous replication. Mutations that increase the stability of the lower portion of the stem have been shown to be lethal or to lead to compensatory destabilising mutations in MHV (Li et al., 2008b). Concurrent mutations were also seen in the 3′ UTR. This supported a previously proposed hypothesis that long range interactions between SL1 and the 3′ UTR promote the discontinuous replication required for sgRNA synthesis (Zuniga et al., 2004). The lower part of the SL1 stem may have an optimal level of stability that allows it to unfold to form transient long-range interactions. However, no long range interaction was reported for the SL1 of SARS-CoV-2 in a study using the RNA-RNA interaction capture technique COMRADES (Ziv et al., 2020) nor in this study by *in virio* SPLASH. It is possible that this interaction is a rare event and so was not captured by COMRADES/SPLASH. SL1 is also essential in ensuring that expression of viral proteins is not inhibited by NSP1 (which suppresses expression of host proteins) (Yuan et al., 2021). This is mediated by an interaction between the NSP1-ribosome complex and SL1 (Tidu et al., 2021).

SL2 is the most conserved structure in the 5′ UTR of the *Coronavirinae* subfamily (Chen and Olsthoorn, 2010) (Madhugiri et al., 2016). Mutational studies in MHV indicate stem mutations can be tolerated as long as base pairing is maintained (Liu et al., 2009). Mutations of the loop in MHV were found to be tolerated to varying degrees, with the exception of the guanine (nucleotide 53), mutation of which was lethal to the virus. As with SL1, lethal mutations in SL2 prevent sgRNA production (Liu et al., 2007), though the nature of its role in replication is equally unclear. COMRADES identified an alternative conformation for this region in cells in which SL2 and SL3 unfold to pair with a region of ORF1a almost 4,000 nucleobases distant (Ziv et al., 2020). The COMRADES study also found that SL3 can unfold to interact with the 3′ termini, suggesting that genome circularisation may be mediated by SL3 in SARS-CoV-2 instead of SL1. Given that SL3 contains the TRS-L sequence required for discontinuous transcription, it is possible that the 3 different states of the SL3 region (SL3/interaction with ORF1a/interaction with the 3′ termini) may act to regulate production of sgRNAs.

SL4 contains a short upstream ORF (uORF) encoding an 8-residue polypeptide. This polypeptide is not essential for replication in MHV or BCoV, but is positively selected for during passage in tissue culture (Raman et al., 2003) (Wu et al., 2014). Disruption of the SL4 stem in BCoV DIs resulted in reduced RNA replication that could be partially recovered by double mutants with restored base pairing (Raman et al., 2003). Experiments in MHV suggest that SL4 can tolerate extensive mutation and viable mutants can be produced when it is replaced by a shorter stem loop of different sequence (Yang et al., 2011). The authors suggested that SL4 acts as a spacer that ensures the optimal orientation of the upstream stem-loops for discontinuous transcription.

SL5 encompasses the start codon for ORF1ab and its presence has been to be required for replication of BCoV DI's (Raman and Brian, 2005). It forms a large, partite, structure with multiple bulges and 3 loops that are often used to subdivide the structure into SL5a, SL5b and SL5c (Fig. 40). The loops of SL5a and b in many *alpha* and *betacoronaviruses* have the consensus sequence UUYCGU, which is supported by the structure present here (loop

sequence is UUUCGU) (Chen and Olsthoorn, 2010). Mutations that disrupt SL5a reduce replicative efficiency in MHV (Guan et al., 2012). It has been reported that 6 cellular proteins may bind to the BCoV SL5 (Raman and Brian, 2005) and it is possible that this consensus loop sequence may be involved in protein recruitment.

SL5c contains a GNRA tetraloop (GAAA). GNRA tetraloops are highly stable structures due to an unusual non-canonical base pair forming between the 5′ G and the 3′ most A residue of the loop, favourable base stacking, and possible base-phosphate hydrogen bonds (Heus and Pardi, 1991). They are highly conserved in nature and are suggested to act as nucleation sites to ensure the correct folding of RNA, which may be the case in SL5. It is also possible SL5c may have a role in protein recruitment, though this seems less likely as the loop is only present in a sub-group of *Betacoronaviruses* (Chen and Olsthoorn, 2010). A reverse genetics study in MHV showed that silent mutations (i.e. not changing the NSP1 protein sequence) disrupting SL5c resulted in only slight reduction in replicative efficiency (Yang et al., 2015). SL6 and SL7 are poorly studied, likely due to their lower conservation amongst betacoronaviruses compared to SL1-5 (Madhugiri et al., 2016). Mutational analysis has indicated that SL6 is not essential for MHV replication (Yang et al., 2015).

The 3′ region of the coronaviruses also contain conserved structural features. The first of these is the BSL-PK which has been suggested to operate as a molecular switch, though its function is not known (Goebel et al., 2004). The region has been shown to be required for MHV DI replication (Hsue et al., 2000). The results of this investigation suggest that *in virio* the BSL is the dominant structure. This is supported by *in vivo* SHAPE analysis (Sun et al., 2021) (Huston et al., 2021), whilst *ex virio* SHAPE analysis led to prediction of the PK structure. Thermal unfolding experiments of the MHV sequence have shown that the PK is not very stable and is only likely to form in a situation where the BSL cannot form (Stammler et al., 2011). In addition, the existence of the PK was not observed by COMRADES (Ziv et al., 2020). This suggests that the PK structure either does not form, or is only present as a low occupancy alternative structure.

The 3′ most structure in the SARS-CoV-2 genome is the HVR, a large hairpin containing the S2M and the extremely highly conserved octanucleotide sequence. A reverse genetics in MHV found mutations targeting the stems in this region resulted in reduced replicative efficiency, but were not lethal (Goebel et al., 2007). The octanucleotide sequence (5'-GGAAGAGC-3') is largely invariant across *Coronavirinae*, with only a few viruses reported that contain single nucleotide changes. The sequence is not required for *in vitro* replication of the MHV genome (Liu et al., 2001) but single point mutations were shown to lead to slower viral growth kinetics (Goebel et al., 2007). UV crosslinking experiments suggest that multiple different host proteins interact with the 3′ UTR (Sola et al., 2011), so it is possible the octanucleotide promotes one or more of these interactions. Ultimately, the function of the octanucleotide is unknown and requires further investigation.

The HVR also encompasses the S2M. The structure of the SARS-CoV-1 S2M has been determined by X-ray crystallography and was found to form a helical structure with a 90° kink, which facilitates the formation of an enclosed, negatively charged region, that binds two magnesium ions (Robertson et al., 2004). The structure is highly similar to that formed by the 530 loop of the 16S bacterial ribosomal subunit. This has led to the suggestion that it may play a role in hijacking the host translational machinery and/or forming an interaction with the NSP9 protein, which contains an oligomer binding fold similar to that of proteins bound by the 530 loop of 16S RNA (Robertson et al., 2004). The SHAPE determined *in virio* structure of the SARS-CoV-2 S2M presented here forms a similar structure to that in the SARS-CoV-1 crystal structure. The bottom most stem and bulge are the same but the top of the structure differs forming a 9-nucleotide loop, as opposed to the 5-nucleotide loop in the SARS-CoV-1 crystal structure (the sequence of the S2M of SARS-CoV-1 differs by 2 nucleotides). Of the 9 nucleotides in the loop SARS-CoV-2 S2M loop, 7 have high SHAPE reactivity values (0.8<), and the other 2 have moderate SHAPE reactivity values (0.4<X<0.8) strongly supporting the existence of this conformation. *In vivo* probing of SARS-CoV-2 suggests a slightly different arrangement of the top of the S2M, with a 6-nucleotide bulge (Huston et al., 2021). This could

indicate that there is a change in RNA conformation in cells. It may also be that a protein binds to the loop in cells, reducing accessibility of the SHAPE reagent, and that the structure does not change.

The SARS-CoV-2 FSE is responsible for the differential expression of ORF1a and ORF1b. This region has been predicted to adopt a pseudoknot conformation in many coronaviruses including SARS-CoV-2 (Rangan et al., 2020). This has been supported by Cryo-EM and SHAPE analysis of an 88-nucleotide *in vitro* transcribed RNA fragment (Zhang et al., 2021). A pseudoknot conformation was also proposed by the authors of one in cell SHAPE experiment (Huston et al., 2021). However, the *in virio* SHAPE data presented here suggests that this region forms a stem loop with the slippery sequence in a single stranded bulge. The lack of a pseudoknot structure as the predominant conformation is also supported in cells by SHAPE (Huston et al., 2021) (Sun et al., 2021) and DMS probing (Lan et al., 2021). *In silico* structure prediction also predicted that the non-pseudoknot conformation was the most stable structure for this region (Andrews et al., 2020).

The ORF1b start codon and slippery sequence are located in a large bulge of a bulged stem loop structure (Fig. 44). The stem-loop structure above this bulge is known as the attenuator stem. Placement of this stem relative to the ORF1b start codon has been shown to be critical to frameshifting in SARS-CoV-1 (Cho et al., 2013). The lower most stem beneath the bulge is predicted to be 4 base pairs longer based on in cell DMS probing (Lan et al., 2021). The same study used DREEM (Tomezsko et al., 2020) to look for alternative conformations based on correlated mutations. An alternative stem-loop structure was predicted to form in the FSE region, raising the possibility of a functionally relevant conformational switch.

In vitro DMS probing of shorter RNAs containing the SARS-CoV-2 FSE region supported formation of the pseudoknot structure (Lan et al., 2021). This resulted in much lower frameshifting rates compared to when a longer ~3,000 nucleotide sequence was used that formed the non-pseudoknot structure (17% vs 40%). This suggests the FSE structure tunes the differential expression of ORF1a and b, a factor which has previously been shown to be

important in coronavirus replicative efficiency (Plant et al., 2010). COMRADES indicated that the FSE is located within a region that is enclosed by base pairs spanning almost 1,500 nucleotides, which was termed the FSE arch (Ziv et al., 2020). This structure was not predicted by the SHAPE informed structure prediction even when the maximum pairing distance was set to 2,000 nucleotides, nor was it supported by *in virio* SPLASH. This may indicate that formation of the FSE arch is triggered by cellular factors.

The structures for the regions encompassing the TRS sequences were found to be the same in cells, with the exception of the N TRS-B which formed a slightly different hairpin (Lan et al., 2021). It is possible that the structure encompassing the TRS-L sequences help to determine differential expression of the different sgRNAs. It has been suggested that there is a correlation between single strandedness of a TRS sequence and the levels of expression of the corresponding sgRNA (Sun et al., 2021). It would be interesting to perform mutational analysis in which these structures were disrupted, or swapped, and to measure the subsequent effect on sgRNA levels and viral replication. Several host proteins from the heterogeneous nuclear ribonucleoproteins family, which are important for viral replication have been shown to be able to bind to TRS sequences in *vitro* (Shi et al., 2003). It is not clear whether structure may play a role in any possible protein recruitment.

The in cell COMRADES investigation found the presence of 41 interactions spanning more than 500 nucleotides that had greater than 500 reads mapping to them (a total of 246 interactions had more than 500 reads mapping to them) (Ziv et al., 2020). This is different to the results of the *in virio* probing in this investigation where no specific long-range interaction loci were identified. The lack of specific interaction loci may be partially supported by electron tomography data on virions showing (at low resolution) that there appear to be multiple conformations of the genomic RNA in virions (Yao et al., 2020). It is possible that this difference is biological and that more long-range interactions are present in cells. This could be due to interactions with host proteins, interactions with the viral NSPs, the phase separation of the genomic RNA at 37°C (Iserman et al., 2020), or an organisation of the genomic RNA

inside virions that makes these interactions less likely to form. It would be interesting to perform SPLASH in cells to see if it supported the findings of COMRADES and vice versa. SPLASH was recently performed on SARS-CoV-2 infected cells (Yang et al., 2021). The majority of reads mapped to close range interactions, suggesting the longer-range interactions were more transient in nature. They did identify a number of long-range interactions (23 interactions spanning 500< nucleotides with more than 50 reads mapping to them). The SPLASH study did not support formation of the FSE-arch seen in COMRADES.

Choosing the maximum allowed pairing distance for structure prediction is challenging for SARS-CoV-2. Increasing the maximum distance may allow capture of longer-range interactions but at the cost of increasing the false positive rate. A pairing distance of 500 was chosen to try to balance these two factors. The SPLASH data indicated that there were not stable long-range interactions *in virio*, which would provide support for using a shorter maximum pairing distance (~250). However, it seemed best to consider the SHAPE and SPLASH datasets independently. In addition, regions with seemingly higher propensity for long range interaction were observed in the SPLASH dataset. If these regions have resulting low SHAPE reactivity it may actually improve the accuracy of prediction if they are able to find more distant partners instead of being constrained to pair more locally. Using a pairing distance of 200 led to the maintenance of the vast majority of interactions predicted (Fig. S3).

