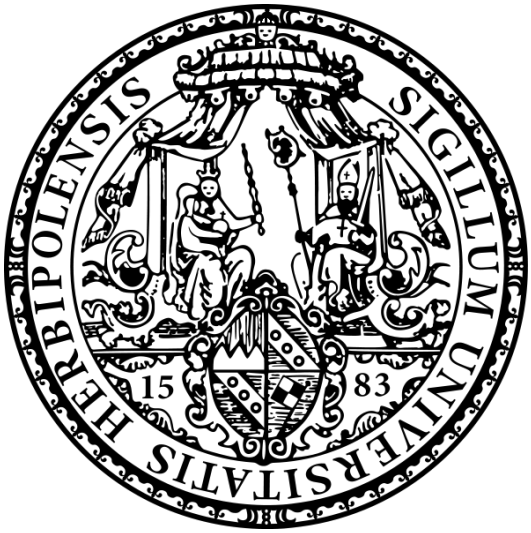


RNA-RNA interactions in viral genome packaging

RNA-RNA-Interaktionen bei der viralen Genomverpackung

Doctoral thesis for a doctoral degree at the Graduate School of Life Sciences,

Section Infection and Immunity,

Julius-Maximilians-Universität Würzburg

submitted by

Liqing Ye

from China

Würzburg 2022

Submitted on: 22nd July 2022

Members of the Thesis Committee

Chairperson: Prof. Dr. Christian Janzen

Primary Supervisor: Jun Prof. Dr. Redmond Smyth

Supervisor (Second): Prof. Dr. Sibylle Schneider-Schaulies

Supervisor (Third): Prof. Dr. Cynthia Sharma

Date of Public Defence:

Date of Receipt of Certificates:

Affidavit/ Eidesstattliche Erklärung

I hereby confirm that my thesis entitled “RNA-RNA interactions in viral genome packaging” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date Signature

Hiermit erkläre ich an Eides statt, die Dissertation “RNA-RNA-Interaktionen bei der viralen Genomverpackung” eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ort, Datum Unterschrift

Summary

RNA is one of the most abundant macromolecules and plays essential roles in numerous biological processes. This doctoral thesis consists of two projects focusing on RNA structure and RNA-RNA interactions in viral genome packaging. In the first project I developed a method called Functional Analysis of RNA Structure (FARS-seq) to investigate structural features regulating genome dimerization within the HIV-1 5'UTR. Genome dimerization is a conserved feature of retroviral replication and is thought to be a prerequisite for binding to the viral structural protein Pr55Gag during genome packaging. It also plays a role in genome integrity and evolution through recombination, and is linked to a structural switch that may regulate genome packaging and translation within cells. Despite its importance for HIV-1 replication, the RNA signals regulating genome dimerization, and the molecular mechanism leading to the selection of the genome dimer over the monomer for packaging are incompletely understood. The FARS-seq method combines RNA structural information obtained by chemical probing with single nucleotide resolution profiles of RNA function obtained by mutational interference. In this way, we found nucleotides that were critical for dimerization, especially within the well-characterized dimerization motif within stem-loop 1 (SL1). We also found stretches of nucleotides that enhanced genome dimerization upon mutation, suggesting their role in negatively regulating dimerization. A structural analysis identified distinct structural signatures within monomeric and dimeric RNA. The dimeric conformation displayed the canonical transactivation response (TAR), PolyA, primer binding site (PBS), and SL1-SL3 stem-loops, and contained a long range U5-AUG interaction. Unexpectedly, in monomeric RNA, SL1 was reconfigured into long- and short-range base-pairings with PolyA and PBS, respectively. Intriguingly, these base pairings concealed the palindromic sequence needed for dimerization and disrupted the internal loop in SL1 previously shown to contain the major packaging motif for Pr55Gag. We therefore rationally introduced mutations into PolyA and PBS, and showed how these regions regulate genome dimerization, and the binding of Pr55Gag *in vitro*, as well as genome packaging into virions. These findings give insights into late stages of the HIV-1 life cycle and a mechanistic explanation for the link between RNA dimerization and packaging.

In the second project, I developed a proximity ligation and high-throughput sequencing-based method, RNA-RNA seq, which can measure direct (RNA-RNA) and indirect (protein-mediated) interactions. In contrast to existing methods, RNA-RNA seq is not limited by specific protein or RNA baits, nor to a particular crosslinking reagent. The genome of influenza A virus contains eight segments, which assemble into a "7+1" supramolecular complex. However, the molecular details of genome assembly are poorly understood. Our goal is to use RNA-RNA seq to identify the sites of interaction between the eight genomic RNAs of influenza, and to use this information to define the quaternary RNA architecture of the genome. We showed that RNA-RNA seq worked on model substrates, like the HIV-1 Dimerization Initiation Site (DIS) RNA and purified ribosome, as well as influenza A virus infected cells.

Zusammenfassung

RNA ist eines der am häufigsten vorkommenden Makromoleküle und spielt bei allen biologischen Prozessen eine wesentliche Rolle. Diese Doktorarbeit besteht aus zwei Projekten, die sich auf die RNA-Struktur und RNA-RNA-Interaktionen bei der viralen Genomverpackung konzentrieren. Im ersten Projekt habe ich eine Methode namens Functional Analysis of RNA Structure (FARS-seq) entwickelt, um strukturelle Merkmale zu untersuchen, die die Genom-Dimerisierung innerhalb des HIV-1 5'UTR regulieren. Die Genomdimerisierung ist ein konserviertes Merkmal der retroviralen Replikation und gilt als Voraussetzung für die Bindung an das virale Strukturprotein Pr55Gag während der Genomverpackung. Sie spielt auch eine Rolle bei der Genomintegrität und -evolution durch Rekombination und ist mit einem strukturellen Schalter verbunden, der die Genomverpackung und -translation in Zellen regulieren kann. Trotz der Bedeutung für die HIV-1-Replikation sind die RNA-Signale welche die Genom-Dimerisierung regulieren, und der molekulare Mechanismus der zur Auswahl des Genom-Dimers gegenüber dem Monomer für die Verpackung führt, nur unvollständig verstanden. FARS-seq kombiniert RNA-Strukturinformationen, die durch chemisches Sondieren gewonnen werden, mit Profilen der RNA-Funktion in Einzelnukleotid-Auflösung, die durch Mutationsinterferenz gewonnen werden. Auf diese Weise fanden wir Nukleotide, die für die Dimerisierung kritisch sind, insbesondere innerhalb des gut charakterisierten Dimerisierungsmotivs von stem-loop 1 (SL1). Wir fanden auch Nukleotidabschnitte, die bei Mutation die Dimerisierung des Genoms verstärkten, was auf ihre Rolle bei der negativen Regulierung der Dimerisierung hindeutet. Eine Strukturanalyse ergab zudem unterschiedliche strukturelle Signaturen innerhalb der RNA von Monomeren und Dimeren. Die dimere Konformation wies die kanonische Transaktivierungsantwort (TAR), PolyA, die Primerbindestelle (PBS) und SL1-SL3-stem loops auf und enthielt eine weitreichende U5-AUG-Interaktion. Unerwarteterweise interagierte SL1 in monomerer RNA mit dem weit entfernten PolyA Signal und der nahegelegenen PBS. Interessanterweise verbargen diese Basenpaarungen die für die Dimerisierung erforderliche palindromische Sequenz und unterbrachen die interne Schleife in SL1, von der zuvor gezeigt wurde, dass sie das Hauptverpackungsmotiv für Pr55Gag enthält. Wir haben daher auf rationale Weise Mutationen in PolyA und PBS eingeführt und gezeigt, wie diese Regionen die Dimerisierung des Genoms und die Bindung von Pr55Gag *in vitro* sowie die Verpackung des Genoms in Virionen regulieren. Diese Ergebnisse geben Einblicke in späte Stadien des HIV-1-Lebenszyklus und eine mechanistische Erklärung für die Verbindung zwischen RNA-Dimerisierung und Verpackung.

Im zweiten Projekt entwickelte ich eine auf Proximity Ligation und Hochdurchsatz-Sequenzierung basierende Methode, RNA-RNA seq, mit der direkte (RNA-RNA) und indirekte (proteinvermittelte) Wechselwirkungen gemessen werden können. Im Gegensatz zu bestehenden Methoden ist RNA-RNA seq nicht durch spezifische Protein- oder RNA Crosslink-Reagenzien eingeschränkt. Das Genom des Influenza-A-Virus besteht aus acht Segmenten, die zu einem supramolekularen "7+1"-Komplex assemblieren. Die molekularen Details des Genomaufbaus sind jedoch kaum bekannt. Unser Ziel ist es, mit Hilfe von RNA-RNA seq die Interaktionsstellen zwischen den acht genomischen RNAs des Influenza-Virus zu identifizieren und somit die quaternäre RNA-Architektur des Genoms zu definieren.

Wir haben gezeigt, dass RNA-RNA seq an Modellsubstraten wie der HIV Dimerization Initiation Site (DIS) RNA und gereinigtem Ribosom sowie an mit dem Influenza-A-Virus infizierten Zellen funktioniert.

List of Contents

Chapter 1 General introduction

1.1 RNA, RNA structure and RNA-RNA interaction in virus infection

RNAs are unique and dynamic molecules, playing essential roles in numerous cellular processes beyond the central dogma of molecular biology[1][2][3]. They can not only incorporate information within the nucleotide sequence, but also fold into a variety of structures through intra- and inter-molecular RNA-RNA interactions, such as stem-loops, pseudoknot structures, and kissing loop complexes. These structures often associate with various RNA binding proteins (RBP), which enable their diverse functions in cells. Most RNAs functions depend on RNA secondary and tertiary structure[3][4][5]. For example, ribozymes, which form specific structures to catalyse cis- or trans-RNA cleavage and splicing[6]. Another long and well-known example is ribosomal RNAs (rRNAs), which form the framework of the machinery that catalyses protein synthesis during translation[7][8].

RNA-RNA interactions and RNA structures play key roles in every part of the RNA virus infection[2][9][10][11][12][13][14][15], including viral protein translation, viral replication, messenger RNA (mRNA) splicing regulation, and RNA packaging. For instance, hepatitis C virus (HCV) has a positive-strand RNA genome that lacks a 5' cap and a 3' poly(A) tail. Instead of being mediated by host eukaryotic initiation factors (eIFs), HCV uses internal ribosome entry sites (IRES) to translate viral proteins[11][13][16]. IRES are highly structured RNA elements within the 5’ untranslated region (UTR), which recruit ribosomes and mediate cap-independent translation initiation. The HCV IRES adopts a complex structure containing two major domains: II and III (**Figure 1.1a**). Domain II is located upstream of domain III and consists of two sub-domains, IIa, a small internal loop, and IIb, an apical loop and internal loop. The larger domain III displays branching hairpin stem-loops: IIIa/b/c/d/e/f. The basal part of domain III is composed of a predicted pseudoknot (IIIf) and a stem-loop (IIIe), which display as a 4-way junction. The middle part of domain III comprises the stem-loop (IIId) incorporated into a 3-way junction. And the upper part of domain III contains a 4-way junction consisting of three stem-loops, IIIa/b/c (**Figure 1.1a**). HCV IRES translation initiation begins with the 40S ribosomal subunit interacting with pseudoknot (IIIf), stem-loops IIId, IIIe, IIb, and IIIc. This promotes the binding of eIF3 and the Met-tRNAiMet-eIF2-GTP ternary complex to form a 48S particle with the AUG initiator codon positioned in the ribosomal P-site without ribosomal scanning. Consequently, the active 80S is assembled with eIF5, eIF5B, GTP and the 60S ribosomal subunit, and translation starts. Furthermore, it has been shown that the long-distance interaction between the 5’UTR with the 3’UTR, which circularizes the HCV mRNA, enhances translation efficiency due to avoidance of mRNA degradation by exonucleases and encourages ribosome recycling[17].

RNA structure and RNA-RNA interaction also regulate viral genome replication. For example, it is reported that the replication of dengue virus (DENV), a member of genus Flavivirus with a positive-strand RNA genome, requires genome circularization, which is mediated by base-pairing interactions between sequences in its genomic termini (**Figure 1.1b**)[14][18][19]. It was demonstrated that there were three complementary regions involved in these interactions, including the cyclization sequence

(CS), the upstream of AUG region (UAR) and the downstream of AUG region (DAR). After genomic RNA termini base-pairing and circularization, the RNA-dependent RNA polymerase (RdRp), which binds to the 5' termini, can reposition to the 3' end of the genome to start negative-strand RNA synthesis. The circularized RNA has been observed by atomic force microscopy (AFM) in the absence of proteins[20]. Chemical probing results showed the circularization was not a protein-dependent interaction, suggesting that this process was mainly driven by RNA-RNA interactions. In addition, the circularization is negatively regulated by the short internal loop region between 3'UAR and 3'DAR. If this region forms a small stem-loop with the partial left and right region, which disrupts the base pairing of 5'UAR-3'UAR and 5'DAR-3'DAR, the RNA circularization is down regulated[21].

Viral mRNA splicing is also regulated by RNA-RNA interactions and RNA secondary structures. Human immunodeficiency virus 1 (HIV-1) transcribes full length RNA, which can either function as the genomic RNA, or as translation template for the Gag-Pol protein, or can be alternatively spliced into mRNAs that are used for translation of the other essential regulatory and accessory proteins[22][23][24][25]. There are at least 5 splicing donor sites (5'splice site, 5' ss) and 8 to 9 splicing acceptor sites (3'splice site, 3'ss) on the HIV-1 RNA, which makes the splicing process very complicated and results in more than 40 different spliced mRNA species[26]. The major 5'ss plays a key role for HIV-1 mRNA splicing, because it is involved in the production of all spliced mRNAs. The 11nt 5'ss (5'CUGGUGAGUAC3') is complementary to the 5' end of U1 small nuclear RNA (snRNA) (3'GUCCAUUCAUA5'), which mediates U1 snRNA annealing (U1 snRNP binding). A stem-loop structure containing the 5'ss region is reported to influence splicing by hindering annealing between the 5'ss and U1 snRNA (**Figure 1.1c**). For example, mutations that stabilize the stem-loop but have no effect on U1 snRNA annealing, or extensions to the stem, reduced HIV-1 splicing efficiency[23][27]. Furthermore, it was also shown that the conformation of the HIV-1 5'UTR affected the splicing process. The splicing of HIV-1 RNA is regulated by RNA dimerization, which is the noncovalent association of two HIV-1 genomes. The unspliced full length HIV-1 RNA can form either dimer, which functions as genomic RNA and will be packaged into virion, or monomer, which is retained in cells and functions as mRNA. When the HIV-1 RNA forms into monomer, splicing is favoured because the monomer conformation exposes the major 5'ss; while if it folds into dimer, the 5'ss is sequestered, which leads to less splicing[22][28]. Another example is influenza A virus[12][29], whose genome is comprised of eight single-stranded negative RNA segments, which encode more than 11 viral proteins. The mRNA transcribed from the M and NS segments can be spliced and translated to different proteins. The M segment encodes matrix protein M1, and translated from un-spliced mRNA. The M2 protein, function as ion channel and is translated from spliced mRNA. It was reported that M encodes other small polypeptides with unknown function[12]. The 3'ss of the M segment mRNA, where SF2/ASF splicing factor binds, can fold into two different structures: a pseudoknot or a hairpin (**Figure 1.1d**). In the pseudoknot conformation, the splice site is hidden within a helix, whereas in the hairpin shape, it is accessible within an internal loop[12]. Similar to the M segment, the mRNA of the NS segment also undergoes splicing. This process is also regulated by the 3'ss RNA structure. The NS segment encodes the NS1 protein, which inhibits host immune response to influenza infection, and the spliced mRNA of NS is translated into the NS2 protein, which helps vRNP nuclear export[30]–[32].

It was reported that the 5'ss and 3'ss of pre-mRNA of NS can fold into different secondary structures, which regulate the equilibrium of NS1 mRNA and spliced mRNA (NS2). (Reviewed in[33]).

RNA structures and RNA-RNA interactions regulate virus genome packaging is further reviewed in Chapter 2[34].

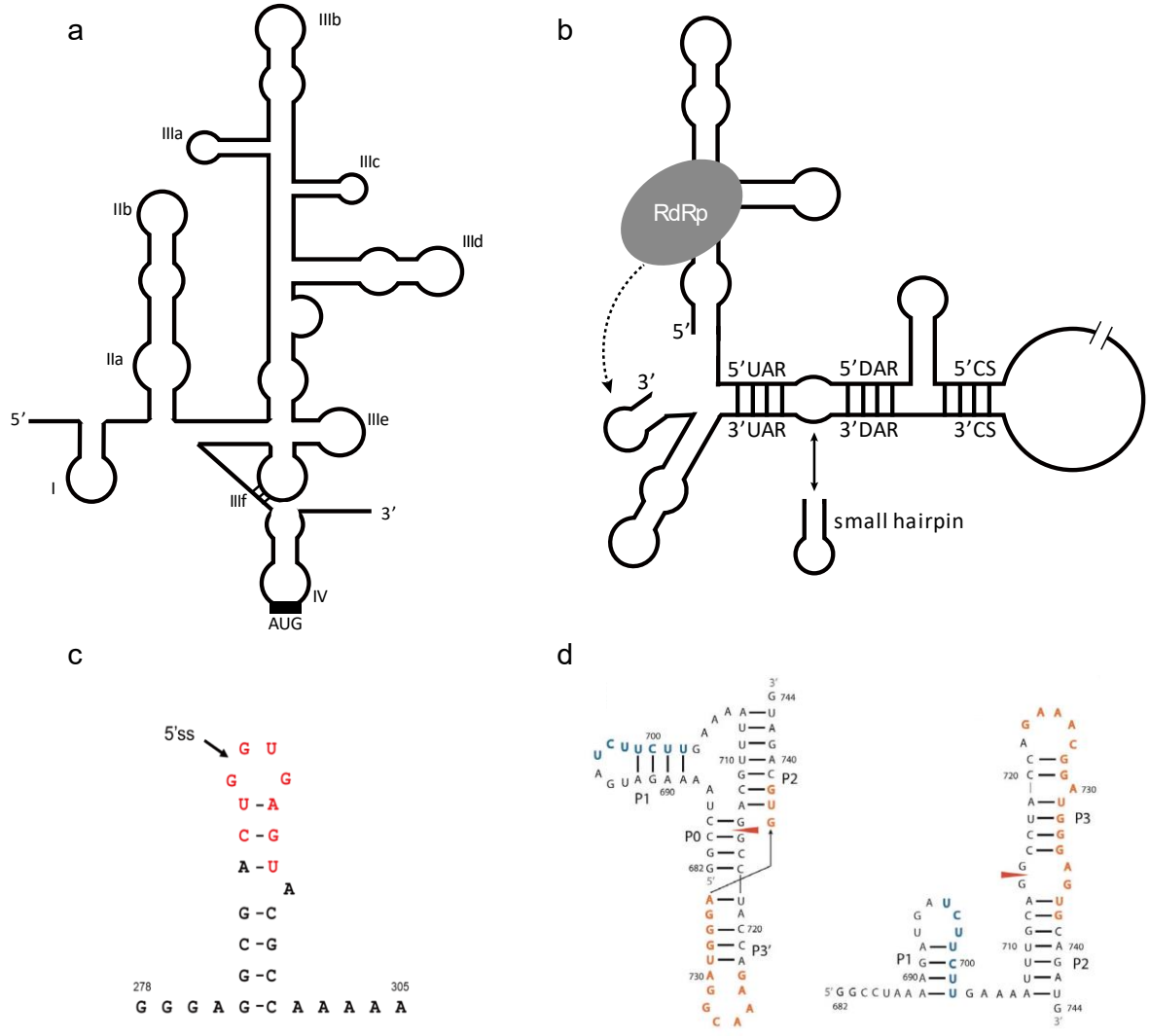


Figure 1.1| RNA structures and RNA-RNA interactions play important roles in virus infection. (a) HCV IRES structure regulates HCV RNA translation initiation; **(b)** DENV genomic RNA circularization regulation by 5' and 3' RNA ends interactions; **(c)** HIV-1 major 5'splice site forms a stem-loop structure[23]; **(d)** Influenza A virus M segment 3' splicing site can fold into a pseudoknot (left) or a hairpin (right) secondary structure, and regulate M segment mRNA splicing[12] (Adapted from https://doi.org/10.1371/journal.pone.0038323.g001).

1.2 RNA structure probing and RNA-RNA interaction methods

Many biophysical methods have been developed to determine RNA structure [35] (**Table 1.1**). The first yeast phenylalanine transfer RNA (tRNA) crystal structure from X-ray crystallography at atomic resolution was reported in 1970s[36][37]. From then on, X-ray crystallography became an essential tool to study RNA structures. However, many biological important RNAs are flexible, which makes crystallization difficult. Nuclear magnetic resonance (NMR) spectroscopy is an alternative powerful tool to analyse conformations of dynamic and flexible RNAs, but NMR is limited to the analysis of short RNAs [38][39][40][41]. For large RNAs, small angle X-ray scattering (SAXS)[42][43][44] and cryogenic electron microscopy (cryo-EM)[45][46][47] are more tractable. SAXS can also be applied on relatively large RNAs, but it only provides low information on global structure and cannot access the influence of the environment onto the RNA structure. Cryo-EM captures frozen-hydrated particles in their native state, and is an emerging technique for solving the structures of large RNA domains to near atomic resolution[48], [49]. Alternative to the biophysical methods, are the sequencing-based methods to investigate RNA secondary structure. These methods have become more popular with the rise of next generation sequencing, and can rapidly obtain secondary structure on a genome wide level. Such methods use enzymatic cleavage and chemical modification of RNA structures, or are based on crosslinking and proximity ligation [4][50][51][52][53].

Enzyme-based methods depend on the selectivity of nucleases to cut either single-stranded or double-stranded RNA regions, including Fragmentation sequencing (Frag-Seq)[54], Parallel Analysis of RNA Structure (PARS)[55], Parallel Analysis of RNA Structures with Temperature Elevation (PARTE)[56], and Protein Interaction Profile sequencing (PIP-seq)[57]. The commonly used enzymes include RNase V1, which cleaves double-stranded or structured regions within RNAs without base specificity; S1 nuclease and RNase P1, which are zinc-dependent endonucleases that cut RNA in single-stranded regions without base specificity; RNase T1 cuts single-stranded RNA region at unpaired guanosine, while RNase T2 prefers to cleave at unpaired adenosine; RNase A also cuts single-stranded regions but only at pyrimidine residues. After the specific enzyme cleavage, the RNA fragments are prepared for sequencing through the ligation of sequencing adaptors. By combining the sequencing information from different enzyme treatments, one obtains a view of the single- and double-stranded regions of an RNA, which can be used to predict and remodel the RNA secondary structure. A limitation of enzyme-based approaches is that it can be only applied *in vitro*, because the large size of RNases makes them impermeable to the cell membrane.

Chemical-based methods utilize small molecules to probe RNA structure. The advantage of chemical probing is that they are small, which makes them permeable to cell membranes and suitable for *in vivo* study[53][58][59][60]. Besides, many different chemicals can be used for RNA structure probing (**Figure 1.2**). For example, some chemicals modify the functional groups on the Watson–Crick or Hoogsteen face of the base: Dimethyl sulfate (DMS), which methylates the N1 position of adenine (N1A) and the N3 position of cytosine (N3C), can be used to identify unpaired adenosine and cytosine nucleotides. DMS also methylates the N7 position of guanine (N7G), but it needs additional biochemical steps to detect the modification because it is located at the Hoogsteen face. So, it is not

commonly used for guanine probing except G-quartets. 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metha-p-toluene sulfonate (CMCT) reacts with the N3 position of unpaired uridine (N3U) and the N1 position of unpaired guanine (N1G) under slightly basic condition, and the modifications can be reversed under slightly acidic conditions or stabilised in the presence of borate ions. Diethylpyrocarbonate (DEPC) modifies the N7 position of adenine (N7A) under neutral pH and it can also detect the adenosine implicated in tertiary interactions after aniline treatment. Another class of chemicals used for RNA structure probing is ribose-specific molecule probes which react with the RNA backbone. 2'-hydroxyl acylation analysed by primer extension (SHAPE) reagents selectively acylate the 2'hydroxyl group of the ribose of all four nucleotides in flexible (normally single-stranded) regions. In addition, Lead (II) can be used to probe RNA secondary structure because it cleaves the single-stranded RNA region by hydrolyzing the phosphodiester backbone[61] (**Figure 1.2**). There are different chemical-based approaches have been developed and used for RNA structure probing, such as dimethyl sulfate sequencing (DMS-seq)[62], dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq)[63], *in vivo* click selective 2-hydroxyl acylation and profiling experiment (icSHAPE)[64], selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP)[65][66], Structure-seq[67], [68], high-throughput sequencing for chemical probing of RNA structure (Mod-seq)[69]. The workflow of these methods is similar: (i) chemical probing of RNA (*in vivo* or *in vitro*), (ii) reverse transcription of modified RNAs, (iii) sequencing. Reverse transcription will either stop or introduce mutations into cDNAs at sites of modification, and by analyzing the sequencing data, we can profile RNA secondary structure.

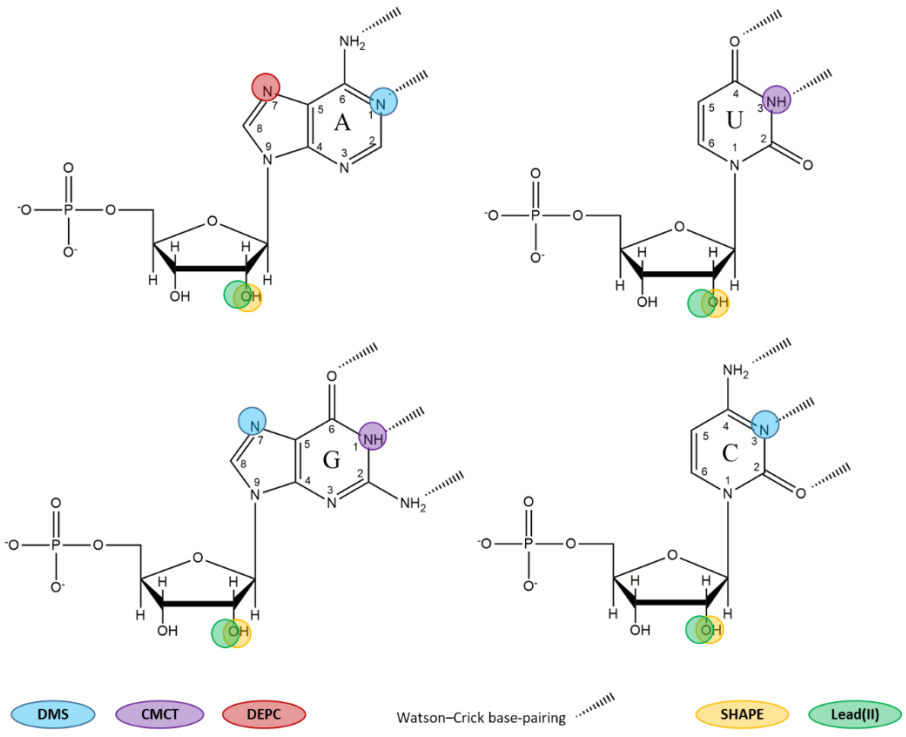


Figure 1.2| RNA probing reagents and their modification position. Base-specific probes like DMS, CMCT, and DEPC. DMS methylates the N1 position of adenine (N1A), the N3 position of cytosine (N3C) and the N7 position of guanine (N7G); CMCT modifies the N3 position of uridine (N3U) and the N1 position of guanine (N1G); and DEPC reacts with N7A in single-stranded region. Ribose/phosphate target probes like SHAPE reagents and Lead(II), acylate the 2'hydroxyl group of the ribose of all four nucleotides and hydrolyze the phosphodiester backbone in single-stranded regions, respectively.

Table 1.1 Methods for Mapping RNA Structure

CLASS	METHODS		ADVANTAGES		LIMITATIONS
BIOPHYSICAL TECHNIQUE	X-ray	✓	Atomic resolution	○	Difficult to resolve the flexible region
				○	Influence of environment cannot be assessed
	NMR	✓	Atomic resolution	○	Only for small RNA motifs
				○	Restrictions on the buffer composition
				○	Influence of environment cannot be assessed
	cryo-EM	✓	Large RNAs	○	Still moderate resolution
				○	Limited to large structure
				○	Influence of environment cannot be assessed
	SAXS	✓	Large RNAs	○	Low information on global structure
				○	Influence of environment cannot be assessed
ENZYME-BASED APPROACH	Frag-seq	✓	Simple and fast protocol	○	Need endogenous controls
		✓	Accompanied with modifiable software	○	Potential for contamination between samples and

			controls
	PARS	✓ Increased sensitivity by sequencing both single- and double-stranded regions	○ RNA was folded *in vitro*
	PARTE	✓ Measures melting temperature ✓ Single-nucleotide resolution ✓ Preserves in vivo RNA modifications ✓ Can infer RNA regulatory motifs	○ RNA was folded *in vitro* ○ RNase V1 treatment in different temperature and concentration
	PIP-seq	✓ Reveals both protein-bound RNA regions and RNA secondary structure ✓ Provides strand-specific information	○ Limited resolution at small nucleotide bulges and loops
CHEMICAL-BASED APPROACH	DMS-seq	✓ Identifies RNA structure in native conditions ✓ Single-nucleotide resolution	○ Limited to the analysis of two bases (As and Cs) ○ RNA-binding proteins can block DMS activity
	DMS-MaPseq	✓ Identifies RNA structure in native conditions ✓ Single-nucleotide resolution	○ Limited by reverse transcription length
	icSHAPE	✓ Measures base flexibility ✓ Single-nucleotide resolution	○ Limited to the analysis of relatively short (~300 nt) *in vitro* transcribed RNAs
	SHAPE-MaP	✓ Can be customized for	○ Length of the RNA must be

		different applications ✓ Applicable to analysis of long RNAs ✓ Can infer structural changes of single-nucleotide and other allelic polymorphisms	at least 150 nt for the randomer and native workflow, and at least 40 nt for the small-RNA workflow

RNA structure is dynamic and highly modulated by diverse environmental factors, such as interactions with other macromolecules, pH, temperature, RNA modification, and cell stress. Many biological processes involve multiple RNAs working in harmony. To integrate RNA sequences, RNA secondary structure and global RNA-RNA interaction in their native environment, many crosslinking and proximity ligation methods have been developed, such as cross-linking, ligation and sequencing of hybrid (CLASH)[70][71], hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation (hiCLIP)[72], [73], Psoralen Analysis of RNA Interactions and Structures (PARIS)[74]–[76], Sequencing of Psoralen crosslinked, Ligated, And Selected Hybrids (SPLASH)[77], [78], Mapping RNA interactome in vivo (MARIO)[79] and RNA *In situ* Conformation sequencing (RIC-seq)[80]. These methods will be illustrated and discussed in Chapter 4.

My PhD thesis focuses on RNA structure and RNA-RNA interactions within the HIV-1 and Influenza A genomes, and specifically how these structures relate to their viral genome packaging.

In Chapter 2, I reviewed the general role of RNA structures and RNA-RNA interactions in selective viral genome packaging. In Chapter 3, I investigated structural features within the HIV-1 5'UTR regulating genome dimerization by combining RNA structural information obtained by chemical probing with single nucleotide resolution profiles of RNA function obtained by mutational interference, which has been published in Nature Structural & Molecular Biology on March 2022 (DOI: 10.1038/s41594-022-00746-2). In Chapter 4, I developed a high-throughput sequencing-based method to study the interactions between the eight segments of influenza A virus genome.

1.3 References

[1] S. Hombach and M. Kretz, "Non-coding RNAs: Classification, Biology and Functioning.," *Adv. Exp. Med. Biol.*, vol. 937, pp. 3–17, 2016, doi: 10.1007/978-3-319-42059-2_1.

[2] S. T. Cross, D. Michalski, M. R. Miller, and J. Wilusz, "RNA regulatory processes in RNA virus biology," *Wiley Interdiscip. Rev. RNA*, vol. 10, no. 5, pp. 1–24, 2019, doi: 10.1002/wrna.1536.

[3] T. C. Nguyen, K. Zaleta-Rivera, X. Huang, X. Dai, and S. Zhong, "RNA, Action through Interactions," *Trends Genet.*, vol. 34, no. 11, pp. 867–882, 2018, doi: 10.1016/j.tig.2018.08.001.

[4] Z. Lu and H. Y. Chang, "The RNA base-pairing problem and base-pairing solutions," *Cold Spring Harb. Perspect. Biol.*, vol. 10, no. 12, 2018, doi: 10.1101/cshperspect.a034926.

[5] E. J. Strobel, A. M. Yu, and J. B. Lucks, "High-throughput determination of RNA structures," *Nat. Rev. Genet.*, vol. 19, no. 10, pp. 615–634, 2018, doi: 10.1038/s41576-018-0034-x.

[6] A. Ren, R. Micura, and D. J. Patel, "Structure-based mechanistic insights into catalysis by small self-cleaving ribozymes," *Curr. Opin. Chem. Biol.*, vol. 41, pp. 71–83, 2017, doi: 10.1016/j.cbpa.2017.09.017.

[7] G. Yusupova and M. Yusupov, "High-Resolution Structure of the Eukaryotic 80S Ribosome," *Annu. Rev. Biochem.*, vol. 83, pp. 467–486, 2014, doi: 10.1146/annurev-biochem-060713-035445.

[8] H. Khatter, A. G. Myasnikov, S. K. Natchiar, and B. P. Klaholz, "Structure of the human 80S ribosome," *Nature*, vol. 520, no. 7549, pp. 640–645, 2015, doi: 10.1038/nature14427.

[9] P. J. Lukavsky, "Structure and function of HCV IRES domains," *Virus Res.*, vol. 139, no. 2, pp. 166–171, 2009, doi: 10.1016/j.virusres.2008.06.004.

[10] W. K. Dawson, M. Lazniewski, and D. Plewczynski, "RNA structure interactions and ribonucleoprotein processes of the influenza A virus," *Brief. Funct. Genomics*, vol. 17, no. 6, pp. 402–414, 2018, doi: 10.1093/bfgp/elx028.

[11] C. Romero-López and A. Berzal-Herranz, "The role of the RNA-RNA interactome in the hepatitis C virus life cycle," *Int. J. Mol. Sci.*, vol. 21, no. 4, 2020, doi: 10.3390/ijms21041479.

[12] W. N. Moss, L. I. Dela-Moss, E. Kierzek, R. Kierzek, S. F. Priore, and D. H. Turner, "The 3′ splice site of influenza A segment 7 mRNA can exist in two conformations: A pseudoknot and a hairpin," *PLoS One*, vol. 7, no. 6, pp. 1–11, 2012, doi: 10.1371/journal.pone.0038323.

[13] M. Niepmann, L. A. Shalamova, G. K. Gerresheim, and O. Rossbach, "Signals involved in regulation of hepatitis C virus RNA genome translation and replication," *Front. Microbiol.*, vol. 9, no. MAR, pp. 1–14, 2018, doi: 10.3389/fmicb.2018.00395.

[14] B. L. Nicholson and K. A. White, "Functional long-range RNA-RNA interactions in positive-strand RNA viruses," *Nat. Rev. Microbiol.*, vol. 12, no. 7, pp. 493–504, 2014, doi: 10.1038/nrmicro3288.

[15] R. G. Huber *et al.*, "Structure mapping of dengue and Zika viruses reveals functional long-range interactions," *Nat. Commun.*, vol. 10, no. 1, 2019, doi: 10.1038/s41467-019-09391-8.

[16] J. Pérard, C. Leyrat, F. Baudin, E. Drouet, and M. Jamin, "Structure of the full-length HCV IRES in solution," *Nat. Commun.*, vol. 4, 2013, doi: 10.1038/ncomms2611.

[17] M. E. Filbin and J. S. Kieft, “Linking A to Ω: diverse and dynamic RNA-based mechanisms to regulate gene expression by 5’-to-3’ communication.,” *F1000Research*, vol. 5, pp. 1–10, 2016, doi: 10.12688/f1000research.7913.1.

[18] T. M. Block, S. Rawat, C. L. Brosgart, and S. Francisco, “Craving and subsequent opioid use among opioid dependent patients who initiate treatment with buprenorphine,” *Am J Drug Alcohol Abus.*, vol. 40, no. 2, pp. 163–169, 2014, doi: 10.1016/j.virusres.2008.07.016.Genome.

[19] D. E. Alvarez, C. V. Filomatori, and A. V. Gamarnik, “Functional analysis of dengue virus cyclization sequences located at the 5′ and 3’UTRs,” *Virology*, vol. 375, no. 1, pp. 223–235, 2008, doi: 10.1016/j.virol.2008.01.014.

[20] D. E. Alvarez, M. F. Lodeiro, S. J. Ludueña, L. I. Pietrasanta, and A. V. Gamarnik, “Long-Range RNA-RNA Interactions Circularize the Dengue Virus Genome,” *J. Virol.*, vol. 79, no. 11, pp. 6631–6643, 2005, doi: 10.1128/jvi.79.11.6631-6643.2005.

[21] S. M. Villordo, D. E. Alvarez, and A. V. Gamarnik, “A balance between circular and linear forms of the dengue virus genome is crucial for viral replication,” *Rna*, vol. 16, no. 12, pp. 2325–2335, 2010, doi: 10.1261/rna.2120410.

[22] A. Emery and R. Swanstrom, “HIV-1: To Splice or Not to Splice, That Is the Question,” *Viruses*, vol. 13, no. 2, 2021, doi: 10.3390/v13020181.

[23] N. Mueller, N. van Bel, B. Berkhout, and A. T. Das, “HIV-1 splicing at the major splice donor site is restricted by RNA structure,” *Virology*, vol. 468, pp. 609–620, 2014, doi: 10.1016/j.virol.2014.09.018.

[24] N. Mueller, B. Berkhout, and A. T. Das, “HIV-1 splicing is controlled by local RNA structure and binding of splicing regulatory proteins at the major 5′ splice site,” *J. Gen. Virol.*, vol. 96, no. 7, pp. 1906–1917, 2015, doi: 10.1099/vir.0.000122.

[25] C. Martin Stoltzfus, *Chapter 1 Regulation of HIV-1 Alternative RNA Splicing and Its Role in Virus Replication*, 1st ed., vol. 74, no. 09. Elsevier Inc., 2009.

[26] N. Nguyen Quang *et al.*, “Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection,” *Retrovirology*, vol. 17, no. 1, pp. 1–24, 2020, doi: 10.1186/s12977-020-00533-1.

[27] T. E. M. Abbink and B. Berkhout, “RNA Structure Modulates Splicing Efficiency at the Human Immunodeficiency Virus Type 1 Major Splice Donor,” *J. Virol.*, vol. 82, no. 6, pp. 3090–3098, 2008, doi: 10.1128/jvi.01479-07.

[28] I. Boeras, B. Seufzer, S. Brady, A. Rendahl, X. Heng, and K. Boris-Lawrie, “The basal translation rate of authentic HIV-1 RNA is regulated by 5’UTR nt-pairings at junction of R and U5,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017, doi: 10.1038/s41598-017-06883-9.

[29] L. I. Dela-Moss, W. N. Moss, and D. H. Turner, “Identification of conserved RNA secondary structures at influenza B and C splice sites reveals similarities and differences between

influenza A, B, and C," *BMC Res. Notes*, vol. 7, no. 1, 2014, doi: 10.1186/1756-0500-7-22.

[30] E. C. Hutchinson and E. Fodor, "Transport of the influenza virus genome from nucleus to nucleus," *Viruses*, vol. 5, no. 10, pp. 2424–2446, 2013, doi: 10.3390/v5102424.

[31] G. Neumann, M. T. Hughes, and Y. Kawaoka, "Influenza A virus NS2 protein mediates vRNP nuclear export through NES-independent interaction with hCRM1," *EMBO J.*, vol. 19, no. 24, pp. 6751–6758, 2000, doi: 10.1093/emboj/19.24.6751.

[32] R. E. O'Neill, J. Talon, and P. Palese, "The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins," *EMBO J.*, vol. 17, no. 1, pp. 288–296, 1998, doi: 10.1093/emboj/17.1.288.

[33] D. Ferhadian, M. Contrant, A. Printz-Schweigert, R. P. Smyth, J. C. Paillart, and R. Marquet, "Structural and functional motifs in influenza virus RNAs," *Front. Microbiol.*, vol. 9, no. MAR, pp. 1–11, 2018, doi: 10.3389/fmicb.2018.00559.

[34] L. Ye *et al.*, "RNA Structures and Their Role in Selective Genome Packaging," *Viruses*, vol. 13, no. 9, 2021, doi: 10.3390/v13091788.

[35] E. Mailler, J. C. Paillart, R. Marquet, R. P. Smyth, and V. Vivet-Boudou, "The evolution of RNA structural probing methods: From gels to next-generation sequencing," *Wiley Interdiscip. Rev. RNA*, vol. 10, no. 2, pp. 1–20, 2019, doi: 10.1002/wrna.1518.

[36] S. H. Kim *et al.*, "Three-dimensional structure of yeast phenylalanine transfer RNA: Folding of the polynucleotide chain," *Science (80-.).*, vol. 179, no. 4070, pp. 285–288, 1973, doi: 10.1126/science.179.4070.285.

[37] J. D. Robertus *et al.*, "Structure of yeast phenylalanine tRNA at 3 Å resolution," *Nature*, vol. 250, no. 5467, pp. 546–551, 1974, doi: 10.1038/250546a0.

[38] S. C. Keane *et al.*, "Structure of the HIV-1 RNA packaging signal," *Science (80-.).*, vol. 348, no. 6237, pp. 917–921, 2015, doi: 10.1126/science.aaa9266.

[39] S. Kharytonchyk *et al.*, "Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 47, pp. 13378–13383, 2016, doi: 10.1073/pnas.1616627113.

[40] J. D. Brown *et al.*, "Structural basis for transcriptional start site control of HIV-1 RNA fate.," *Science*, vol. 368, no. 6489, pp. 413–417, 2020, doi: 10.1126/science.aaz7959.

[41] M. Marušič, J. Schlagnitweit, and K. Petzold, "RNA Dynamics by NMR Spectroscopy," *ChemBioChem*, vol. 20, no. 21, pp. 2685–2710, 2019, doi: 10.1002/cbic.201900072.

[42] W. A. Cantara, E. D. Olson, and K. Musier-Forsyth, "Analysis of RNA structure using small-angle X-ray scattering," *Methods*, vol. 113, pp. 46–55, 2017, doi: 10.1016/j.ymeth.2016.10.008.

[43] Y. R. Bhandari *et al.*, "Topological Structure Determination of RNA Using Small-Angle X-Ray

Scattering," *J. Mol. Biol.*, vol. 429, no. 23, pp. 3635–3649, Nov. 2017, doi: 10.1016/j.jmb.2017.09.006.

[44] Y. X. Wang, X. Zuo, J. Wang, P. Yu, and S. E. Butcher, "Rapid global structure determination of large RNA and RNA complexes using NMR and small-angle X-ray scattering," *Methods*, vol. 52, no. 2, pp. 180–191, 2010, doi: 10.1016/j.ymeth.2010.06.009.

[45] H. Ma, X. Jia, K. Zhang, and Z. Su, "Cryo-EM advances in RNA structure determination," *Signal Transduct. Target. Ther.*, vol. 7, no. 1, pp. 1–6, 2022, doi: 10.1038/s41392-022-00916-0.

[46] L. Yan *et al.*, "Cryo-EM Structure of an Extended SARS-CoV-2 Replication and Transcription Complex Reveals an Intermediate State in Cap Synthesis," *Cell*, vol. 184, no. 1, pp. 184-193.e10, 2021, doi: 10.1016/j.cell.2020.11.016.

[47] Z. Su *et al.*, "Cryo-EM structures of full-length Tetrahymena ribozyme at 3.1 Å resolution," *Nature*, vol. 596, no. 7873, pp. 603–607, 2021, doi: 10.1038/s41586-021-03803-w.

[48] K. Zhang *et al.*, "Cryo-EM and antisense targeting of the 28-kDa frameshift stimulation element from the SARS-CoV-2 RNA genome," *Nat. Struct. Mol. Biol.*, vol. 28, no. 9, pp. 747–754, 2021, doi: 10.1038/s41594-021-00653-y.

[49] K. Kappel *et al.*, "Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures.," *Nat. Methods*, vol. 17, no. 7, pp. 699–707, 2020, doi: 10.1038/s41592-020-0878-9.

[50] T. C. Nguyen, K. Zaleta-Rivera, X. Huang, X. Dai, and S. Zhong, "RNA, Action through Interactions," *Trends Genet.*, vol. 34, no. 11, pp. 867–882, Nov. 2018, doi: 10.1016/j.tig.2018.08.001.

[51] X. W. Wang, C. X. Liu, L. L. Chen, and Q. C. Zhang, "RNA structure probing uncovers RNA structure-dependent biological functions," *Nat. Chem. Biol.*, vol. 17, no. 7, pp. 755–766, 2021, doi: 10.1038/s41589-021-00805-7.

[52] X. Dai, S. Zhang, and K. Zaleta-Rivera, "RNA: interactions drive functionalities," *Mol. Biol. Rep.*, vol. 47, no. 2, pp. 1413–1434, 2020, doi: 10.1007/s11033-019-05230-7.

[53] P. C. Bevilacqua and S. M. Assmann, "Technique development for probing RNA structure in vivo and genome-wide," *Cold Spring Harb. Perspect. Biol.*, vol. 10, no. 10, 2018, doi: 10.1101/cshperspect.a032250.

[54] J. G. Underwood *et al.*, "FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing," *Nat. Methods*, vol. 7, no. 12, pp. 995–1001, 2010, doi: 10.1038/nmeth.1529.

[55] F. Righetti *et al.*, "Temperature-responsive in vitro RNA structurome of Yersinia pseudotuberculosis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 26, pp. 7237–7242, 2016, doi: 10.1073/pnas.1523004113.

[56] Y. Wan *et al.*, "Genome-wide Measurement of RNA Folding Energies," *Mol. Cell*, vol. 48, no. 2, pp. 169–181, 2012, doi: 10.1016/j.molcel.2012.08.008.

[57] S. W. Foley and B. D. Gregory, "Protein interaction profile sequencing (PIP-seq)," *Curr. Protoc. Mol. Biol.*, vol. 2016, no. October, pp. 1–15, 2016, doi: 10.1002/cpmb.21.

[58] M. Kubota, C. Tran, and R. C. Spitale, "Progress and challenges for chemical probing of RNA structure inside living cells," *Nat. Chem. Biol.*, vol. 11, no. 12, pp. 933–941, 2015, doi: 10.1038/nchembio.1958.

[59] D. Mitchell, S. M. Assmann, and P. C. Bevilacqua, "Probing RNA structure in vivo," *Curr. Opin. Struct. Biol.*, vol. 59, no. 814, pp. 151–158, 2019, doi: 10.1016/j.sbi.2019.07.008.

[60] W. E. England, C. M. Garfio, and R. C. Spitale, "Chemical Approaches To Analyzing RNA Structure Transcriptome-Wide," *ChemBioChem*, vol. 22, no. 7, pp. 1114–1121, 2021, doi: 10.1002/cbic.202000340.

[61] C. Twittenhoff *et al.*, "Lead-seq: Transcriptome-wide structure probing in vivo using lead(II) ions," *Nucleic Acids Res.*, vol. 48, no. 12, pp. E71–E71, 2020, doi: 10.1093/nar/gkaa404.

[62] T. Umeyama and T. Ito, "DMS-Seq for In Vivo Genome-wide Mapping of Protein-DNA Interactions and Nucleosome Centers," *Cell Rep.*, vol. 21, no. 1, pp. 289–300, 2017, doi: 10.1016/j.celrep.2017.09.035.

[63] P. Tomezsko, H. Swaminathan, and S. Rouskin, "Viral RNA structure analysis using DMS-MaPseq," *Methods*, vol. 183, no. August 2019, pp. 68–75, 2020, doi: 10.1016/j.ymeth.2020.04.001.

[64] R. C. Spitale *et al.*, "Structural imprints in vivo decode RNA regulatory mechanisms," *Nature*, vol. 519, no. 7544, pp. 486–490, 2015, doi: 10.1038/nature14263.

[65] M. J. Smola and K. M. Weeks, "In-cell RNA structure probing with SHAPE-MaP," *Nat. Protoc.*, vol. 13, no. 6, pp. 1181–1195, 2018, doi: 10.1038/nprot.2018.010.

[66] S. Martin, C. Blankenship, J. W. Rausch, and J. Sztuba-Solinska, "Using SHAPE-MaP to probe small molecule-RNA interactions," *Methods*, vol. 167, no. December 2018, pp. 105–116, 2019, doi: 10.1016/j.ymeth.2019.04.009.

[67] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann, "In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features," *Nature*, vol. 505, no. 7485, pp. 696–700, 2014, doi: 10.1038/nature12756.

[68] Y. Ding, C. K. Kwok, Y. Tang, P. C. Bevilacqua, and S. M. Assmann, "Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq," *Nat. Protoc.*, vol. 10, no. 7, pp. 1050–1066, 2015, doi: 10.1038/nprot.2015.064.

[69] J. Talkish, G. May, Y. Lin, J. L. Woolford, and C. J. McManus, "Mod-seq: High-throughput sequencing for chemical probing of RNA structure," *Rna*, vol. 20, no. 5, pp. 713–720, 2014, doi:

10.1261/rna.042218.113.

[70] G. Kudla, S. Granneman, D. Hahn, J. D. Beggs, and D. Tollervey, “Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 24, pp. 10010–5, Jun. 2011, doi: 10.1073/pnas.1017386108.

[71] A. Helwak and D. Tollervey, “Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH),” *Nat. Protoc.*, vol. 9, no. 3, pp. 711–728, Mar. 2014, doi: 10.1038/nprot.2014.043.

[72] Y. Sugimoto *et al.*, “HiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1,” *Nature*, vol. 519, no. 7544, pp. 491–494, 2015, doi: 10.1038/nature14280.

[73] Y. Sugimoto, A. M. Chakrabarti, N. M. Luscombe, and J. Ule, “Using hiCLIP to identify RNA duplexes that interact with a specific RNA-binding protein,” *Nat. Protoc.*, vol. 12, no. 3, pp. 611–637, 2017, doi: 10.1038/nprot.2016.188.

[74] Z. Lu, J. Gong, and Q. C. Zhang, “PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution,” vol. 1649, I. Gaspar, Ed. New York, NY: Springer New York, 2018, pp. 59–84.

[75] Z. Lu *et al.*, “RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure.,” *Cell*, vol. 165, no. 5, pp. 1267–1279, May 2016, doi: 10.1016/j.cell.2016.04.028.

[76] M. Zhang *et al.*, “Optimized photochemistry enables efficient analysis of dynamic RNA structuromes and interactomes in genetic and infectious diseases,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–14, 2021, doi: 10.1038/s41467-021-22552-y.

[77] J. G. A. Aw *et al.*, “In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation,” *Mol. Cell*, vol. 62, no. 4, pp. 603–617, May 2016, doi: 10.1016/j.molcel.2016.04.028.

[78] J. G. A. Aw, Y. Shen, N. Nagarajan, and Y. Wan, “Mapping RNA-RNA interactions globally using biotinylated psoralen,” *J. Vis. Exp.*, vol. 2017, no. 123, pp. 1–10, 2017, doi: 10.3791/55255.

[79] T. C. Nguyen *et al.*, “Mapping RNA–RNA interactome and RNA structure in vivo by MARIO,” *Nat. Commun.*, vol. 7, no. 1, p. 12023, Nov. 2016, doi: 10.1038/ncomms12023.

[80] Z. Cai *et al.*, “RIC-seq for global in situ profiling of RNA–RNA spatial interactions,” *Nature*, vol. 582, no. 7812, pp. 432–437, 2020, doi: 10.1038/s41586-020-2249-1.

Chapter 2 Review: RNA Structures and Their Role in Selective Genome Packaging

Modified from a manuscript published on Viruses (Doi: 10.3390/v13091788)

Liqing Ye [1], Uddhav B. Ambi [1], Marco Olguin-Nava [1], Anne-Sophie Gribling-Burrer [1], Shazeb Ahmad [1], Patrick Bohn [1], Melanie M. Weber [1] and Redmond P. Smyth [1,2,*]

[1] Genome Architecture and Evolution of RNA Viruses, Helmholtz Institute for RNA-Based Infection Research, Helmholtz-Center for Infection Research, 97080 Würzburg, Germany; liqing.ye@helmholtz-hiri.de (L.Y.); uddhav.ambi@helmholtz-hiri.de (U.B.A.); marco.olguin@helmholtz-hiri.de (M.O.-N.); anne-sophie.gribling@helmholtz-hiri.de (A.-S.G.-B.); shazeb.ahmad@helmholtz-hiri.de (S.A.); patrick.bohn@helmholtz-hiri.de (P.B.); melanie.weber@helmholtz-hiri.de (M.M.W.)

[2] Faculty of Medicine, University of Würzburg, 97080 Würzburg, Germany

[*] Correspondence: redmond.smyth@helmholtz-hiri.de; Tel.: +49-(0)931-318-9152

Abstract

To generate infectious viral particles, viruses must specifically select their genomic RNA from milieu that contains a complex mixture of cellular or non-genomic viral RNAs. In this review, we focus on the role of viral encoded RNA structures in genome packaging. We first discuss how packaging signals are constructed from local and long-range base pairings within viral genomes, as well as inter-molecular interactions between viral and host RNAs. Then, how genome packaging is regulated by the biophysical properties of RNA. Finally, we examine the impact of RNA packaging signals on viral evolution.

Keywords: RNA virus; RNA; RNA structure; genome packaging; viral assembly; evolution

2.1. Introduction

Genome packaging is the process whereby viruses assemble their genomes into capsids[1]. The primary purpose of the capsid is to protect the genome from a hostile cellular and extracellular environment until its cargo can be released into a new host for a further round of replication. For faithful replication it is essential that genome packaging occurs with high fidelity. In the case of RNA viruses, this is particularly challenging because viral RNA must be specifically selected from a complex mix of cellular RNA, which often includes non-genomic viral RNA. Furthermore, genome packaging must be tightly regulated as it is often in competition with other essential functions, such as genome replication or translation. RNA viruses have solved this problem by exploiting the capacity of RNA to fold into three-dimensional structures that are recognised by the viral packaging machinery[2–4]. RNA structures are formed from local intra-molecular and long-range base pairings within the same molecule, as well as inter-molecular RNA-RNA interactions[5–8]. Because RNA structures are rarely static, packaging can be dynamically regulated by intrinsic RNA structural switches, binding of viral factors, or in some cases, inter-molecular interactions with host RNAs.

Interestingly, many RNA viruses have evolved genome organizations that greatly complicate viral assembly and packaging. Segmented viruses, such as influenza and rotavirus, need to incorporate multiple genome segments for their virions to be infectious[9]. On the other hand, retroviruses package two copies of their genome, even though the total genetic material of only one genome is replicated[10]. In exchange for this increased complexity, RNA viruses enhance their evolvability through recombination or reassortment. Here, genome packaging exploits specific inter-molecular RNA interactions that bring together different segments or genomes for assembly. In this review, we discuss how RNA viruses exploit the proprieties of RNA structure to regulate their packaging and explain how RNA based packaging mechanisms can influence viral evolution.

2.2. Packaging Signals in RNA Viruses

RNA viruses distinguish their genomes from cellular RNAs using *cis* acting packaging signals that serve as high affinity binding sites for the viral capsid (or nucleocapsid) proteins[11]. RNA readily folds back on itself to form secondary and tertiary structures, which are complex enough to enable specific selection of genomic viral RNA from diverse pools of cellular RNA[12]. A canonical example is the 19-nucleotide stem-loop, also known as TR, in the genome of the MS2 bacteriophage (**Figure 2.1a**)[13,14]. The MS2 coat protein (CP) dimer specifically recognizes this short stem loop structure to initiate assembly. Despite the simplicity of the structure, which is only formed from local base-pairings, MS2 CP binds with high affinity and specificity. This has led to the extensive repurposing of the TR-CP interaction in applications such as single molecule live cell imaging[15,16]. Remarkably, stem-loops having a C at position −5 in the loop have a higher affinity for CP than the wild-type U[17]. Whilst this is useful for biotechnology purposes, it also neatly demonstrates that increased affinity is not always beneficial for viruses, presumably because genomes must eventually be released during the early steps of the next replication cycle. Interestingly, high-throughput RNA structure-function analyses reveal that nucleotides in the stem can be exchanged without impairing binding to CP. In

contrast, specific single stranded residues were required for function[18,19]. This is because, for the most part, RNA binding proteins (RBPs) make non-specific interactions with double stranded RNA (dsRNA) through generic contacts with the 2'hydroxyl groups of the ribose or the phosphodiester backbone[20]. On the other hand, unique structural features created by loops and mismatches can be more readily recognized through extensive sequence-specific contacts [21].

MS2 bacteriophage genome packaging depends not only on the high affinity TR interaction site, but also on lower affinity pseudo-packaging sites that are dispersed throughout the genome (**Figure 2.1b**)[22]. This strategy is proposed to enhance the specificity and efficiency of assembly through cooperative interactions with viral capsid proteins[23]. This can be seen in other RNA viruses, such as tobacco necrosis virus (TNV) [24], hepatitis B virus (HBV) [25], and alphavirus[26]. On the contrary, HIV-1 has no described packaging signals outside of the 5' end of the genome. Instead, assembly on viral genomic RNA is driven by changes in the specificity of the structural protein Gag during its multimerization at the plasma membrane[27] (**Figure 2.1c**). This change in specificity enhances its affinity for A-rich sequences that are enriched in the HIV-1 genome compared to cellular RNA[27,28]. Evidently, pairing a limited number of high affinity binding sites with lower affinity RNA structures or sequences throughout viral genomes is a robust mechanism of packaging. Indeed, similar features emerged in directed evolution experiments that successfully converted a bacterial enzyme into a nucleocapsid that packages and protects its own encoding mRNA[11].

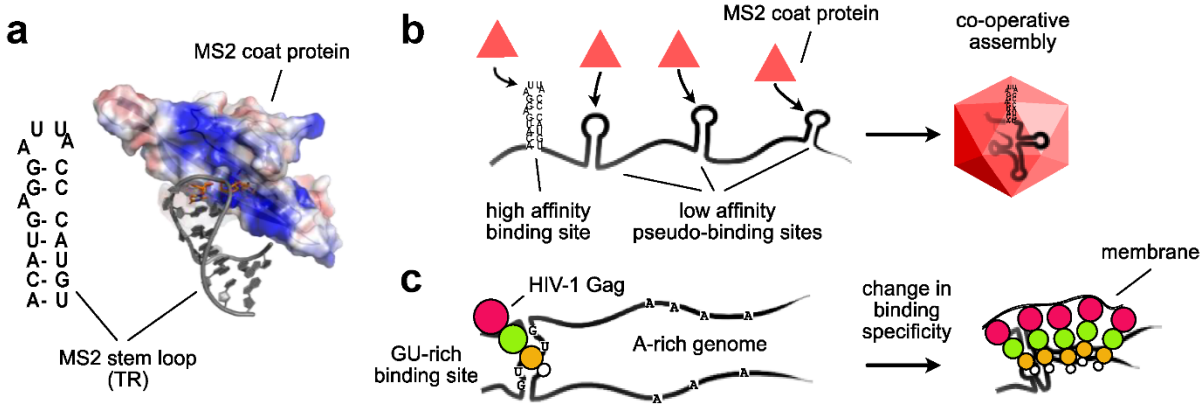


Figure 2.1| RNA packaging signals. (a) MS2 bacteriophage encodes a coat protein (CP) that binds to a 19-nucleotide stem loop structure, known as TR. Secondary structure model of the TR stem loop and three-dimensional structure of the coat protein-TR interaction; (PDB: 1AQ3) **(b)** MS2 bacteriophage genome is encapsidated through cooperative interactions between coat protein that are bound to a high affinity binding site (TR) and multiple low affinity binding sites encoded throughout the genome; **(c)** HIV-1 Gag recognizes the genomic RNA through a GU-rich high affinity binding site in the 5' untranslated region (5'UTR). During assembly at the plasma membrane Gag switches its specificity from GU-rich sequences to A-rich sequences, which is proposed to favour assembly of Gag on its cognate genomic RNA

Complex viruses that express sub-genomic or spliced viral RNAs have an additional challenge: they must not only distinguish their genomic RNA from cellular RNA, but also from non-genomic viral RNAs (**Figure 2.2**). One simple way to achieve this selectivity is the removal of the packaging signal from the non-genomic RNA during its production. This mechanism occurs in certain retroviruses, such as Moloney murine leukemia virus (MoMLV), which contains a packaging signal with high affinity binding sites for viral nucleocapsid (NC) composed of three stem-loop structures (DIS-2, SL-C and SL-D)[29,30]. All these RNA structures lie downstream of the major splice donor site and are thus removed from spliced viral RNAs (**Figure 2.2a**). Another retrovirus, HIV-1, recognizes its genomic RNA through specific interactions between the viral Gag protein and packaging signals present at the 5' end of the genome[31–37]. Early deletion mutagenesis studies identified SL3 (Ψ), which lies downstream of the major splice donor SL2, as the major packaging motif[36–38]. This genome organization was originally thought to explain the selectivity for genomic over spliced viral RNA[39,40]. However, an abundance of evidence has now revised this picture. Specifically, the basal part and internal loop of SL1, which lies upstream of SL2, is now recognized as the primary Gag binding site[19,41–44]. Notably, deletion or mutagenesis of SL1 has a more drastic effect on Gag binding and genome packaging compared to SL3, and deletion of sequences downstream of SL2 has only modest effects on binding[42,43]. This revision in understanding resurrected the problem of how HIV-1 discriminates between spliced viral RNA and genomic RNA. Surprisingly, genome fragments from the first nucleotide through to SL3 – containing SL1 – are not efficiently bound by Gag unless they contain sequences downstream of SL3[41]. A model was proposed whereby a long-range interaction between sequences downstream of the splice donor site counteracts a negative regulatory element upstream of the high affinity binding site in SL1[41] (**Figure 2.2b**). As this interaction can only be formed in genomic RNA, it enables the selectivity of Gag for genomic RNA over spliced viral RNA at the initial binding step[41,44–47].

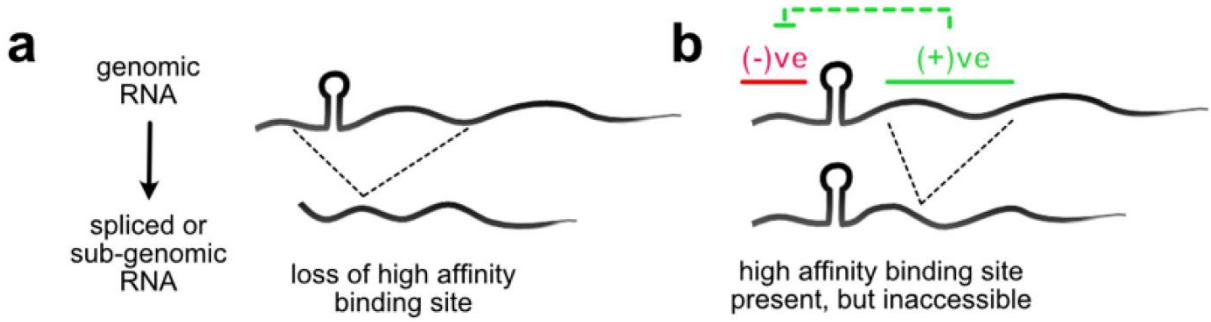


Figure 2.2| Selection of genomic RNA from non-genomic viral RNA. Dotted lines signify splicing or production of sub-genomic RNA. **(a)** Non-genomic RNA cannot be packaged because it does not contain the high affinity binding site for the (nucleo)capsid protein; **(b)** Non-genomic RNA contains the sequences for high affinity binding site, but it is not presented for binding. In HIV-1, negative regulatory elements (−ve) upstream of the splice donor (SD) conceal the Gag binding site unless counteracted by positive regulatory (+ve) sequences downstream of the SD.

Coronaviruses (CoV) have extraordinarily large genomes (~30 kb) that presumably pose additional difficulties for packaging, yet genomic RNA is efficiently and selectively incorporated into virions[5,48–50]. Accumulating evidence suggests that SARS-CoV-2 exploits liquid-liquid phase separation (LLPS) during its replication[51–59] (**Figure 2.3**). LLPS occurs when biological molecules condense into a phase resembling a liquid droplet, and is an emerging paradigm for organizing membrane-less viral factories[60,61]. It is a common property of RBPs containing intrinsically disordered regions (IDR), such as the SARS-CoV-2 nucleocapsid (N)[53,56,57,62]. LLPS of N protein is enhanced in the presence of viral RNA[51,54], and even though N protein binds throughout the genomic RNA[51,59], LLPS is specifically promoted by RNA sequences at the 5' and 3' of the SARS-CoV-2 genome (**Figure 2.3**)[51]. Interestingly, other sequences, such as the CoV frameshift site, were found to disperse condensates[51], and importantly, sub-genomic RNA was efficiently excluded from preformed droplets[51]. This demonstrates that for LLPS mediated packaging, the biophysical properties of RNA-protein interaction are as important as the protein-RNA affinities. LLPS likely promotes viral assembly by enhancing interaction between RNA and N protein within a privileged site[51], but may have other roles in viral replication, such as hiding viral RNA from cellular immune sensors[63,64].

Packaging sites may also include motifs necessary for the correct presentation of the RNA molecule in time and space. For example, influenza viruses have a segmented genome of negative sense viral RNAs (vRNAs) that are replicated in the nucleus *via* complementary RNA (cRNA) intermediates. Long-range interactions between the 5' and 3' termini, in some cases over distances of thousands of nucleotides, construct the promoter structure that is involved in transcription, replication and packaging[65]. Interestingly, cRNAs and vRNAs are both complexed into ribonucleoproteins (vRNPs) with very similar protein compositions, but only vRNPs are packaged into virions. Slight differences in promoter structures between vRNPs and cRNPs, due to imperfect complementarity between the terminal sequences, affect its interaction with the viral M1 protein that acts as a bridge between vRNPs and the nuclear export machinery. This structural difference allows the virus to discriminate between cRNPs and vRNPs by either preventing nuclear export of the cRNP[66] or by changing nuclear export pathways[67] (**Figure 2.3**). In the same vein, several studies show that the binding of the HIV-1 Rev protein to its cognate RNA structure, the Rev Response Element (RRE), enhances genome packaging[68–70]. Surprisingly, this enhancement effect seems to be unrelated to the role of Rev/RRE in increasing cytoplasmic RNA levels. Rather, the Rev/RRE is proposed to enhance packaging by defining the correct nuclear export pathway and subcellular localization of the genomic RNA[68–71] (**Figure 2.3**).

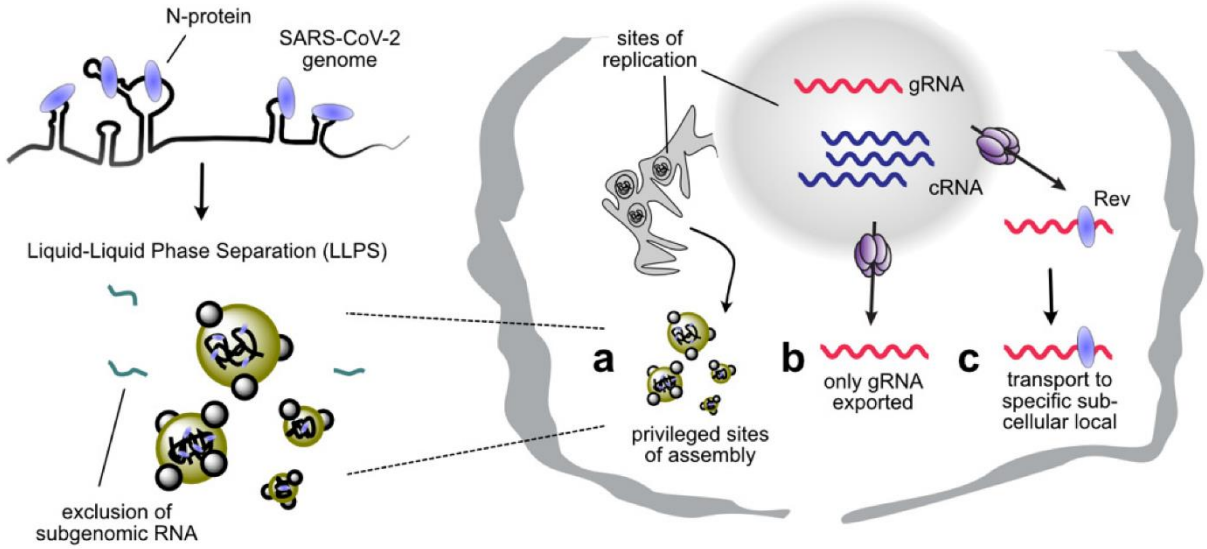


Figure 2.3| Successful genome packaging requires RNA signals to direct the genome from sites of replication to sites of assembly. (a) Sites of assembly generated by liquid-liquid phase separation (LLPS). Specific interactions between the SARS Coronavirus-2 (SARS-CoV-2) nucleoprotein (N) and its genome induce LLPS. Sub-genomic RNA is excluded. **(b)** Influenza complementary RNA (cRNA) replication intermediates are not correctly exported from the nucleus and are therefore not packaged. **(c)** HIV-1 genomic RNA contains the Rev Response Element (RRE) which binds to the viral Rev protein needed to export the RNA from the nucleus. Rev binding is proposed to enhance packaging by transporting RNA to the correct sub-cellular location.

2.3. RNA Structure as a Regulator of Genome Packaging

Genome packaging occurs during the late stages of replication when sufficient genomes and structural proteins have been replicated and produced to ensure effective viral assembly. Sometimes even, the same viral RNA molecule must carry out several competing functions. It is not surprising therefore that viruses heavily regulate the translation, replication, and packaging of their genomes. RNA viruses achieve this, in part, by exploiting the dynamic and flexible properties of RNA. Namely, RNA molecules can spontaneously fold into multiple, mutually exclusive structures, acting as riboswitches with each structure having a different function[72]. RNA can also respond to the binding of cellular or viral biomolecules, which can act as a regulatory trigger for further remodelling of ribonucleoprotein complexes[73].