Overall the SARS-CoV-2 genomic RNA is extensively structured *in virio*. Many of these structures are highly conserved across coronaviruses suggesting that they may be functionally important. Future research should focus on characterising these structures, as they present potential drug targets.

7. Discussion

The objective of this thesis was to investigate genome structure in RNA viruses. In influenza viruses extensive, redundant networks of interactions were identified between the different genomic segments. It has been demonstrated that these interaction networks can affect reassortment, a factor that may now be considered when trouble-shooting or optimising vaccine production. Intra-segment structure in IAVs was also investigated. Many structures were identified across all of the segments, though there was very limited conservation of these structures between viruses from different sub-types. The structure of a H3N2 influenza virus NP was also presented at 2.2 Å resolution. This protein appears to be highly structurally conserved across IAVs making it a prime target for universal interventions against influenza. Finally, extensive RNA structure was identified in the SARS-CoV-2 genome. Many of these structures are highly conserved amongst coronaviruses and are likely to be functionally important elements.

One of the things highlighted by this investigation is how differently SARS-CoV-2 and IAVs utilise RNA structure. The segmented influenza virus genome provides advantages to the virus in its ability to reassort and evade populational immunity. However, this comes at the cost of complicating genome packaging. The evidence presented here suggests that influenza has overcome this problem by mediating bundling through inter-segment interactions. There appears to be great flexibility in these interaction networks, likely to accommodate mutations and reassortment events.

SARS-CoV-2 does not have a segmented genome, but has a large number of RNA structures that are widely conserved amongst coronaviruses. This includes several structures in the terminal UTRs, as well as the FSE. This is in contrast to influenza where conserved intra-segment structural features appear to be few in number. It is possible that this relative lack of conserved elements may be a direct consequence of the need to maintain flexibility in the inter-segment interaction networks that facilitate genome bundling. The conserved structural

features identified in IAVs in this investigation are confined to segments that are also highly conserved at a protein level. It would be interesting to extend SHAPE analysis to influenza B, C, and D viruses to see if any of these structural features are conserved amongst the different *Orthomyxoviridae* genera, as well as to more distantly-related segmented RNA viruses (e.g. *Bunyavirales*).

Another interesting difference between SARS-CoV-2 and influenza is in their nucleoproteins. Studies determining RNA structure of *in vitro* transcribed SARS-CoV-2 viral RNA have shown that it forms largely the same structure as it does when coated with N (Manfredonia et al., 2020). By contrast the influenza NP protein seems to melt much of the structure that is present in *in vitro* transcribed influenza vRNAs (Dadonaite et al., 2019). The reason for this difference is not clear, but it is possible that the NP-mediated melting of local RNA structure is designed to increase the probability of inter-segment bundling interactions forming (which would be less likely if RNA is locally well folded). This is somewhat supported by a study on influenza inter-segment interactions suggesting that regions of the genome with high NP binding are more likely to be involved in inter-segment interactions (Le Sage et al., 2020) (though this finding was not supported by the data in this investigation).

The SHAPE directed structure prediction gave significantly lower Shannon entropy for SARS-CoV-2 (median = 0.04) compared to influenza (median = 0.11) (Mann-Whitney $P < 0.0001$). This suggests as a whole that the influenza genome is less likely to form a single defined structure. It would be interesting to see how the Shannon entropy would compare for in vitro transcripts of the influenza vRNAs to see if this is an NP-dependent phenomenon. It would also be interesting to perform SHAPE experiments on *in vitro* transcripts of the two viral genomes in the presence of the nucleoprotein of the other virus (i.e. influenza vRNA with SARS-CoV-2 N and SARS-CoV-2 genomic RNA with influenza NP).

The lack of perturbation of the SARS-CoV-2 RNA structure may equally be an intrinsic feature of the N protein. There have been suggestions that the amount of structure in the SARS-CoV-2 genome is in itself regulated in order to favour phase separation of the genome in infected

cells (Iserman et al., 2020). This raises the question of what the expected amount of structure is in an RNA of a given length. Performing MFE structural prediction on all possible RNAs of a given length would allow a probability distribution to be produced of the likelihood of any given amount of double strandedness. While not perfect, this may allow better comparison between the amount of RNA structure in different viruses (or a regions within a genome) and give clues as to whether overall single vs double strandedness exerts evolutionary pressure on a virus.

There is high likelihood that many of the RNA structures presented in this investigation do not represent the true predominant structures present *in virio* and there is still great room for improvement in the field of RNA structure determination. A major problem in this field is the lack of 'ground truth', that is RNAs of which we already know the structure against which to benchmark our techniques. This is limited to a relatively small number of RNAs for which crystal structures have been produced. In addition, crystal structures of potentially dynamic RNA molecules may in themselves be misleading. It seems that there are many different techniques, chemical probing reagents, analysis methods, and structure prediction algorithms which all claim the greatest accuracy. This is currently very hard to assess in an objective manner. For chemical probing techniques validation and folding parameter optimisation is most often based on the ribosomal RNA structure. This is not necessarily an ideal model, as it is often done in a different organism (where membrane differences may affect effective reagent concentration) and ribosomal RNA has methylations and protein interactions that may be quite different to the RNA being investigated. There can also be more specific issues such as protein binding (like NP binding of viral RNA), phase separation (as with SARS-CoV-2 genomic RNA in cells), or sub-cellular location, that may affect optimal folding parameters and/or the best reagent to use.

In the future, advances in Cryo-EM may help increase the number of RNA structures to allow us to improve chemical probing prediction accuracy. In the short term, techniques such as FISH should be used to compliment RNA structure prediction to try to eliminate some of the

assumptions that are made. Direct RNA-RNA interaction capture techniques are also improving with the development of new techniques such as COMRADES (Ziv et al., 2018). There are a number of improvements that could be made to SPLASH. For example, RNA could be fragmented to a much greater degree and shorter reads used. This would require the addition of barcodes to reads to allow de-duplication, but would facilitate both the identification of much closer ranger interactions and could improve the identification of the exact interaction location within a loci. This may then compliment chemical probing data and allow better local structure prediction when combined.

With the recent updates to AlphaFold the world of protein structure is now moving closer to accurate sequence-informed structure prediction (Jumper et al., 2021). The field of RNA structure seems further from this point, as these prediction programs rely upon training sets (a ground truth) which are currently very limited. Despite this, rapid advances have recently been made in the field by utilising machine learning (Townshend et al., 2021). Such approaches can also predict RNA tertiary structure, a factor that is not currently addressed in the majority of chemical probing based predictions.

The disruption of the current SARS-CoV-2 pandemic shows the need to be able to rapidly characterise new viruses and our current understanding is highly focused on viruses that are of current (or past) economic importance. There is a great diversity of functional RNA structure not just in SARS-CoV-2 and IAV, but other viruses such as poliovirus and HIV. Characterising a greater diversity of viruses may help us to be more prepared for future pandemics. It seems likely that many undiscovered RNA functions exist that could form the basis for drug targets or as useful molecular biology tools (such as IRES).

References

AFONINE, P. V., GROSSE-KUNSTLEVE, R. W., ECHOLS, N., HEADD, J. J., MORIARTY, N. W., MUSTYAKIMOV, M., TERWILLIGER, T. C., URZHUMTSEV, A., ZWART, P. H. & ADAMS, P. D. 2012. Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallographica Section D: Biological Crystallography,* 68**,** 352-367.

ANDREWS, R. J., PETERSON, J. M., HANIFF, H. S., CHEN, J., WILLIAMS, C., GREFE, M., DISNEY, M. D. & MOSS, W. N. 2020. An in silico map of the SARS-CoV-2 RNA Structurome. *BioRxiv*.

ANGELINI, M. M., AKHLAGHPOUR, M., NEUMAN, B. W. & BUCHMEIER, M. J. 2013. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *MBio,* 4.

ARRANZ, R., COLOMA, R., CHICHÓN, F. J., CONESA, J. J., CARRASCOSA, J. L., VALPUESTA, J. M., ORTÍN, J. & MARTÍN-BENITO, J. 2012. The structure of native influenza virion ribonucleoproteins. *Science,* 338**,** 1634-1637.

AW, J. G. A., SHEN, Y., WILM, A., SUN, M., LIM, X. N., BOON, K.-L., TAPSIN, S., CHAN, Y.-S., TAN, C.-P. & SIM, A. Y. 2016. In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Molecular cell,* 62**,** 603-617.

BARANOV, P. V., HENDERSON, C. M., ANDERSON, C. B., GESTELAND, R. F., ATKINS, J. F. & HOWARD, M. T. 2005. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology,* 332**,** 498-510.

BARBEZANGE, C., JONES, L., BLANC, H., ISAKOV, O., CELNIKER, G., ENOUF, V., SHOMRON, N., VIGNUZZI, M. & VAN DER WERF, S. 2018. Seasonal genetic drift of human influenza A virus quasispecies revealed by deep sequencing. *Frontiers in microbiology,* 9**,** 2596.

BECKHAM, S. A., MATAK, M. Y., BELOUSOFF, M. J., VENUGOPAL, H., SHAH, N., VANKADARI, N., ELMLUND, H., NGUYEN, J. H., SEMLER, B. L. & WILCE, M. C. 2020. Structure of the PCBP2/stem–loop IV complex underlying translation initiation mediated by the poliovirus type I IRES. *Nucleic acids research,* 48**,** 8006-8021.

BEIGEL, J. H., VOELL, J., MUÑOZ, P., KUMAR, P., BROOKS, K. M., ZHANG, J., IVERSEN, P., HEALD, A., WONG, M. & DAVEY, R. T. 2018. Safety, tolerability, and pharmacokinetics of radavirsen (AVI-7100), an antisense oligonucleotide targeting influenza a M1/M2 translation. *British journal of clinical pharmacology,* 84**,** 25-34.

BERGERON, C., VALETTE, M., LINA, B. & OTTMANN, M. 2010. Genetic content of influenza H3N2 vaccine seeds. *PLoS currents,* 2.

BOND, P. 2017. *JTSA* [Online]. Available: http://paulsbond.co.uk/jtsa [Accessed 5/3/2021].

BONI, M. F., GALVANI, A. P., WICKELGREN, A. L. & MALANI, A. 2013. Economic epidemiology of avian influenza on smallholder poultry farms. *Theoretical population biology,* 90**,** 135-144.

BORAU, M. S. & STERTZ, S. 2021. Entry of influenza A virus into host cells—recent progress and remaining challenges. *Current Opinion in Virology,* 48**,** 23-29.

BRIERLEY, I., DIGARD, P. & INGLIS, S. C. 1989. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell,* 57**,** 537-547.

BRUCE, E. A., DIGARD, P. & STUART, A. D. 2010. The Rab11 pathway is required for influenza A virus budding and filament formation. *Journal of virology,* 84**,** 5848-5859.

BUSAN, S., WEIDMANN, C. A., SENGUPTA, A. & WEEKS, K. M. 2019. Guidelines for SHAPE reagent choice and detection strategy for RNA structure probing studies. *Biochemistry,* 58**,** 2655-2664.

BUSCH, A., RICHTER, A. S. & BACKOFEN, R. 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics,* 24**,** 2849-2856.

BYRD-LEOTIS, L., JIA, N., DUTTA, S., TROST, J. F., GAO, C., CUMMINGS, S. F., BRAULKE, T., MÜLLER-LOENNIES, S., HEIMBURG-MOLINARO, J. & STEINHAUER, D. A. 2019. Influenza binds phosphorylated glycans from human lung. *Science advances,* 5**,** eaav2554.

CHAN, C.-M., CHU, H., ZHANG, A. J., LEUNG, L.-H., SZE, K.-H., KAO, R. Y.-T., CHIK, K. K.-H., TO, K. K.-W., CHAN, J. F.-W. & CHEN, H. 2016. Hemagglutinin of influenza A virus binds specifically to cell surface nucleolin and plays a role in virus internalization. *Virology,* 494**,** 78-88.

CHAN, W.-H., NG, A. K.-L., ROBB, N. C., LAM, M. K.-H., CHAN, P. K.-S., AU, S. W.-N., WANG, J.-H., FODOR, E. & SHAW, P.-C. 2010. Functional analysis of the influenza virus H5N1 nucleoprotein tail loop reveals amino acids that are crucial for oligomerization and ribonucleoprotein activities. *Journal of virology,* 84**,** 7337-7345.