Hepatitis C virus (HCV) is a model of such complex RNA based regulation[74–76]. The *cis*-acting replicating element (CRE) in the coding region of the NS5B protein forms a long-distance base pairing with the highly conserved X-region in 3’ UTR[77–81] (**Figure 2.4a**). This interaction is required for replication, but also acts as a regulatory switch between replication and packaging by masking the core protein binding sites present in the 3’UTR[81] (**Figure 2.4a**). At the same time, the CRE regulates HCV genome translation *via* a long-range intra-molecular interaction with the internal ribosome

entry site (IRES) in the 5'UTR[75,82] (**Figure 2.4a**). Finally, the 3'UTR X-region contains a palindromic sequence that promotes homo-dimerization of the HCV genome *via* a kissing loop inter-molecular RNA-RNA interaction[74,83,84] (**Figure 2.4a**). Since homo-dimerization is incompatible with the CRE-X interaction, and because it is likely tied to the concentration of genomes and viral chaperones in the cell, this mechanism is predicted to inhibit genome replication in favour of packaging late in the replication cycle. In this way, HCV elegantly fine tunes its replication using a complex network of dynamic and mutually exclusive RNA-RNA interactions[74–76].

Unsurprisingly, other viruses use similar principles to regulate their replication. HIV-1 genomic RNA is transcribed by the host cell and exported into the cytoplasm as a single pool of RNA that can be either selected by the viral Gag protein for packaging into viral particles or translated by host cell ribosomes[6]. A long-standing hypothesis is that the HIV-1 5'UTR adopts two alternative structural conformations to regulate the balance between genome translation and packaging (**Figure 2.4b**). Many structural models have been proposed, but all of them have the common feature that SL1 is presented in one conformation and sequestered in another[85–92]. As previously noted, SL1 is a key packaging motif in HIV-1 because the stem of SL1 contains the major Gag binding motif[41,93]. In addition to the Gag binding site, SL1 contains a six-nucleotide palindromic loop sequence that mediates an inter-molecular kissing loop interaction leading to the formation of genome dimers[94–96]. Unlike HCV, which produces homodimers that remain in the cell, HIV-1 dimers are packaged into virions[97]. This process, known as dimerization, is a conserved feature of retroviral replication that is assumed, but not formally proven, to be a pre-requisite for packaging[98]. A series of NMR studies have identified a region, U5, in the 5'UTR that base pairs with the loop sequence of SL1, or alternatively with a region surrounding the AUG start site[87,89,90,99,100]. When the SL1 loop sequence is base paired with U5, genomic RNA is monomeric, which promotes translation (**Figure 2.4b**). When U5 is base paired with a region surrounding the AUG start codon, the SL1 loop is available for dimerization and packaging[101] (**Figure 2.4b**). Remarkably, transcription start site heterogeneity inherent to the HIV-1 promotor strongly influences the equilibrium between these two structures[87,102]. HIV-1 genomes transcribed with a single guanosine favour the dimer conformation and are packaged into viral particles, while genomes transcribed with two or three guanosines form monomers that are preferentially translated[87,102] (**Figure 2.4b**). The fact that a single GC base-pair perturbs the monomer-dimer equilibrium provides striking proof that viruses exploit metastable RNA structures in their regulation.

Added complexity comes from the fact that RNA viruses also regulate their replication using inter-molecular interactions between host RNAs and their genomes. The HCV 5'UTR contains binding sites for the host micro-RNA miR-122[103–105] (**Figure 2.4a**). miR-122 is essential for the stability of HCV genomic RNA by inhibiting RNA decay by Xrn exonucleases[106–108]. Binding of miR-122 also increases HCV genome translation[109–113] and replication[114,115] by other mechanisms. Several lines of evidence suggest that miR-122 can act by inducing RNA structural changes in the 5'UTR. Specifically, miR-122 either alone or in partnership with Ago, enhances translation by promoting the folding of a functional IRES and suppressing alternative folds of the 5' UTR that interfere with IRES function[111,113]. Others have proposed that miR-122 enhances translation by facilitating cyclization

of the genome, by promoting stranded separation of the replication intermediates, or by bringing or displacing protein co-factors to the genome[114]. Similarly, the HIV-1 5'UTR contains a binding site for a host cellular tRNALys3, which is used as a primer for reverse transcription[116,117] (**Figure 2.4b**). Its binding results in RNA conformational changes that favour dimerization, and presumably packaging[92,118].

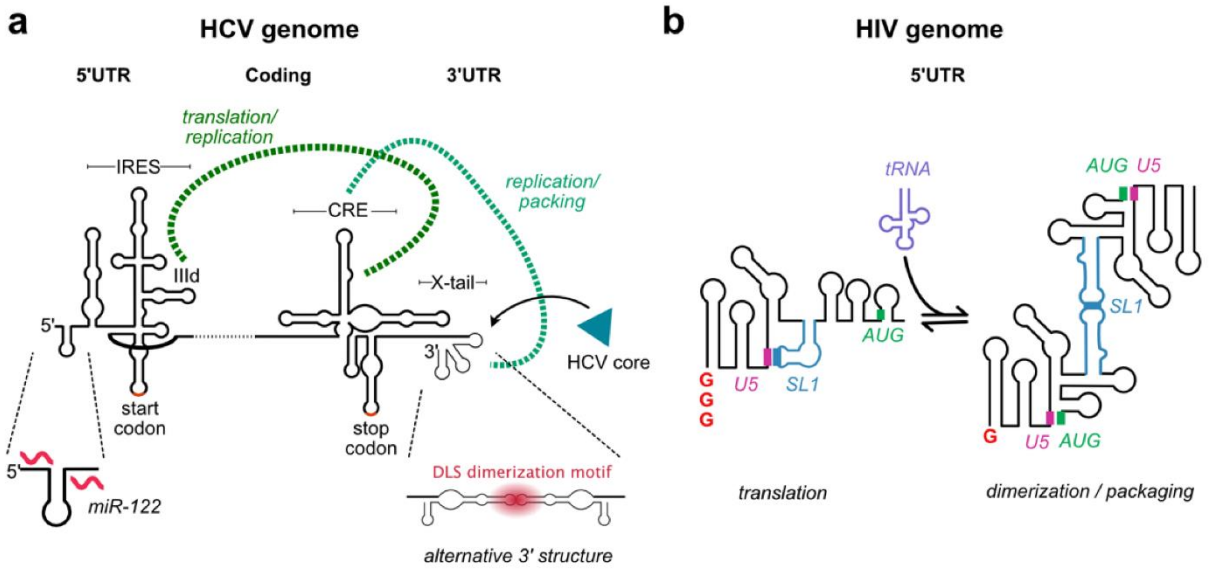


Figure 2.4| RNA structural switches and long-distance interactions regulate the balance between genome replication, packaging, and translation. (a) The HCV life cycle is regulated by a complex network of long-distance intra-molecular interactions and inter-molecular interactions. The packaging site which binds to the HCV core protein resides in the 3' untranslated region (3'UTR). A long-distance base pairing between the cis-acting element (CRE) in the coding region and the X-tail in the 3'UTR regulates the balance between replication and packaging (light green dotted line). A long-distance interaction between the CRE and the internal ribosome entry site (IRES) regulates the balance between replication and translation (dark green dotted line). The 3'UTR is alternatively structured, leading to the formation of homodimers through an intermolecular interaction. The HCV 5'UTR binds the host microRNA miR-122 to regulate different aspects of HCV replication; **(b)** A structural switch in the HIV-1 5'UTR regulates the balance between genome translation and packaging. Transcripts beginning with three G residues fold into a monomer conformation and are preferentially translated. Transcripts beginning with one G residue fold into a dimer conformation. Structural switching is mediated by mutually exclusion interactions between regions U5 (pink), SL1 (blue), the AUG region (green), and a host tRNA (purple).

2.4. Intermolecular RNA-RNA Interactions in Segmented Viruses

Many viruses split their genome into smaller independent segments. This causes problems for genome assembly, which is solved using one of two strategies: random *vs* selective packaging (**Figure 2.5**). Tri-segmented bunyaviruses, such as the rift valley fever virus (RVFV) and Schmallenberg virus (SBV), use the simpler strategy of random incorporation[119,120] (**Figure 2.5a**). Single molecule

fluorescent in situ hybridisation (smFISH) revealed only 1 in 10 RVFV particles contain the full complement of genome segments due to the inherent heterogeneity in this packaging strategy[119]. Intuitively, as the number of segments increase, the probability of packaging one copy per particle decreases rapidly unless a large number of genome segments are incorporated per particle[121]. As this is not very efficient, many segmented viruses have overcome the genome assembly problem with specific packaging signals, which allow each distinct segment to be identified and packaged (**Figure 2.5b**).

A well-studied example is influenza A virus of the *Orthomyxoviridae* family. Its genome consists of 8 negative-sense viral RNAs (vRNAs) that are packaged into viral particles as viral ribonucleoprotein (vRNP) complexes[121]. Because each vRNA encodes for an essential protein, every infectious viral particle must contain at least one copy of each segment. Indeed, smFISH experiments prove that most viral particles contain precisely one of each segment[122,123]. Furthermore, numerous electron tomography studies demonstrate that influenza vRNPs in budding viruses adopt an arrangement, also known as '1+7' conformation, in which seven vRNPs surround a central one[124–126]. Altogether, these data argue for a selective packaging process. Defective interfering (DI) RNA, which naturally arise in cell culture at high multiplicity of infection (MOI), retain 100-300 nucleotides from their terminal sequences indicating that these regions contain packaging signals[127–130]. Indeed, deletion and mutagenesis studies have grossly defined terminal packaging regions within all eight vRNAs[131–140]. Terminal packaging signals are proposed to be bipartite, containing a non-specific "incorporation signal" in the UTR/promoter region, and a specific "bundling signal" in the terminal coding regions[140]. The hypothesized incorporation signal directs vRNP packaging into virions, whereas the bundling signal allows discrimination between vRNPs. The mechanism mediating this phenomenon is still not completely understood, but the most attractive explanation is that packaging signals discriminate between segments by defining direct and segment specific inter-molecular RNA-RNA interactions (**Figure 2.5c**). In support of this idea, electron microscopy studies show frequent physical contacts through the entire length of each vRNP with a string-like form reminiscent of RNA[125,126,141,142], and vRNAs are able to form RNA-RNA interactions *in vitro*[126,143,144]. The prevailing model is that influenza vRNPs are packaged as a supramolecular complex that held together through a network of interactions where each vRNA contacts at least one other vRNA[8]. This would help explain why mutations to packaging signals in one vRNA often affected the packaging of other vRNAs[136,139,145]. Furthermore, the capacity of RNA to tolerate mutations without disrupting structure and function would explain why packaging site mutations do not always give rise to phenotypic effects. Importantly, several vRNA-vRNA interactions have been characterised at the nucleotide level proving that at least some packaging signals define direct RNA-RNA contacts[143,146,147]. These recent results have spurred efforts to map more completely inter-segment interactions in influenza using high-throughput sequencing and RNA proximity ligation technologies[147,148]. Collectively, these studies have revealed that the inter-molecular RNA-RNA interactions are extensive, with frequent contacts seen throughout vRNAs, including in the central coding regions. Nevertheless, comprehensive maps of direct vRNA-vRNA contacts have been surprisingly difficult to interpret, with many interactions having no apparent functional role. One major conclusion could be that vRNA packaging signals are complex and redundant, but it could also

reflect biases in contact map technology. Furthermore, the role of protein-RNA and protein-protein interactions in this process is not excluded. As a matter of fact, influenza nucleoprotein (NP) provides an additional layer of complexity to this process as it incompletely coats the vRNA and helps to define which vRNA sequences are available to form inter-molecular interactions[149–151].

Similar principles seem to apply to dsRNA segmented viruses of the *Reoviridae* family, which includes rotavirus and bluetongue virus. Rotavirus has a genome composed of eleven dsRNA segments of different sequences and lengths (0.7 to 3.1 kb)[152]. Paralleling recent results in influenza, inter-molecular interactions between segments 9, 10 and 11 of the rotavirus RNAs were observed *in vitro*[153]. Disruption of the putative interaction sites by mutation or with oligoribonucleotides inhibited complex formation and viral replication in cell culture[153]. Studies with bluetongue virus, which has ten dsRNA segments, suggest a model whereby assembly begins with the formation of an initial complex built of the small RNA segments[154–156]. This complex would then serve as a base for sequentially recruiting the remaining RNA segments to ultimately generate a complete complex that is packaged into virions. In the case of influenza, smFISH studies reveal that sequential vRNP-vRNP interactions occur *en route* to the plasma membrane where packaging takes place[123,157,158]. However, current evidence indicates that there is not a single assembly pathway, but a number of alternative preferred pathways that nevertheless prevent the incorporation of more than one copy of each segment[123]. How this is achieved at the mechanistic level is still an open question.

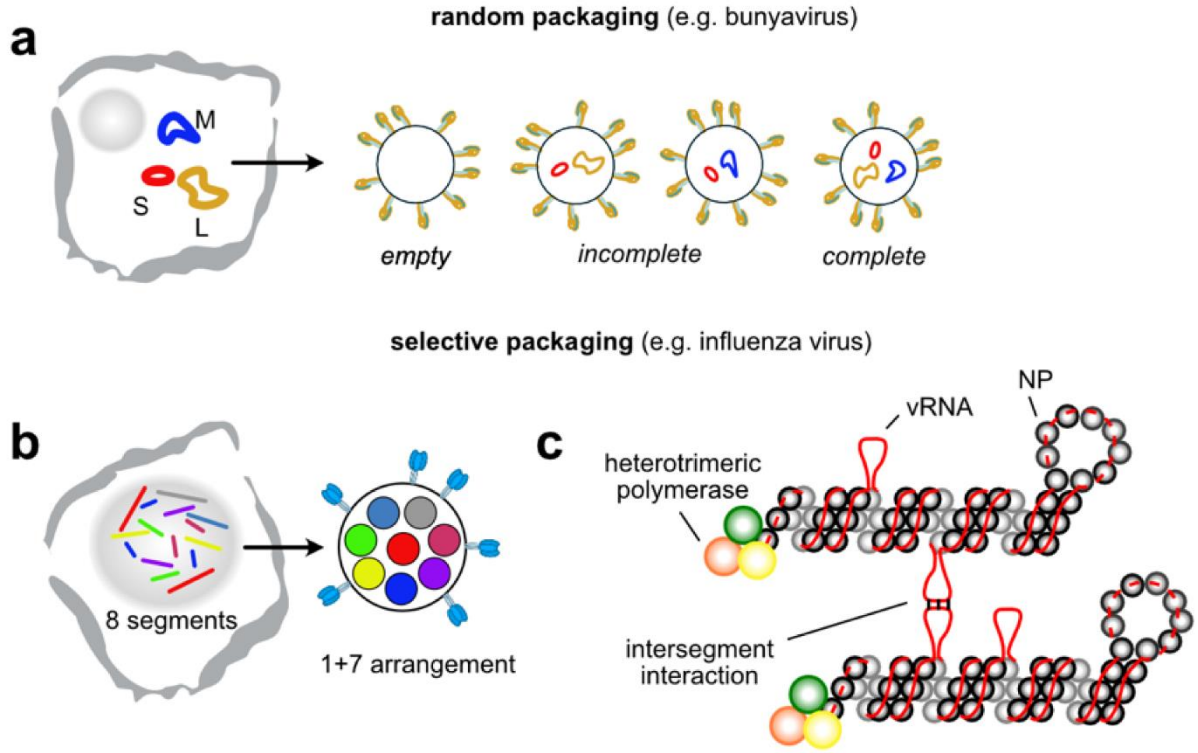


Figure 2.5| Segmented viruses package their genomes randomly or by using segment specific packaging signals. (a) Bunyavirus randomly package three genome segments, small (S), medium (M) and large (L), such that many progeny virions are empty or incomplete; **(b)** Influenza A virus package 8 genome segments selectively. Genome segments within budding virions are organized into a 1 + 7 arrangement with one central segment surrounded by seven others; **(c)** Influenza vRNAs are packaged as viral ribonucleoproteins (vRNPs). vRNAs are bound by nucleoprotein (NP) and a heterotrimeric polymerase. vRNA is incompletely coated by NP allowing for inter-segment vRNA-vRNA interactions to occur as a possible mechanism underlying the selective packaging process.

2.5. RNA Packaging and Evolution

RNA based packaging signals play a much broader role in viral life cycles than assembly, and it is now appreciated that RNA virus genome structures are optimized to facilitate viral evolution during co-infection. One widespread strategy is template switching during replication leading to recombination and the formation of genome chimeras, which is a conserved phenomenon in retroviruses[10,159] (**Figure 2.6a**). Another common strategy is genome segmentation leading to reassortment, which can be seen in rotaviruses and influenza viruses[160] (**Figure 2.6b**). Recombination and reassortment are both non-random processes that are heavily biased by RNA sequence and structure.

Retroviruses package two near identical copies of the genome as a non-covalently associated dimer[97]. One evolutionary advantage for this dimeric genome organization is that it brings together two templates for packaging into virions. Template switching during subsequent infection and reverse transcription generates a recombinant virus that is genetically distinct from the two parental viruses[161–165] (**Figure 2.6a**). Retroviral recombination is a major mechanism by which retroviruses escape selective pressures imposed by the immune system or antiretroviral therapy[10]. As previously noted, HIV-1 dimerization is mediated by the palindromic loop sequence of SL1[94–96]. Sequence variations that are unable to form the kissing loop interaction are also defective in recombination due to their inability to be co-packaged into virions[166]. Indeed, the loop sequences in subtype B (GCGCGC), and subtypes A, C and G(GUGCAC) are incompatible. Inter-subtype recombination is thus much lower compared to intra-subtype recombination[94,166,167]. Interestingly, HIV-1 genomes containing deletions in SL1 are still packaged into virions as dimers, albeit at a lower level than wild-type viruses[168,169]. This provides strong evidence that so-far undetected inter-molecular interaction exist throughout the HIV-1 genome that may enable the formation inter-subtype recombinants even viruses are unable to form the kissing-loop interaction at SL1.

Packaging signal incompatibilities are also thought to be a major restriction to reassortment in segmented viruses[146,147,170–173]. This is especially important for influenza where introductions of sequence variation from animal reservoirs have led to pandemics in the past[174]. Fortunately, divergence in packaging signals between human and animal viruses is one of many steps that may block reassortment[171,172]. The molecular mechanism restricting reassortment probably lies in the

inability of divergent sequences to form intermolecular vRNA-vRNA interactions required for packaging[146,147,170]. This provides hope that in the future, better knowledge of viral structures may be repurposed to predict or even direct viral evolution to combat both emerging and endemic RNA viruses.

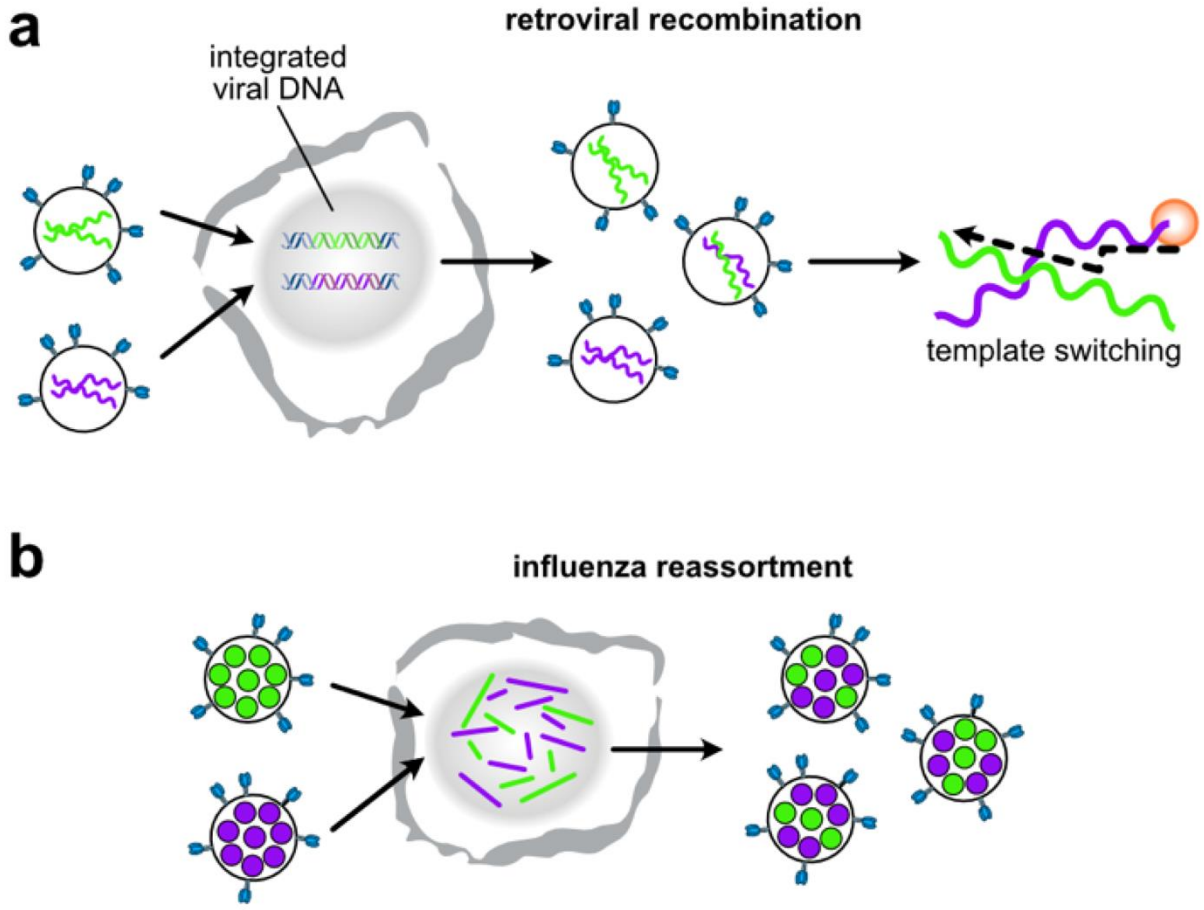


Figure 2.6| Viral RNA packaging influences viral evolution. (a) HIV-1 virions contain two copies of the genome. After co-infection, up to 50% of progeny co-package different genomes. Template switching during reverse transcription produce cDNA that is a chimera of the two genomes; **(b)** Influenza A virions contain eight different vRNA segments. After co-infection, reassortant progeny virions are produced containing a mixture of segments from each parental virus.

2.6 Outlook

Generally, viral RNA packaging is assumed to be a process dependent on a few clearly defined RNA structural motifs specifically recognized by a viral protein. However, when evaluating binding affinities and specificities of those RNA-protein complexes *in vitro*, they often don't show the specificity that is observed for the packaging process *in vivo*[23]. Thus, it may be reasonable to think

of packaging as an integrative process that involves multiple co-occurrent interactions that must also take place at the correct time and subcellular localization for genome packaging to occur.

Excitingly, new methods to characterize RNA virus packaging signals are being developed that may help to resolve the details of these integrative processes. Approaches to study viral RNA packaging spans disciplines and can now shed light on this process across multiple scales. For example, advanced cryo-EM techniques promise to determine RNA structures at high resolution in three-dimensions[175–177]. Furthermore, cryo-EM[178–181] and X-ray scattering[182,183] may reveal RNA-protein interaction sites inside of viral capsids. In parallel, RNA structural probing techniques are being developed that enable the detection of structural changes in RNA that may be the result of RNA packaging, e.g. by identifying alternative structures[184,185] and/or mapping RNA-protein interaction sites on the RNA[186,187]. Continual improvements in quantitative live, super resolution, and expansion microscopy will be key for understanding mechanisms of viral assembly in cells[123,188–192]. These improvements are beginning to reveal how inherent variability in viral assembly allow viruses to replicate and evolve in the face of complex and unpredictable environments[123,190]. Finally, comparative high throughput sequencing can identify RNA packaging signals, be it historically from identifying genomic constraints of packaging-competent defective viral genomes[193–197], or more recently by reverse genetics systems that quantify relative packaging efficiencies of large pools of mutants in parallel[43]. Together, these technical revolutions are sure to dramatically improve our understanding of the molecular mechanisms of viral RNA genome packaging across virus families and scales. In the near future, these insights can be pivoted into novel antiviral drugs and vaccines for controlling these important human pathogens.

Author Contributions: Conceptualization, L.Y. and R.S.; writing—original draft preparation, L.Y. and R.S.; writing—review and editing, all authors.; visualization, all authors; supervision, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Helmholtz Association (VH-NG-1347 to R.S.) and the Bundesministerium für Bildung und Forschung (BMBF) (COMPLS-182 to R.S.). A.S.G. was sup-ported with a fellowship from the Peter und Traudl Engelhorn Stiftung. U.A. was supported by a fellowship from the German Academic Exchange Service (DAAD).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the writing of the manuscript, or in the decision to publish.

2.7 References

[1] S. Sun, V. B. Rao, and M. G. Rossmann, “Genome packaging in viruses,” Curr. Opin. Struct. Biol., vol. 20, no. 1, pp. 114–120, 2010, doi: 10.1016/j.sbi.2009.12.006.

[2] R. L. Adams, N. Pirakitikulr, and A. M. Pyle, “Functional RNA structures throughout the Hepatitis C Virus genome.,” Curr. Opin. Virol., vol. 24, pp. 79–86, 2017, doi: 10.1016/j.coviro.2017.04.007.

[3] K. a Wilkinson et al., “High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states.,” PLoS Biol., vol. 6, no. 4, p. e96, Apr. 2008, doi: 10.1371/journal.pbio.0060096.

[4] R. Rangan et al., “RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: A first look,” Rna, vol. 26, no. 8, pp. 937–959, 2020, doi: 10.1261/RNA.076141.120.

[5] P. S. Masters, “Coronavirus genomic RNA packaging,” Virology, vol. 537, no. August, pp. 198–207, 2019, doi: 10.1016/j.virol.2019.08.031.

[6] E. Mailler et al., “The Life-Cycle of the HIV-1 Gag-RNA Complex.,” Viruses, vol. 8, no. 9, p. 248, Sep. 2016, doi: 10.3390/v8090248.

[7] A. Borodavka, U. Desselberger, and J. T. Patton, “Genome packaging in multi-segmented dsRNA viruses: distinct mechanisms with similar outcomes,” Curr. Opin. Virol., vol. 33, pp. 106–112, 2018, doi: 10.1016/j.coviro.2018.08.001.

[8] M. Gerber, C. Isel, V. Moules, and R. Marquet, “Selective packaging of the influenza A genome and consequences for genetic reassortment,” Trends Microbiol., vol. 22, no. 8, pp. 446–455, 2014, doi: 10.1016/j.tim.2014.04.001.

[9] L. R. Newburn and K. A. White, “Trans-Acting RNA-RNA Interactions in Segmented RNA Viruses.,” Viruses, vol. 11, no. 8, 2019, doi: 10.3390/v11080751.

[10] R. P. Smyth, M. P. Davenport, and J. Mak, “The origin of genetic diversity in HIV-1.,” Virus Res., vol. 169, no. 2, pp. 415–29, Nov. 2012, doi: 10.1016/j.virusres.2012.06.015.

[11] S. Tetter et al., “Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein.,” Science, vol. 372, no. 6547, pp. 1220–1224, 2021, doi: 10.1126/science.abg2822.

[12] N. Sanchez de Groot et al., “RNA structure drives interaction with proteins,” Nat. Commun., vol. 10, no. 1, pp. 1–13, 2019, doi: 10.1038/s41467-019-10923-5.

[13] D. S. Peabody, “The RNA binding site of bacteriophage MS2 coat protein.,” EMBO J., vol. 12, no. 2, pp. 595–600, Feb. 1993.

[14] C. Z. Ni, R. Syed, R. Kodandapani, J. Wickersham, D. S. Peabody, and K. R. Ely, “Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly.,” Structure, vol. 3, no. 3, pp. 255–63, Mar. 1995, doi: 10.1016/S0969-2126(01)00156-3.

[15] J. M. Halstead et al., “Translation. An RNA biosensor for imaging the first round of translation from single cells to living animals.,” Science, vol. 347, no. 6228, pp. 1367–671, 2015, doi:

10.1126/science.aaa3380.

[16] E. Bertrand, P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer, and R. M. Long, “Localization of ASH1 mRNA particles in living yeast.,” Mol. Cell, vol. 2, no. 4, pp. 437–45, Oct. 1998, doi: 10.1016/s1097-2765(00)80143-4.

[17] W. T. Horn et al., “The crystal structure of a high affinity RNA stem-loop complexed with the bacteriophage MS2 capsid: further challenges in the modeling of ligand-RNA interactions.,” RNA, vol. 10, no. 11, pp. 1776–82, Nov. 2004, doi: 10.1261/rna.7710304.

[18] J. D. Buenrostro et al., “Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes,” Nat. Biotechnol., vol. 32, no. 6, pp. 562–8, Jun. 2014, doi: 10.1038/nbt.2880.

[19] R. P. Smyth et al., “Mutational interference mapping experiment (MIME) for studying RNA structure and function,” Nat. Methods, vol. 12, no. 9, pp. 866–872, 2015, doi: 10.1038/nmeth.3490.

[20] G. Masliah, P. Barraud, and F. H.-T. Allain, “RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence.,” Cell. Mol. Life Sci., vol. 70, no. 11, pp. 1875–95, Jun. 2013, doi: 10.1007/s00018-012-1119-x.

[21] M. Treger and E. Westhof, “Statistical analysis of atomic contacts at RNA-protein interfaces.,” J. Mol. Recognit., vol. 14, no. 4, pp. 199–214, doi: 10.1002/jmr.534.

[22] Ó. Rolfsson et al., “Direct Evidence for Packaging Signal-Mediated Assembly of Bacteriophage MS2,” J. Mol. Biol., vol. 428, no. 2, pp. 431–448, 2016, doi: 10.1016/j.jmb.2015.11.014.

[23] R. Twarock and P. G. Stockley, “RNA-Mediated Virus Assembly: Mechanisms and Consequences for Viral Evolution and Therapy.,” Annu. Rev. Biophys., vol. 48, pp. 495–514, 2019, doi: 10.1146/annurev-biophys-052118-115611.

[24] N. Patel et al., “Rewriting nature’s assembly manual for a ssRNA virus.,” Proc. Natl. Acad. Sci. U. S. A., vol. 114, no. 46, pp. 12255–12260, 2017, doi: 10.1073/pnas.1706951114.

[25] N. Patel et al., “HBV RNA pre-genome encodes specific motifs that mediate interactions with the viral core protein that promote nucleocapsid assembly.,” Nat. Microbiol., vol. 2, p. 17098, Jun. 2017, doi: 10.1038/nmicrobiol.2017.98.

[26] R. S. Brown, D. G. Anastasakis, M. Hafner, and M. Kielian, “Multiple capsid protein binding sites mediate selective packaging of the alphavirus genomic RNA,” Nat. Commun., vol. 11, no. 1, pp. 1–16, 2020, doi: 10.1038/s41467-020-18447-z.

[27] S. B. Kutluay et al., “Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis.,” Cell, vol. 159, no. 5, pp. 1096–109, Nov. 2014, doi: 10.1016/j.cell.2014.09.057.

[28] C. P. C. P. C. P. Keating et al., “The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA,” Nucleic Acids Res., vol. 37, no. 3, pp. 945–956, Feb. 2009, doi:

10.1093/nar/gkn1015.

[29] M. Mougel, Y. Zhang, and E. Barklis, “cis-active structural motifs involved in specific encapsidation of Moloney murine leukemia virus RNA.,” J. Virol., vol. 70, no. 8, pp. 5043–50, Aug. 1996, doi: 10.1128/JVI.70.8.5043-5050.1996.

[30] J. E. Murphy and S. P. Goff, “Construction and analysis of deletion mutations in the U5 region of Moloney murine leukemia virus: effects on RNA packaging and reverse transcription.,” J. Virol., vol. 63, no. 1, pp. 319–27, Jan. 1989, doi: 10.1128/JVI.63.1.319-327.1989.

[31] P. Ding et al., “Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal.,” Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 30, pp. 17737–17746, 2020, doi: 10.1073/pnas.2008519117.

[32] K. Lu, X. Heng, and M. F. Summers, “Structural determinants and mechanism of HIV-1 genome packaging.,” J. Mol. Biol., vol. 410, no. 4, pp. 609–33, Jul. 2011, doi: 10.1016/j.jmb.2011.04.029.

[33] J. Luban and S. P. Goff, “Binding of human immunodeficiency virus type 1 (HIV-1) RNA to recombinant HIV-1 gag polyprotein.,” J. Virol., vol. 65, no. 6, pp. 3203–12, Jun. 1991, doi: 10.1128/JVI.65.6.3203-3212.1991.

[34] J. Luban and S. P. Goff, “Mutational analysis of cis-acting packaging signals in human immunodeficiency virus type 1 RNA.,” J. Virol., vol. 68, no. 6, pp. 3784–93, Jun. 1994, doi: 10.1128/JVI.68.6.3784-3793.1994.

[35] M. S. McBride and A. T. Panganiban, “The human immunodeficiency virus type 1 encapsidation site is a multipartite RNA element composed of functional hairpin structures.,” J. Virol., vol. 70, no. 5, pp. 2963–73, May 1996, doi: 10.1128/JVI.70.5.2963-2973.1996.

[36] A. Lever, H. Gottlinger, W. Haseltine, and J. Sodroski, “Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions.,” J. Virol., vol. 63, no. 9, pp. 4085–7, Sep. 1989, doi: 10.1128/JVI.63.9.4085-4087.1989.

[37] A. Aldovini and R. A. Young, “Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus,” J. Virol., vol. 64, no. 5, pp. 1920–1926, 1990, doi: 10.1128/jvi.64.5.1920-1926.1990.

[38] F. Clavel and J. M. Orenstein, “A mutant of human immunodeficiency virus with reduced RNA packaging and abnormal particle morphology,” J. Virol., vol. 64, no. 10, pp. 5230–5234, 1990, doi: 10.1128/jvi.64.10.5230-5234.1990.

[39] R. N. De Guzman, Z. R. Wu, C. C. Stalling, L. Pappalardo, P. N. Borer, and M. F. Summers, “Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element.,” Science, vol. 279, no. 5349, pp. 384–8, Jan. 1998, doi: 10.1126/science.279.5349.384.

[40] G. K. Amarasinghe, R. N. De Guzman, R. B. Turner, K. J. Chancellor, Z. R. Wu, and M. F. Summers, “NMR structure of the HIV-1 nucleocapsid protein bound to stem-loop SL2 of the

psi-RNA packaging signal. Implications for genome recognition.," J. Mol. Biol., vol. 301, no. 2, pp. 491–511, Aug. 2000, doi: 10.1006/jmbi.2000.3979.

[41] E. W. Abd El-Wahab et al., "Specific recognition of the HIV-1 genomic RNA by the Gag precursor.," Nat. Commun., vol. 5, no. 1, p. 4304, Jul. 2014, doi: 10.1038/ncomms5304.

[42] L. Houzet et al., "HIV controls the selective packaging of genomic, spliced viral and cellular RNAs into virions through different mechanisms.," Nucleic Acids Res., vol. 35, no. 8, pp. 2695–704, Jan. 2007, doi: 10.1093/nar/gkm153.

[43] R. P. Smyth et al., "In cell mutational interference mapping experiment (in cell MIME) identifies the 5' polyadenylation signal as a dual regulator of HIV-1 genomic RNA production and packaging," Nucleic Acids Res., vol. 46, no. 9, pp. 1–16, 2018, doi: 10.1093/nar/gky152.

[44] S. Bernacchi et al., "HIV-1 Pr55Gag binds genomic and spliced RNAs with different affinity and stoichiometry," RNA Biol., vol. 14, no. 1, pp. 90–103, 2017, doi: 10.1080/15476286.2016.1256533.

[45] M. Comas-Garcia, S. A. Datta, L. Baker, R. Varma, P. R. Gudla, and A. Rein, "Dissection of specific binding of HIV-1 Gag to the 'packaging signal' in viral RNA.," Elife, vol. 6, 2017, doi: 10.7554/eLife.27055.

[46] A. Rein, "RNA Packaging in HIV," Trends Microbiol., vol. 27, no. 8, pp. 715–723, 2019, doi: 10.1016/j.tim.2019.04.003.

[47] J. A. Webb, C. P. Jones, L. J. Parent, I. Rouzina, and K. Musier-Forsyth, "Distinct binding interactions of HIV-1 Gag to Psi and non-Psi RNAs: Implications for viral genomic RNA packaging," Rna, vol. 19, no. 8, pp. 1078–1088, 2013, doi: 10.1261/rna.038869.113.

[48] W.-C. Hsin et al., "Nucleocapsid protein-dependent assembly of the RNA packaging signal of Middle East respiratory syndrome coronavirus.," J. Biomed. Sci., vol. 25, no. 1, p. 47, May 2018, doi: 10.1186/s12929-018-0449-x.

[49] P.-K. Hsieh et al., "Assembly of severe acute respiratory syndrome coronavirus RNA packaging signal into virus-like particles is nucleocapsid dependent.," J. Virol., vol. 79, no. 22, pp. 13848–55, Nov. 2005, doi: 10.1128/JVI.79.22.13848-13855.2005.

[50] S. Klein et al., "SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography.," Nat. Commun., vol. 11, no. 1, p. 5885, 2020, doi: 10.1038/s41467-020-19619-7.

[51] C. Iserman et al., "Genomic RNA Elements Drive Phase Separation of the SARS-CoV-2 Nucleocapsid.," Mol. Cell, vol. 80, no. 6, pp. 1078-1091.e6, 2020, doi: 10.1016/j.molcel.2020.11.041.

[52] J. Wang, C. Shi, Q. Xu, and H. Yin, "SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation into stress granules through its N-terminal intrinsically disordered region.," Cell Discov., vol. 7, no. 1, p. 5, Jan. 2021, doi: 10.1038/s41421-020-00240-3.

[53] A. Savastano, A. Ibáñez de Opakua, M. Rankovic, and M. Zweckstetter, “Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates.,” Nat. Commun., vol. 11, no. 1, p. 6041, 2020, doi: 10.1038/s41467-020-19843-1.

[54] H. Chen et al., “Liquid-liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA.,” Cell Res., vol. 30, no. 12, pp. 1143–1145, 2020, doi: 10.1038/s41422-020-00408-2.

[55] C. R. Carlson et al., “Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions.,” Mol. Cell, vol. 80, no. 6, pp. 1092-1103.e4, 2020, doi: 10.1016/j.molcel.2020.11.025.

[56] T. M. Perdikari, A. C. Murthy, V. H. Ryan, S. Watters, M. T. Naik, and N. L. Fawzi, “SARS-CoV-2 nucleocapsid protein phase-separates with RNA and with human hnRNPs.,” EMBO J., vol. 39, no. 24, p. e106478, 2020, doi: 10.15252/embj.2020106478.

[57] Y. Wu et al., “RNA-induced liquid phase separation of SARS-CoV-2 nucleocapsid protein facilitates NF-κB hyper-activation and inflammation.,” Signal Transduct. Target. Ther., vol. 6, no. 1, p. 167, 2021, doi: 10.1038/s41392-021-00575-7.

[58] J. Cubuk et al., “The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA.,” Nat. Commun., vol. 12, no. 1, p. 1936, 2021, doi: 10.1038/s41467-021-21953-3.

[59] S. Lu et al., “The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein.,” Nat. Commun., vol. 12, no. 1, p. 502, 2021, doi: 10.1038/s41467-020-20768-y.

[60] S. Guseva et al., “Measles virus nucleo- and phosphoproteins form liquid-like phase-separated compartments that promote nucleocapsid assembly.,” Sci. Adv., vol. 6, no. 14, p. eaaz7095, 2020, doi: 10.1126/sciadv.aaz7095.

[61] T. A. Etibor, Y. Yamauchi, and M. J. Amorim, “Liquid Biomolecular Condensates and Viral Lifecycles: Review and Perspectives.,” Viruses, vol. 13, no. 3, 2021, doi: 10.3390/v13030366.

[62] D. S. W. Protter et al., “Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly.,” Cell Rep., vol. 22, no. 6, pp. 1401–1412, 2018, doi: 10.1016/j.celrep.2018.01.036.

[63] S. Wang et al., “Targeting liquid-liquid phase separation of SARS-CoV-2 nucleocapsid protein promotes innate antiviral immunity by elevating MAVS activity.,” Nat. Cell Biol., Jul. 2021, doi: 10.1038/s41556-021-00710-0.

[64] M. Zhao et al., “GCG inhibits SARS-CoV-2 replication by disrupting the liquid phase condensation of its nucleocapsid protein.,” Nat. Commun., vol. 12, no. 1, p. 2114, 2021, doi: 10.1038/s41467-021-22297-8.

[65] A. Pflug, D. Guilligay, S. Reich, and S. Cusack, “Structure of influenza A polymerase bound to the viral RNA promoter.,” Nature, vol. 516, no. 7531, pp. 355–60, Dec. 2014, doi:

10.1038/nature14008.

[66] S. Tchatalbachev, R. Flick, and G. Hobom, “The packaging signal of influenza viral RNA molecules.,” RNA, vol. 7, no. 7, pp. 979–89, Jul. 2001, doi: 10.1017/s1355838201002424.

[67] C. Chaimayo, T. Hayashi, A. Underwood, E. Hodges, and T. Takimoto, “Selective incorporation of vRNP into influenza A virions determined by its specific interaction with M1 protein.,” Virology, vol. 505, pp. 23–32, 2017, doi: 10.1016/j.virol.2017.02.008.

[68] S. Brandt, M. Blissenbach, B. Grewe, R. Konietzny, T. Grunwald, and K. Uberla, “Rev proteins of human and simian immunodeficiency virus enhance RNA encapsidation.,” PLoS Pathog., vol. 3, no. 4, p. e54, Apr. 2007, doi: 10.1371/journal.ppat.0030054.

[69] M. Blissenbach, B. Grewe, B. Hoffmann, S. Brandt, and K. Uberla, “Nuclear RNA export and packaging functions of HIV-1 Rev revisited.,” J. Virol., vol. 84, no. 13, pp. 6598–604, Jul. 2010, doi: 10.1128/JVI.02264-09.

[70] G. M. Pocock, J. T. Becker, C. M. Swanson, P. Ahlquist, and N. M. Sherer, “HIV-1 and M-PMV RNA Nuclear Export Elements Program Viral Genomes for Distinct Cytoplasmic Trafficking Behaviors.,” PLoS Pathog., vol. 12, no. 4, p. e1005565, Apr. 2016, doi: 10.1371/journal.ppat.1005565.

[71] B. Grewe, K. Ehrhardt, B. Hoffmann, M. Blissenbach, S. Brandt, and K. Uberla, “The HIV-1 Rev protein enhances encapsidation of unspliced and spliced, RRE-containing lentiviral vector RNA.,” PLoS One, vol. 7, no. 11, p. e48688, 2012, doi: 10.1371/journal.pone.0048688.

[72] A. Serganov and E. Nudler, “A decade of riboswitches.,” Cell, vol. 152, no. 1–2, pp. 17–24, Jan. 2013, doi: 10.1016/j.cell.2012.12.024.

[73] S. R. Lee and J. Lykke-Andersen, “Emerging roles for ribonucleoprotein modification and remodeling in controlling RNA fate.,” Trends Cell Biol., vol. 23, no. 10, pp. 504–10, Oct. 2013, doi: 10.1016/j.tcb.2013.05.001.

[74] S. Shetty, S. Stefanovic, and M. R. Mihailescu, “Hepatitis C virus RNA: molecular switches mediated by long-range RNA-RNA interactions?,” Nucleic Acids Res., vol. 41, no. 4, pp. 2526–40, Feb. 2013, doi: 10.1093/nar/gks1318.

[75] C. Romero-López, A. Barroso-Deljesus, A. García-Sacristán, C. Briones, and A. Berzal-Herranz, “End-to-end crosstalk within the hepatitis C virus genome mediates the conformational switch of the 3’X-tail region.,” Nucleic Acids Res., vol. 42, no. 1, pp. 567–82, Jan. 2014, doi: 10.1093/nar/gkt841.

[76] N. Pirakitikulr, A. Kohlway, B. D. Lindenbach, and A. M. Pyle, “The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures.,” Mol. Cell, vol. 62, no. 1, pp. 111–20, Apr. 2016, doi: 10.1016/j.molcel.2016.01.024.

[77] P. Friebe, J. Boudet, J.-P. Simorre, and R. Bartenschlager, “Kissing-loop interaction in the 3’ end of the hepatitis C virus genome essential for RNA replication.,” J. Virol., vol. 79, no. 1, pp.

380–92, Jan. 2005, doi: 10.1128/JVI.79.1.380-392.2005.

[78] S. You, D. D. Stump, A. D. Branch, and C. M. Rice, “A cis-acting replication element in the sequence encoding the NS5B RNA-dependent RNA polymerase is required for hepatitis C virus RNA replication.,” J. Virol., vol. 78, no. 3, pp. 1352–66, Feb. 2004, doi: 10.1128/jvi.78.3.1352-1366.2004.

[79] H. Lee, H. Shin, E. Wimmer, and A. V Paul, “cis-acting RNA signals in the NS5B C-terminal coding sequence of the hepatitis C virus genome.,” J. Virol., vol. 78, no. 20, pp. 10865–77, Oct. 2004, doi: 10.1128/JVI.78.20.10865-10877.2004.

[80] A. Tuplin, M. Struthers, P. Simmonds, and D. J. Evans, “A twist in the tail: SHAPE mapping of long-range interactions and structural rearrangements of RNA elements involved in HCV replication.,” Nucleic Acids Res., vol. 40, no. 14, pp. 6908–21, Aug. 2012, doi: 10.1093/nar/gks370.

[81] G. Shi et al., “Involvement of the 3’ Untranslated Region in Encapsidation of the Hepatitis C Virus.,” PLoS Pathog., vol. 12, no. 2, p. e1005441, Feb. 2016, doi: 10.1371/journal.ppat.1005441.

[82] C. Romero-López and A. Berzal-Herranz, “The functional RNA domain 5BSL3.2 within the NS5B coding sequence influences hepatitis C virus IRES-mediated translation.,” Cell. Mol. Life Sci., vol. 69, no. 1, pp. 103–13, Jan. 2012, doi: 10.1007/s00018-011-0729-z.

[83] S. Shetty, S. Kim, T. Shimakami, S. M. Lemon, and M.-R. Mihailescu, “Hepatitis C virus genomic RNA dimerization is mediated via a kissing complex intermediate.,” RNA, vol. 16, no. 5, pp. 913–25, May 2010, doi: 10.1261/rna.1960410.

[84] R. Ivanyi-Nagy et al., “Analysis of hepatitis C virus RNA dimerization and core-RNA interactions.,” Nucleic Acids Res., vol. 34, no. 9, pp. 2618–33, 2006, doi: 10.1093/nar/gkl240.

[85] H. Huthoff and B. Berkhout, “Two alternating structures of the HIV-1 leader RNA.,” RNA, vol. 7, no. 1, pp. 143–57, Jan. 2001, doi: 10.1017/s1355838201001881.

[86] M. Ooms, H. Huthoff, R. Russell, C. Liang, and B. Berkhout, “A Riboswitch Regulates RNA Dimerization and Packaging in Human Immunodeficiency Virus Type 1 Virions,” J. Virol., vol. 78, no. 19, pp. 10814–10819, 2004, doi: 10.1128/jvi.78.19.10814-10819.2004.

[87] J. D. Brown et al., “Structural basis for transcriptional start site control of HIV-1 RNA fate.,” Science, vol. 368, no. 6489, pp. 413–417, 2020, doi: 10.1126/science.aaz7959.

[88] T. E. M. Abbink and B. Berkhout, “A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon.,” J. Biol. Chem., vol. 278, no. 13, pp. 11601–11, Mar. 2003, doi: 10.1074/jbc.M210291200.

[89] S. C. Keane et al., “RNA structure. Structure of the HIV-1 RNA packaging signal.,” Science, vol. 348, no. 6237, pp. 917–21, May 2015, doi: 10.1126/science.aaa9266.

[90] K. Lu et al., “NMR detection of structures in the HIV-1 5’-leader RNA that regulate genome packaging.,” Science, vol. 334, no. 6053, pp. 242–5, Oct. 2011, doi: 10.1126/science.1210460.

[91] T. E. M. Abbink, M. Ooms, P. C. J. Haasnoot, and B. Berkhout, “The HIV-1 Leader RNA Conformational Switch Regulates RNA Dimerization but Does Not Regulate mRNA Translation †,” Biochemistry, vol. 44, no. 25, pp. 9058–9066, Jun. 2005, doi: 10.1021/bi0502588.

[92] B. S. Brigham, J. P. Kitzrow, J.-P. C. Reyes, K. Musier-Forsyth, and J. B. Munro, “Intrinsic conformational dynamics of the HIV-1 genomic RNA 5’UTR,” Proc. Natl. Acad. Sci., vol. 116, no. 21, pp. 10372–10381, May 2019, doi: 10.1073/pnas.1902271116.

[93] K. L. K. L. Jones et al., “Early Events of HIV-1 Infection: Can Signaling be the Next Therapeutic Target?,” J. Neuroimmune Pharmacol., vol. 6, no. 2, pp. 269–283, Mar. 2011, doi: 10.1007/s11481-011-9268-5.

[94] E. Skripkin et al., “Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro.,” Proc. Natl. Acad. Sci. U. S. A., vol. 91, no. 11, pp. 4945–9, May 1994, doi: 10.1073/pnas.91.11.4945.

[95] J. C. Paillart, R. Marquet, E. Skripkin, B. Ehresmann, and C. Ehresmann, “Mutational analysis of the bipartite dimer linkage structure of human immunodeficiency virus type 1 genomic RNA.,” J. Biol. Chem., vol. 269, no. 44, pp. 27486–93, Nov. 1994.

[96] J. C. Paillart et al., “A dual role of the putative RNA dimerization initiation site of human immunodeficiency virus type 1 in genomic RNA packaging and proviral DNA synthesis.,” J. Virol., vol. 70, no. 12, pp. 8348–54, Dec. 1996.

[97] J. C. Paillart, M. Shehu-Xhilaga, R. Marquet, and J. Mak, “Dimerization of retroviral RNA genomes: An inseparable pair,” Nat. Rev. Microbiol., vol. 2, no. 6, pp. 461–472, 2004, doi: 10.1038/nrmicro903.

[98] R. S. Russell, C. Liang, and M. A. Wainberg, “Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably?,” Retrovirology, vol. 1, no. Mlv, pp. 1–14, 2004, doi: 10.1186/1742-4690-1-23.

[99] S. C. Keane et al., “Structure of the HIV-1 RNA packaging signal,” Science (80-.)., vol. 348, no. 6237, pp. 917–921, 2015, doi: 10.1126/science.aaa9266.

[100] S. C. Keane et al., “NMR detection of intermolecular interaction sites in the dimeric 5’-leader of the HIV-1 genome,” Proc. Natl. Acad. Sci., p. 201614785, 2016, doi: 10.1073/pnas.1614785113.

[101] P. S. Boyd et al., “NMR studies of retroviral genome packaging,” Viruses, vol. 12, no. 10, pp. 1–52, 2020, doi: 10.3390/v12101115.

[102] S. Kharytonchyk et al., “Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome.,” Proc. Natl. Acad. Sci. U. S. A., p. 201616627, 2016, doi: 10.1073/pnas.1616627113.

[103] C. L. Jopling, M. Yi, A. M. Lancaster, S. M. Lemon, and P. Sarnow, “Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA.,” Science, vol. 309, no. 5740, pp. 1577–81, Sep. 2005, doi: 10.1126/science.1113329.

[104] J. M. Luna et al., “Hepatitis C virus RNA functionally sequesters miR-122.,” Cell, vol. 160, no. 6, pp. 1099–110, Mar. 2015, doi: 10.1016/j.cell.2015.02.025.

[105] C. L. Jopling, S. Schütz, and P. Sarnow, “Position-dependent function for a tandem microRNA miR-122-binding site located in the hepatitis C virus RNA genome.,” Cell Host Microbe, vol. 4, no. 1, pp. 77–85, Jul. 2008, doi: 10.1016/j.chom.2008.05.013.

[106] Y. Li, T. Masaki, D. Yamane, D. R. McGivern, and S. M. Lemon, “Competing and noncompeting activities of miR-122 and the 5’ exonuclease Xrn1 in regulation of hepatitis C virus replication.,” Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 5, pp. 1881–6, Jan. 2013, doi: 10.1073/pnas.1213515110.

[107] C. D. Sedano and P. Sarnow, “Hepatitis C virus subverts liver-specific miR-122 to protect the viral genome from exoribonuclease Xrn2.,” Cell Host Microbe, vol. 16, no. 2, pp. 257–264, Aug. 2014, doi: 10.1016/j.chom.2014.07.006.

[108] T. Shimakami et al., “Stabilization of hepatitis C virus RNA by an Ago2-miR-122 complex.,” Proc. Natl. Acad. Sci. U. S. A., vol. 109, no. 3, pp. 941–6, Jan. 2012, doi: 10.1073/pnas.1112263109.

[109] M. Niepmann, “Activation of hepatitis C virus translation by a liver-specific microRNA.,” Cell Cycle, vol. 8, no. 10, pp. 1473–7, May 2009, doi: 10.4161/cc.8.10.8349.

[110] A. P. E. Roberts, A. P. Lewis, and C. L. Jopling, “miR-122 activates hepatitis C virus translation by a specialized mechanism requiring particular RNA components.,” Nucleic Acids Res., vol. 39, no. 17, pp. 7716–29, Sep. 2011, doi: 10.1093/nar/gkr426.

[111] P. Schult et al., “microRNA-122 amplifies hepatitis C virus translation by shaping the structure of the internal ribosomal entry site.,” Nat. Commun., vol. 9, no. 1, p. 2613, 2018, doi: 10.1038/s41467-018-05053-3.

[112] J. I. Henke et al., “microRNA-122 stimulates translation of hepatitis C virus RNA.,” EMBO J., vol. 27, no. 24, pp. 3300–10, Dec. 2008, doi: 10.1038/emboj.2008.244.

[113] R. D. Kunden, S. Ghezelbash, J. Q. Khan, and J. A. Wilson, “Location specific annealing of miR-122 and other small RNAs defines an Hepatitis C Virus 5’ UTR regulatory element with distinct impacts on virus translation and genome stability.,” Nucleic Acids Res., vol. 48, no. 16, pp. 9235–9249, 2020, doi: 10.1093/nar/gkaa664.

[114] T. Masaki et al., “miR-122 stimulates hepatitis C virus RNA synthesis by altering the balance of viral RNAs engaged in replication versus translation.,” Cell Host Microbe, vol. 17, no. 2, pp. 217–28, Feb. 2015, doi: 10.1016/j.chom.2014.12.014.

[115] T. Fukuhara et al., “Expression of microRNA miR-122 facilitates an efficient replication in

nonhepatic cells upon infection with hepatitis C virus.," J. Virol., vol. 86, no. 15, pp. 7918–33, Aug. 2012, doi: 10.1128/JVI.00567-12.

[116] M. Jiang, J. Mak, M. A. Wainberg, M. A. Parniak, E. Cohen, and L. Kleiman, "Variable tRNA content in HIV-1IIIB.," Biochem. Biophys. Res. Commun., vol. 185, no. 3, pp. 1005–15, Jun. 1992, doi: 10.1016/0006-291x(92)91727-8.

[117] M. Jiang et al., "Identification of tRNAs incorporated into wild-type and mutant human immunodeficiency virus type 1.," J. Virol., vol. 67, no. 6, pp. 3246–53, Jun. 1993, doi: 10.1128/JVI.67.6.3246-3253.1993.

[118] E. Seif, M. Niu, and L. Kleiman, "Annealing to sequences within the primer binding site loop promotes an HIV-1 RNA conformation favoring RNA dimerization and packaging.," RNA, vol. 19, no. 10, pp. 1384–93, Oct. 2013, doi: 10.1261/rna.038497.113.

[119] P. J. Wichgers Schreur and J. Kortekaas, "Single-Molecule FISH Reveals Non-selective Packaging of Rift Valley Fever Virus Genome Segments.," PLoS Pathog., vol. 12, no. 8, p. e1005800, 2016, doi: 10.1371/journal.ppat.1005800.

[120] E. Bermúdez-Méndez, E. A. Katrukha, C. M. Spruit, J. Kortekaas, and P. J. Wichgers Schreur, "Visualizing the ribonucleoprotein content of single bunyavirus virions reveals more efficient genome packaging in the arthropod host.," Commun. Biol., vol. 4, no. 1, p. 345, Mar. 2021, doi: 10.1038/s42003-021-01821-y.

[121] E. C. Hutchinson, J. C. von Kirchbach, J. R. Gog, and P. Digard, "Genome packaging in influenza A virus," J. Gen. Virol., vol. 91, no. 2, pp. 313–328, 2010, doi: 10.1099/vir.0.017608-0.

[122] Y.-Y. Chou et al., "One influenza virus particle packages eight unique viral RNAs as shown by FISH analysis.," Proc. Natl. Acad. Sci. U. S. A., vol. 109, no. 23, pp. 9101–6, Jun. 2012, doi: 10.1073/pnas.1206069109.

[123] I. Haralampiev et al., "Selective flexible packaging pathways of the segmented genome of influenza A virus," Nat. Commun., vol. 11, no. 1, pp. 1–13, 2020, doi: 10.1038/s41467-020-18108-1.

[124] T. Noda et al., "Architecture of ribonucleoprotein complexes in influenza A virus particles," Nature, vol. 439, no. 7075, pp. 490–492, 2006, doi: 10.1038/nature04378.

[125] T. Noda et al., "Importance of the 1+7 configuration of ribonucleoprotein complexes for influenza A virus genome packaging," Nat. Commun., vol. 9, no. 1, pp. 1–10, 2018, doi: 10.1038/s41467-017-02517-w.

[126] E. Fournier et al., "A supramolecular assembly formed by influenza A virus genomic RNA segments," Nucleic Acids Res., vol. 40, no. 5, pp. 2197–2209, 2012, doi: 10.1093/nar/gkr985.

[127] S. D. Duhaut and J. W. McCauley, "Defective RNAs inhibit the assembly of influenza virus genome segments in a segment-specific manner," Virology, vol. 216, no. 2, pp. 326–337, 1996, doi: 10.1006/viro.1996.0068.

[128] S. D. Duhaut and N. J. Dimmock, “Heterologous protection of mice from a lethal human H1N1 influenza A virus infection by H3N8 equine defective interfering virus: comparison of defective RNA sequences isolated from the DI inoculum and mouse lung.,” Virology, vol. 248, no. 2, pp. 241–53, Sep. 1998, doi: 10.1006/viro.1998.9267.

[129] P. A. Jennings, J. T. Finch, G. Winter, and J. S. Robertson, “Does the higher order structure of the influenza virus ribonucleoprotein guide sequence rearrangements in influenza viral RNA?,” Cell, vol. 34, no. 2, pp. 619–27, Sep. 1983, doi: 10.1016/0092-8674(83)90394-x.

[130] S. Noble and N. J. Dimmock, “Characterization of putative defective interfering (DI) A/WSN RNAs isolated from the lungs of mice protected from an otherwise lethal respiratory infection with influenza virus A/WSN (H1N1): a subset of the inoculum DI RNAs.,” Virology, vol. 210, no. 1, pp. 9–19, Jun. 1995, doi: 10.1006/viro.1995.1312.

[131] E. Dos Santos Afonso, N. Escriou, I. Leclercq, S. van der Werf, and N. Naffakh, “The generation of recombinant influenza A viruses expressing a PB2 fusion protein requires the conservation of a packaging signal overlapping the coding and noncoding regions at the 5’ end of the PB2 segment.,” Virology, vol. 341, no. 1, pp. 34–46, Oct. 2005, doi: 10.1016/j.virol.2005.06.040.

[132] Y. Liang, Y. Hong, and T. G. Parslow, “cis-Acting packaging signals in the influenza virus PB1, PB2, and PA genomic RNA segments.,” J. Virol., vol. 79, no. 16, pp. 10348–55, Aug. 2005, doi: 10.1128/JVI.79.16.10348-10355.2005.

[133] Y. Liang, T. Huang, H. Ly, T. G. Parslow, and Y. Liang, “Mutational analyses of packaging signals in influenza virus PA, PB1, and PB2 genomic RNA segments.,” J. Virol., vol. 82, no. 1, pp. 229–36, Jan. 2008, doi: 10.1128/JVI.01541-07.

[134] K. Fujii et al., “Importance of both the coding and the segment-specific noncoding regions of the influenza A virus NS segment for its efficient incorporation into virions.,” J. Virol., vol. 79, no. 6, pp. 3766–74, 2005, doi: 10.1128/JVI.79.6.3766-3774.2005.

[135] Y. Fujii, H. Goto, T. Watanabe, T. Yoshida, and Y. Kawaoka, “Selective incorporation of influenza virus RNA segments into virions.,” Proc. Natl. Acad. Sci. U. S. A., vol. 100, no. 4, pp. 2002–7, Feb. 2003, doi: 10.1073/pnas.0437772100.

[136] G. A. Marsh, R. Hatami, and P. Palese, “Specific residues of the influenza A virus hemagglutinin viral RNA are important for efficient packaging into budding virions,” J Virol, vol. 81, no. 18, pp. 9727–9736, 2007, doi: 10.1128/jvi.01144-07.

[137] M. Ozawa et al., “Nucleotide sequence requirements at the 5’ end of the influenza A virus M RNA segment for efficient virus replication.,” J. Virol., vol. 83, no. 7, pp. 3384–8, Apr. 2009, doi: 10.1128/JVI.02513-08.

[138] M. Ozawa et al., “Contributions of two nuclear localization signals of influenza A virus nucleoprotein to viral replication.,” J. Virol., vol. 81, no. 1, pp. 30–41, Jan. 2007, doi: 10.1128/JVI.01434-06.

[139] E. C. Hutchinson, H. M. Wise, K. Kudryavtseva, M. D. Curran, and P. Digard, “Characterisation

of influenza A viruses with mutations in segment 5 packaging signals.," Vaccine, vol. 27, no. 45, pp. 6270–5, Oct. 2009, doi: 10.1016/j.vaccine.2009.05.053.

[140] H. Goto, Y. Muramoto, T. Noda, and Y. Kawaoka, "The genome-packaging signal of the influenza A virus genome comprises a genome incorporation signal and a genome-bundling signal.," J. Virol., vol. 87, no. 21, pp. 11316–22, 2013, doi: 10.1128/JVI.01301-13.

[141] T. Noda et al., "Three-dimensional analysis of ribonucleoprotein complexes in influenza A virus," Nat. Commun., vol. 3, 2012, doi: 10.1038/ncomms1647.

[142] E. Fournier et al., "Interaction network linking the human H3N2 influenza A virus genomic RNA segments," Vaccine, vol. 30, no. 51, pp. 7359–7367, 2012, doi: 10.1016/j.vaccine.2012.09.079.

[143] C. Gavazzi et al., "A functional sequence-specific interaction between influenza A virus genomic RNA segments," Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 41, pp. 16604–16609, 2013, doi: 10.1073/pnas.1314419110.

[144] C. Gavazzi et al., "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 Res., vol. 41, no. 2, pp. 1241–1254, 2013, doi: 10.1093/nar/gks1181.

[145] E. C. Hutchinson, M. D. Curran, E. K. Read, J. R. Gog, and P. Digard, "Mutational analysis of cis-acting RNA signals in segment 7 of influenza A virus.," J. Virol., vol. 82, no. 23, pp. 11869–79, Dec. 2008, doi: 10.1128/JVI.01634-08.

[146] B. Gilbertson et al., "Influenza NA and PB1 Gene Segments Interact during the Formation of Viral Progeny: Localization of the Binding Region within the PB1 Gene.," Viruses, vol. 8, no. 8, p. 238, 2016, doi: 10.3390/v8080238.

[147] B. Dadonaite et al., "The structure of the influenza A virus genome," Nat. Microbiol., vol. 4, no. 11, pp. 1781–1789, 2019, doi: 10.1038/s41564-019-0513-7.

[148] V. Le Sage, J. P. Kanarek, D. J. Snyder, V. S. Cooper, S. S. Lakdawala, and N. Lee, "Mapping of Influenza Virus RNA-RNA Interactions Reveals a Flexible Network," Cell Rep., vol. 31, no. 13, p. 107823, 2020, doi: 10.1016/j.celrep.2020.107823.

[149] H. Bolte, M. E. Rosu, E. Hagelauer, A. García-Sastre, and M. Schwemmle, "Packaging of the influenza A virus genome is governed by a plastic network of RNA/protein interactions," J. Virol., no. November, 2018, doi: 10.1128/JVI.01861-18.

[150] G. D. Williams et al., "Nucleotide resolution mapping of influenza A virus nucleoprotein-RNA interactions reveals RNA features required for replication.," Nat. Commun., vol. 9, no. 1, p. 465, Jan. 2018, doi: 10.1038/s41467-018-02886-w.

[151] É. A. Moreira et al., "A conserved influenza A virus nucleoprotein code controls specific viral genome packaging.," Nat. Commun., vol. 7, p. 12861, 2016, doi: 10.1038/ncomms12861.

[152] H. Jayaram, M. K. Estes, and B. V. V. Prasad, “Emerging themes in rotavirus cell entry, genome organization, transcription and replication.,” Virus Res., vol. 101, no. 1, pp. 67–81, Apr. 2004, doi: 10.1016/j.virusres.2003.12.007.

[153] T. Fajardo, P. Y. Sung, C. C. Celma, and P. Roy, “Rotavirus genomic RNA complex forms via specific RNA–RNA interactions: Disruption of RNA complex inhibits virus infectivity,” Viruses, vol. 9, no. 7, pp. 1–15, 2017, doi: 10.3390/v9070167.

[154] T. Fajardo, P. Y. Sung, and P. Roy, “Disruption of Specific RNA-RNA Interactions in a Double-Stranded RNA Virus Inhibits Genome Packaging and Virus Infectivity,” PLoS Pathog., vol. 11, no. 12, 2015, doi: 10.1371/journal.ppat.1005321.

[155] P.-Y. Sung and P. Roy, “Sequential packaging of RNA genomic segments during the assembly of Bluetongue virus.,” Nucleic Acids Res., vol. 42, no. 22, pp. 13824–38, Dec. 2014, doi: 10.1093/nar/gku1171.

[156] K. AlShaikhahmed, G. Leonov, P.-Y. Sung, R. J. Bingham, R. Twarock, and P. Roy, “Dynamic network approach for the modelling of genomic sub-complexes in multi-segmented viruses.,” Nucleic Acids Res., vol. 46, no. 22, pp. 12087–12098, 2018, doi: 10.1093/nar/gky881.

[157] S. S. Lakdawala et al., “Influenza A Virus Assembly Intermediates Fuse in the Cytoplasm,” PLoS Pathog., vol. 10, no. 3, 2014, doi: 10.1371/journal.ppat.1003971.

[158] Y. Chou et al., “Colocalization of different influenza viral RNA segments in the cytoplasm before viral budding as shown by single-molecule sensitivity FISH analysis.,” PLoS Pathog., vol. 9, no. 5, p. e1003358, Jan. 2013, doi: 10.1371/journal.ppat.1003358.

[159] R. P. Smyth and M. Negroni, “A step forward understanding HIV-1 diversity.,” Retrovirology, vol. 13, no. 1, p. 27, Apr. 2016, doi: 10.1186/s12977-016-0259-8.

[160] S. M. McDonald, M. I. Nelson, P. E. Turner, and J. T. Patton, “Reassortment in segmented RNA viruses: Mechanisms and outcomes,” Nat. Rev. Microbiol., vol. 14, no. 7, pp. 448–460, 2016, doi: 10.1038/nrmicro.2016.46.

[161] R. P. Smyth et al., “Identifying Recombination Hot Spots in the HIV-1 Genome.,” J. Virol., vol. 88, no. 5, pp. 2891–902, Mar. 2014, doi: 10.1128/JVI.03014-13.

[162] T. E. Schlub, R. P. Smyth, A. J. Grimm, J. Mak, and M. P. Davenport, “Accurately measuring recombination between closely related HIV-1 genomes.,” PLoS Comput. Biol., vol. 6, no. 4, p. e1000766, Apr. 2010, doi: 10.1371/journal.pcbi.1000766.

[163] D. Cromer et al., “HIV-1 Mutation and Recombination Rates Are Different in Macrophages and T-cells.,” Viruses, vol. 8, no. 4, p. 118, Apr. 2016, doi: 10.3390/v8040118.

[164] T. E. Schlub et al., “Fifteen to twenty percent of HIV substitution mutations are associated with recombination.,” J. Virol., vol. 88, no. 7, pp. 3837–49, Apr. 2014, doi: 10.1128/JVI.03136-13.

[165] D. N. Levy, G. M. Aldrovandi, O. Kutsch, and G. M. Shaw, “Dynamics of HIV-1 recombination in its natural target cells.,” Proc. Natl. Acad. Sci. U. S. A., vol. 101, no. 12, pp. 4204–9, Mar. 2004, doi: 10.1073/pnas.0306764101.

[166] M. P. S. Chin, T. D. Rhodes, J. Chen, W. Fu, and W.-S. Hu, “Identification of a major restriction in HIV-1 intersubtype recombination.,” Proc. Natl. Acad. Sci. U. S. A., vol. 102, no. 25, pp. 9002–7, Jun. 2005, doi: 10.1073/pnas.0502522102.

[167] J. C. Paillart, E. Skripkin, B. Ehresmann, C. Ehresmann, and R. Marquet, “A loop-loop ‘kissing’ complex is the essential part of the dimer linkage of genomic HIV-1 RNA.,” Proc. Natl. Acad. Sci. U. S. A., vol. 93, no. 11, pp. 5572–7, May 1996.