CHANG, R.-Y., HOFMANN, M. A., SETHNA, P. B. & BRIAN, D. A. 1994. A cis-acting function for the coronavirus leader in defective interfering RNA replication. *Journal of Virology,* 68**,** 8223-8231.

CHATHAPPADY HOUSE, N. N., PALISSERY, S. & SEBASTIAN, H. 2021. Corona Viruses: A Review on SARS, MERS and COVID-19. *Microbiology Insights,* 14**,** 11786361211002481.

CHEN, S.-C. & OLSTHOORN, R. C. 2010. Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. *Virology,* 401**,** 29-41.

CHENAVAS, S., ESTROZI, L. F., SLAMA-SCHWOK, A., DELMAS, B., DI PRIMO, C., BAUDIN, F., LI, X., CRÉPIN, T. & RUIGROK, R. W. 2013. Monomeric nucleoprotein of influenza A virus. *PLoS Pathog,* 9**,** e1003275.

CHO, C.-P., LIN, S.-C., CHOU, M.-Y., HSU, H.-T. & CHANG, K.-Y. 2013. Regulation of programmed ribosomal frameshifting by co-translational refolding RNA hairpins. *PloS one,* 8**,** e62283.

CHOU, Y.-Y., HEATON, N. S., GAO, Q., PALESE, P., SINGER, R. & LIONNET, T. 2013. Colocalization of different influenza viral RNA segments in the cytoplasm before viral budding as shown by single-molecule sensitivity FISH analysis. *PLoS pathogens,* 9**,** e1003358.

CHOU, Y.-Y., VAFABAKHSH, R., DOĞANAY, S., GAO, Q., HA, T. & PALESE, P. 2012. One influenza virus particle packages eight unique viral RNAs as shown by FISH analysis. *Proceedings of the National Academy of Sciences,* 109**,** 9101-9106.

CIANCI, C., GERRITZ, S. W., DEMINIE, C. & KRYSTAL, M. 2013. Influenza nucleoprotein: promising target for antiviral chemotherapy. *Antiviral Chemistry and Chemotherapy,* 23**,** 77-91.

CIMINO, G. D., GAMPER, H. B., ISAACS, S. T. & HEARST, J. E. 1985. Psoralens as photoactive probes of nucleic acid structure and function: organic chemistry, photochemistry, and biochemistry. *Annual review of biochemistry,* 54**,** 1151-1193.

COBBIN, J. C., ONG, C., VERITY, E., GILBERTSON, B. P., ROCKMAN, S. P. & BROWN, L. E. 2014. Influenza virus PB1 and neuraminidase gene segments can cosegregate during vaccine reassortment driven by interactions in the PB1 coding region. *Journal of virology,* 88**,** 8971-8980.

COBBIN, J. C., VERITY, E. E., GILBERTSON, B. P., ROCKMAN, S. P. & BROWN, L. E. 2013. The source of the PB1 gene in influenza vaccine reassortants selectively alters the hemagglutinin content of the resulting seed virus. *Journal of virology,* 87**,** 5577-5585.

COLOMA, R., ARRANZ, R., JOSÉ, M., SORZANO, C. O., MUNIER, S., CARLERO, D., NAFFAKH, N., ORTÍN, J. & MARTÍN-BENITO, J. 2020. Structural insights into influenza A virus ribonucleoproteins reveal a processive helical track as transcription mechanism. *Nature microbiology,* 5**,** 727-734.

CORDERO, P., KLADWANG, W., VANLANG, C. C. & DAS, R. 2012. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry,* 51**,** 7037-7039.

DADONAITE, B., GILBERTSON, B., KNIGHT, M. L., TRIFKOVIC, S., ROCKMAN, S., LAEDERACH, A., BROWN, L. E., FODOR, E. & BAUER, D. L. 2019. The structure of the influenza A virus genome. *Nature microbiology,* 4**,** 1781-1789.

DE CASTRO MARTIN, I. F., FOURNIER, G., SACHSE, M., PIZARRO-CERDA, J., RISCO, C. & NAFFAKH, N. 2017. Influenza virus genome reaches the plasma membrane via a modified endoplasmic reticulum and Rab11-dependent vesicles. *Nature communications,* 8**,** 1-12.

DE GRAAF, M. & FOUCHIER, R. A. 2014. Role of receptor binding specificity in influenza A virus transmission and pathogenesis. *The EMBO journal,* 33**,** 823-841.

DE HAAN, C. A., VOLDERS, H., KOETZNER, C. A., MASTERS, P. S. & ROTTIER, P. J. 2002. Coronaviruses maintain viability despite dramatic rearrangements of the strictly conserved genome organization. *Journal of virology,* 76**,** 12491-12502.

DE VRIES, E., TSCHERNE, D. M., WIENHOLTS, M. J., COBOS-JIMÉNEZ, V., SCHOLTE, F., GARCÍA-SASTRE, A., ROTTIER, P. J. & DE HAAN, C. A. 2011. Dissection of the influenza A virus endocytic routes reveals macropinocytosis as an alternative entry pathway. *PLoS Pathog,* 7**,** e1001329.

DEIGAN, K. E., LI, T. W., MATHEWS, D. H. & WEEKS, K. M. 2009. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences,* 106**,** 97-102.

DESFOSSES, A., MILLES, S., JENSEN, M. R., GUSEVA, S., COLLETIER, J.-P., MAURIN, D., SCHOEHN, G., GUTSCHE, I., RUIGROK, R. W. & BLACKLEDGE, M. 2019. Assembly and cryo-EM structures of RNA-specific measles virus nucleocapsids provide mechanistic insight into paramyxoviral replication. *Proceedings of the National Academy of Sciences,* 116**,** 4256-4264.

DIBBEN, O., CROWE, J., COOPER, S., HILL, L., SCHEWE, K. E. & BRIGHT, H. 2021. Defining the root cause of reduced H1N1 live attenuated influenza vaccine effectiveness: low viral fitness leads to inter-strain competition. *NPJ vaccines,* 6**,** 1-12.

DOBIN, A., DAVIS, C., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics29, 15–21.

DONCHET, A., OLIVA, J., LABARONNE, A., TENGO, L., MILOUDI, M., GERARD, F. C., MAS, C., SCHOEHN, G., RUIGROK, R. W. & DUCATEZ, M. 2019. The structure of the nucleoprotein of Influenza D shows that all Orthomyxoviridae nucleoproteins have a similar NP CORE, with or without a NP TAIL for nuclear transport. *Scientific reports,* 9**,** 1-14.

DOSHI, K. J., CANNONE, J. J., COBAUGH, C. W. & GUTELL, R. R. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics,* 5**,** 1-22.

DOSZTÁNYI, Z. 2018. Prediction of protein disorder based on IUPred. *Protein Science,* 27**,** 331-340.

DOU, D., REVOL, R., ÖSTBYE, H., WANG, H. & DANIELS, R. 2018. Influenza A virus cell entry, replication, virion assembly and movement. *Frontiers in immunology,* 9**,** 1581.

DUHAUT, S. & DIMMOCK, N. 2002. Defective segment 1 RNAs that interfere with production of infectious influenza A virus require at least 150 nucleotides of 5′ sequence: evidence from a plasmid-driven system. *Journal of General Virology,* 83**,** 403-411.

DUPAI, C. D., MCWHITE, C. D., SMITH, C. B., GARTEN, R., MAURER-STROH, S. & WILKE, C. O. 2019. Influenza passaging annotations: what they tell us and why we should listen. *Virus evolution,* 5**,** vez016.

DUPEUX, F., RÖWER, M., SEROUL, G., BLOT, D. & MÁRQUEZ, J. A. 2011. A thermal stability assay can help to estimate the crystallization likelihood of biological samples. *Acta Crystallographica Section D: Biological Crystallography,* 67**,** 915-919.

EDDY, S. R. 2014. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annual review of biophysics,* 43**,** 433-456.

EHRESMANN, C., BAUDIN, F., MOUGEL, M., ROMBY, P., EBEL, J.-P. & EHRESMANN, B. 1987. Probing the structure of RNAs in solution. *Nucleic acids research,* 15**,** 9109-9128.

EIERHOFF, T., HRINCIUS, E. R., RESCHER, U., LUDWIG, S. & EHRHARDT, C. 2010. The epidermal growth factor receptor (EGFR) promotes uptake of influenza A viruses (IAV) into host cells. *PLoS Pathog,* 6**,** e1001099.

EISFELD, A. J., NEUMANN, G. & KAWAOKA, Y. 2015. At the centre: influenza A virus ribonucleoproteins. *Nature Reviews Microbiology,* 13**,** 28-41.

ELTON, D., MEDCALF, L., BISHOP, K., HARRISON, D. & DIGARD, P. 1999. Identification of amino acid residues of influenza virus nucleoprotein essential for RNA binding. *Journal of Virology,* 73**,** 7357-7367.

EMSLEY, P., LOHKAMP, B., SCOTT, W. G. & COWTAN, K. 2010. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography,* 66**,** 486-501.

ESSERE, B., YVER, M., GAVAZZI, C., TERRIER, O., ISEL, C., FOURNIER, E., GIROUX, F., TEXTORIS, J., JULIEN, T. & SOCRATOUS, C. 2013. Critical role of segment-specific packaging signals in genetic reassortment of influenza A viruses. *Proceedings of the National Academy of Sciences,* 110**,** E3840-E3848.

FAN, H., WALKER, A. P., CARRIQUE, L., KEOWN, J. R., MARTIN, I. S., KARIA, D., SHARPS, J., HENGRUNG, N., PARDON, E. & STEYAERT, J. 2019. Structures

of influenza A virus RNA polymerase offer insight into viral genome replication. *Nature,* 573**,** 287-290.
FINKEL, Y., MIZRAHI, O., NACHSHON, A., WEINGARTEN-GABBAY, S., MORGENSTERN, D., YAHALOM-RONEN, Y., TAMIR, H., ACHDOUT, H., STEIN, D. & ISRAELI, O. 2021. The coding capacity of SARS-CoV-2. *Nature,* 589**,** 125-130.
FODOR, E. & TE VELTHUIS, A. J. 2020. Structure and function of the influenza virus transcription and replication machinery. *Cold Spring Harbor perspectives in medicine,* 10**,** a038398.
FOURNIER, E., MOULES, V., ESSERE, B., PAILLART, J.-C., SIRBAT, J.-D., ISEL, C., CAVALIER, A., ROLLAND, J.-P., THOMAS, D. & LINA, B. 2012. A supramolecular assembly formed by influenza A virus genomic RNA segments. *Nucleic acids research,* 40**,** 2197-2209.
FREIER, S. M., KIERZEK, R., JAEGER, J. A., SUGIMOTO, N., CARUTHERS, M. H., NEILSON, T. & TURNER, D. H. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences,* 83**,** 9373-9377.
FRENSING, T., KUPKE, S. Y., BACHMANN, M., FRITZSCHE, S., GALLO-RAMIREZ, L. E. & REICHL, U. 2016. Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in MDCK cells. *Applied microbiology and biotechnology,* 100**,** 7181-7192.
FUJIOKA, Y., NISHIDE, S., OSE, T., SUZUKI, T., KATO, I., FUKUHARA, H., FUJIOKA, M., HORIUCHI, K., SATOH, A. O. & NEPAL, P. 2018. A sialylated voltage-dependent Ca2+ channel binds hemagglutinin and mediates influenza a virus entry into mammalian cells. *Cell host & microbe,* 23**,** 809-818. e5.
FULVINI, A. A., RAMANUNNINAIR, M., LE, J., POKORNY, B. A., ARROYO, J. M., SILVERMAN, J., DEVIS, R. & BUCHER, D. 2011. Gene constellation of influenza A virus reassortants with high growth phenotype prepared as seed candidates for vaccine production. *PloS one,* 6**,** e20823.
GABRIEL, G., HERWIG, A. & KLENK, H.-D. 2008. Interaction of polymerase subunit PB2 and NP with importin α1 is a determinant of host range of influenza A virus. *PLoS pathogens,* 4**,** e11.
GALLAGHER, J. R., TORIAN, U., MCCRAW, D. M. & HARRIS, A. K. 2017. Structural studies of influenza virus RNPs by electron microscopy indicate molecular contortions within NP supra-structures. *Journal of structural biology,* 197**,** 294-307.
GAO, J., GUI, M. & XIANG, Y. 2020a. Structural intermediates in the low pH-induced transition of influenza hemagglutinin. *PLoS Pathogens,* 16**,** e1009062.
GAO, Y., YAN, L., HUANG, Y., LIU, F., ZHAO, Y., CAO, L., WANG, T., SUN, Q., MING, Z. & ZHANG, L. 2020b. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science,* 368**,** 779-782.
GAVAZZI, C., ISEL, C., FOURNIER, E., MOULES, V., CAVALIER, A., THOMAS, D., LINA, B. & MARQUET, R. 2013. An in vitro network of intermolecular interactions between viral RNA segments of an avian H5N2 influenza A virus: comparison with a human H3N2 virus. *Nucleic acids research,* 41**,** 1241-1254.
GERBER, M., ISEL, C., MOULES, V. & MARQUET, R. 2014. Selective packaging of the influenza A genome and consequences for genetic reassortment. *Trends in microbiology,* 22**,** 446-455.

GHOSH, S., DELLIBOVI-RAGHEB, T., PAK, E., QIU, Q., FISHER, M., TAKVORIAN, P., BLECK, C., HSU, V., FEHR, A. & PERLMAN, S. 2020. β-Coronaviruses use lysosomal organelles for cellular egress.

GILBERTSON, B., ZHENG, T., GERBER, M., PRINTZ-SCHWEIGERT, A., ONG, C., MARQUET, R., ISEL, C., ROCKMAN, S. & BROWN, L. 2016. Influenza NA and PB1 gene segments interact during the formation of viral progeny: localization of the binding region within the PB1 gene. *Viruses,* 8**,** 238.

GODDARD, T. D., HUANG, C. C., MENG, E. C., PETTERSEN, E. F., COUCH, G. S., MORRIS, J. H. & FERRIN, T. E. 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science,* 27**,** 14-25.

GOEBEL, S. J., HSUE, B., DOMBROWSKI, T. F. & MASTERS, P. S. 2004. Characterization of the RNA components of a putative molecular switch in the 3′ untranslated region of the murine coronavirus genome. *Journal of Virology,* 78**,** 669-682.

GOEBEL, S. J., MILLER, T. B., BENNETT, C. J., BERNARD, K. A. & MASTERS, P. S. 2007. A hypervariable region within the 3′ cis-acting element of the murine coronavirus genome is nonessential for RNA synthesis but affects pathogenesis. *Journal of virology,* 81**,** 1274-1287.

GOULD, P. S., EASTON, A. J. & DIMMOCK, N. J. 2017. Live attenuated influenza vaccine contains substantial and unexpected amounts of defective viral genomic RNA. *Viruses,* 9**,** 269.

GREENHALGH, T., JIMENEZ, J. L., PRATHER, K. A., TUFEKCI, Z., FISMAN, D. & SCHOOLEY, R. 2021. Ten scientific reasons in support of airborne transmission of SARS-CoV-2. *The Lancet,* 397**,** 1603-1605.

GRUBER, A. R., BERNHART, S. H. & LORENZ, R. 2015. The ViennaRNA web services. *RNA bioinformatics.* Springer.

GU, Z., GU, L., EILS, R., SCHLESNER, M. & BRORS, B. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics,* 30**,** 2811-2812.

GUAN, B.-J., SU, Y.-P., WU, H.-Y. & BRIAN, D. A. 2012. Genetic evidence of a long-range RNA-RNA interaction between the genomic 5′ untranslated region and the nonstructural protein 1 coding region in murine and bovine coronaviruses. *Journal of virology,* 86**,** 4631-4643.

GULTYAEV, A. P., SPRONKEN, M. I., RICHARD, M., SCHRAUWEN, E. J., OLSTHOORN, R. C. & FOUCHIER, R. A. 2016. Subtype-specific structural constraints in the evolution of influenza A virus hemagglutinin genes. *Scientific reports,* 6**,** 1-15.

GULTYAEV, A. P., TSYGANOV-BODOUNOV, A., SPRONKEN, M. I., VAN DER KOOIJ, S., FOUCHIER, R. A. & OLSTHOORN, R. C. 2014. RNA structural constraints in the evolution of the influenza A virus genome NP segment. *RNA biology,* 11**,** 942-952.

HAAGMANS, B. L., AL DHAHIRY, S. H., REUSKEN, C. B., RAJ, V. S., GALIANO, M., MYERS, R., GODEKE, G.-J., JONGES, M., FARAG, E. & DIAB, A. 2014. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *The Lancet infectious diseases,* 14**,** 140-145.

HAJDIN, C. E., BELLAOUSOV, S., HUGGINS, W., LEONARD, C. W., MATHEWS, D. H. & WEEKS, K. M. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences,* 110**,** 5498-5503.

HAMMING, I., TIMENS, W., BULTHUIS, M., LELY, A., NAVIS, G. V. & VAN GOOR, H. 2004. Tissue distribution of ACE2 protein, the functional receptor for SARS

coronavirus. A first step in understanding SARS pathogenesis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland,* 203**,** 631-637.
HARALAMPIEV, I., PRISNER, S., NITZAN, M., SCHADE, M., JOLMES, F., SCHREIBER, M., LOIDOLT-KRÜGER, M., JONGEN, K., CHAMIOLO, J. & NILSON, N. 2020. Selective flexible packaging pathways of the segmented genome of influenza A virus. *Nature communications,* 11**,** 1-13.
HAWKSWORTH, A., LOCKHART, R., CROWE, J., MAESO, R., RITTER, L., DIBBEN, O. & BRIGHT, H. 2020. Replication of live attenuated influenza vaccine viruses in human nasal epithelial cells is associated with H1N1 vaccine effectiveness. *Vaccine,* 38**,** 4209-4218.
HEUS, H. A. & PARDI, A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science,* 253**,** 191-194.
HILLEN, H. S., KOKIC, G., FARNUNG, L., DIENEMANN, C., TEGUNOV, D. & CRAMER, P. 2020. Structure of replicating SARS-CoV-2 polymerase. *Nature,* 584**,** 154-156.
HOFFMANN, E., KRAUSS, S., PEREZ, D., WEBBY, R. & WEBSTER, R. G. 2002. Eight-plasmid system for rapid generation of influenza virus vaccines. *Vaccine,* 20**,** 3165-3170.
HOFFMANN, M., KLEINE-WEBER, H. & PÖHLMANN, S. 2020. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular cell,* 78**,** 779-784. e5.
HOUSER, K. & SUBBARAO, K. 2015. Influenza vaccines: challenges and solutions. *Cell host & microbe,* 17**,** 295-300.
HSUE, B., HARTSHORNE, T. & MASTERS, P. S. 2000. Characterization of an essential RNA secondary structure in the 3′ untranslated region of the murine coronavirus genome. *Journal of virology,* 74**,** 6911-6921.
HU, Y., SNEYD, H., DEKANT, R. & WANG, J. 2017. Influenza A virus nucleoprotein: a highly conserved multi-functional viral protein as a hot antiviral drug target. *Current topics in medicinal chemistry,* 17**,** 2271-2285.
HUSTON, N. C., WAN, H., STRINE, M. S., TAVARES, R. D. C. A., WILEN, C. B. & PYLE, A. M. 2021. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Molecular cell,* 81**,** 584-598. e5.
HUTCHINSON, E. C., CHARLES, P. D., HESTER, S. S., THOMAS, B., TRUDGIAN, D., MARTÍNEZ-ALONSO, M. & FODOR, E. 2014. Conserved and host-specific features of influenza virion architecture. *Nature communications,* 5**,** 1-11.
HUTCHINSON, E. C., CURRAN, M. D., READ, E. K., GOG, J. R. & DIGARD, P. 2008. Mutational analysis of cis-acting RNA signals in segment 7 of influenza A virus. *Journal of virology,* 82**,** 11869-11879.
HUYNEN, M., GUTELL, R. & KONINGS, D. 1997. Assessing the reliability of RNA folding using statistical mechanics. *Journal of molecular biology,* 267**,** 1104-1112.
INAGAKI, A., GOTO, H., KAKUGAWA, S., OZAWA, M. & KAWAOKA, Y. 2012. Competitive incorporation of homologous gene segments of influenza A virus into virions. *Journal of virology,* 86**,** 10200-10202.
INCARNATO, D., MORANDI, E., SIMON, L. M. & OLIVIERO, S. 2018. RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic acids research,* 46**,** e97-e97.

ISERMAN, C., RODEN, C. A., BOERNEKE, M. A., SEALFON, R. S., MCLAUGHLIN, G. A., JUNGREIS, I., FRITCH, E. J., HOU, Y. J., EKENA, J. & WEIDMANN, C. A. 2020. Genomic RNA elements drive phase separation of the SARS-CoV-2 nucleocapsid. *Molecular cell,* 80**,** 1078-1091. e6.

ISON, M. G. 2011. Antivirals and resistance: influenza virus. *Current opinion in virology,* 1**,** 563-573.

IULIANO, A. D., ROGUSKI, K. M., CHANG, H. H., MUSCATELLO, D. J., PALEKAR, R., TEMPIA, S., COHEN, C., GRAN, J. M., SCHANZER, D. & COWLING, B. J. 2018. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet,* 391**,** 1285-1300.

JAAFAR, Z. A. & KIEFT, J. S. 2019. Viral RNA structure-based strategies to manipulate translation. *Nature Reviews Microbiology,* 17**,** 110-123.

JACKSON, S., VAN HOEVEN, N., CHEN, L.-M., MAINES, T. R., COX, N. J., KATZ, J. M. & DONIS, R. O. 2009. Reassortment between avian H5N1 and human H3N2 influenza viruses in ferrets: a public health risk assessment. *Journal of virology,* 83**,** 8131-8140.

JANG, J. & BAE, S.-E. 2018. Comparative co-evolution analysis between the HA and NA genes of influenza A virus. *Virology: research and treatment,* 9**,** 1178122X18788328.

JIANG, H., LEI, R., DING, S.-W. & ZHU, S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics,* 15**,** 182.

JONASSEN, C. M., JONASSEN, T. & GRINDE, B. 1998. A common RNA motif in the 3'end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. *Journal of General Virology,* 79**,** 715-718.

JONES, J. E., LE SAGE, V., PADOVANI, G. H., CALDERON, M., WRIGHT, E. S. & LAKDAWALA, S. 2021. Parallel evolution between genomic segments of seasonal human influenza viruses reveals RNA-RNA relationships. *bioRxiv*.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A. & POTAPENKO, A. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature***,** 1-11.

KAKISAKA, M., SASAKI, Y., YAMADA, K., KONDOH, Y., HIKONO, H., OSADA, H., TOMII, K., SAITO, T. & AIDA, Y. 2015. A novel antiviral target structure involved in the RNA binding, dimerization, and nuclear export functions of the influenza A virus nucleoprotein. *PLoS Pathog,* 11**,** e1005062.

KARAKUS, U., THAMAMONGOOD, T., CIMINSKI, K., RAN, W., GÜNTHER, S. C., POHL, M. O., ELETTO, D., JENEY, C., HOFFMANN, D. & REICHE, S. 2019. MHC class II proteins mediate cross-species entry of bat influenza viruses. *Nature,* 567**,** 109-112.

KER, D.-S., JENKINS, H. T., GREIVE, S. J. & ANTSON, A. A. 2020. CryoEM structure of the Nipah virus nucleocapsid assembly. *BioRxiv*.

KIM, D., LEE, J.-Y., YANG, J.-S., KIM, J. W., KIM, V. N. & CHANG, H. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell,* 181**,** 914-921. e10.

KNIGHT, M. L., FAN, H., BAUER, D. L., GRIMES, J. M., FODOR, E. & KEOWN, J. R. 2021. Structure of an H3N2 influenza virus nucleoprotein. *Acta Crystallographica Section F: Structural Biology Communications,* 77**,** 208-214.