[168] M. K. Hill, M. Shehu-Xhilaga, S. M. Campbell, P. Poumbourios, S. M. Crowe, and J. Mak, “The Dimer Initiation Sequence Stem-Loop of Human Immunodeficiency Virus Type 1 Is Dispensable for Viral Replication in Peripheral Blood Mononuclear Cells,” J. Virol., vol. 77, no. 15, pp. 8329–8335, 2003, doi: 10.1128/jvi.77.15.8329-8335.2003.

[169] K. L. Jones, S. Sonza, and J. Mak, “Primary T-lymphocytes rescue the replication of HIV-1 DIS RNA mutants in part by facilitating reverse transcription.,” Nucleic Acids Res., vol. 36, no. 5, pp. 1578–88, 2008, doi: 10.1093/nar/gkm1149.

[170] J. C. a Cobbin, C. Ong, E. Verity, B. P. Gilbertson, S. P. Rockman, and L. E. Brown, “Influenza virus PB1 and neuraminidase gene segments can cosegregate during vaccine reassortment driven by interactions in the PB1 coding region.,” J. Virol., vol. 88, no. 16, pp. 8971–80, Aug. 2014, doi: 10.1128/JVI.01022-14.

[171] B. Essere et al., “Critical role of segment-specific packaging signals in genetic reassortment of influenza A viruses.,” Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 40, pp. E3840-8, 2013, doi: 10.1073/pnas.1308649110.

[172] M. C. White, H. Tao, J. Steel, and A. C. Lowen, “H5N8 and H7N9 packaging signals constrain HA reassortment with a seasonal H3N2 influenza A virus.,” Proc. Natl. Acad. Sci. U. S. A., vol. 0, p. 201818494, 2019, doi: 10.1073/pnas.1818494116.

[173] Q. Gao and P. Palese, “Rewiring the RNAs of influenza virus to prevent reassortment,” Proc. Natl. Acad. Sci., vol. 106, no. 37, pp. 15891–15896, 2009, doi: 10.1073/pnas.0908897106.

[174] M. I. Nelson and E. C. Holmes, “The evolution of epidemic influenza.,” Nat. Rev. Genet., vol. 8, no. 3, pp. 196–205, 2007, doi: 10.1038/nrg2053.

[175] R. Rangan et al., “De novo 3D models of SARS-CoV-2 RNA elements from consensus experimental secondary structures.,” Nucleic Acids Res., vol. 49, no. 6, pp. 3092–3108, 2021, doi: 10.1093/nar/gkab119.

[176] P. R. Bhatt et al., “Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome.,” Science, vol. 372, no. 6548, pp. 1306–1313, 2021, doi: 10.1126/science.abf3546.

[177] K. Kappel et al., “Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures.,” Nat. Methods, vol. 17, no. 7, pp. 699–707, 2020, doi: 10.1038/s41592-020-0878-9.

[178] M. Tihova et al., “Nodavirus coat protein imposes dodecahedral RNA structure independent of nucleotide sequence and length.,” J. Virol., vol. 78, no. 6, pp. 2897–905, Mar. 2004, doi: 10.1128/jvi.78.6.2897-2905.2004.

[179] E. L. Hesketh et al., “Mechanisms of assembly and genome packaging in an RNA virus revealed by high-resolution cryo-EM.,” Nat. Commun., vol. 6, p. 10113, Dec. 2015, doi: 10.1038/ncomms10113.

[180] C. Beren et al., “Genome organization and interaction with capsid protein in a multipartite RNA virus.,” Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 20, pp. 10673–10680, 2020, doi: 10.1073/pnas.1915078117.

[181] R. Chandler-Bostock et al., “Assembly of infectious enteroviruses depends on multiple, conserved genomic RNA-coat protein contacts.,” PLoS Pathog., vol. 16, no. 12, p. e1009146, 2020, doi: 10.1371/journal.ppat.1009146.

[182] J. San Emeterio and L. Pollack, “Visualizing a viral genome with contrast variation small angle X-ray scattering.,” J. Biol. Chem., vol. 295, no. 47, pp. 15923–15932, 2020, doi: 10.1074/jbc.RA120.013961.

[183] T. Lin, J. Cavarelli, and J. E. Johnson, “Evidence for assembly-dependent folding of protein and RNA in an icosahedral virus.,” Virology, vol. 314, no. 1, pp. 26–33, Sep. 2003, doi: 10.1016/s0042-6822(03)00457-4.

[184] E. Morandi et al., “Genome-scale deconvolution of RNA structure ensembles.,” Nat. Methods, vol. 18, no. 3, pp. 249–252, 2021, doi: 10.1038/s41592-021-01075-w.

[185] P. J. Tomezsko et al., “Determination of RNA structural diversity and its role in HIV-1 RNA splicing,” Nature, vol. 582, no. 7812, pp. 438–442, 2020, doi: 10.1038/s41586-020-2253-5.

[186] N. Schmidt et al., “The SARS-CoV-2 RNA-protein interactome in infected human cells.,” Nat. Microbiol., vol. 6, no. 3, pp. 339–353, 2021, doi: 10.1038/s41564-020-00846-z.

[187] C. A. Weidmann, A. M. Mustoe, P. B. Jariwala, J. M. Calabrese, and K. M. Weeks, “Analysis of RNA-protein networks with RNP-MaP defines functional hubs on RNA.,” Nat. Biotechnol., Oct. 2020, doi: 10.1038/s41587-020-0709-7.

[188] R. Gao et al., “A highly homogeneous polymer composed of tetrahedron-like monomers for high-isotropy expansion microscopy.,” Nat. Nanotechnol., vol. 16, no. 6, pp. 698–707, Jun. 2021, doi: 10.1038/s41565-021-00875-7.

[189] M. Ferrer et al., “Imaging HIV-1 RNA dimerization in cells by multicolor super-resolution and fluctuation microscopies.,” Nucleic Acids Res., p. gkw511, Jun. 2016, doi: 10.1093/nar/gkw511.

[190] M. D. Vahey and D. A. Fletcher, “Low-Fidelity Assembly of Influenza A Virus Promotes Escape from Host Cells.,” Cell, vol. 180, no. 1, p. 205, Jan. 2020, doi: 10.1016/j.cell.2019.12.028.

[191] L. Sardo et al., “Dynamics of HIV-1 RNA Near the Plasma Membrane during Virus Assembly,” J. Virol., vol. 89, no. 21, pp. 10832–10840, 2015, doi: 10.1128/jvi.01146-15.

[192] N. Jouvenet, P. D. Bieniasz, and S. M. Simon, “Imaging the biogenesis of individual HIV-1 virions in live cells.,” Nature, vol. 454, no. 7201, pp. 236–40, Jul. 2008, doi: 10.1038/nature06998.

[193] Z. Pénzes, C. Wroe, T. D. Brown, P. Britton, and D. Cavanagh, “Replication and packaging of coronavirus infectious bronchitis virus defective RNAs lacking a long open reading frame.,” J. Virol., vol. 70, no. 12, pp. 8660–8, Dec. 1996, doi: 10.1128/JVI.70.12.8660-8668.1996.

[194] Q. Li, Y. Tong, Y. Xu, J. Niu, and J. Zhong, “Genetic Analysis of Serum-Derived Defective Hepatitis C Virus Genomes Revealed Novel Viral cis Elements for Virus Replication and Assembly.,” J. Virol., vol. 92, no. 7, 2018, doi: 10.1128/JVI.02182-17.

[195] J. A. Fosmire, K. Hwang, and S. Makino, “Identification and characterization of a coronavirus packaging signal.,” J. Virol., vol. 66, no. 6, pp. 3522–30, Jun. 1992, doi: 10.1128/JVI.66.6.3522-3530.1992.

[196] R. G. van der Most, P. J. Bredenbeek, and W. J. Spaan, “A domain at the 3’ end of the polymerase gene is essential for encapsidation of coronavirus defective interfering RNAs.,” J. Virol., vol. 65, no. 6, pp. 3219–26, Jun. 1991, doi: 10.1128/JVI.65.6.3219-3226.1991.

[197] S. Makino, K. Yokomori, and M. M. Lai, “Analysis of efficiently packaged defective interfering RNAs of murine coronavirus: localization of a possible RNA-packaging signal.,” J. Virol., vol. 64, no. 12, pp. 6045–53, Dec. 1990, doi: 10.1128/JVI.64.12.6045-6053.1990.

Chapter 3 Short and long-range interactions in the HIV-1 5'UTR regulate genome dimerization and packaging

Modified from the manuscript published in Nature Structural & Molecular Biology

(DOI: 10.1038/s41594-022-00746-2)

Note: the results in 3.3.8 Dimerization regulation in HIV-1 strain, Mal are not published yet

Liqing Ye[1], Anne-Sophie Gribling-Burrer[1,4], Patrick Bohn[1,4], Anuja Kibe[1], Charlene Börtlein[1], Uddhav B Ambi[1], Shazeb Ahmad[1], Marco Olguin-Nava[1], Maureen Smith[3], Neva Caliskan[1,2], Max von Kleist[3], Redmond P. Smyth[1,2]

[1]Helmholtz Institute for RNA-based Infection Research, Helmholtz Centre for Infection Research, Würzburg, Germany

[2]Faculty of Medicine, University of Würzburg, Würzburg, Germany

[3]P5 Systems Medicine of Infectious Disease, Robert Koch-Institute, Berlin, Germany

[4]These authors contributed equally

Corresponding author: Redmond P. Smyth; redmond.smyth@helmholtz-hiri.de

Abstract

RNA dimerization is the noncovalent association of two human immunodeficiency virus-1 (HIV-1) genomes. It is a conserved step in the HIV-1 life cycle and assumed to be a prerequisite for binding to the viral structural protein Pr55Gag during genome packaging. Here, we developed functional analysis of RNA structure-sequencing (FARS-seq) to comprehensively identify sequences and structures within the HIV-1 5' untranslated region (UTR) that regulate this critical step. Using FARS-seq, we found nucleotides important for dimerization throughout the HIV-1 5' UTR and identified distinct structural conformations in monomeric and dimeric RNA. In the dimeric RNA, key functional domains, such as stem-loop 1 (SL1), polyadenylation signal (PolyA) and primer binding site (PBS), folded into independent structural motifs. In the monomeric RNA, SL1 was reconfigured into long- and short-range base pairings with polyA and PBS, respectively. We show that these interactions disrupt genome packaging, and additionally show that the PBS-SL1 interaction unexpectedly couples the PBS with dimerization and Pr55Gag binding. Altogether, our data provide insights into late stages of HIV-1 life cycle and a mechanistic explanation for the link between RNA dimerization and packaging.

3.1 Introduction

3.1.1 Overview of human immunodeficiency virus (HIV)

Human immunodeficiency virus (HIV) belongs to Retroviridae, which attacks human immune system and causes acquired immunodeficiency syndrome (AIDS)[1][2]. Since the 1980s, about 55.9-110 million people have been infected and about 27.2-47.8 million have died of HIV (https://www.who.int/data/gho/data/themes/hiv-aids). It can be divided into two major types, HIV-1 and HIV-2. HIV-2 mainly causes epidemic in west Africa, while HIV-1 is more infective and virulent, and is the cause of the majority of HIV infections globally.

HIV can infect different immune cells, such as helper T cells (specifically CD4+ T cells), dendritic cells, microglial cells, and macrophages[3]. HIV weakens the human immune system by killing immune cells through several mechanisms, like apoptosis[4] and pyroptosis[5]. Patients usually die from severe viral, bacterial or fungal infections or cancers. Even though there is no effective cure for HIV infection, it can be controlled with proper medical treatment.

HIV-1 viral particles have lipid membranes derived from the host cell harbouring the surface glycoproteins, gp120 and gp41. The matrix protein lies beneath the membrane, and the capsid core of the virus particle contains two copies of positive full length single-stranded RNA, which are tightly bound to nucleocapsid, as well as the proteins necessary for HIV-1 replication, including reverse transcriptase, protease, and integrase[6][7]. HIV-1 enters cells using CD4 as a receptor and CCR5 or CXCR4 as co-receptors. Upon binding to the host cell, the viral and cell membrane fuse, releasing the capsid into cytoplasm[8][9]. After removal of the capsid protein by uncoating[10], the genomic RNA is reverse transcribed into cDNA then viral dsDNA by carried reverse transcriptase. Next, the viral dsDNA is transported to cell nucleus, and integrated into the host cell genome by the viral integrase. The resulting provirus might be dormant and under latency, or be transcribed by host cell RNA polymerase II into full-length viral RNA. The full-length positive RNA can be either directly translated into Gag or Gag-Pol proteins, or be packaged with the Gag precursor proteins into viral particles to function as the genomic RNA (**Figure 3.1a**).

3.1.2 HIV-1 genome dimerization and packaging

Like other retrovirus members, HIV-1 packages two copies of its genome into viral particles[11][12][13]. These genomes are non-covalently associated at an RNA motif called the dimerization initiation site (DIS)[14][15]. This association, known as dimerization, impacts multiple steps of the HIV-1 life cycle. Packaging two complete RNAs as genome has the advantage for frequent template switching events during reverse transcription, which results in recombinant viruses that have distinct heredity from the two parental viruses. Besides, allowing strand transfer also act as a rescue mechanism to recover genetic information from an incompletely integrated RNA

molecule[16]. It is also linked to a structural switch that may regulate genome packaging and translation within cells.

Many studies have shown that dimerization is highly regulated by the conserved HIV-1 5'UTR[17][18][19][20][21]. The most essential region is DIS, which is a six-nucleotide long GC-rich palindromic sequence within stem-loop 1 (SL1) that initiates dimerization through an inter-molecular "kissing loop" interaction[16][22]–[24]. Although SL1 is widely considered the primary dimerization motif, numerous studies indicate that genome dimerization is also modulated by sequences outside of SL1[25]–[31]. For example, dimerization is promoted by a long-range base pairing between nucleotides overlapping the gag start codon (AUG) and the unique 5' element (U5)[31]–[33] (**Figure 3.1b**). Alternatively, it is inhibited when the region containing the AUG folds into a small hairpin, in turn freeing U5 to form a pseudoknot interaction with SL1[32][34] (**Figure 3.1b**). The U5-SL1 pseudoknot interaction was originally proposed as a liable interaction between the loop region of SL1 and U5, but a recent NMR study uncovered a more extensive base-pairing between U5 and SL1[34]–[36]. Furthermore, intrinsic transcriptional start site heterogeneity, which produces transcript variants beginning with different counts of G residues (1G, 2G or 3G), has been shown to regulate dimerization by shifting the equilibrium between mutually exclusive structures containing either an U5-AUG, or a U5-SL1 interaction[35]–[37]: 1G transcripts expose the DIS for dimerization and sequester the 5' cap, whereas 3G variants conceal the DIS whilst exposing the cap to enhance translation[36]. In addition to the U5-AUG and U5-SL1 conformations, over 20 structural models of the HIV-1 genome have been proposed, suggesting that the 5'UTR may dynamically adopt multiple conformational states[17][38]. It seems therefore likely that other structural forms of the HIV-1 genome exist to regulate genome dimerization, or other critical aspects of HIV-1 biology. Dimerization is assumed to be a prerequisite for genome packaging into virions, although the mechanistic relationship between dimerization and packaging is still under debate[19], [25], [39]–[41].

As it is discussed in Chapter 2[42], HIV-1 recognizes the genomic RNA through specific interactions between the viral Gag protein and packaging signals present at the 5' end of the genome, which was initially identified as SL3[39], [43]–[45]. More recent studies proved that SL1 is the primary Gag binding site and there was a more significant packaging deficiency when it was deleted or mutated[18], [46]–[48]. Later studies demonstrated that U5-AUG interactions also played key roles in the Gag selectively binding with genomic RNAs[31], [47], [49]–[51].

Sequences required for dimerization largely overlap with other conserved functional elements, such as those involved with genome packaging. Indeed, this genetic overlap between dimerization and packaging signals is a major reason why dimerization is considered to be a pre-requisite for packaging, even though the precise molecular mechanism underlying this phenomenon is unclear.

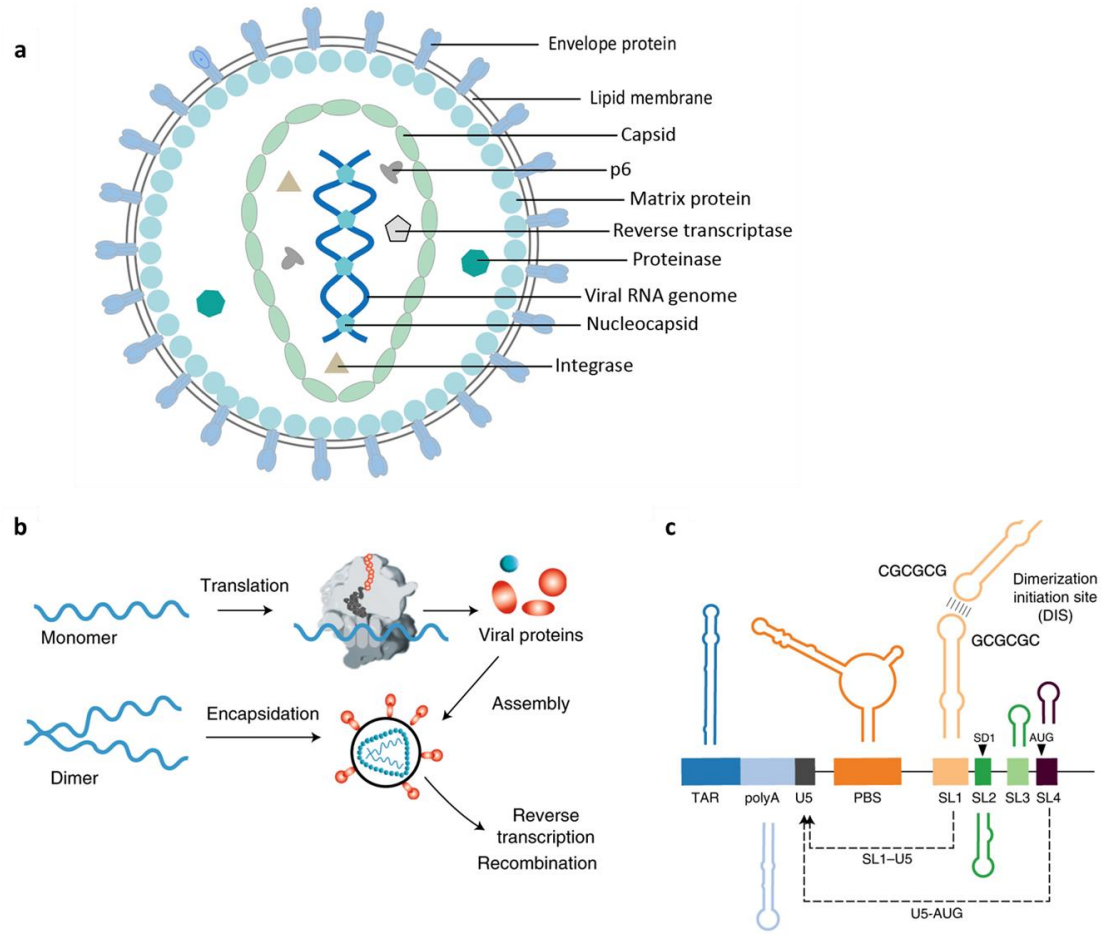


Figure 3.1| HIV-1 virion structure and genome dimerization and regulation. (a) HIV-1 virion diagram. **(b)** Dimerization is a key step in the HIV-1 life cycle. Monomeric RNA is thought to be preferentially translated, in contrast to dimeric RNA, which is a prerequisite for packaging into virions. Dimeric RNA helps maintain genome integrity through recombination. **(c)** The HIV-1 5′ UTR is composed of distinct structural domains linked to different functions in the HIV-1 life cycle. TAR stands for transcription. PolyA stands for polyadenylation that is inactive in the 5′ UTR. U5 in unique 5 region or PBS, stands for annealing of the host tRNA for initiating reverse transcription. SL1–SL3 contain the packaging signal. SL2 contains the splice donor site. Dimerization occurs through a kissing loop interaction at a sequence in SL1. LDIs/ alternative folds involving LDIs, such as between SL1–U5 and U5–AUG may regulate dimerization.

3.1.3 Methodology to study RNA structure and function

Mutational interference mapping experiment (MIME) is an unbiased, quantitative single-nucleotide resolution method to identify the RNA primary sequence and secondary structures of an RNA molecule that are crucial for its function[48][52]. The work flow of MIME includes introducing mutations into RNAs randomly, functional selection and next-generation sequencing. In the end, the sequences important for the function will be identified. Therefore, it can be applied to identify the sequences regulate HIV-1 dimerization easily *in vitro*.

As discussed in chapter 1, chemical probing is a powerful approach to determine RNA secondary structure[53]–[55]. DMS-MaPseq has the advantage that the DMS modifications can be read out as mutations under certain reverse transcription conditions [56]–[58]. Another method called M2-seq[59], which is also based on DMS probing and mutation read out, overcomes the limitation of DMS-Mapseq that predicts RNA structure based on DMS activity by introducing low frequent mutations following DMS probing and identifying the co-related mutations to remodel the RNA structure.

By combining the methods above, we developed a novel approach that we call Functional Analysis of RNA Structure (FARS-seq) to disentangle genome dimerization from other steps of the viral life-cycle and comprehensively mapped structure determinants of HIV-1 genome dimerization[60].

The workflow of FARS-seq includes: RNA mutagenesis, function selection, DMS probing, mutational profiling (reverse transcription which encodes DMS modifications as mismatches), mutation co-relation analysis and RNA structure remodelling.

3.2 Material and methods

Plasmid

NL43 sequences were obtained from pDRNL43 ΔEnv plasmid, which is containing full-length NL43 but without flanking cellular sequences[61] and contains a deletion in Env for biosafety.

Protein expression and purification

Expression, purification and characterization of NL4.3 Pr55Gag with an appended C-terminal His6-tag was performed as described by McKinstry et al[62].

Mutant library preparation

DNA templates were prepared by PCR using Taq DNA polymerase (NEB). For NL43, using RNA expression plasmid pDRNL43- ΔEnv and forward primers containing T7 RNA polymerase promoter and 3G/2G/1G at 5'end AAAgaagacTTggggTAATACGACTCACTATAGGGTCTCTCTGGTTAGACCAG / AAAgaagacTTggggTAATACGACTCACTATAGGTCTCTCTGGTTAGACCAG / AAAgaagacTTggggTAATACGACTCACTATAGTCTCTCTGGTTAGACCAG and reverse primer mGmATCTAAGTTCTTCTGATCCTGTCTG. And for Mal, using plasmid puc19_HIV-1 Mal_5'UTR and forward primers containing T7 RNA polymerase promoter and 3G/1G at 5'end TAATACGACTCACTATAGGGTCTCTCTTGTTAGACCAG / TAATACGACTCACTATAGTCTCTCTTGTTAGACCAG and reverse primer mGmATTTAATTTCTTCTGATCCTGTCTTG. PCR amplifications were performed in 1X reaction buffer, 0.2 mM dNTPs, 250 nM forward primer and reverse primer, 1 ng plasmid as template, 1.25 U Taq DNA polymerase (NEB) using the PCR cycling conditions: 98°C for 30 s, followed by 32 cycles of 98°C for 10

s, 60°C for 30s, and 68°C for 1 min. Products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer and column purified with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). The purified PCR products were used as template for error prone PCR using the Mutazyme II DNA polymerase (Agilent) and forward primer TAATACGACTCACTATA and the same reverse primers above. The PCR reaction volume was 50 µl and consisted of 2 ng of template DNA, 1X buffer, 200 µM dNTPs, 0.25 mM of each primer, 2.5 U of Mutazyme II DNA polymerase. PCR cycling conditions were 95 °C for 2 min followed by 35 cycles of 95°C for 30 s, 35-42°C for 30 s and 72°C for 1 min. Products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer. A final column purification was carried out with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) kit.

RNA preparation

Purified WT and mutated PCR products (900 ng) were used as templates for RNA *in vitro* transcription with a homemade T7 RNA polymerase. Reaction contained 1X reaction buffer (40 mM Tris pH 7.5, 18 mM $MgCl_2$, 10 mM DDT, 1 mM Spermidine), 5 mM NTPs, 40 U RNasin (Molox), 900 ng DNA template, and 0.05 U of Pyrophosphatase, (NEB) and 5 µl of homemade T7 RNA polymerase. The reaction was incubated at 37°C for 3 h, followed by DNase I treatment for 30 min at 37 °C. RNA was gel purified after electrophoresis on 1% agarose gels in 1X TAE buffer using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) with NTC buffer (Macherey-Nagel). Half of the purified RNA was capped with Vaccinia Capping System (NEB). Briefly, 10 µg of RNA was mixed with nuclease-free H_2O in a 1.5 ml microfuge tube to a final volume of 15 µl. The sample was heated at 65°C for 5 min, then placed on ice for 5 min. 2 µl of 10x capping buffer, 1 µl 10 mM GTP, 1 µl 2 mM SAM, and 1 µl of Vaccinia Capping Enzyme were added and the sample was incubated at 37°C for 30 min. Capped RNA was column purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) with NTC buffer (Macherey-Nagel).

Native agarose gel electrophoresis

RNA (600 ng) was denatured at 90°C for 2 min followed by chilling on ice for 2 min. RNA was incubated at 37°C for overnight (15-17 h) in high salt buffer (10 mM KH_2PO_4, pH 7.4, 1 mM $MgCl_2$, 122 mM KCl) or low salt buffer (10 mM NaCl, 10 mM Tris, pH 7.4). Samples were loaded with native loading dye (0.17% Bromophenol Blue and 40% (vol/vol) sucrose) on 1% agarose gel prepared with 1X Tris-Borate Magnesium (TBM) buffer (89 mM Tris base, 89 mM boric acid and 0.2 mM $MgCl_2$) and fractionated at 100 V for 85 min at room temperature. In some experiments, 12 pmol of oligos cPBS (182-199) GTCCCTGTTCGGGCGCCA and/or cPBS (199-216) TTCCCTTTCGCTTTCAAG were added to the RNA before denaturing to assess the effect of disrupting the PBS on dimerization.

Native polyacrylamide gel electrophoresis

RNA (800 ng) was denatured at 90°C for 2 min followed by chilling on ice for 2 min. RNA was then incubated at 37°C for 30 min in high salt buffer (50 mM sodium cacodylate, pH 7.5, 300 mM, KCl, and 5 mM $MgCl_2$) or low salt buffer (50 mM sodium cacodylate, pH 7.5, 40 mM KCl, and 0.1 mM $MgCl_2$). Samples were loaded with native loading dye (0.17 % Bromophenol Blue and 40 % (vol/vol) sucrose)

on 4 % acrylamide non-denaturing gel prepared with 1X Tris-Borate Magnesium (TBM) (89 mM Tris base, 89 mM boric acid and 0.1 mM $MgCl_2$) and fractionated at 150 V for 4 h at 4°C, including 2 reference samples with SYBR gold (Invitrogen), which could be visualized under blue LED light. The dimer and monomer bands in samples were cut from the gel according to the position of reference samples by scalpel.

In gel DMS probing

Each gel piece from the polyacrylamide gel was divided into 2 parts. Half was soaked in 1X TBM containing 170 mM DMS (dissolved in EtOH), incubated at 37°C for 15 min, followed by quenching with 50% (final) β-mercaptoethanol. The other half was soaked in 1X TBM (89 mM Tris base, 89 mM boric acid and 0.1 mM $MgCl_2$) containing the equivalent volume of EtOH as the DMS treated sample, and incubated at 37°C for 15 min. Gel slices were crushed into small pieces, soaked in 1X TBM (89 mM Tris base, 89 mM boric acid and 0.1 mM $MgCl_2$) buffer at 4°C overnight. RNA was extracted using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) with NTC buffer (Macherey-Nagel).

Reverse transcription

35 ng of DMS modified RNA or 25 ng of control RNA was performed with 200 U SuperScript II reverse transcriptase (Invitrogen), 0.1 µM of reverse transcription primer mGmATCTAAGTTCTTCTGATCCTGTCTG for NL43 and mGmATTTAATTTCTTCTGATCCTGTCTTG for Mal, 0.5 mM dNTPs, 50 mM Tris·HCl, pH 8.0, 75 mM KCl, 6 mM $MnCl_2$, 10 mM DTT in 20 µl reactions. The RT reaction was incubated at 42°C for 3 h.

Library preparation

For the functional probing MIME experiments, reverse transcribed cDNAs were amplified with 250 nM primers Fw- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTCTCTCTGGTTAGACC, Rv-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATGGTTGTAGCTGTCCCAG, 200 µM dNTPs, 1X Q5 reaction buffer, Q5 polymerase (NEB) using the PCR cycling conditions: 98°C for 30 s, followed by 32 cycles of 98°C for 10 s, 55°C for 30 s, and 72°C for 30 s. The PCR products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer and column purified (using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). 25 ng purified products were used in the final sequencing library preparation with Nextera DNA Flex Library Prep (Illumina) and Nextera DNA CD Indexes (96 Indexes, 96 Samples, Illumina), according to the manufacturer's instructions. For structural profiling by DMS, we performed amplicon sequencing. PCR reaction volume was 25 µl, 200 µM dNTPs, 250 nM primer pair 1 for NL43 (Fw- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGggtctctctggttagacc and Rv-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCGTACTCACCAGTCGCC) or primer pair 2 for NL43 (Fw-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcgaaagtaaagccagaggag and Rv-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCCCTGCTTGCCCATAC), or 250 nM primer pair 3 for Mal (Fw- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTCTCTCTTGTTAGACC and Rv-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGCTCTCGCCTCTTGCTG) or primer pair 4 for Mal (Fw-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGGCGCCCGAACAGGG and Rv-

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTTCCCCCTGGCCTTAACC), 1X GXL reaction buffer, 0.625 U of PrimeSTAR GXL DNA Polymerase (Takara Bio). Two PCR amplifications were performed using the PCR cycling conditions: 98°C for 30 s, followed by 34 cycles of 98°C for 10 s, 60°C for 15 s, and 68°C for 30 s. Amplified libraries were column purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). Paired end PE150 sequencing was carried out on an Illumina Novaseq instrument (Novogene).

Data analysis

Sequencing data relating to MIME functional profiling experiments was first preprocessed using automated python scripts. First, sequencing reads were quality trimmed and stripped of adaptors with CutAdapt with the parameters "--nextseq-trim 35 – max-n 0 -A CTGTCTCTTATA -a CTGTCTCTTATA". Second, reads were aligned to the HIV-1 5'UTR using Novoalign with the parameters "-o SAM -o SoftClip". Sam files were then analyzed using MIMEAnTo 73 to generate Kdimer, which is a quantitative metric relating the effect of a mutation on dimerization (derivation in supplementary note). Statistical methods used in MIMEAnTo are described in detail elsewhere[52][63].

Sequencing data relating to DMS structural probing was first preprocessed with ShapeMapper2 using parameters "--output-parsed-mutations --output-counted-mutations --render-mutation". EtOH treated and DMS treated raw sequencing reads were passed to ShapeMapper2 via the modified and unmodified parameters, respectively[64]. DMS reactivities were calculated from ShapeMapper2 mutation rates using 90% Winsoring[65]. DMS reactivities were saved as XML files for processing with rf-fold module of the RNA Framework software package[58][66]. rf-fold was used to calculate Shannon entropies and base pairing probabilities with the parameters "-ow -dp -KT -sh -g". Initial RNA structure predictions of monomer and dimer conformations, using DMS reactivities as soft constraints, were performed with rf-fold using the RNA folding algorithms in the Vienna RNA 2.0 package[67]. Refined RNA structure predictions using multidimensional probing results as additional hard constraints were performed using RNAfold of the Vienna RNA 2.0 package[67]. Cluster maps of DMS reactivities were generated using the clustermap function of the python Seaborn data visualization library using 'kendall' correlation method and 'average' cluster method (v 0.11.1). Principal Component Analysis was carried out using the PCA function of the python scikit-learn library (v 0.23.2). Variances in DMS reactivities were calculated using the var function from the python NumPy library (v 1.19.2). Pairwise comparison of DMS reactivities were carried out using a modified deltaSHAPE calculation[68]. This modified deltaSHAPE (v 1.0) analysis uses several criteria to identify statistically significant changes in reactivities. First, a z-factor test identifies nucleotides where DMS reactivities change by > 1.96 standard deviations of the DMS errors. Second, a standard score threshold of 1.5 is applied, meaning that delta reactivity values are at least 1.5 standard deviations away from the mean reactivity change. To filter these statistically significant sites for biological meaning, we next applied an absolute and a relative threshold filter. The absolute difference threshold ensures that a minimum reactivity change of 0.2 is needed for the site to be considered biologically relevant. The relative threshold filter was set so that a relative change of at least 0.75-

fold was needed to remove false positives where DMS reactivities are high in both conditions such that a large change in reactivity is unlikely to affect RNA structure. RNA structures were visualized using Visualization Applet for RNA (VARNA)[69].

RNA structural interference by multi-dimensional structural probing was carried out using the M2-seq pipeline[59]. Briefly, data was preprocessed by ShapeMapper into simple files which are string representations of mutations in each read. Simple files were converted into the rich and compact rdat format specific for RNA structure mapping experiments[70]. A two-dimensional matrix containing mutation rates at pairs of nucleotide positions was constructed. Mutation counts were subsequently normalized for total number of mutations along each row to give a true modification frequency. RNA structure signatures were further refined by calculating z-scores. A thresholding of 0 was applied to remove negative values, and a convolution filter was applied to enhance cross diagonal features. RNA helixes were finally identified in an unbiased manner by applying a filter for stems of Watson-Crick and G-U wobble base pairs of at least 3 base-pairs in length. Best stems were predicted by eliminating conflicting stems through selecting the highest scoring stem. Bootstrapping analyses were performed using the rna_structure function of the Basic Inference Engine for RNA Structure (Biers) (https://ribokit.github.io/Biers/) using the default parameters (100 bootstrapping iterations).

Microscale Thermophoresis

RNA was labeled at the 3' end using pCp-Cy5 (Cytidine-5'-phosphate-3'-(6-aminohexyl) phosphate) (Jena Biosciences) with T4 RNA ligase (NEB) overnight at 16°C, followed by RNA Clean & Concentrator Kits (ZYMO). 500 nM labeled and purified RNA was denatured at 90°C for 2 min followed by chilling on ice for 2 min. RNA was folded at 37°C for overnight (15-17 h) in high salt buffer (10 mM KH_2PO_4, pH 7.4, 1 mM $MgCl_2$, 122 mM KCl). For each binding experiment, RNA was diluted to 10 nM in high salt buffer (10 mM KH_2PO_4, pH 7.4, 1 mM $MgCl_2$, 122 mM KCl, 0.01% Triton X-100, 10mM DTT and 0.02% BSA). A series of 16 tubes with Pr55Gag dilutions were prepared in high salt buffer, producing Pr55Gag ligand concentrations ranging from 30 pM to 1 µM. For measurements, each ligand dilution was mixed with one volume of labeled RNA, which led to a final concentration of 5 nM labeled RNA. The reaction was mixed by pipetting, incubated for 30 min at 37°C, followed by 30 min on ice. Samples were then centrifuged at 10,000 x g for 5 min. Capillary forces were used to load the samples into Monolith NT.115 Premium Capillaries (NanoTemper Technologies). Measurements were performed using a Monolith Pico instrument (NanoTemper Technologies) at an ambient temperature of 25°C. Instrument parameters were adjusted to 5% LED power, medium MST power, and MST on-time of 1.5 seconds. An initial fluorescence scan was performed across the capillaries to determine the sample quality and afterward, 16 subsequent thermophoresis measurements were performed. Data of three independently pipetted measurements were analyzed for the ΔFnorm values, and binding affinities were determined by the MO. Affinity Analysis software (v 2.3 NanoTemper Technologies). Graphs were plotted using GraphPad Prism 8.4.3 software.

In solution DMS-MaPseq (*in vitro*)

RNA (300 nM) was denatured at 90°C for 2 min followed by chilling on ice for 2 min. Next, RNA was refolded at 37°C for overnight (15-17 h) in high salt buffer (10 mM KH_2PO_4, pH 7.4, 1 mM $MgCl_2$, 122 mM KCl). DMS was added to the RNA solution to final concentration 170 mM, incubated at 37 °C for 6 min, followed by quenching with β-mercaptoethanol and purification with ethanol precipitation. The purified DMS probed RNAs followed the same reverse transcription and library preparation process as the in gel DMS probed RNA samples.

In solution DMS-MaPseq (in cells)

24 h prior transfection, 107 HEK293T cells were plated in 10 mL DMEM media containing 10% FBS. 4 µg of plasmids expressing HIV-1 WT or HIV-1 mutants were mixed with 48 µl Polyethylenimine (1mg/mL, Max 40k, Polysciences) and 500 µl DMEM and incubated for 10 min at room temperature before being added dropwise on the cells. 24hr post-transfection the cells were probed by replacing the media with 3 mL DMEM containing 170mM DMS and incubated at 37°C for 6 min. Cells were then washed with 5 mL PBS buffer containing 140 mM β-mercaptoethanol to quench the DMS. 1 mL TRI-reagent (Sigma-Aldrich) is directly added on the cells to extract RNA according to the manufacturer's instructions. DNA contaminants were removed by TurboDNase (Invitrogen) treatment for 30 min at 37°C. RNA was purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) with NTC buffer (Macherey-Nagel) and eluted in 20 µl of 5 mM Tris-HCl pH 8.0. 7 µl of the purified DMS probed RNAs were used for the reverse transcription following the same reverse transcription and library preparation process as for the in gel DMS probed RNA samples.

Competition Assay

24 h prior transfection, $7*10^5$ HEK293T cells were plated in 2 mL DMEM media containing 10% FBS. For the co-transfection experiments, equal amount of plasmids expressing WT or mutants (600 ng total) were mixed with 7.2 µl Polyethylenimine (1mg/mL, Max 40k, Polysciences) and 100 µl DMEM and incubated for 10 min at room temperature before being added dropwise on the cells. Cells and viral supernatant were collected at 24 h post-transfection. Cells were washed with 2 mL PBS buffer and RNA was extracted with 1000 µl TRI-reagent (Sigma-Aldrich) according to the manufacturer's instructions. Viral supernatant was first clarified for 2 min at 17000 x g, followed by a filtration step through 0.45-micron filter. The filtrate was then transferred into a new tube and the virus was pelleted for 2 h at 17000 x g. The viral pellet was extracted with 500 µl TRI-reagent (Sigma-Aldrich) and DNA contaminants were removed by TurboDNase (Invitrogen) treatment for 30 min at 37°C. RNA was purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) with NTC buffer (Macherey-Nagel) and eluted in 20 µl of 5 mM Tris-HCl pH 8.0. 10 µl purified RNA was heat denatured at 65°C for 5 min together with 0.67 µM reverse primer (GATGGTTGTAGCTGTCCCAGTATTTGCC) and 1.67 mM dNTPs in 15 µl total volume, then chilled on ice for 2 min. RNA was then reverse transcribed by adding 1X SSIV buffer, 5 mM DTT, 20U RNasin, 100U SSIV in 25 µl total volume and incubating at 52°C for 1 h. cDNAs were amplified with 250 nM primers

Fw- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGggtctctctggttagacc, Rv- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCGTACTCACCAGTCGCC, 200 µM dNTPs, 1X Q5 reaction buffer, and 0.02 U/uL of Q5 polymerase (NEB) using the PCR cycling conditions: 98 °C for 1 min, followed by 22 cycles of 98 °C for 10 s, 55 °C for 20 s, and 72 °C for 30 s. The PCR products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer and column purified (using the NucleoSpin Gel and PCR Clean-up kit, Macherey-Nagel). 40 ng purified products were used in the final indexing PCR using 2.5 µl of Nextera DNA CD Indexes (96 Indexes, 96 Samples, Illumina) in a 14 uL reaction (200 µM dNTPs, 1X Q5 reaction buffer, and 0.02 U/µl of Q5 polymerase (NEB)). The PCR cycling conditions were 98 °C for 2 min, followed by 5 cycles of 98 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min. Paired end PE150 sequencing was carried out on an Miniseq instrument (Illumina) according to the manufacturer's instructions.

3.3 Results

3.3.1 Functional and structural analysis of RNA dimerization

HIV-1 genome dimerization largely depends on the stem of SL1 and its GC-rich palindromic loop sequence. Nevertheless, evidence suggests that RNA sequences and structures outside of SL1 also play a role[17], [28], [36], [39], [71]–[73]. We therefore devised a strategy to exhaustively survey the 5'UTR for nucleotides influencing dimerization whilst at the same time generating information about RNA structure. We call this approach the Functional Analysis of RNA Structure (FARS-seq) (**Figure 3.2**). Fundamentally, FARS-seq uses mutational interference to generate complete, unbiased, quantitative profiles of RNA function at single nucleotide resolution[48], [52], (**Figure 3.2b**). These functional profiles are generated by physical separation of mutant RNA populations according to functionality followed by next generation sequencing and the analysis of mutation frequencies in the 'functional' and 'non-functional' populations. Simultaneously, structural profiles are obtained by treating the fractions with dimethyl sulfate (DMS), which is a chemical widely used for probing RNA structure[74], [75]. DMS reacts with unpaired adenosine and cytosine bases to form adducts that can be read out as mutations on next generation sequencing machines[53], [56], [65]. Normally, DMS only provides information on whether a nucleotide is base paired, and not the identity of the base paring partner. However, when DMS modification is performed on mutational libraries it enables the direct detection of RNA stems[59], [76] (**Figure 3.2c-d**). That is, when a mutation in the library occurs within a stem, it creates an unpaired nucleotide at the position facing the mutation. This newly unpaired residue becomes more accessible for DMS modification leading to correlated mutations in the sequencing data. Thus, FARS-seq combines two different mutational read outs to experimentally couple RNA structural and functional information.

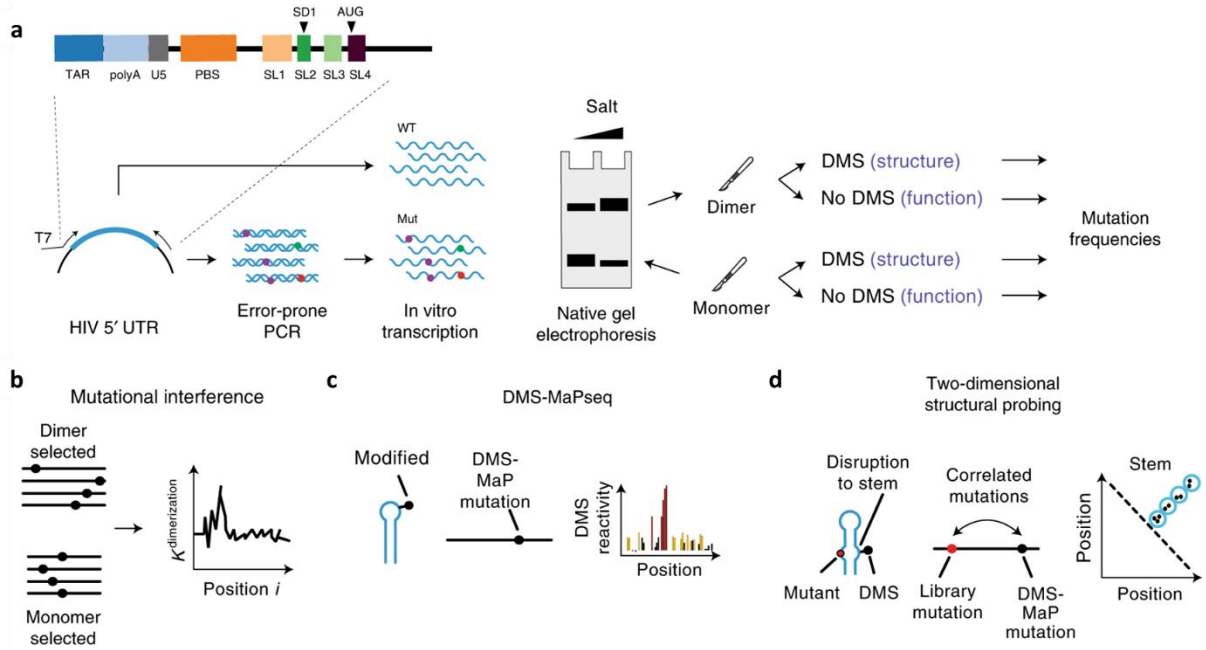


Figure 3.2| Functional and structural analysis of RNA dimerization. (a) FARS-seq. Mutant RNA sequences are generated by mutagenic PCR and *in vitro* transcription. Mutant populations are physically separated into monomer and dimer fractions and probed with DMS or left untreated. Mutation frequencies are analyzed by next generation sequencing. **(b)** Functional profiles are obtained by mutational interference. K^{dimer} is a quantitative measure of dimerization based on the ratio of mutations in the dimer selected versus monomer selected population, corrected for mutations introduced during the library preparation and sequencing. **(c)** Structural profiles are obtained by DMS that specifically reacts with unpaired A and C residues. DMS-MaPseq measures DMS reactivities as mutation rates in DMS treated versus untreated controls. **(d)** Two-dimensional analysis identifies RNA stems through correlations between stem-disrupting mutations and mutations induced by DMS.

To physically separate mutants according to their effects on dimerization, we took advantage of the observation that RNA transcripts containing dimerization signals spontaneously associate *in vitro*, producing a dimeric RNA species that can be physically separated from the monomeric species on native agarose gels (**Figure 3.3a**). Similar gel-based assays have been instrumental in the discovery of dimerization motifs in HIV-1[26], [27], [77] and other viruses[78]–[80]. This setup also disentangles the effect of RNA structure on dimerization from other factors, such as the binding of protein or other co-factors. To assess the effect of transcription start site heterogeneity on the dimerization properties of the HIV-1 genome we tested three transcript variants beginning with 1G, 2G or 3G[35], [81] (**Figure 3.3a**). For each of these transcript variants, we also tested whether capping affected dimerization and assessed their dimerization properties under low salt and high salt buffers favoring monomerization and dimerization, respectively (**Figure 3.3a**). After physical separation on a native gel, bands corresponding to monomeric and dimeric RNA populations were excised and either left untreated or soaked in DMS. For the DMS sample (and its control), RNA was reverse transcribed in

the presence of Mn^{2+} to allow mutagenic bypass of the modified nucleotides by the reverse transcriptase[56][64]. In the absence of DMS, mutation frequencies in the mutated and non-mutated control library were $5.4x10^{-3}$ and $3.7x10^{-4}$, respectively, and the mutational interference libraries with a signal to noise $D_m(i) > 2$ (**Figure 3.3b;** details of signal to noise in https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM1_ESM.pdf). In the DMS treated samples, we saw an additional increase in mutation frequencies at the expected A and C residues indicating a successful modification of RNA (3.4- and 7.8-fold increase at C and A, respectively, **Figure 3.3c**).

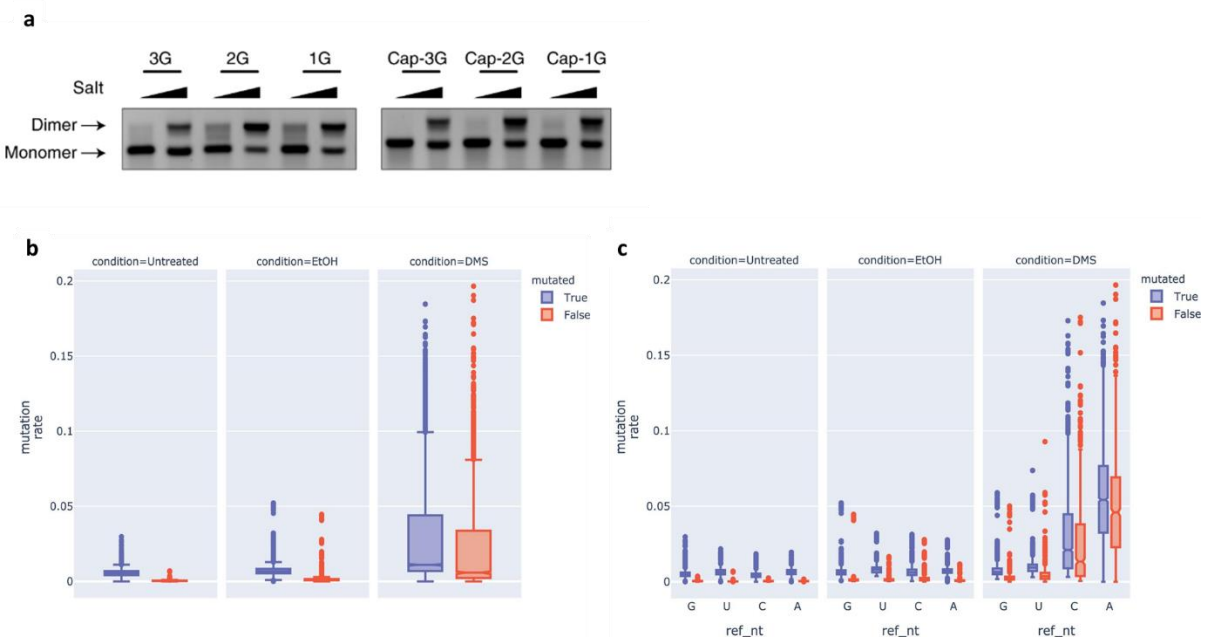


Figure 3.3| Functional and structural analysis of RNA dimerization. (a) 1G, 2G and 3G capped and uncapped transcript variants migrate as distinct monomer and dimer bands on native agarose gels in both low and high salt buffers. Experiments were performed four times and representative data shown. **(b)** Global mutation rates for mutated (blue) and unmutated (red) samples that were untreated (left panel), ethanol treated (middle panel) and DMS treated samples (right panel). Mutation rates are higher in mutated compared to unmutated samples. Untreated samples, and samples treated as DMS control (EtOH) have similar mutation rates. DMS treated samples show a greatly increase mutation rate in both mutated and unmutated samples compared to the controls. **(c)** Nucleotide specific mutation rates (A, C, G, U) for mutated (blue) and unmutated (red) samples that were untreated (left panel), ethanol treated (middle panel) and DMS treated samples (right panel). Mutation frequencies in the mutated samples are consistently higher at all nucleotides in the mutated compared to unmutated samples. In the DMS treated samples, mutations are greatly enriched at C and A residues, as expected by the selectivity of the DMS chemical. Box plots show quartile 1 (Q1) to quartile 3 (Q3). The second quartile (Q2) is marked by a line inside the box. Whiskers correspond to the box' edges + /− 1.5 times the interquartile range (IQR: Q3-Q1). Outliers are shown as points. Data are pooled from two independent experiments, each consisting of 32 independent samples.

3.3.2 RNA dimerization is regulated by the HIV-1 5'UTR

We first asked which regions of the RNA were required for dimerization using mutational interference mapping (MIME) to calculate K^{dimer} values for each nucleotide position. This metric is related to the ratio of mutation frequencies in the monomer vs dimer RNA. For computational analysis however, these ratios are corrected for errors introduced during library preparation and sequencing (mechanistic derivation in https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM1_ESM.pdf). Thus, K^{dimer} is a quantitative measure of the relative effect of each mutation on dimerization. Across all samples and conditions, median $\log_2(K^{dimer})$ values were heavily skewed towards positive values indicating that most mutations inhibited, rather than enhanced, dimerization indicating that the HIV-1 genome is highly optimized to dimerize as a key part of its life cycle **(Figure 3.4a).** By segregating K^{dimer} values by structural domain we found that most dimerization inhibiting mutations mapped to SL1 **(Figure 3.4a).** Although less prominent than SL1, many other domains exhibited skewed distributions. Mutations to SL3, SL4, and polyA were biased towards inhibiting dimerization whereas mutations to TAR and SL2 preferentially enhanced dimerization. In contrast, mutations to the inter-domain regions were largely neutral with a narrow distribution centered around zero **(Figure 3.4a)**.

We next plotted median $\log_2(K^{dimer})$ values at each nucleotide position for capped and uncapped transcript variants measured under the two buffer conditions (**Figure 3.4b-e**). All conditions exhibited a very large peak that localized to SL1, as well as a smaller double peak mapping to SL3 (**Figure 3.4b-e**). In high salt buffer most mutations inhibited dimerization, whereas under low salt conditions it was possible to distinguish additional dimerization enhancing or inhibiting regions (**Figure 3.4b-e**). Notably, sequences surrounding the AUG start codon and mapping to U5 were both required for dimerization in low salt buffer, suggestive of a functionally important U5-AUG interaction (**Figure 3.4c, e).** A double peak also emerged within the region 122-141 in low salt buffer (**Figure 3.4c, e)**. This region contains the primer activation sequence (PAS) which hints that structural changes in the PBS domain may regulate RNA dimerization[82][83]. Conversely, we found regions within TAR, polyA, PBS and SL2 that enhanced dimerization upon mutation (**Figure 3.4c, e)**. The strongest of these regions mapped to the 3' end of PBS and SL2. Taken together, these data reinforce the key importance of SL1 for genome dimerization, but also reveal sequences outside of SL1 participate in the dimerization process.

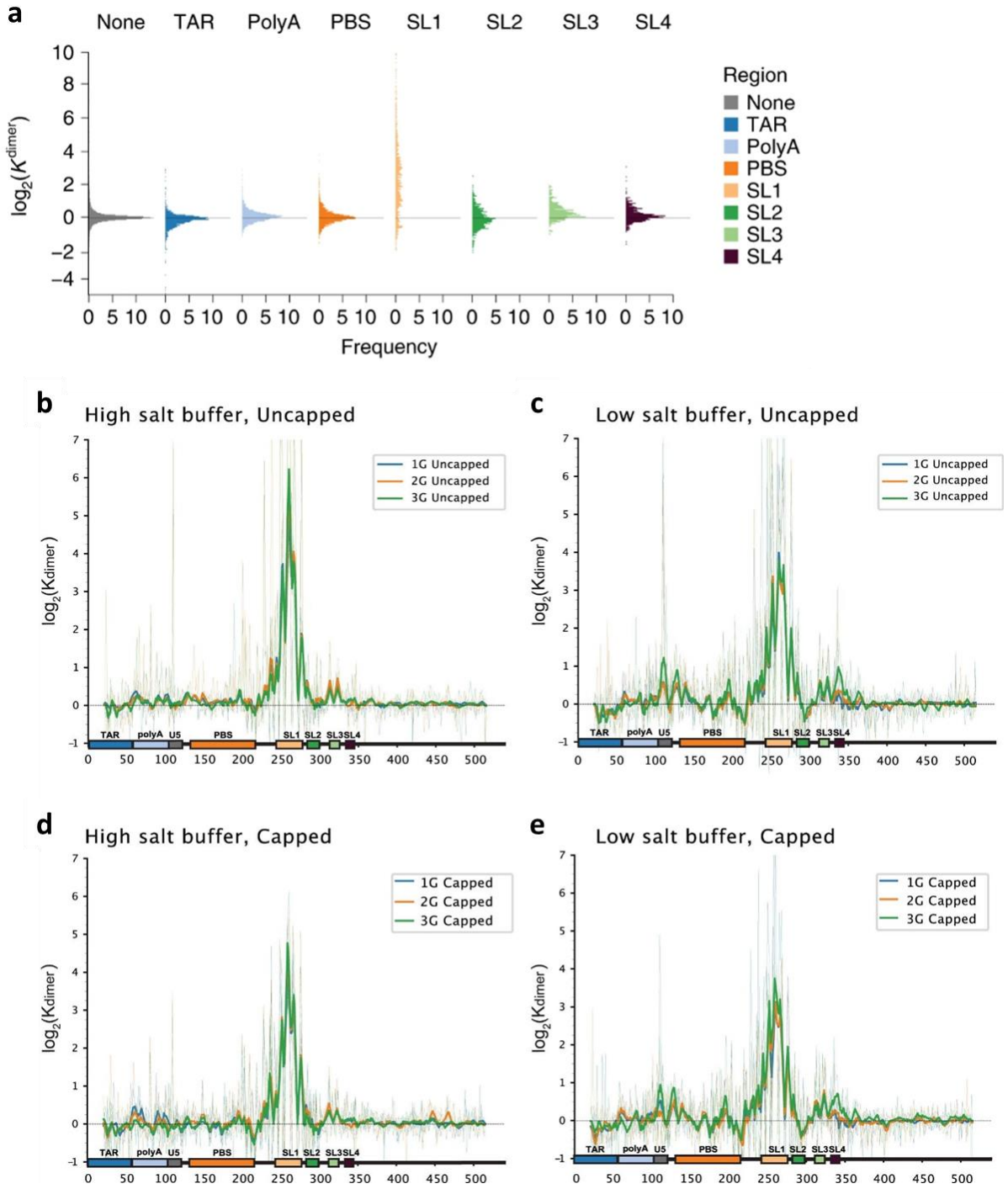


Figure 3.4| Functional profiling of sequences involved in dimerization. Functional profiling of sequences involved in dimerization by analysed by mutational interference. k^{dimer} is a relative measure of the effects of a mutation on dimerization. **(a)**The log2(K^{dimer}) values binned according to functional domain in the 5' UTR: TAR, U5, PBS, SL1, SL2, SL3 and SL4. None refers to nucleotide positions that do not fall into any structural domain. **(b-e)** Median $log_2(k^{dimer})$ values for each genome position for all three uncapped transcript variants in high and low salt buffers. Thin lines are unsmoothed data, whereas thick lines are smoothed with a window size of 5 nt.

$log_2(k^{dimer})$ values for **(b)** high salt uncapped transcripts **(c)** low salt uncapped transcripts **(d)** high salt capped transcripts **(e)** low salt capped transcripts.

3.3.3 1G and 3G RNAs have different dimerization properties

Because the HIV-1 transcription start site has been reported to alter the structure of the HIV-1 5'UTR, we next tested which RNA sequences were important for dimerization within the 1G, 2G, and 3G uncapped variants (**Figure 3.5a-d**). We did this by plotting the absolute difference between the median $log_2(K^{dimer})$ values of each variant to the mean values of the three transcripts. In high salt buffer, most positions were unchanged in the 1G, 2G and 3G variants (less than D0.25 $log_2(K^{dimer})$ variant – mean) (**Figure 3.5a, c**). The only exception was nucleotides mapping to the SL1, which were functionally more important in the 3G variant, and less important in the 1G variant. Upon performing a similar analysis for the low salt condition, distinct functional profiles for the 1G and 3G transcript variants emerged, with divergence across regions compared to the mean of the three transcripts (**Figure 3.5b, d**). The 3G variant had increased dependence on a region spanning the U5 and PAS (nts 105-117 and nts 125-131) and sequences surrounding the AUG start site (nts 335-344). Increased dependencies of smaller magnitudes were also observed in the tRNA primer binding site (nts 182-200), the anti-PAS (nts 217-223), regions flanking SL1 such as the CU rich motif (nts 228-247), a region in SL2 (nts 299-300), and a G rich region downstream of the AUG start codon (nts 360-366). We note that the regions in TAR, PBS, and SL2 that enhanced dimerization upon mutation in low salt conditions behaved identically in 1G, 2G, and 3G variants, meaning that they impact dimerization in a way that is unrelated to transcription start site selection. We also remarked that the 1G and 2G transcripts variants behaved similar in both buffer conditions with a reduced dependency on regions external to SL1 for dimerization (**Figure 3.5**). Our interpretation is that the 1G and 2G transcripts readily fold into a dimer promoting conformation, whereas the 3G variant has a reduced capacity to dimerize. Capped and uncapped transcripts had near identical functional profiles (**Figure 3.4b-e**). The only region that differed in capped and uncapped transcripts mapped to polyA, providing indirect evidence of a functional interaction between the 5' cap structure and polyA (**Figure 3.5**).

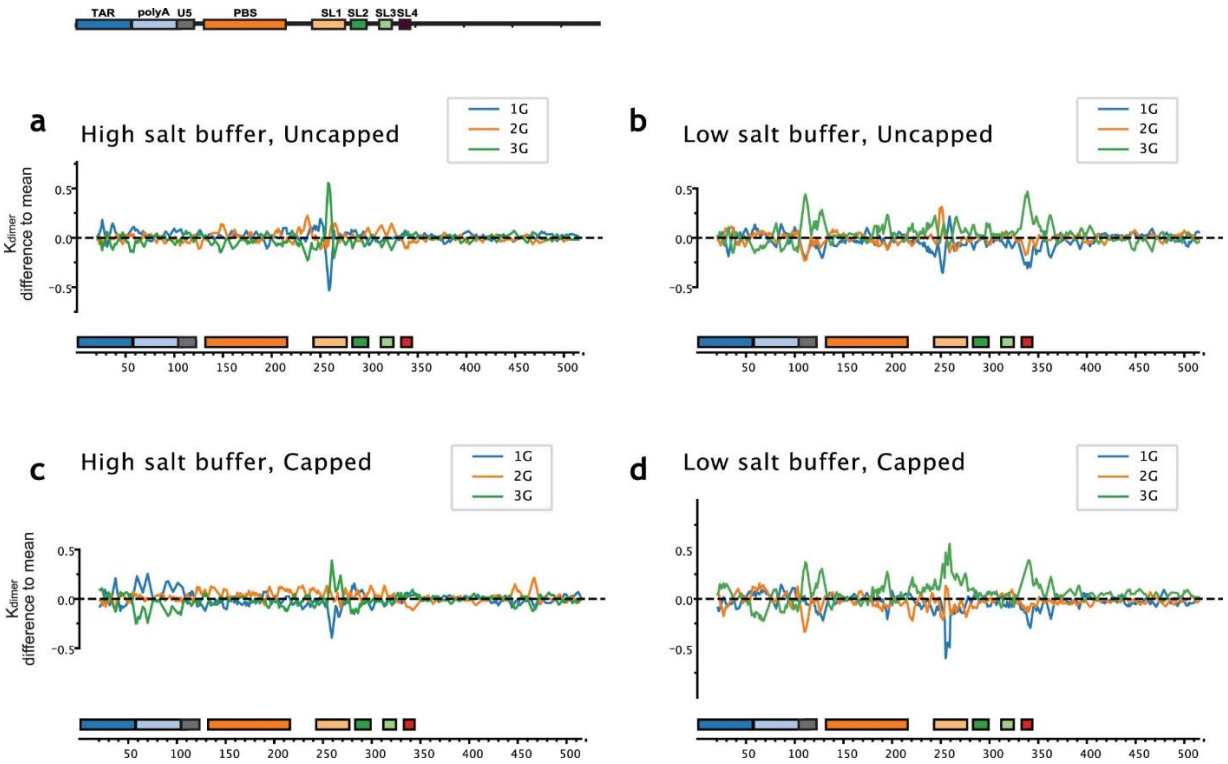


Figure 3.5| Relative dimerization properties of 1G, 2G, 3G transcripts. log2(k^{dimer}) values of the 1G, 2G, and 3G transcript variants compared to the mean of the 1G, 2G and 3G transcripts for **(a)** high salt uncapped transcripts **(b)** low salt uncapped transcripts **(c)** high salt capped transcripts (d) low salt capped transcripts. In all conditions, regions within SL1 are more important for dimerization in 3G compared to 1G samples. In low salt conditions, the 3G variant had increased dependence on regions outside of SL1. Capped and uncapped RNAs show very similar profiles, with the exception of a region in polyA in high salt buffer, which was more important for dimerization in the 1G sample compared to 3G.

3.3.4 Distinct structural signals in monomeric and dimeric RNA

So far, the analysis of the functional profiles demonstrate that sequences involved in genome dimerization map to distinct regions of the HIV-1 5'UTR. These sequences may fold into RNA structures that are necessary for genome dimerization itself, or indirectly regulate genome dimerization by altering folding pathways. We therefore next determined RNA structural motifs present in monomers and dimers by analyzing the DMS reactivities of the FARS-seq data.

As before, we analyzed capped and uncapped 1G, 2G, 3G transcript variants in both monomer and dimer buffers. Correlations between DMS reactivities at each position amongst all conditions were very high (**Figure 3.6a**; Kendall rank correlation coefficients, mean 0.84, min 0.70, max 1.0) suggesting that significant portion of the 5'UTR was folded into a similar conformation under all conditions. Nevertheless, hierarchical clustering of the DMS reactivities revealed a clear structural

distinction between monomer and dimer, as well as between the 1G/2G and 3G transcript variants (**Figure 3.6a**). In contrast to the functional profiling, where buffer conditions had a very large effect on the functional profiles, structural information obtained under both conditions were highly correlated (correlation coefficients; low salt 0.84, high salt 0.85), as were uncapped and capped RNAs (correlation coefficients; capped 0.85, uncapped 0.84). The first branchpoint separated 1G/2G dimer structures from the 1G/2G monomer and 3G structures. Subsequent branching grouped 1G/2G monomer structures away from the 3G structures. Finally, 3G structures separated into monomer and dimer subclusters. These four structural groupings were also supported by principal component analysis (PCA) of DMS reactivities, which separated monomer from dimer, and 3G variants from 1G/2G variants (**Figure 3.6b**). Guided by the PCA and hierarchical clustering, we pooled DMS reactivity data into 4 structural groups: 3G dimer, 3G monomer, 1G/2G monomer, 1G/2G dimer. Interestingly, across all samples, variance in DMS reactivities localized mainly to polyA and SL1 (**Figure 3.6c**). To further explore this, we used a statistical approach to compare DMS reactivities in the 1G/2G dimer cluster with the 3G monomer cluster as these were the most structurally divergent samples (correlation coefficient 0.740) (**Figure 3.6d**). Between these clusters, we found statistically significant changes in reactivity that again remained localized to polyA and SL1 (**Figure 3.6d**).

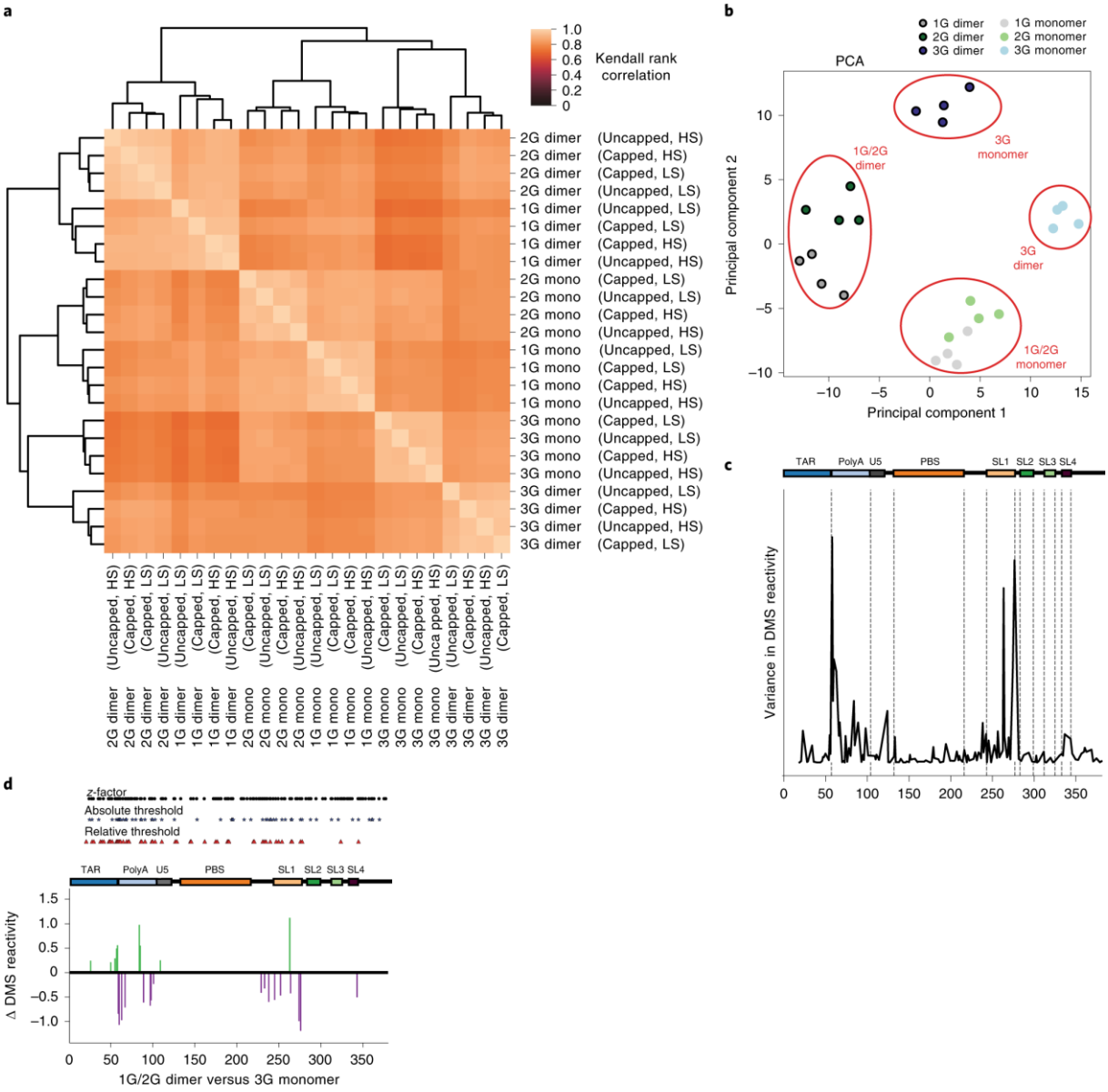


Figure 3.6| Structural profiling identifies distinct structural conformations of the HIV-1 5′ UTR. (a) Clustering of Kendal rank correlations of DMS reactivities across all positions reveals structural relationships between monomer and dimer isolated populations from uncapped and capped transcript variants in high and low salt buffer. Relationships between sample DMS reactivities was determined by hierarchical clustering using the 'average' linking method. **(b)** PCA of DMS reactivities identifies structural four structural classes of the HIV-1 5′ UTR. **(c)** Variance in DMS reactivities across genome positions from all samples is enriched at the SL1 and TAR/polyA boundary. **(d)** Statistical analysis of DMS reactivities in 1G/2G and 3G structural classes finds that significant differences in reactivities are mainly localized to polyA and SL1. A z-factor test identifies nucleotides where DMS reactivities change by >1.96 standard deviations of the DMS errors. An absolute difference threshold ensures that a minimum reactivity change of 0.2 is needed for the site to be considered biologically relevant. The relative threshold of 0.75-fold is used to remove false positives where DMS reactivities are high in both conditions such that a large change in reactivity is unlikely to affect RNA structure.