KOBAYASHI, Y., DADONAITE, B., VAN DOREMALEN, N., SUZUKI, Y., BARCLAY, W. S. & PYBUS, O. G. 2016. Computational and molecular analysis of

conserved influenza A virus RNA secondary structures involved in infectious virion production. *RNA biology,* 13**,** 883-894.
KRANZ, J. K. & SCHALK-HIHI, C. 2011. Protein thermal shifts to identify low molecular weight fragments. *Methods in enzymology,* 493**,** 277-298.
KRYAZHIMSKIY, S., DUSHOFF, J., BAZYKIN, G. A. & PLOTKIN, J. B. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet,* 7**,** e1001301.
KUKOL, A. & HUGHES, D. J. 2014. Large-scale analysis of influenza A virus nucleoprotein sequence conservation reveals potential drug-target sites. *Virology,* 454**,** 40-47.
KUTCHKO, K. M. & LAEDERACH, A. 2017. Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *Wiley Interdisciplinary Reviews: RNA,* 8**,** e1374.
LABARONNE, A., SWALE, C., MONOD, A., SCHOEHN, G., CRÉPIN, T. & RUIGROK, R. W. 2016. Binding of RNA by the nucleoproteins of influenza viruses A and B. *Viruses,* 8**,** 247.
LAKDAWALA, S. S., WU, Y., WAWRZUSIN, P., KABAT, J., BROADBENT, A. J., LAMIRANDE, E. W., FODOR, E., ALTAN-BONNET, N., SHROFF, H. & SUBBARAO, K. 2014. Influenza a virus assembly intermediates fuse in the cytoplasm. *PLoS Pathog,* 10**,** e1003971.
LAN, T. C., ALLAN, M. F., MALSICK, L., KHANDWALA, S., NYEO, S. S., SUN, Y., GUO, J. U., BATHE, M., GRIFFITHS, A. & ROUSKIN, S. 2021. Insights into the secondary structural ensembles of the full SARS-CoV-2 RNA genome in infected cells. *Biorxiv***,** 2020.06. 29.178343.
LANGE, S. J., MATICZKA, D., MÖHL, M., GAGNON, J. N., BROWN, C. M. & BACKOFEN, R. 2012. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic acids research,* 40**,** 5215-5226.
LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods,* 9**,** 357-359.
LE SAGE, V., KANAREK, J. P., SNYDER, D. J., COOPER, V. S., LAKDAWALA, S. S. & LEE, N. 2020. Mapping of influenza virus RNA-RNA interactions reveals a flexible network. *Cell reports,* 31**,** 107823.
LEE, C. W., LI, L. & GIEDROC, D. P. 2011. The solution structure of coronaviral stem-loop 2 (SL2) reveals a canonical CUYG tetraloop fold. *FEBS letters,* 585**,** 1049-1053.
LEE, N., LE SAGE, V., NANNI, A. V., SNYDER, D. J., COOPER, V. S. & LAKDAWALA, S. S. 2017. Genome-wide analysis of influenza viral RNA and nucleoprotein association. *Nucleic acids research,* 45**,** 8968-8977.
LENARTOWICZ, E., NOGALES, A., KIERZEK, E., KIERZEK, R., MARTÍNEZ-SOBRIDO, L. & TURNER, D. H. 2016. Antisense oligonucleotides targeting influenza A segment 8 genomic RNA inhibit viral replication. *Nucleic acid therapeutics,* 26**,** 277-285.
LEWIS, G. D. & METCALF, T. G. 1988. Polyethylene glycol precipitation for recovery of pathogenic viruses, including hepatitis A virus and human rotavirus, from oyster, water, and sediment samples. *Applied and environmental microbiology,* 54**,** 1983-1988.
LI, C., HATTA, M., WATANABE, S., NEUMANN, G. & KAWAOKA, Y. 2008a. Compatibility among polymerase subunit proteins is a restricting factor in reassortment between equine H7N7 and human H3N2 influenza viruses. *Journal of virology,* 82**,** 11880-11888.

LI, L., KANG, H., LIU, P., MAKKINJE, N., WILLIAMSON, S. T., LEIBOWITZ, J. L. & GIEDROC, D. P. 2008b. Structural lability in stem–loop 1 drives a 5′ UTR–3′ UTR interaction in coronavirus replication. *Journal of molecular biology,* 377**,** 790-803.

LIANG, Y., HONG, Y. & PARSLOW, T. G. 2005. cis-Acting packaging signals in the influenza virus PB1, PB2, and PA genomic RNA segments. *Journal of virology,* 79**,** 10348-10355.

LIU, P., LI, L., KEANE, S. C., YANG, D., LEIBOWITZ, J. L. & GIEDROC, D. P. 2009. Mouse hepatitis virus stem-loop 2 adopts a uYNMG (U) a-like tetraloop structure that is highly functionally tolerant of base substitutions. *Journal of virology,* 83**,** 12084-12093.

LIU, P., LI, L., MILLERSHIP, J. J., KANG, H., LEIBOWITZ, J. L. & GIEDROC, D. P. 2007. A U-turn motif-containing stem–loop in the coronavirus 5′ untranslated region plays a functional role in replication. *Rna,* 13**,** 763-780.

LIU, Q., JOHNSON, R. F. & LEIBOWITZ, J. L. 2001. Secondary structural elements within the 3′ untranslated region of mouse hepatitis virus strain JHM genomic RNA. *Journal of virology,* 75**,** 12105-12113.

LO, M.-C., AULABAUGH, A., JIN, G., COWLING, R., BARD, J., MALAMAS, M. & ELLESTAD, G. 2004. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical biochemistry,* 332**,** 153-159.

LORENZ, R., BERNHART, S. H., ZU SIEDERDISSEN, C. H., TAFER, H., FLAMM, C., STADLER, P. F. & HOFACKER, I. L. 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology,* 6**,** 1-14.

LORENZ, R., WOLFINGER, M. T., TANZER, A. & HOFACKER, I. L. 2016. Predicting RNA secondary structures from sequence and probing data. *Methods,* 103**,** 86-98.

LU, Z., ZHANG, Q. C., LEE, B., FLYNN, R. A., SMITH, M. A., ROBINSON, J. T., DAVIDOVICH, C., GOODING, A. R., GOODRICH, K. J. & MATTICK, J. S. 2016. RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell,* 165**,** 1267-1279.

LUYTJES, W., GERRITSMA, H. & SPAAN, W. J. 1996. Replication of synthetic defective interfering RNAs derived from coronavirus mouse hepatitis virus-A59. *Virology,* 216**,** 174-183.

MADHUGIRI, R., FRICKE, M., MARZ, M. & ZIEBUHR, J. 2014. RNA structure analysis of alphacoronavirus terminal genome regions. *Virus research,* 194**,** 76-89.

MADHUGIRI, R., FRICKE, M., MARZ, M. & ZIEBUHR, J. 2016. Coronavirus cis-acting RNA elements. *Advances in virus research,* 96**,** 127-163.

MANFREDONIA, I., NITHIN, C., PONCE-SALVATIERRA, A., GHOSH, P., WIRECKI, T. K., MARINUS, T., OGANDO, N. S., SNIJDER, E. J., VAN HEMERT, M. J. & BUJNICKI, J. M. 2020. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic acids research,* 48**,** 12436-12452.

MARINUS, T., FESSLER, A. B., OGLE, C. A. & INCARNATO, D. 2021. A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic acids research,* 49**,** e34-e34.

MARSH, G. A., RABADÁN, R., LEVINE, A. J. & PALESE, P. 2008. Highly conserved regions of influenza a virus polymerase gene segments are critical for efficient viral RNA packaging. *Journal of virology,* 82**,** 2295-2304.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal,* 17**,** 10-12.

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.
MATHEWS, D. H., SABINA, J., ZUKER, M. & TURNER, D. H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology,* 288**,** 911-940.
MATLIN, K. S., REGGIO, H., HELENIUS, A. & SIMONS, K. 1981. Infectious entry pathway of influenza virus in a canine kidney cell line. *The Journal of cell biology,* 91**,** 601-613.
MATROSOVICH, M. N., MATROSOVICH, T. Y., GRAY, T., ROBERTS, N. A. & KLENK, H.-D. 2004. Human and avian influenza viruses target different cell types in cultures of human airway epithelium. *Proceedings of the National Academy of Sciences,* 101**,** 4620-4624.
MAUGER, D. M., GOLDEN, M., YAMANE, D., WILLIFORD, S., LEMON, S. M., MARTIN, D. P. & WEEKS, K. M. 2015. Functionally conserved architecture of hepatitis C virus RNA genomes. *Proceedings of the National Academy of Sciences,* 112**,** 3692-3697.
MAZZETTO, E., BORTOLAMI, A., FUSARO, A., MAZZACAN, E., MANIERO, S., VASCELLARI, M., BEATO, M. S., SCHIAVON, E., CHIAPPONI, C. & TERREGINO, C. 2020. Replication of Influenza D Viruses of Bovine and Swine Origin in Ovine Respiratory Explants and Their Attachment to the Respiratory Tract of Bovine, Sheep, Goat, Horse, and Swine. *Frontiers in microbiology,* 11**,** 1136.
MCCASKILL, J. S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers: Original Research on Biomolecules,* 29**,** 1105-1119.
MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, L. C. & READ, R. J. 2007. Phaser crystallographic software. *Journal of applied crystallography,* 40**,** 658-674.
MCMAHON, M., ASTHAGIRI ARUNKUMAR, G., LIU, W.-C., STADLBAUER, D., ALBRECHT, R. A., PAVOT, V., ARAMOUNI, M., LAMBE, T., GILBERT, S. C. & KRAMMER, F. 2019. Vaccination with viral vectors expressing chimeric hemagglutinin, NP and M1 antigens protects ferrets against influenza virus challenge. *Frontiers in immunology,* 10**,** 2005.
MENA, I., NELSON, M. I., QUEZADA-MONROY, F., DUTTA, J., CORTES-FERNÁNDEZ, R., LARA-PUENTE, J. H., CASTRO-PERALTA, F., CUNHA, L. F., TROVAO, N. S. & LOZANO-DUBERNARD, B. 2016. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. *Elife,* 5**,** e16777.
MENDONÇA, L., HOWE, A., GILCHRIST, J. B., SHENG, Y., SUN, D., KNIGHT, M. L., ZANETTI-DOMINGUES, L. C., BATEMAN, B., KREBS, A.-S. & CHEN, L. 2021. Correlative multi-scale cryo-imaging unveils SARS-CoV-2 assembly and egress. *Nature Communications,* 12**,** 1-10.
METZ, D. H. & BROWN, G. L. 1969. Investigation of nucleic acid secondary structure by means of chemical modification with a carbodiimide reagent. I. Reaction between N-cyclohexyl-N'-β-(4-methylmorpholinium) ethylcarbodiimide and model nucleotides. *Biochemistry,* 8**,** 2312-2328.
MICHALAK, P., SOSZYNSKA-JOZWIAK, M., BIALA, E., MOSS, W. N., KESY, J., SZUTKOWSKA, B., LENARTOWICZ, E., KIERZEK, R. & KIERZEK, E. 2019. Secondary structure of the segment 5 genomic RNA of influenza A virus and

its application for designing antisense oligonucleotides. *Scientific reports,* 9**,** 1-16.
MITCHELL, D., RENDA, A. J., DOUDS, C. A., BABITZKE, P., ASSMANN, S. M. & BEVILACQUA, P. C. 2019. In vivo RNA structural probing of uracil and guanine base-pairing by 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC). *RNA,* 25**,** 147-157.
MITCHELL III, D., ASSMANN, S. M. & BEVILACQUA, P. C. 2019. Probing RNA structure in vivo. *Current opinion in structural biology,* 59**,** 151-158.
MORANDI, E., MANFREDONIA, I., SIMON, L. M., ANSELMI, F., VAN HEMERT, M. J., OLIVIERO, S. & INCARNATO, D. 2021. Genome-scale deconvolution of RNA structure ensembles. *Nature Methods,* 18**,** 249-252.
MOSS, W. N., PRIORE, S. F. & TURNER, D. H. 2011. Identification of potential conserved RNA secondary structure throughout influenza A coding regions. *Rna,* 17**,** 991-1011.
MOULÈS, V., TERRIER, O., YVER, M., RITEAU, B., MORISCOT, C., FERRARIS, O., JULIEN, T., GIUDICE, E., ROLLAND, J.-P. & ERNY, A. 2011. Importance of viral genomic composition in modulating glycoprotein content on the surface of influenza virus particles. *Virology,* 414**,** 51-62.
MUNDIGALA, H., MICHAUX, J. B., FEIG, A. L., ENNIFAR, E. & RUEDA, D. 2014. HIV-1 DIS stem loop forms an obligatory bent kissing intermediate in the dimerization pathway. *Nucleic acids research,* 42**,** 7281-7289.
MUSTOE, A. M., LAMA, N. N., IRVING, P. S., OLSON, S. W. & WEEKS, K. M. 2019. RNA base-pairing complexity in living cells visualized by correlated chemical probing. *Proceedings of the National Academy of Sciences,* 116**,** 24574-24582.
NEVEROV, A. D., LEZHNINA, K. V., KONDRASHOV, A. S. & BAZYKIN, G. A. 2014. Intrasubtype reassortments cause adaptive amino acid replacements in H3N2 influenza genes. *PLoS Genet,* 10**,** e1004037.
NEWCOMB, L. L., KUO, R.-L., YE, Q., JIANG, Y., TAO, Y. J. & KRUG, R. M. 2009. Interaction of the influenza a virus nucleocapsid protein with the viral RNA polymerase potentiates unprimed viral RNA replication. *Journal of virology,* 83**,** 29-36.
NG, A. K.-L., LAM, M. K.-H., ZHANG, H., LIU, J., AU, S. W.-N., CHAN, P. K.-S., WANG, J. & SHAW, P.-C. 2012. Structural basis for RNA binding and homo-oligomer formation by influenza B virus nucleoprotein. *Journal of virology,* 86**,** 6758-6767.
NG, A. K. L., ZHANG, H., TAN, K., LI, Z., LIU, J. H., CHAN, P. K. S., LI, S. M., CHAN, W. Y., AU, S. W. N. & JOACHIMIAK, A. 2008. Structure of the influenza virus A H5N1 nucleoprotein: implications for RNA binding, oligomerization, and vaccine design. *The FASEB journal,* 22**,** 3638-3647.
NODA, T., SAGARA, H., YEN, A., TAKADA, A., KIDA, H., CHENG, R. H. & KAWAOKA, Y. 2006. Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature,* 439**,** 490-492.
NODA, T., SUGITA, Y., AOYAMA, K., HIRASE, A., KAWAKAMI, E., MIYAZAWA, A., SAGARA, H. & KAWAOKA, Y. 2012. Three-dimensional analysis of ribonucleoprotein complexes in influenza A virus. *Nature communications,* 3**,** 1-6.
NOTON, S. L., MEDCALF, E., FISHER, D., MULLIN, A. E., ELTON, D. & DIGARD, P. 2007. Identification of the domains of the influenza A virus M1 matrix protein required for NP binding, oligomerization and incorporation into virions. *The Journal of general virology,* 88**,** 2280.

NOTON, S. L., SIMPSON-HOLLEY, M., MEDCALF, E., WISE, H. M., HUTCHINSON, E. C., MCCAULEY, J. W. & DIGARD, P. 2009. Studies of an influenza A virus temperature-sensitive mutant identify a late role for NP in the formation of infectious virions. *Journal of virology,* 83**,** 562-571.

O'NEILL, R. E., JASKUNAS, R., BLOBEL, G., PALESE, P. & MOROIANU, J. 1995. Nuclear import of influenza virus RNA can be mediated by viral nucleoprotein and transport factors required for protein import. *Journal of Biological Chemistry,* 270**,** 22701-22704.

OCTAVIANI, C. P., GOTO, H. & KAWAOKA, Y. 2011. Reassortment between seasonal H1N1 and pandemic (H1N1) 2009 influenza viruses is restricted by limited compatibility among polymerase subunits. *Journal of virology,* 85**,** 8449-8452.

ORTEGA, J., MARTÍN-BENITO, J., ZÜRCHER, T., VALPUESTA, J. M., CARRASCOSA, J. L. & ORTÍN, J. 2000. Ultrastructural and functional analyses of recombinant influenza virus ribonucleoproteins suggest dimerization of nucleoprotein during virus amplification. *Journal of virology,* 74**,** 156-163.

OZAWA, M., FUJII, K., MURAMOTO, Y., YAMADA, S., YAMAYOSHI, S., TAKADA, A., GOTO, H., HORIMOTO, T. & KAWAOKA, Y. 2007. Contributions of two nuclear localization signals of influenza A virus nucleoprotein to viral replication. *Journal of virology,* 81**,** 30-41.

OZAWA, M., MAEDA, J., IWATSUKI-HORIMOTO, K., WATANABE, S., GOTO, H., HORIMOTO, T. & KAWAOKA, Y. 2009. Nucleotide sequence requirements at the 5′ end of the influenza A virus M RNA segment for efficient virus replication. *Journal of virology,* 83**,** 3384-3388.

PAILLART, J.-C., SKRIPKIN, E., EHRESMANN, B., EHRESMANN, C. & MARQUET, R. 1996. A loop-loop" kissing" complex is the essential part of the dimer linkage of genomic HIV-1 RNA. *Proceedings of the National Academy of Sciences,* 93**,** 5572-5577.

PANCER, K., MILEWSKA, A., OWCZAREK, K., DABROWSKA, A., KOWALSKI, M., ŁABAJ, P. P., BRANICKI, W., SANAK, M. & PYRC, K. 2020. The SARS-CoV-2 ORF10 is not essential in vitro or in vivo in humans. *PLoS Pathogens,* 16**,** e1008959.

PARK, Y. W., WILUSZ, J. & KATZE, M. G. 1999. Regulation of eukaryotic protein synthesis: selective influenza viral mRNA translation is mediated by the cellular RNA-binding protein GRSF-1. *Proceedings of the National Academy of Sciences,* 96**,** 6694-6699.

PELLETIER, J. & SONENBERG, N. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature,* 334**,** 320-325.

PLANT, E. P., RAKAUSKAITE, R., TAYLOR, D. R. & DINMAN, J. D. 2010. Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. *Journal of virology,* 84**,** 4330-4340.

PLEGUEZUELOS, O., JAMES, E., FERNANDEZ, A., LOPES, V., ROSAS, L. A., CERVANTES-MEDINA, A., CLEATH, J., EDWARDS, K., NEITZEY, D. & GU, W. 2020. Efficacy of FLU-v, a broad-spectrum influenza vaccine, in a randomized phase IIb human influenza challenge study. *NPJ vaccines,* 5**,** 1-9.

PUNJANI, A., RUBINSTEIN, J. L., FLEET, D. J. & BRUBAKER, M. A. 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature methods,* 14**,** 290-296.

PUTRI, W. C., MUSCATELLO, D. J., STOCKWELL, M. S. & NEWALL, A. T. 2018. Economic burden of seasonal influenza in the United States. *Vaccine,* 36**,** 3960-3966.

RAMAN, S., BOUMA, P., WILLIAMS, G. D. & BRIAN, D. A. 2003. Stem-loop III in the 5′ untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *Journal of virology,* 77**,** 6720-6730.

RAMAN, S. & BRIAN, D. A. 2005. Stem-loop IV in the 5′ untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *Journal of virology,* 79**,** 12434-12446.

RANGAN, R., ZHELUDEV, I. N., HAGEY, R. J., PHAM, E. A., WAYMENT-STEELE, H. K., GLENN, J. S. & DAS, R. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *Rna,* 26**,** 937-959.

ROBERTSON, M. P., IGEL, H., BAERTSCH, R., HAUSSLER, D., ARES JR, M. & SCOTT, W. G. 2004. The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol,* 3**,** e5.

SANDERS, W., FRITCH, E. J., MADDEN, E. A., GRAHAM, R. L., VINCENT, H. A., HEISE, M. T., BARIC, R. S. & MOORMAN, N. J. 2020. Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. *BioRxiv*.

SCHROEDER, S. J. 2018. Challenges and approaches to predicting RNA with multiple functional structures. *Rna,* 24**,** 1615-1624.

SEDERDAHL, B. K. & WILLIAMS, J. V. 2020. Epidemiology and clinical characteristics of influenza C virus. *Viruses,* 12**,** 89.

SHANG, J., WAN, Y., LUO, C., YE, G., GENG, Q., AUERBACH, A. & LI, F. 2020. Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences,* 117**,** 11727-11734.

SHARMA, E., STERNE-WEILER, T., O'HANLON, D. & BLENCOWE, B. J. 2016. Global mapping of human RNA-RNA interactions. *Molecular cell,* 62**,** 618-626.

SHI, S. T., YU, G.-Y. & LAI, M. M. 2003. Multiple type A/B heterogeneous nuclear ribonucleoproteins (hnRNPs) can replace hnRNP A1 in mouse hepatitis virus RNA synthesis. *Journal of virology,* 77**,** 10584-10593.

SIEGFRIED, N. A., BUSAN, S., RICE, G. M., NELSON, J. A. & WEEKS, K. M. 2014. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods,* 11**,** 959-965.

SIMON, L. M., MORANDI, E., LUGANINI, A., GRIBAUDO, G., MARTINEZ-SOBRIDO, L., TURNER, D. H., OLIVIERO, S. & INCARNATO, D. 2019. In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic acids research,* 47**,** 7003-7017.

SKELLY, D. T., HARDING, A. C., GILBERT-JARAMILLO, J., KNIGHT, M. L., LONGET, S., BROWN, A., ADELE, S., ADLAND, E., BROWN, H. & TIPTON, T. 2021. Two doses of SARS-CoV-2 vaccination induce robust immune responses to emerging SARS-CoV-2 variants of concern. *Nature communications,* 12**,** 1-12.

SMOLA, M. J., RICE, G. M., BUSAN, S., SIEGFRIED, N. A. & 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. *Nature protocols,* 10**,** 1643-1669.

SNIJDER, E. J., LIMPENS, R. W., DE WILDE, A. H., DE JONG, A. W., ZEVENHOVEN-DOBBE, J. C., MAIER, H. J., FAAS, F. F., KOSTER, A. J. &

BÁRCENA, M. 2020. A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. *PLoS biology,* 18**,** e3000715.
SOLA, I., MATEOS-GOMEZ, P. A., ALMAZAN, F., ZUNIGA, S. & ENJUANES, L. 2011. RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. *RNA biology,* 8**,** 237-248.
SOLA, I., MORENO, J. L., ZÚÑIGA, S., ALONSO, S. & ENJUANES, L. 2005. Role of nucleotides immediately flanking the transcription-regulating sequence core in coronavirus subgenomic mRNA synthesis. *Journal of virology,* 79**,** 2506-2516.
SPRONKEN, M., VAN DE SANDT, C., DE JONGH, E., VUONG, O., VAN DER VLIET, S., BESTEBROER, T., OLSTHOORN, R., RIMMELZWAAN, G., FOUCHIER, R. & GULTYAEV, A. 2017. A compensatory mutagenesis study of a conserved hairpin in the M gene segment of influenza A virus shows its role in virus replication. *RNA biology,* 14**,** 1606-1616.
STAMMLER, S. N., CAO, S., CHEN, S.-J. & GIEDROC, D. P. 2011. A conserved RNA pseudoknot in a putative molecular switch domain of the 3′-untranslated region of coronaviruses is only marginally stable. *Rna,* 17**,** 1747-1759.
STOLOFF, G. A. & CAPARROS-WANDERLEY, W. 2007. Synthetic multi-epitope peptides identified in silico induce protective immunity against multiple influenza serotypes. *European journal of immunology,* 37**,** 2441-2449.
SÜKÖSD, Z., SWENSON, M. S., KJEMS, J. & HEITSCH, C. E. 2013. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic acids research,* 41**,** 2807-2816.
SUN, L., LI, P., JU, X., RAO, J., HUANG, W., REN, L., ZHANG, S., XIONG, T., XU, K. & ZHOU, X. 2021. In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. *Cell,* 184**,** 1865-1883. e20.
SUN, W., LUO, T., LIU, W. & LI, J. 2020. Progress in the Development of Universal Influenza Vaccines. *Viruses,* 12**,** 1033.
SWANN, H., SHARMA, A., PREECE, B., PETERSON, A., ELDREDGE, C., BELNAP, D. M., VERSHININ, M. & SAFFARIAN, S. 2020. Minimal system for assembly of SARS-CoV-2 virus like particles. *Scientific reports,* 10**,** 1-5.
SZABAT, M., LORENT, D., CZAPIK, T., TOMASZEWSKA, M., KIERZEK, E. & KIERZEK, R. 2020. RNA secondary structure as a first step for rational design of the oligonucleotides towards inhibition of influenza a virus replication. *Pathogens,* 9**,** 925.
TAKIZAWA, N., HIGASHI, K., KAWAGUCHI, R. K., GOTOH, Y., SUZUKI, Y., HAYASHI, T. & KUROKAWA, K. 2020. A functional RNA structure in the influenza A virus ribonucleoprotein complex for segment bundling. *bioRxiv*.
TANG, Y.-S., XU, S., CHEN, Y.-W., WANG, J.-H. & SHAW, P.-C. 2021. Crystal structures of influenza nucleoprotein complexed with nucleic acid provide insights into the mechanism of RNA interaction. *Nucleic Acids Research,* 49**,** 4144-4154.
TARUS, B., BAKOWIEZ, O., CHENAVAS, S., DUCHEMIN, L., ESTROZI, L., BOURDIEU, C., LEJAL, N., BERNARD, J., MOUDJOU, M. & CHEVALIER, C. 2012. Oligomerization paths of the nucleoprotein of influenza A virus. *Biochimie,* 94**,** 776-785.
TAUBENBERGER, J. K. & KASH, J. C. 2010. Influenza virus evolution, host adaptation, and pandemic formation. *Cell host & microbe,* 7**,** 440-451.