To obtain information on RNA secondary structure differences between these structural classes we used pooled DMS reactivities as soft constraints to guide *in silico* RNA folding[58][67] (**Figure 3.7, Figure 3.8**). For the 1G/2G dimer class we obtained an RNA structure that closely resembled the 'canonical' HIV-1 5'UTR (**Figure 3.7** and **Figure 3.10**). This structure contains the TAR, PolyA, PBS, SL1 and SL3 stem loops, as well as the AUG-U5 interaction. The basal portion of SL1 folded into an extended form containing unpaired purines that are important for genome packaging[47][84]. SL2, which can fold into alternative stem loop structures, folded as an imperfect stem-loop that exposes part of the U1snRNA binding site within the loop, and SL3 folded into its canonical short stem loop structure. We then assessed the robustness of this prediction by computing Shannon entropies of base pairing probabilities at each position in the 5'UTR (**Figure 3.7b, Figure 3.8**). Low entropy values throughout the 5'UTR indicated high confidence in the prediction and a well-ordered structure with only some ambiguity in the base pairing at the basal portion of SL1. This was confirmed by dot plots of base pairing probabilities and a bootstrapping analysis showing high confidence stem loop structures for the TAR, PolyA, PBS, SL1 and SL3 stem loops, as well as the AUG/U5 interaction (**Figure 3.7c** and **Figure 3.8**).

We next analyzed the structure of the 3G monomer sample, finding that it was dramatically reorganized (**Figure 3.7d**). The most striking changes were seen in the polyA, AUG-U5, and SL1. PolyA and SL1 no longer folded into their canonical stem-loops. Instead, these stem loops were reorganized into a long-distance interaction, with the GCGCGC palindromic loop of SL1 base pairing with the apical portion of the polyA stem. The AUG-U5 interaction was also no longer present; U5 now base paired with the 5' stem of SL1, and the AUG containing region fold into a stem loop structure also referred to as SL4. Finally, we observed a new long-distance interaction between polyA and a region within the Gag coding sequence (nts 358-367). The SL1-PolyA reorganization was well supported by the DMS reactivity changes (**Figure 3.7** and **Figure 3.10**). In particular, the unpaired adenosine 263A in the SL1 loop, which was highly reactive in the dimer structure, became unreactive in the monomer due to base pairing with U87. Similarly, nucleotides C84 and C85 in polyA, which were reactive in the dimer structure, became unreactive in the monomer due to base pairing with 265G and 266G in the SL1 stem. Finally, A89 in the stem of polyA, which was unreactive in the dimer structure, became unpaired in monomer structure and reactive to DMS. Shannon entropies, base paring and bootstrapping probabilities at the predicted PolyA-SL1 interaction indicated some uncertainty in the prediction, especially within U5 and the 5' portion of SL1 (**Figure 3.7f** and **Figure 3.8**). Remarkably, despite the reorganization of PolyA and SL1, a large proportion of the 5'UTR folded identically in 1G/2G dimer and 3G monomeric populations, with PBS, SL2 and SL3 unchanged. Interestingly, TAR was present in all predictions, but in the 3G monomer the first nucleotides in the base of TAR became single stranded, and potentially more available for the translation machinery.

The 3G dimer and 1G/2G monomer populations folded into the population folded into the canonical 5'UTR structure and the alternative polyA-SL1 containing structure, respectively (**Figure 3.8**). However, these two structural classes showed increased Shannon entropies in polyA, U5, SL1 and the Gag coding sequence when compared to the 1/2G dimer and 3G monomer structures. Thus, 3G dimer and 1/2G monomer populations are structurally less uniform, even though we selected for

pure dimer and monomer structures in the native gels. The most likely explanation is that these structures partially return to equilibrium after isolation, probably during the probing reaction at 37°C.

Altogether, these data support a novel structural rearrangement of the HIV 5'UTR leading to extensive base pairing between SL1 and the polyA-U5 region. This monomeric rearrangement appears to be favored in the 3G populations, whereas the 1G/2G population tend towards the dimer structure.

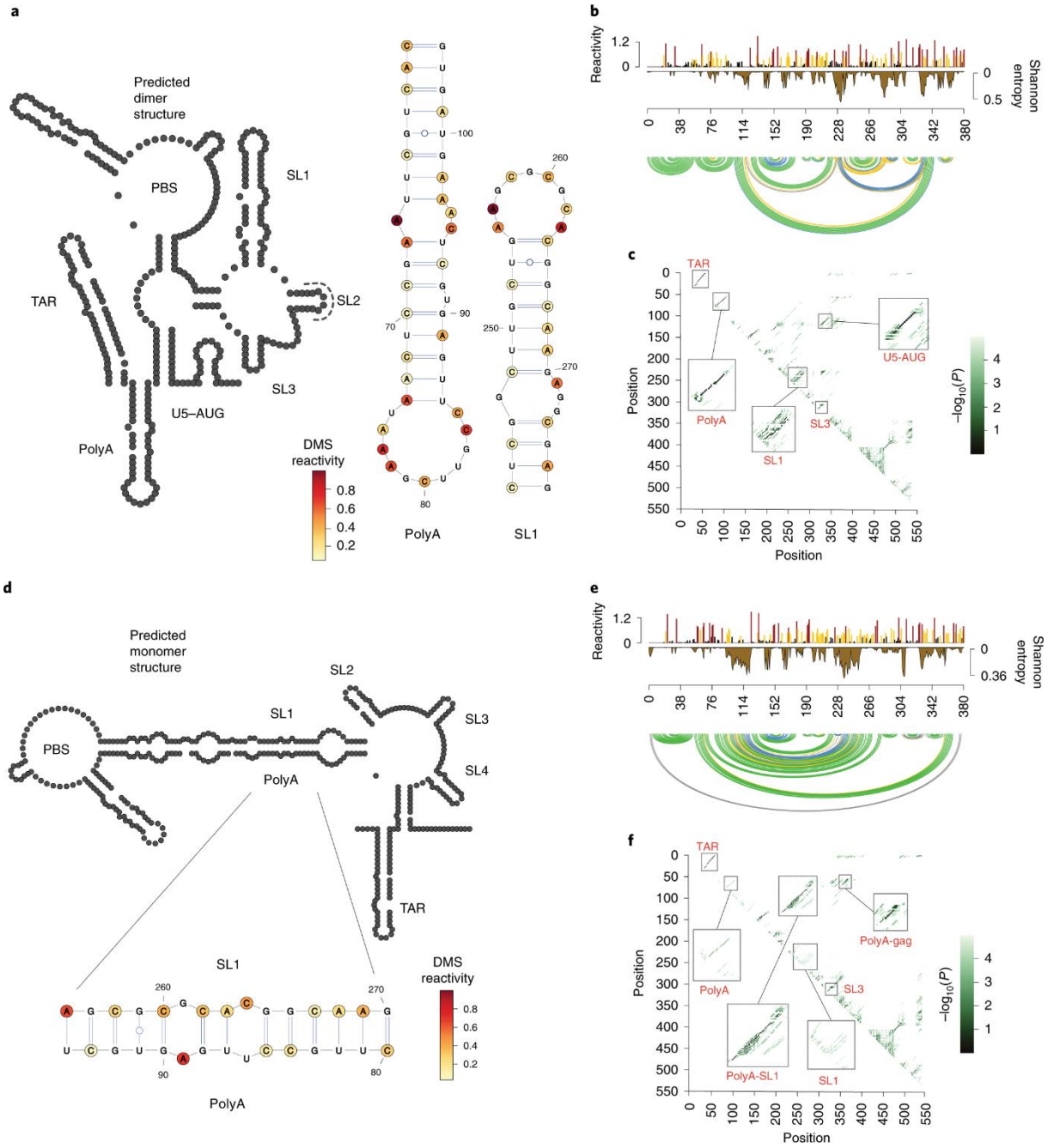


Figure 3.7| Secondary structure model for 1G/2G dimer and 3G monomer populations. (a, d) Secondary structure model of dimer **(a)** and monomer class **(d)**. Models were obtained using DMS reactivities as soft constraints for in silico folding in the Vienna RNA structure package. For the dimer structure, the U1sRNA binding site within SL2 is shown. Structures of polyA and SL1 stem loops and polyA–SL1 interaction are shown. DMS reactivities from dimer samples were mapped to A and C residues of the polyA and SL1 stem-loop structures. DMS reactivities from the monomer samples were mapped to A and C residues of the polyA–SL1 interaction. Red signifies highly reactive positions that are unpaired. Pale yellow signifies unreactive positions that are base paired. b,e, DMS reactivities and Shannon entropies for the 1G/2G dimer (b) and 3G monomer class **(e)**. Arc plots show base-pairing probabilities (green 70–100%; blue 40–70%; yellow 10–40%; gray 5–10%). Gray bar in e signifies the polyA–SL1 interaction. **(c,f)** Dot plots of RNA base-pairing probabilities for the 1G/2G dimer **(c)** and 3G monomer class **(f)**, reveal alternative folding possibilities. RNA stems are shown along the diagonals.

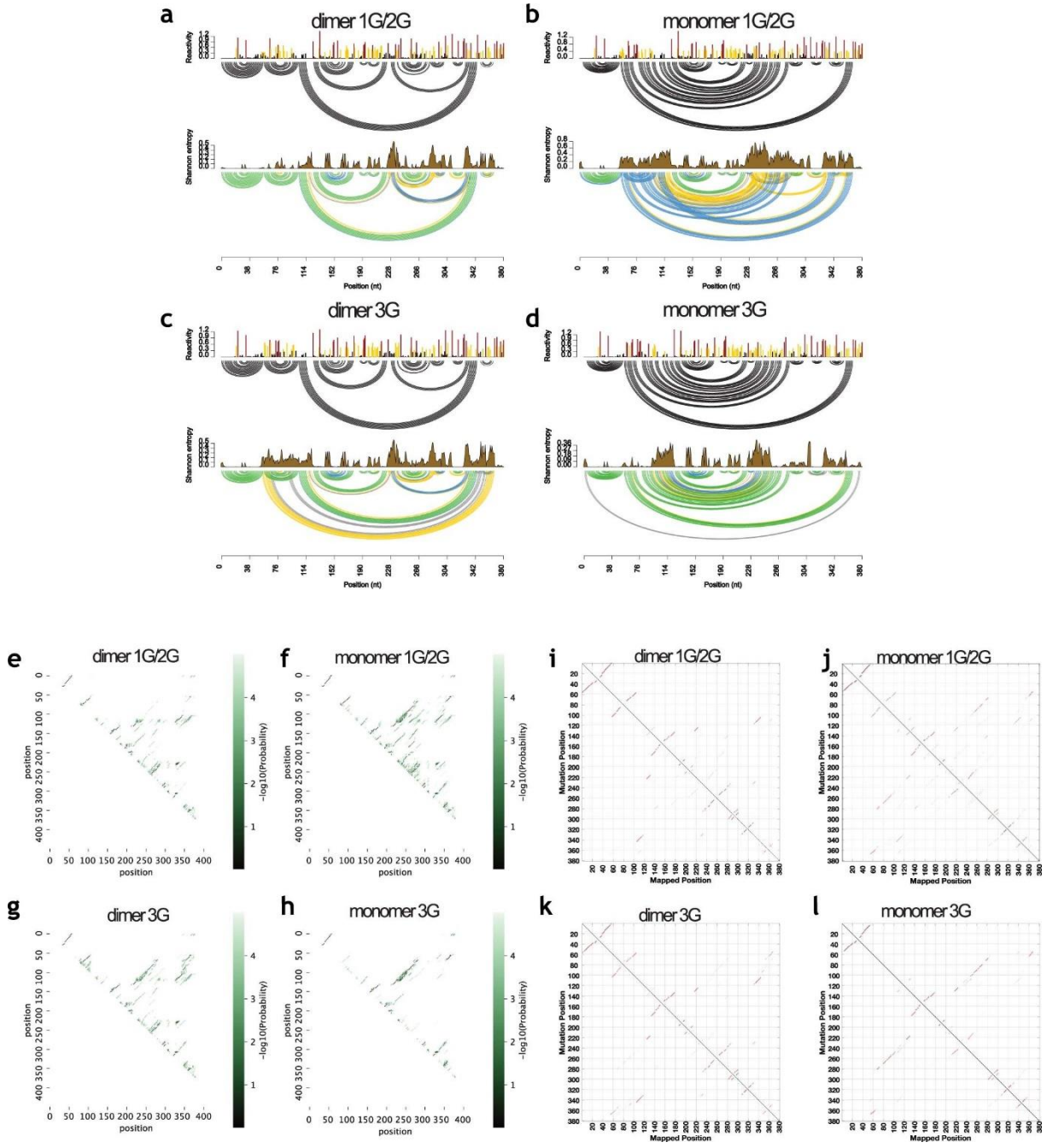


Figure 3.8| DMS reactivities and Shannon entropies. DMS reactivities and Shannon entropies for the **(a)** 1G/2G dimer class **(b)** 1G/2G monomer class **(c)** 3G dimer class and **(d)** 3G monomer class. Arc plots show base pairing probabilities (green = 70-100%; blue=40-70%; yellow=10-40%; gray=5-10%). **(e-h)** Dot plots of RNA base pairing probabilities reveal alternative folding possibilities for the **(e)** 1G/2G dimer class **(f)** 1G/2G monomer class **(g)** 3G dimer class and **(h)** 3G monomer class. RNA stems are shown along the diagonals. **(i-l)** Bootstrapping analysis of the predicted dimer and monomer structure. Predicted structure is shown in red. The bootstrap support is shown in greyscale, with darker greys signifying better bootstrap support for the **(i)** 1G/2G dimer class **(j)** 1G/2G monomer class **(k)** 3G dimer class and **(l)** 3G monomer class.

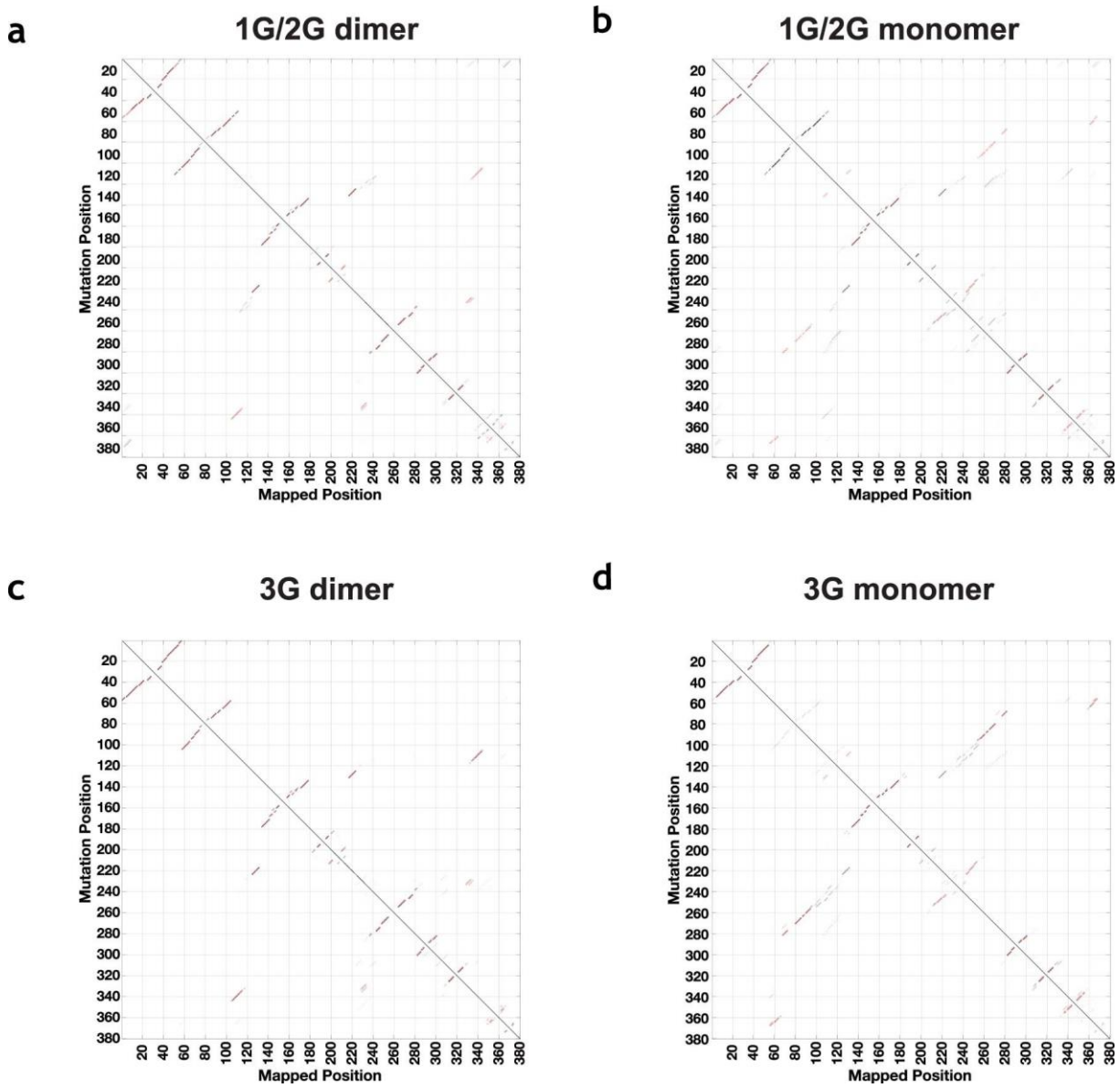
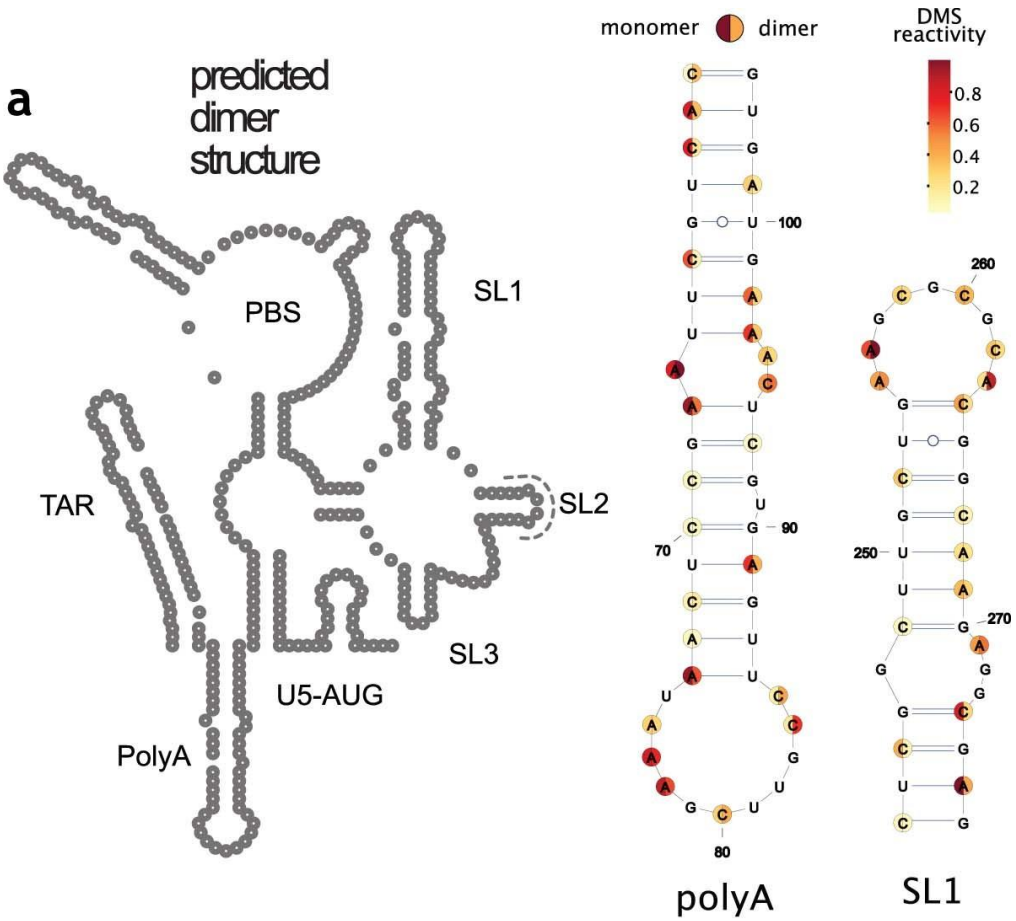


Figure 3.9| Bootstrapping analysis for 2-dimensional structural probing. Bootstrapping analysis for 2-dimensional structural probing. The predicted structure for the enhanced dimer and monomer structures are shown in red. Bootstrap support is shown in greyscale, with darker greys signifying better bootstrap support for the **(a)** 1G/2G dimer class, **(b)** 1G/2G monomer class, **(c)** 3G dimer class, and **(d)** 3G monomer class.


a
predicted
dimer
structure
PBS
SL1
TAR
SL2
SL3
U5-AUG
PolyA
monomer
dimer
DMS
reactivity
0.8
0.6
0.4
0.2
100
90
70
80
polyA
260
270
250
SL1

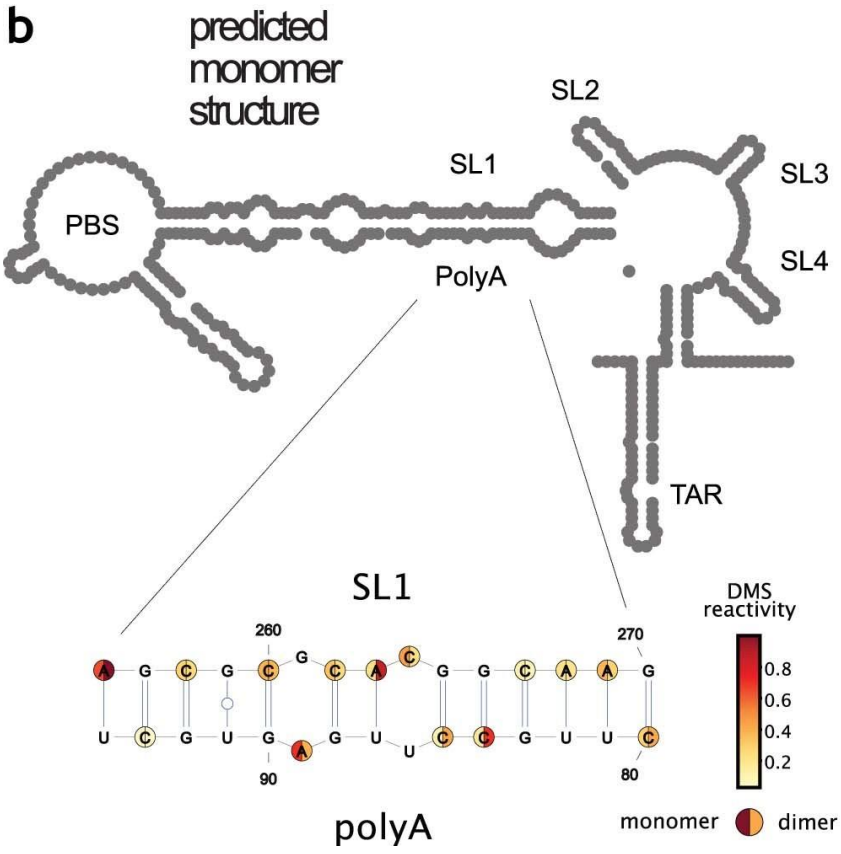

b
predicted
monomer
structure
SL2
SL1
SL3
PBS
SL4
PolyA
TAR
SL1
260
270
90
80
polyA
DMS
reactivity
0.8
0.6
0.4
0.2
monomer
dimer

Figure 3.10| Secondary structure model for 1G/2G dimer and 3G monomer class. Secondary structure model for 1G/2G dimer and 3G monomer class. **(a, b)** Secondary structure model of dimer and monomer class, respectively. Models were obtained using DMS reactivities as soft constraints for in silico folding in the Vienna RNA structure package. For the dimer structure, the U1sRNA binding site within SL2 is shown. For the insets, DMS reactivities from monomer and dimer samples were mapped to A and C residues. Structures of polyA and SL1 stem loops and polyA-SL1 interaction are shown. DMS reactivities for the monomer population are shown on the left hemisphere, and DMS reactivities for the dimer population on the right. Red signifies highly reactive positions that are unpaired. Pale yellow signifies unreactive positions that are base-paired.

3.3.5 Refinement of monomer and dimer structures

The incorporation of information from RNA structural probing experiments improves the accuracy of RNA structure predictions, but structural elements can still be incorrectly predicted because data from chemical probing experiments typically provide information on whether a nucleotide is base-pair or not, but not its base pairing partner[85][86]. FARS-seq enables a more powerful model-free approach to RNA structure determination by exploiting information in the mutation library to identify RNA helices directly (**Figure 3.12a**). When mutating a nucleotide in a stem structure, the base pairing partner, now unpaired, becomes more reactive to the chemical probe leading to correlated mutations in the sequencing data[59][86]. This two-dimensional data can directly detect RNA helices (along the diagonal) as well as non-canonical and tertiary interactions that are otherwise impossible to predict from classical one-dimensional RNA structural probing experiments.

Signals for RNA helices were visible in the raw mutational and z-score normalized data along the diagonals (**Figure 3.12b, e**). These signals were refined by applying convolution and threshold filters to enhance stems as well as tertiary interactions (**Figure 3.12b, e**). Finally, high confidence stems were highlighted by applying a helix filter and algorithm to select the 'best' non-conflicting stems with the highest score (**Figure 3.12c, f)**. Stem signals corresponding to SL1 were systematically present in dimer selected samples and absent in monomer selected samples (**Figure 3.11** and https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM1_ESM.pdf). In the 1G/2G dimer sample, both SL1 and polyA stem signals were observed. In the 3G monomer, polyA and SL1 stems were replaced with a signal matching the long distance SL1-polyA interaction (compare **Figure 3.12b** and **c** with **e** and **f**). Unexpectedly, in the 3G monomer we detected an additional novel interaction between the PBS loop and SL1, as well as a weaker signal between TAR and PBS, both of which were supported by a bootstrapping analysis (**Figure 3.7e, f** and **Figure 3.9**). In the previous structural prediction these regions in PBS and SL1 had high Shannon entropies and were poorly resolved (**Figure 3.7**).

Intriguingly, and uniquely in the 1G/2G structures, the TAR and polyA stem signals in the filtered z-scores were accompanied by punctate signals characteristic of tertiary contacts, alternative folds or non-canonical base-pairings (**Figure 3.11**). Because these contacts were consistently present in the 1G/2G samples and missing from the 3G samples, we speculate that they help to stabilize the 5' end

of the HIV-1 transcript to inhibit the translation of 1G/2G transcripts (**Figure 3.11** and https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM1_ESM.pdf). Additionally, in the 1G/2G monomer, the mutually exclusive polyA stem and the polyA-SL1 interaction were both observed, strengthening the idea that 1G/2G samples are preferentially dimeric and that some interconversion occurs even when monomers are isolated (**Figure 3.11**).

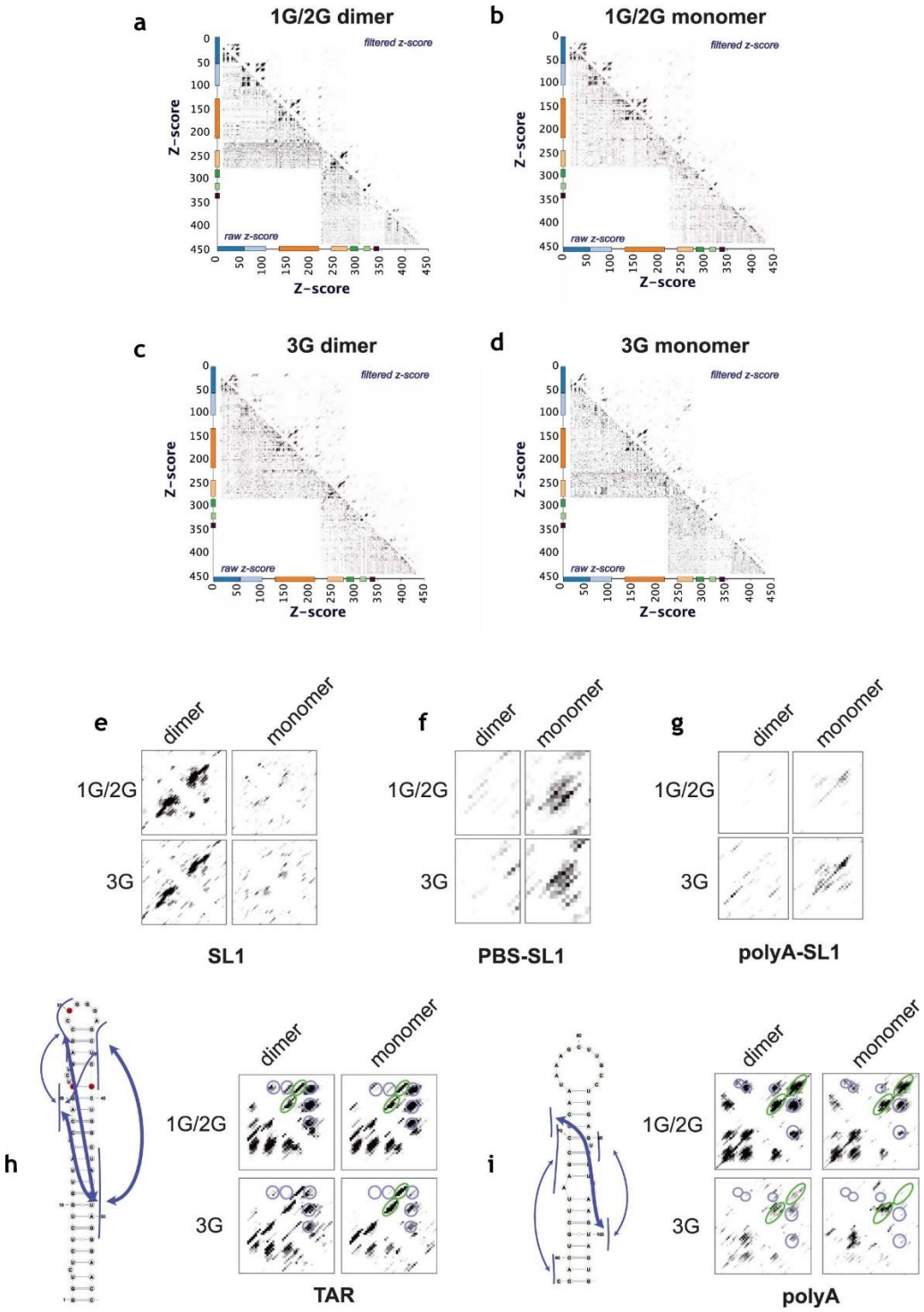


Figure 3.11| Two dimensional plots of mutation frequencies. Two dimensional plots of mutation frequencies for the **(a)** 1G/2G dimer class **(b)** 1G/2G monomer class **(c)** 3G dimer class and **(d)** 3G monomer class. z-scores of two-dimension structural probing data reveals RNA stems along the diagonal, as well as non-canonical or tertiary interactions. Regions in **(e)** SL1, **(f)** PBS-SL1, **(g)** polyA-SL1, **(h)** TAR and **(i)** polyA stem are highlighted. For TAR and polyA, detected stems are highlighted with green circles. Putative tertiary or non-canonical interactions are highlighted with purple circles and arrows.

To obtain enhanced structural models of the dimer and monomer structures we focused on the 1G/2G dimer and 3G monomer samples as these were the most structurally uniform. Here, the best stems obtained by multidimensional chemical probing were used as additional hard constraints in RNA structure prediction (**Figure 3.12d** and **g**). The enhanced 1G/2G dimer structure was nearly identical to that obtained without hard constraints, and contained the TAR, PolyA, PBS, SL1 and SL3 stem loops, as well as the AUG/U5 interaction as previously predicted (**Figure 3.12d**). The enhanced 3G monomer structure contained TAR, polyA-SL1 interaction, SL2, SL3 SL4 and polyA-Gag interaction, as before, but now included a stem loop structure due to base paring between PBS and SL1 (**Figure 3.12g**). A TAR-PBS pseudoknot interaction was added post-hoc, as it was selected by the best stem algorithm and supported by a bootstrapping analysis, although we note that the 2d stem score was relatively weak. All in all, multidimensional chemical probing not only provided direct experimental evidence that 1G/2G dimer and 3G monomer fractions are structurally distinct, but also identified structural features that couldn't be predicted by classical RNA structural probing experiments.