TAUBENBERGER, J. K. & MORENS, D. M. 2006. 1918 Influenza: the mother of all pandemics. *Revista Biomedica,* 17**,** 69-79.

TE VELTHUIS, A. J., ROBB, N. C., KAPANIDIS, A. N. & FODOR, E. 2016. The role of the priming loop in influenza A virus RNA synthesis. *Nature microbiology,* 1**,** 1-7.

THORVALDSDÓTTIR, H., ROBINSON, J. T. & MESIROV, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics,* 14**,** 178-192.

TICKLE, I., FLENSBURG, C., KELLER, P., PACIOREK, W., SHARFF, A., VONRHEIN, C. & BRICOGNE, G. 2018. Staraniso. *Global Phasing Ltd., Cambridge, UK*.

TIDU, A., JANVIER, A., SCHAEFFER, L., SOSNOWSKI, P., KUHN, L., HAMMANN, P., WESTHOF, E., ERIANI, G. & MARTIN, F. 2021. The viral protein NSP1 acts as a ribosome gatekeeper for shutting down host translation and fostering SARS-CoV-2 translation. *Rna,* 27**,** 253-264.

TINOCO, I., UHLENBECK, O. C. & LEVINE, M. D. 1971. Estimation of secondary structure in ribonucleic acids. *Nature,* 230**,** 362-367.

TOMEZSKO, P. J., CORBIN, V. D., GUPTA, P., SWAMINATHAN, H., GLASGOW, M., PERSAD, S., EDWARDS, M. D., MCINTOSH, L., PAPENFUSS, A. T. & EMERY, A. 2020. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature,* 582**,** 438-442.

TOWNSHEND, R. J., EISMANN, S., WATKINS, A. M., RANGAN, R., KARELINA, M., DAS, R. & DROR, R. O. 2021. Geometric deep learning of RNA structure. *Science,* 373**,** 1047-1051.

TURNER, D. H. & MATHEWS, D. H. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research,* 38**,** D280-D282.

TURRELL, L. 2015. *The role of nucleoprotein in transcription and replication of the Influenza virus genome.* DPhil thesis, University of Oxford

UCHAŃSKI, T., MASIULIS, S., FISCHER, B., KALICHUK, V., LÓPEZ-SÁNCHEZ, U., ZARKADAS, E., WECKENER, M., SENTE, A., WARD, P. & WOHLKÖNIG, A. 2021. Megabodies expand the nanobody toolkit for protein structure determination by single-particle cryo-EM. *Nature Methods,* 18**,** 60-68.

V'KOVSKI, P., KRATZEL, A., STEINER, S., STALDER, H. & THIEL, V. 2021. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology,* 19**,** 155-170.

VAN DOREMALEN, N., BUSHMAKER, T., MORRIS, D. H., HOLBROOK, M. G., GAMBLE, A., WILLIAMSON, B. N., TAMIN, A., HARCOURT, J. L., THORNBURG, N. J. & GERBER, S. I. 2020. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England journal of medicine,* 382**,** 1564-1567.

VONRHEIN, C., FLENSBURG, C., KELLER, P., SHARFF, A., SMART, O., PACIOREK, W., WOMACK, T. & BRICOGNE, G. 2011. Data processing and analysis with the autoPROC toolbox. *Acta Crystallographica Section D: Biological Crystallography,* 67**,** 293-302.

WALKER, A. P. & FODOR, E. 2019. Interplay between influenza virus and the host RNA polymerase II transcriptional machinery. *Trends in microbiology,* 27**,** 398-407.

WALTER, T. S., REN, J., TUTHILL, T. J., ROWLANDS, D. J., STUART, D. I. & FRY, E. E. 2012. A plate-based high-throughput assay for virus stability and vaccine formulation. *Journal of virological methods,* 185**,** 166-170.
WANG, M. & KONG, L. 2019. pblat: A multithread blat algorithm speeding up aligning sequences to genomes. *BMC bioinformatics,* 20**,** 1-4.
WANG, P. Y., SEXTON, A. N., CULLIGAN, W. J. & SIMON, M. D. 2019. Carbodiimide reagents for the chemical probing of RNA structure in cells. *RNA,* 25**,** 135-146.
WASHIETL, S., HOFACKER, I. L., STADLER, P. F. & KELLIS, M. 2012. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic acids research,* 40**,** 4261-4272.
WATTS, J. M., DANG, K. K., GORELICK, R. J., LEONARD, C. W., BESS JR, J. W., SWANSTROM, R., BURCH, C. L. & WEEKS, K. M. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature,* 460**,** 711-716.
WEEKS, K. M. 2021. SHAPE directed discovery of new functions in large RNAs. *Accounts of chemical research,* 54**,** 2502-2517.
WEI, Z., MCEVOY, M., RAZINKOV, V., POLOZOVA, A., LI, E., CASAS-FINET, J., TOUS, G. I., BALU, P., PAN, A. A. & MEHTA, H. 2007. Biophysical characterization of influenza virus subpopulations using field flow fractionation and multiangle light scattering: correlation of particle counts, size distribution and infectivity. *Journal of virological methods,* 144**,** 122-132.
WILKINSON, K. A., MERINO, E. J. & WEEKS, K. M. 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols,* 1**,** 1610-1616.
WILLIAMS, C. J., HEADD, J. J., MORIARTY, N. W., PRISANT, M. G., VIDEAU, L. L., DEIS, L. N., VERMA, V., KEEDY, D. A., HINTZE, B. J. & CHEN, V. B. 2018a. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science,* 27**,** 293-315.
WILLIAMS, G. D., TOWNSEND, D., WYLIE, K. M., KIM, P. J., AMARASINGHE, G. K., KUTLUAY, S. B. & BOON, A. C. 2018b. Nucleotide resolution mapping of influenza A virus nucleoprotein-RNA interactions reveals RNA features required for replication. *Nature communications,* 9**,** 1-12.
WU, H.-Y., GUAN, B.-J., SU, Y.-P., FAN, Y.-H. & BRIAN, D. A. 2014. Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5′-untranslated-region mutants. *Journal of virology,* 88**,** 846-858.
WUCHTY, S., FONTANA, W., HOFACKER, I. L. & SCHUSTER, P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers: Original Research on Biomolecules,* 49**,** 145-165.
XU, Y., LI, X., ZHU, B., LIANG, H., FANG, C., GONG, Y., GUO, Q., SUN, X., ZHAO, D. & SHEN, J. 2020. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nature medicine,* 26**,** 502-505.
YANG, D. & LEIBOWITZ, J. L. 2015. The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus research,* 206**,** 120-133.
YANG, D., LIU, P., GIEDROC, D. P. & LEIBOWITZ, J. 2011. Mouse hepatitis virus stem-loop 4 functions as a spacer element required to drive subgenomic RNA synthesis. *Journal of virology,* 85**,** 9199-9209.
YANG, D., LIU, P., WUDECK, E. V., GIEDROC, D. P. & LEIBOWITZ, J. L. 2015. SHAPE analysis of the RNA secondary structure of the Mouse Hepatitis Virus 5'untranslated region and N-terminal nsp1 coding sequences. *Virology,* 475**,** 15-27.

YANG, S. L., DEFALCO, L., ANDERSON, D. E., ZHANG, Y., AW, J. G. A., LIM, S. Y., LIM, X. N., TAN, K. Y., ZHANG, T., CHAWLA, T., SU, Y., LEZHAVA, A., MERITS, A., WANG, L.-F., HUBER, R. G. & WAN, Y. 2021. Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. *Nature Communications,* 12**,** 5113.

YAO, H., SONG, Y., CHEN, Y., WU, N., XU, J., SUN, C., ZHANG, J., WENG, T., ZHANG, Z. & WU, Z. 2020. Molecular architecture of the SARS-CoV-2 virus. *Cell,* 183**,** 730-738. e13.

YE, Q., KRUG, R. M. & TAO, Y. J. 2006. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature,* 444**,** 1078-1082.

YORK, A., HENGRUNG, N., VREEDE, F. T., HUISKONEN, J. T. & FODOR, E. 2013. Isolation and characterization of the positive-sense replicative intermediate of a negative-strand RNA virus. *Proceedings of the National Academy of Sciences,* 110**,** E4238-E4245.

YUAN, S., BALAJI, S., LOMAKIN, I. B. & XIONG, Y. 2021. Coronavirus Nsp1: Immune Response Suppression and Protein Expression Inhibition. *Frontiers in Microbiology,* 12.

ZARRINGHALAM, K., MEYER, M. M., DOTU, I., CHUANG, J. H. & CLOTE, P. 2012. Integrating chemical footprinting data into RNA secondary structure prediction.

ZHANG, P., MENDONCA, L., HOWE, A., GILCHRIST, J., SUN, D., KNIGHT, M., ZANETTI-DOMINGUES, L., BATEMAN, B., KREBS, A.-S. & CHEN, L. 2021. Correlative Multi-scale Cryo-imaging Unveils SARS-CoV-2 Assembly and Egress.

ZIV, O., GABRYELSKA, M. M., LUN, A. T. L., GEBERT, L. F. R., SHEU-GRUTTADAURIA, J., MEREDITH, L. W., LIU, Z.-Y., KWOK, C. K., QIN, C.-F. & MACRAE, I. J. 2018. COMRADES determines in vivo RNA structures and interactions. *Nature methods,* 15**,** 785-788.

ZIV, O., PRICE, J., SHALAMOVA, L., KAMENOVA, T., GOODFELLOW, I., WEBER, F. & MISKA, E. A. 2020. The short-and long-range RNA-RNA Interactome of SARS-CoV-2. *Molecular cell,* 80**,** 1067-1077. e5.

ZOU, X., CHEN, K., ZOU, J., HAN, P., HAO, J. & HAN, Z. 2020. Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Frontiers of medicine***,** 1-8.

ZUNIGA, S., SOLA, I., ALONSO, S. & ENJUANES, L. 2004. Sequence motifs involved in the regulation of discontinuous coronavirus subgenomic RNA synthesis. *Journal of virology,* 78**,** 980-994.

ZWART, P. H., GROSSE-KUNSTLEVE, R. W. & ADAMS, P. D. 2005. Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newsl,* 43**,** 27-35.

Supplementary Material

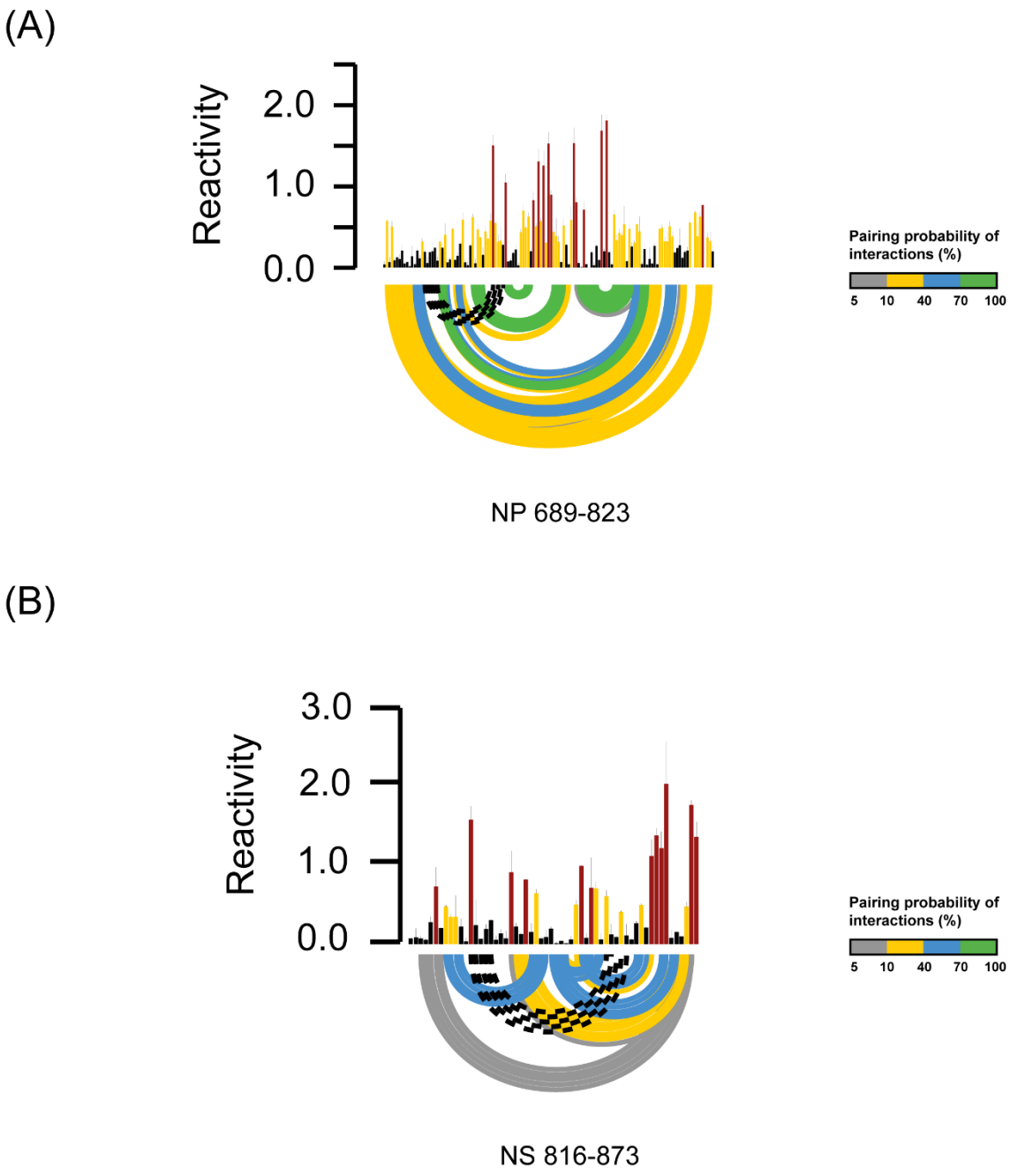


Figure S1: Predicted pseudoknots in the NP (A) and NS (B) segments.

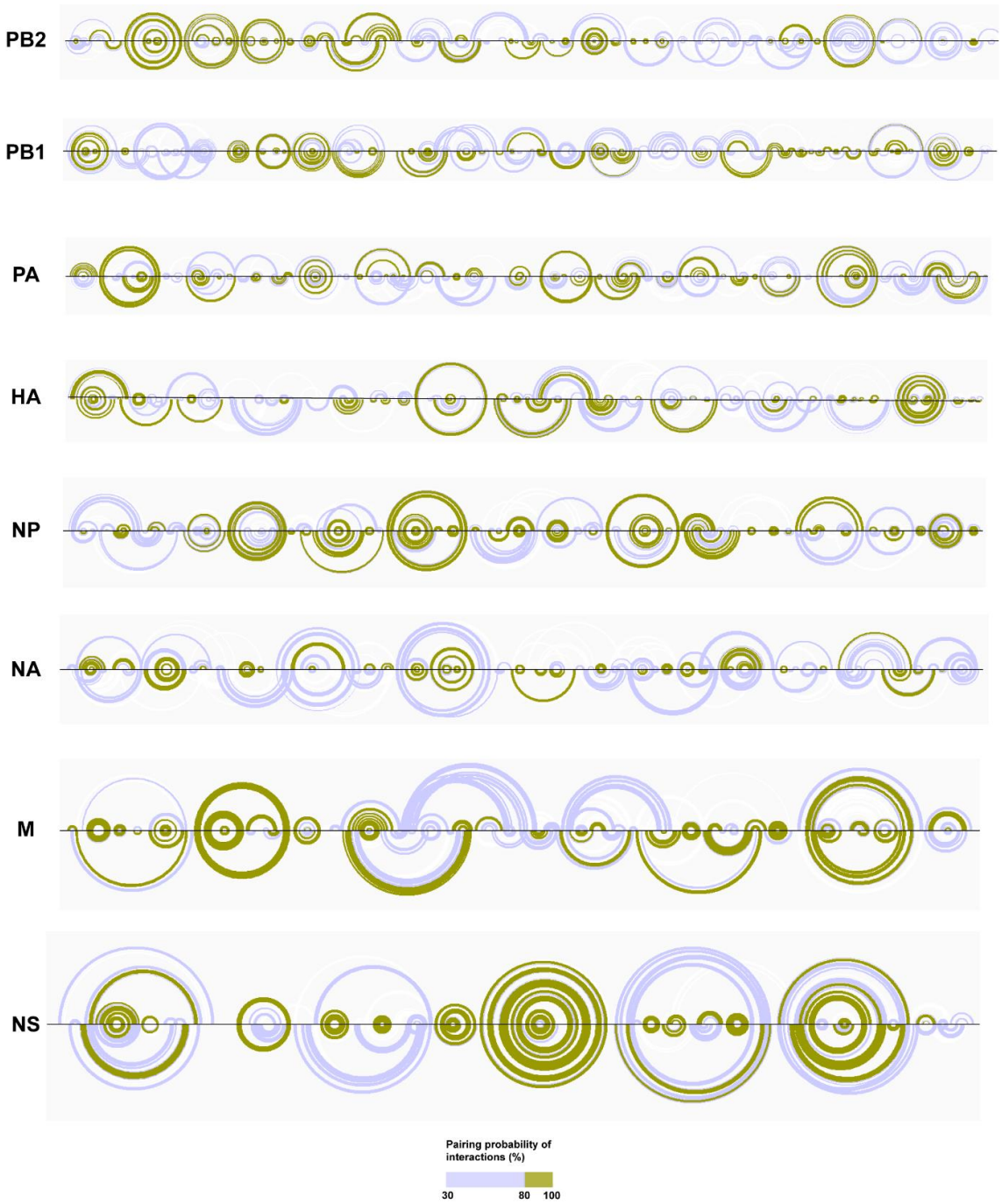


Figure S2: A comparison of the SHAPE informed structure predictions for the PR8 virus performed in this investigation (above the line) and a previous investigation (Dadonaite et al., 2019).

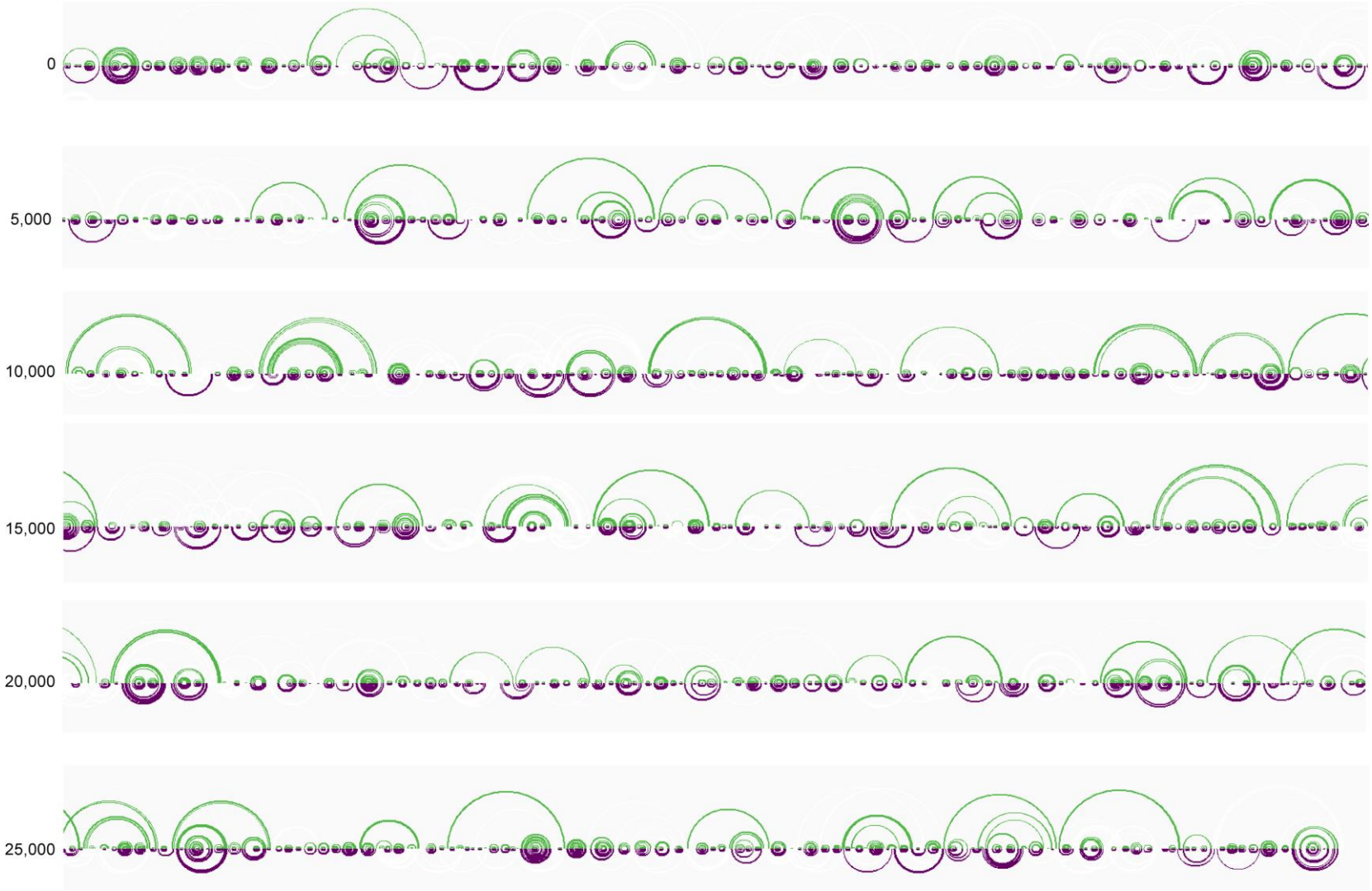


Figure S3: A comparison between the SHAPE guided RNA structures predicted for the SARS-CoV-2 genome when maximum pairing distance is set to 500 (the top arcs in light green) or 200 (the bottom arcs in purple). Numbers indicate the position in the genome.

Sample	Median read depth per base	Median mapping quality (phred scaled)	Median base call quality (phred scaled)	Median Read length	Median base mutation rate (%)
PR8_1M7_1	39,161	43.2	45.9	99.9	0.53
PR8_1M7_2	37,433	42.9	47.8	99.9	0.76
PR8_Untreated_1	39,688	43.2	47.8	99.8	0.22
PR8_Untreated_2	43,835	43.1	47.6	99.9	0.23
PR8-Ud_1M7_1	28,777	42.7	50.7	99.5	0.62
PR8-Ud_1M7_2	38,078	42.8	47.4	100	0.65
PR8-Ud_Untreated_1	41,015	43.1	46.6	100	0.23
PR8-Ud_Untreated_2	39,626	43.1	46.3	100.1	0.2
PR8-Wyo_1M7_1	45,879	43	46.5	100.1	0.6
PR8-Wyo_1M7_2	43,170	43.2	47.1	99.9	0.66
PR8-Wyo_Untreated_1	47,562	43.2	46.1	100.1	0.22
PR8_Wyo_Untreated_2	47,561	43.4	46.9	100	0.29
PR8_Wyo(subs)_1M7_1	38,558	42.6	48.2	99.9	0.65
PR8_Wyo(subs)_1M7_2	43,236	43.2	46.8	100	0.68
PR8_Wyo(subs)_Untreated_1	50,781	43.2	45.8	100.1	0.24
PR8_Wyo(subs)_Untreated_2	48,282	43.4	46.4	100	0.28
Udorn_1M7_1	40,557	43	47.1	99.9	0.79
Udorn_1M7_2	43,829	43	46.9	100.1	0.7
Udorn_Untreated_1	47,164	43.1	47	100	0.23
Udorn_Untreated_2	53,596	43.1	46.1	100.1	0.32
Wyoming_1M7_1	14,204	42.9	48.6	99.9	0.68
Wyoming_1M7_2	15,479	42.9	47.7	99.9	0.68
Wyoming_Untreated_1	17,048	42.8	49.9	99.6	0.28
Wyoming_Untreated_2	21,964	43	48.2	99.9	0.22
WSN_1M7_1	17,103	43.6	37.5	72.6	0.29
WSN_1M7_2	19,741	43.6	38	72.6	0.31
WSN_Untreated_1	19,895	43.6	38.6	72.3	0.16
WSN_Untreated_2	11,729	43.5	42.1	71.6	0.16
WSN_DMS_1	2,096	41.8	43	71.6	1.82
WSN_DMS_2	2,670	41.8	42.2	72	1.89
WSN_EDC_1	17,863	43.5	39	72.3	0.25
WSN_EDC_2	18,473	43.6	38.6	72.4	0.26

Table S1: *Statistics from the sequencing of the chemical probing experiments performed on IAV.*