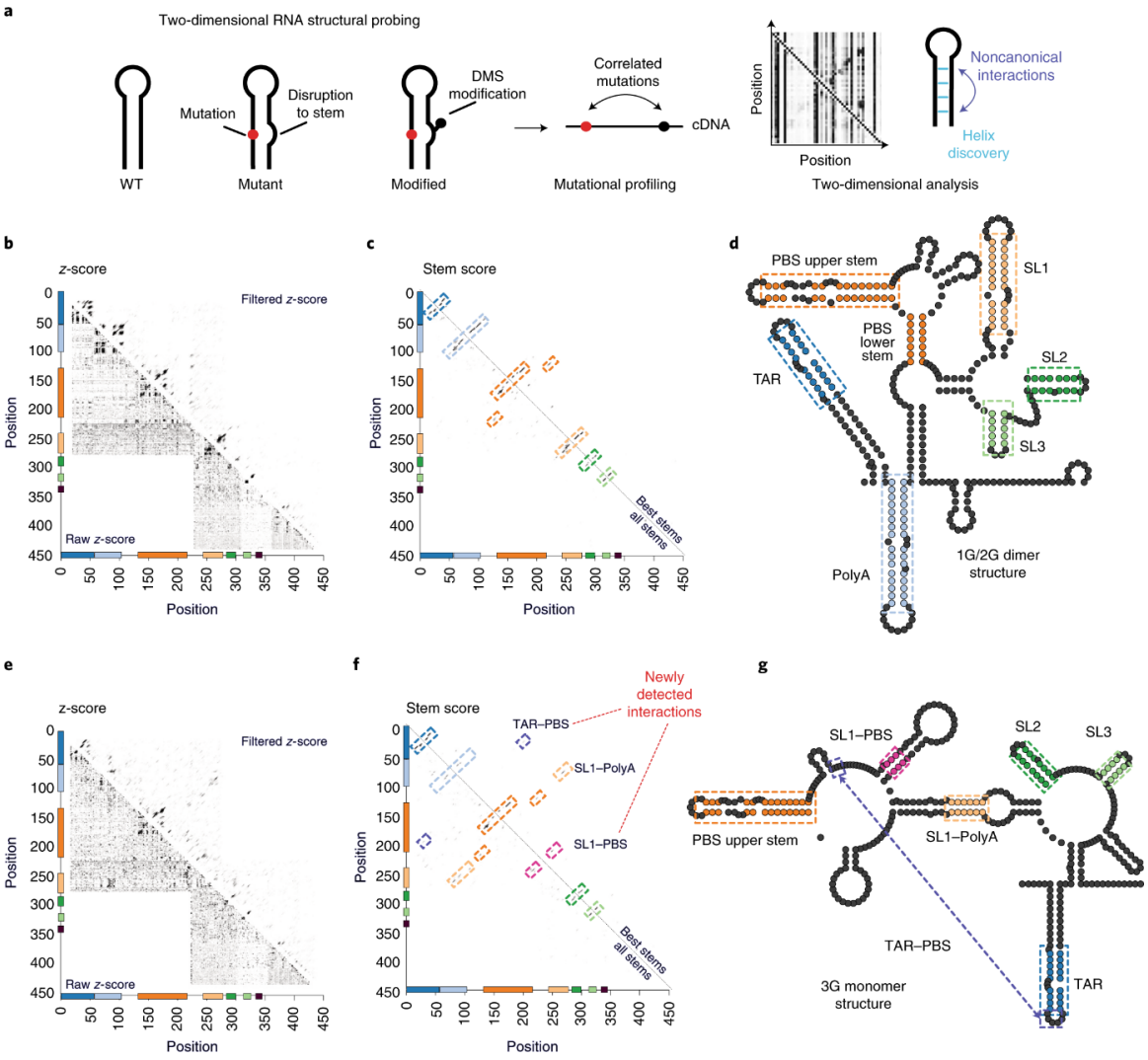


Figure 3.12| Two-dimensional mapping of RNA structure in 1G/2G dimer and 3G monomer populations. (a) Mutations disrupting RNA stems lead to increases in DMS reactivity at positions opposite the mutation. Positions of DMS modification are read out as mutations leading to correlated mutations at pairs of nucleotides involved in RNA structure. RNA secondary structures (blue circles) are identified along the diagonals. Punctate signals (purple circles) can signify noncanonical or tertiary interactions. b,e, The z-score analysis of mutation frequencies from 1G/2G dimer **(b)** and 3G monomer populations **(e)**. Raw z-scores (lower diagonal) reveal pairs of positions enriched with mutations. Filtered z-scores (upper diagonal) enhance stem signals by applying a convolution filter and signal threshold. Insets are zooms of the filtered z-scores for the polyA–SL1, PBS–SL1 and SL1 stems. **(c,f)** Stem detection in 1G/2G dimer **(c)** and 3G monomer populations **(f)**. All stems (lower diagonal) reveal all possible stems of minimum length 3 by applying a filter for Watson–Crick and Wobble base pairs to the filtered z-score. Best stem (upper diagonal) selects the best nonconflicting stems by removing conflicting stems based on filtered z-score. Colored boxes represent regions that are highlighted in enhance RNA secondary structure models. **(d,g)** Enhanced RNA secondary structure models of 1G/2G dimer **(d)** and 3G monomer populations **(g)**. Colored base pairings were detected in multidimensional mapping and used as hard constraints before in RNA secondary structure prediction. Dark blue is TAR. Light blue is polyA. Orange is PBS. Mustard is SL1. Dark green is SL2. Light green is SL3. Red represents the polyA–SL1 interactions, pink shows the new SL1–PBS interaction and purple the TAR–PBS interaction.

3.3.6 SL1 stability is a key element for genome dimerization.

One of the strengths of FARS-seq is the coupling of RNA structural and functional information at single nucleotide resolution. We therefore mapped the K^{dimer} values onto the dimer and monomer structures. In both buffers, the median mutations with the strongest effects mapped to the apical portion of SL1, with mutations to the palindromic loop sequence revealed to be the most destabilizing for dimerization, in agreement with their crucial role in the kissing loop interaction (**Figure 3.13a**, see https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM10_ESM.zip). The unpaired adenosine residues flanking the loop sequence were less important for dimerization than the palindromic sequences, in keeping with the observation that they can be individually mutated without disrupting dimerization[87]. Mutations to the stem of SL1 also strongly inhibited dimerization, with apical stem mutations generally having a stronger effect on dimerization compared to the basal stem mutants [$\log_2(K^{dimer})$ values 0.61-6.85 vs 0.31-3.49] (**Figure 3.13a**). Surprisingly, mutations at several positions within the SL1 internal loop (G247, A271, G272, G273) strongly enhanced dimerization upon mutation (**Figure 3.13a**, and https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM10_ESM.zip). Dimer enhancing mutations at these positions presumably stabilize SL1 by closing or reducing the size of the internal loop, strongly indicating that SL1 stability is a critical parameter for dimerization.

Whilst two-dimensional structural probing identified SL1 as a short stem loop with an apical and basal stem separated by an internal loop (nucleotides 243-277), our data nevertheless reveals structural plasticity in SL1. This realization comes from mapping the functional data to different

extended forms of SL1 that have been proposed in the literature: a two-internal loop model (2IL), a three-internal loop model (3IL), and three-way junction (3WJ) model (**Figure 3.13b)**. Even though these models have mutually exclusive internal loop configurations, mutations that closed or reduced the size of SL1 internal loops were invariably dimerization enhancing (**Figure 3.13b,** green arrows). For example, A235C, A235U or G281U strongly enhanced dimerization by converting the A235-G281 internal loop into a base pair in the three-way junction (3WJ) model, even though these mutations would have no effect on SL1 stability on the other structural models (**Figure 3.13b,** green arrows). Similarly, G282C and G239C would close the internal loop in the 3 internal loop (3IL) model explaining their dimerization enhancing properties (**Figure 3.13b,** green arrows). To confirm the structural plasticity of SL1, we performed in solution DMS-MaPseq analysis of mutants A235C and A239C and showed that they reconfigured the SL1 stem, as predicted (**Figure 3.14**). Interestingly, mutations A242C or A242U reduced the size of an SL1 internal loop in all models but nevertheless disrupted dimerization (**Figure 3.13b** and **c**; red arrows). These functional effects are explained by the fact that A242C or A242U extend the PBS-SL1 interaction to stabilize the monomer structure. Thus, the core dimerization structure in SL1 comprises an apical 7 nt stem and a basal 4 nt stem separated by an internal loop that can be further stabilized by metastable stem extensions or disrupted by a base-pairing interaction with PBS.

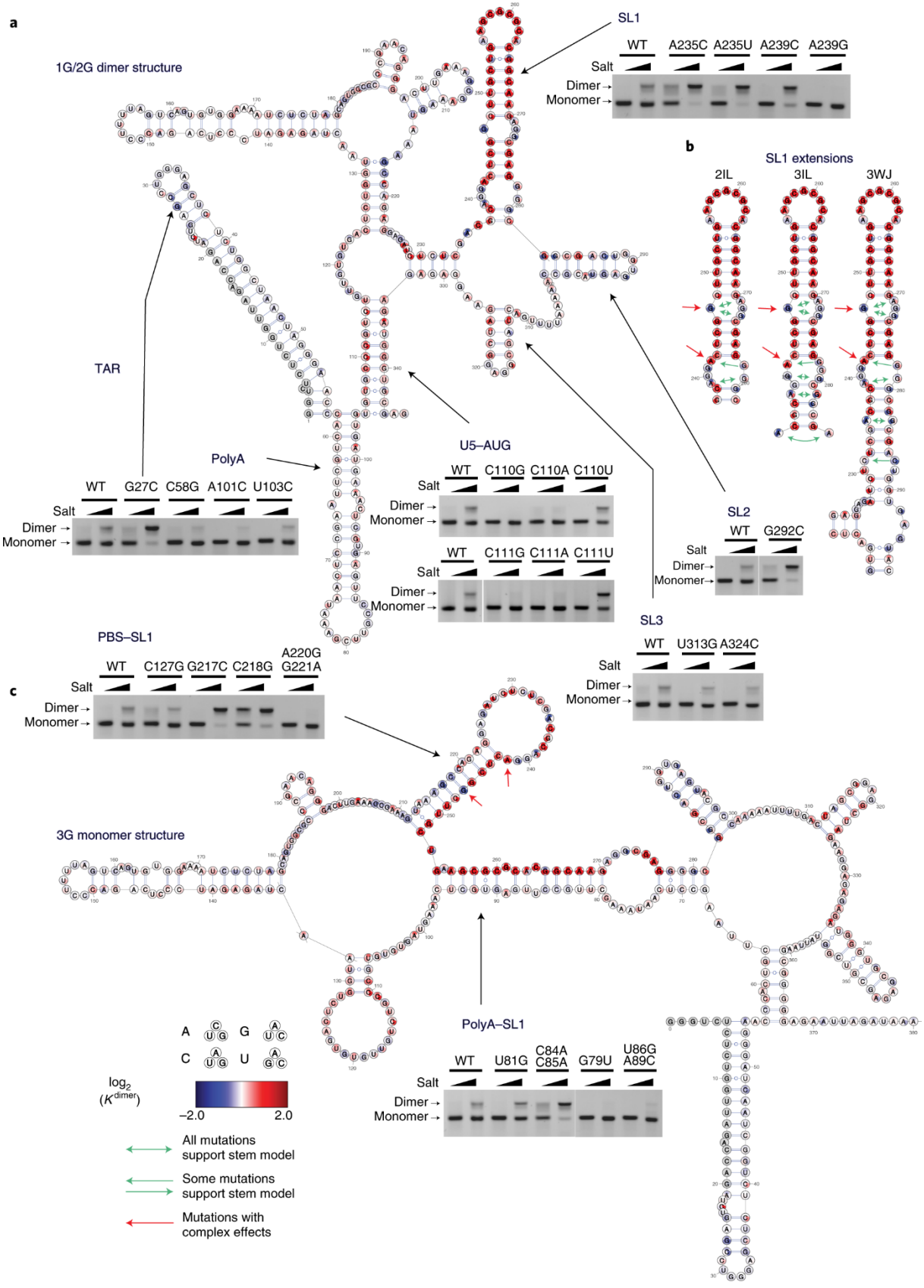

a
1G/2G dimer structure
SL1
WT A235C A235U A239C A239G
Salt
Dimer →
Monomer →
b
SL1 extensions
2IL 3IL 3WJ
TAR
PolyA
WT G27C C58G A101C U103C
U5–AUG
WT C110G C110A C110U
WT C111G C111A C111U
SL2
WT G292C
PBS–SL1
WT C127G G217C C218G A220G G221A
SL3
WT U313G A324C
c
3G monomer structure
PolyA–SL1
WT U81G C84A C85A G79U U86G A89C
log2 (K dimer)
–2.0
2.0
All mutations support stem model
Some mutations support stem model
Mutations with complex effects

Figure 3.13| Structure/function analysis of HIV-1 dimerization. (a,c) Single nucleotide resolution functional profiling data pooled from six low salt samples mapped the dimer **(a)** and monomer **(c)** structures expressed as log2(K^{dimer}) values. Each individual mutant shown as one of three circle in the order A, C,G,U clockwise from upper position (excluding the WT base). Validation of structural models on 3G RNA by point mutagenesis followed by native agarose gel electrophoresis in two different buffer conditions. Experiments were performed at least twice, representative data shown. Red circles show mutations inhibiting dimerization, and blue circles show mutations enhancing dimerization. log2(Kdimer) values above 2 are capped. **(b)** Functional profiling data mapped to different structural models of SL1 containing mutually exclusive internal loop configurations. The two-internal loop (2IL), 3IL and the 3WJ are mutually exclusive models of SL1 structure based on chemical probing or biophysical measurements. Green arrows show mutations that improve dimerization by closing or reducing the size of internal loops, providing evidence that SL1 is metastable and that alternative SL1 conformations can form and dimerize. Red arrows show mutations that have complex effects on dimerization because they affect the new PBS–SL1 interaction.

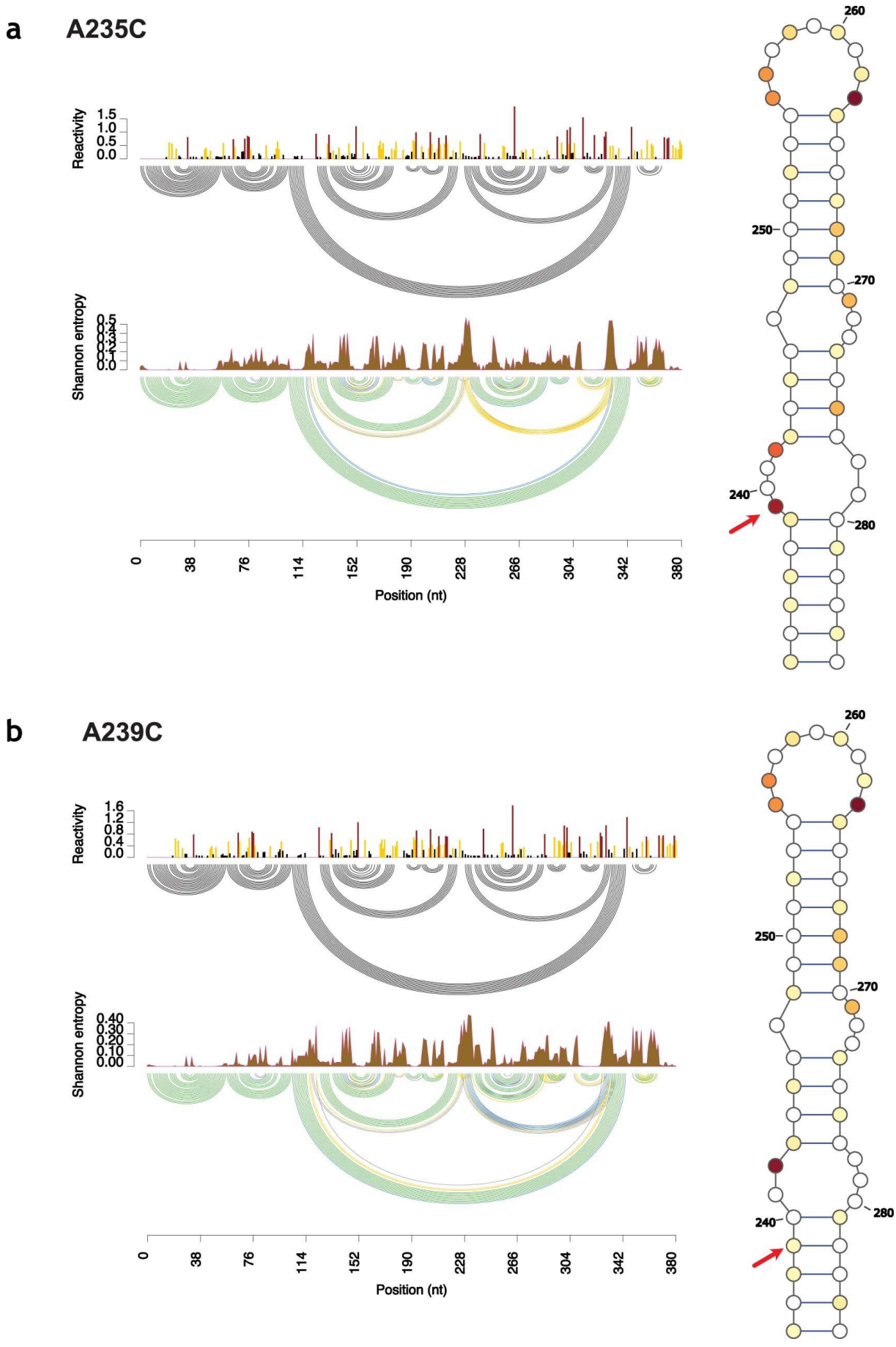

a
A235C
Reactivity
1.5
1.0
0.5
0.0
Shannon entropy
0.5
0.0
0
38
76
114
152
190
228
266
304
342
380
Position (nt)
260
250
270
240
280
b
A239C
Reactivity
1.6
1.2
0.8
0.4
0.0
Shannon entropy
0.40
0.30
0.20
0.10
0.00
0
38
76
114
152
190
228
266
304
342
380
Position (nt)
260
250
270
240
280

Figure 3.14| Secondary structure model for SL1 mutants. Secondary structure model for SL1 mutants. Models were obtained using DMS reactivities as soft constraints for in silico folding in the Vienna RNA structure package. DMS reactivities for each nucleotide position are show in the upper barchart. Shannon entropies are shown in the lower chart. Upper arc plots show consensus structure. Lower arc plots show base pairing probabilities (green = 70-100%; blue = 40-70%; yellow = 10-40%; grey = 5-10%). DMS reactivities from monomer and dimer samples were mapped to A and C residues on SL1. Red signifies highly reactive positions that are unpaired. Pale yellow signifies unreactive positions that are base-paired. **(a)** A235C and **(b)** A239C reconfigure the SL1 lower helix and internal loop to enhance dimerization. Red arrows highlight position 239 showing a reactivity change between the A235C and A239C mutations.

3.3.7 Inter-domain interactions regulate dimerization.

Outside of SL1, we found several structural domains and inter-domain interactions that affected dimerization (**Figure 3.13**). Our data support a role for the AUG-U5 interaction in positively regulating dimerization, as conversion of GU base pairs at U107-G342, G108-U341, G112-U337 to either AU or GC base pairs consistently enhanced dimerization, whereas mutations disrupting the interaction were inhibitory (**Figure 3.13a**). SL3 stem mutations weakly inhibited dimerization, most likely because disruption of SL3 would induce misfolding of the RNA (**Figure 3.13a**). Finally, mutations to SL2 were generally dimerization enhancing, and these types of mutations were especially evident in the 3' SL2 stem (**Figure 3.13a**).

We also validated the novel short- and long-range interactions between polyA-SL1 and PBS-SL1. Mutations to the base of polyA generally inhibited dimerization, indicating that destabilizing the polyA stem favours the formation of the polyA-SL1 interaction (**Figure 3.13a**). On the other hand, mutations to the upper portion of polyA enhanced dimerization by disrupting the polyA-SL1 base pairing (**Figure 3.13c**). In the same vein, we found stretches of nucleotides in PBS that strongly enhanced dimerization upon mutation (**Figure 3.13c**). Functional profiles in the lower PBS stem were particularly interesting as this stem structure is universally found in contemporary models of the HIV-1 5'UTR and contains the primer activation sequence (PAS) known to be important for efficient reverse transcription[88]. We found that mutation of two nucleotides G217 and C218 in the lower PBS stem very strongly enhanced dimerization, even though mutations to this stem were generally inhibitory (**Figure 3.13a**). This can be mechanistically explained because mutation of these nucleotides disrupted a novel base-pairing between PBS and SL1 that stabilizes the monomer structure.

Because these results suggested a functional interaction between primer tRNA binding and dimerization, we next assessed whether disruption of the PBS with tRNA mimic oligos affected dimerization. $cPBS_{182\text{-}199}$ annealed to the loop region disrupted the putative TAR-PBS interaction, whereas $cPBS_{199\text{-}216}$ disrupted the novel PBS-SL1 stem loop (**Figure 3.15a**). Both oligos enhanced dimerization confirming a functional interaction between PBS and dimerization. Surprisingly, annealing the $cPBS_{182\text{-}199}$ oligo also led to the formation of a higher, presumably tetrameric molecular

species. The TAR apical loop contains a 10-nucleotide palindromic sequence that has been proposed to dimerize by a TAR–TAR kissing interaction analogous to the one used by SL1[31]. We therefore postulate that $cPBS_{182-199}$ disrupts the TAR-PBS interaction detected by multi-dimensional structural probing, allowing TAR to dimerize independently of SL1.

Finally, since genome dimerization is thought to be a pre-requisite for genome packaging, we selected mutations in adjacent nucleotides with divergent effects on dimerization and measured their effects on $Pr55^{Gag}$ binding by microscale thermophoresis (MST) (**Figure 3.15b**). Importantly, none of these mutations resided in the HIV-1 packaging domain (SL1-SL3). In PBS, C218G, which strongly enhanced dimerization had higher affinity (K_d 19 nM) to $Pr55^{Gag}$ compared with WT RNA (K_d 38 nM). In contrast, PBS A220G-G221A, which was unable to dimerize, did not bind $Pr55^{Gag}$ at any of the concentrations tested (K_d n.a.). In polyA, dimerization enhancing mutation C84A-C85C bound $Pr55^{Gag}$ with higher affinity (17 nM) than WT, whereas dimerization disrupting mutation U86G-A89C bound $Pr55^{Gag}$ with lower affinity than WT (110 nM). By performing in solution DMS-MaPseq analysis *in vitro*, we established that mutations in polyA-SL1 and PBS-SL1 alter ensemble reactivities towards the profiles seen in the isolated monomer and dimer. (**Figure 3.16**). Thus, the four mutants not only alter the monomer/dimer equilibrium but produce the predicted structural changes that affect $Pr55^{Gag}$ binding. We also introduced these mutations into the full-length HIV-1 genome and assessed their effects on packaging efficiency in competition assays (**Figure 3.15c**). In PBS, dimer promoting mutant C218G was enriched 1.5-fold in virions compared to the monomer promoting mutant A220G-G221A. In polyA, dimer promoting mutant C84A-C85A was enriched 2-fold in virions compared to the monomer promoting mutant U86G-A89C. In a five-way competition assay between wild-type (WT) HIV-1 and the mutants, dimer promoting mutants C218G and C84A-C85A were packaged equivalently or better than WT. Conversely, monomer promoting mutants U86G-A89C and A220G-G221A were deficient in packaging compared to WT. Lastly, we performed in solution DMS-MaPseq analysis of these four mutants directly in cells. Despite complex reactivity changes induced by cellular ligands, dimer promoting mutants folded into structures containing SL1, whereas monomer promoting mutants folded into structures where SL1 was hidden through long- and short-range interactions with polyA and PBS (**Figure 3.17**). Thus, we conclude that the regulatory mechanism we identified *in vitro* also takes place in cells.

Taken together, our results provide a clear mechanistic explanation for the link between dimerization, $Pr55^{Gag}$ binding and packaging. We also show how changes to the PBS functionally link the tRNA binding region to packaging (**Figure 3.15d**).

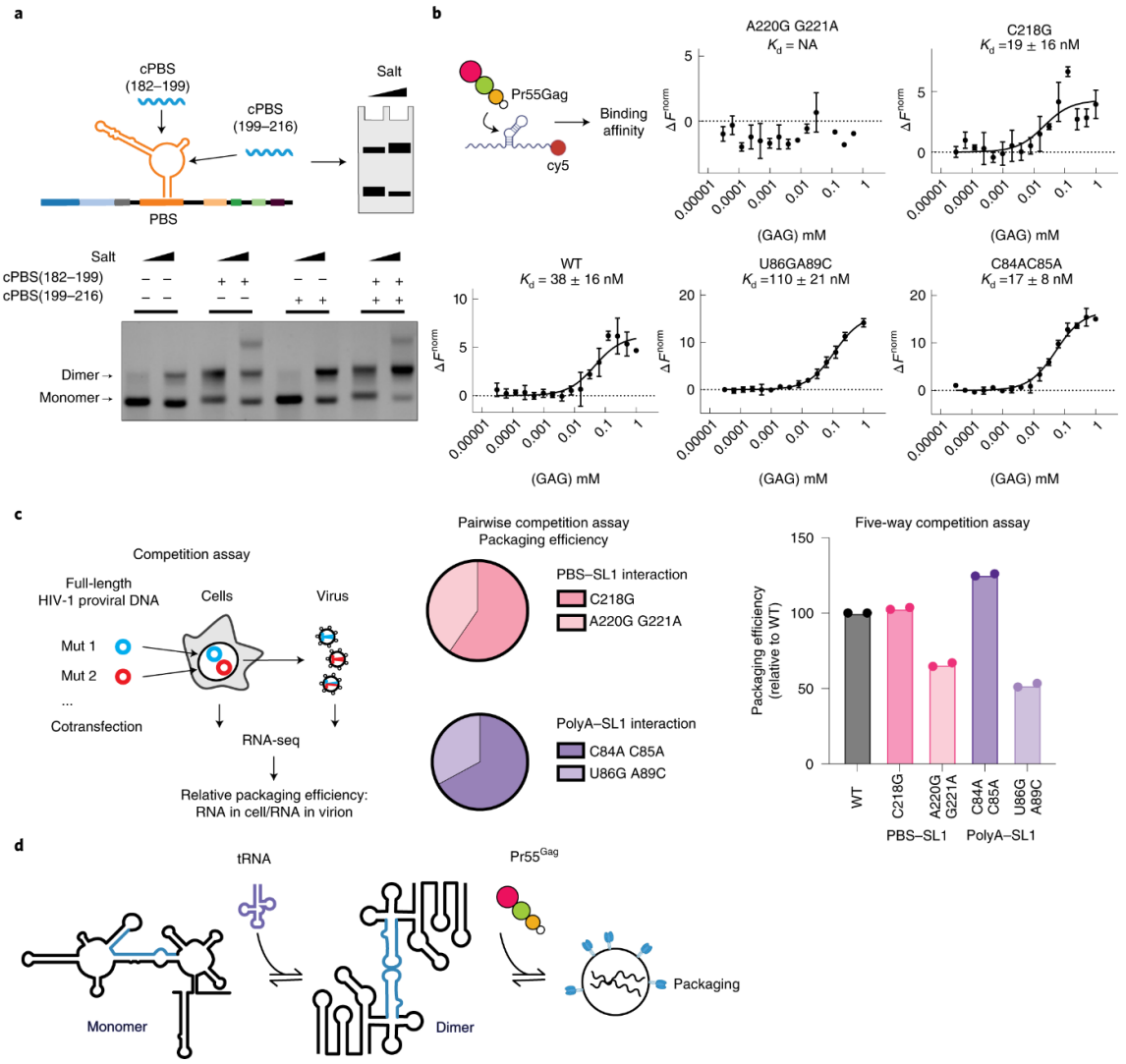
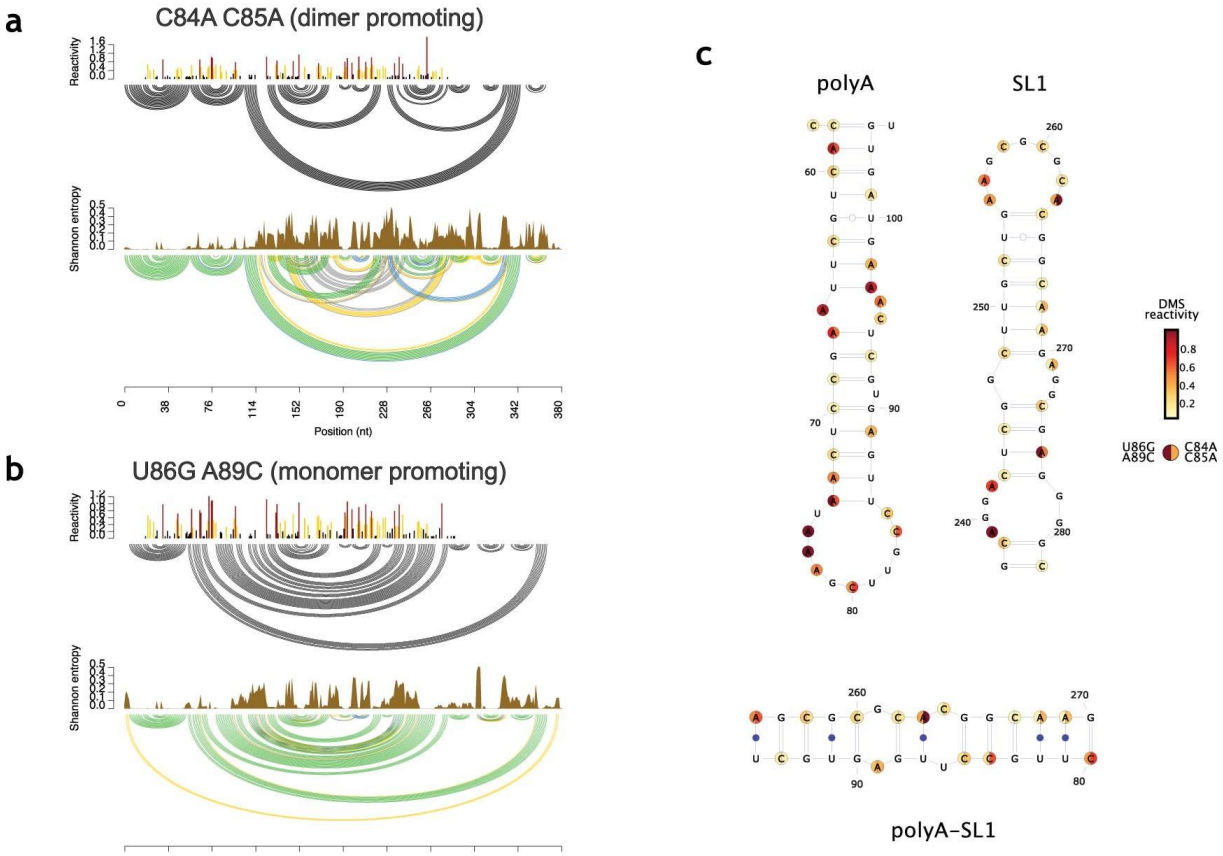


Figure 3.15| PBS and polyA regulate HIV-1 dimerization, Pr55Gag binding and genome packaging. **(a)** PBS targeting oligos can trigger dimerization of a 3G RNA. cPBS(182–199) disrupts the TAR–PBS interaction leading to the formation of a higher order RNA structure. cPBS(199–216) disrupts the PAS-anti-PAS stem and enhances dimerization. The effects of both oligos are additive. Experiments were performed in duplicate, with representative data shown. **(b)** Mutations targeting the polyA–SL1 and PBS–SL1 interaction affect Pr55Gag binding as measured by MST. Data from three independently experiments were analyzed. Data are represented as mean with error bars showing standard deviations. **(c)** Competition assays to measure the relative effects of mutations on genome packaging into virions. Two-way competition assays show that dimer promoting mutations C218G and C84A–C85A are enhanced in genome packaging compared to monomer promoting mutations A220G–G221A and U86G–A89C. Five-way competition assays between WT HIV-1 and mutants show that dimer promoting mutants are packaged similar or better than WT, whereas monomer promoting mutants are packaged less efficiently than WT. Experiments were performed in duplicate. **(d)** Model showing how the binding of host factors can regulate viral replication, in part, through remodeling RNA structure.

polyA-SL1 mutants



PBS-SL1 mutants

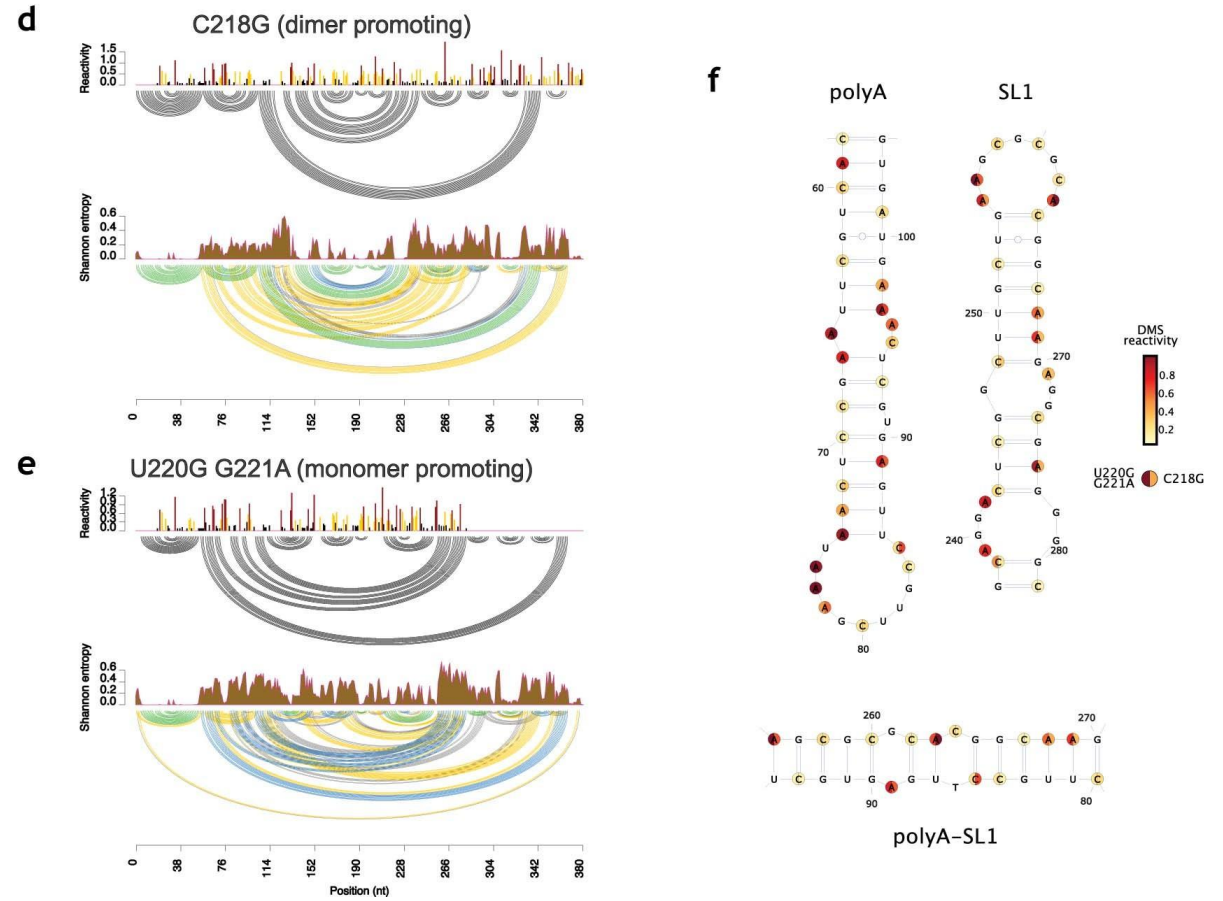


Figure 3.16| Secondary structure predictions for dimer and monomer promoting mutants. Secondary structure predictions for dimer and monomer promoting mutants targeting the polyA-SL1 and PBS-SL1 interactions. DMS reactivities and secondary structure models for polyA, SL1 and polyA-SL1. DMS reactivities for each nucleotide position are show in the upper barchart. Shannon entropies are shown in lower chart. Upper arc plots show consensus structure. Lower arc plots show base pairing probabilities (green = 70-100%; blue = 40-70%; yellow = 10-40%; grey = 5-10%). DMS reactivities from monomer and dimer samples were mapped to A and C residues. Red signifies highly reactive positions that are unpaired. Pale yellow signifies unreactive positions that are base-paired. **(a-c)** DMS reactivities and secondary structure models for polyA-SL1 mutants. **(a)** dimer promoting mutant C84A-C85A folds into the canonical 5'UTR structure **(b)** Monomer promoting mutant U86G–A89C contains the polyA-SL1 interaction. **(c)** Reactivities for both mutants U86G-A89C (left hemisphere) and C84A-C85A (right hemisphere) mapped to the structures polyA, SL1, and polyA-SL1. **(d)** Dimer promoting mutant C218G folds into the structure containing SL1 **(e)** Monomer promoting mutant U220G-G221A folds into a structure containing the PBS-SL1 interaction. **(f)** Reactivities for both mutants U220G-G221A (left hemisphere) and C218G (right hemisphere) mapped to the structures polyA, SL1 and polyA-SL1.

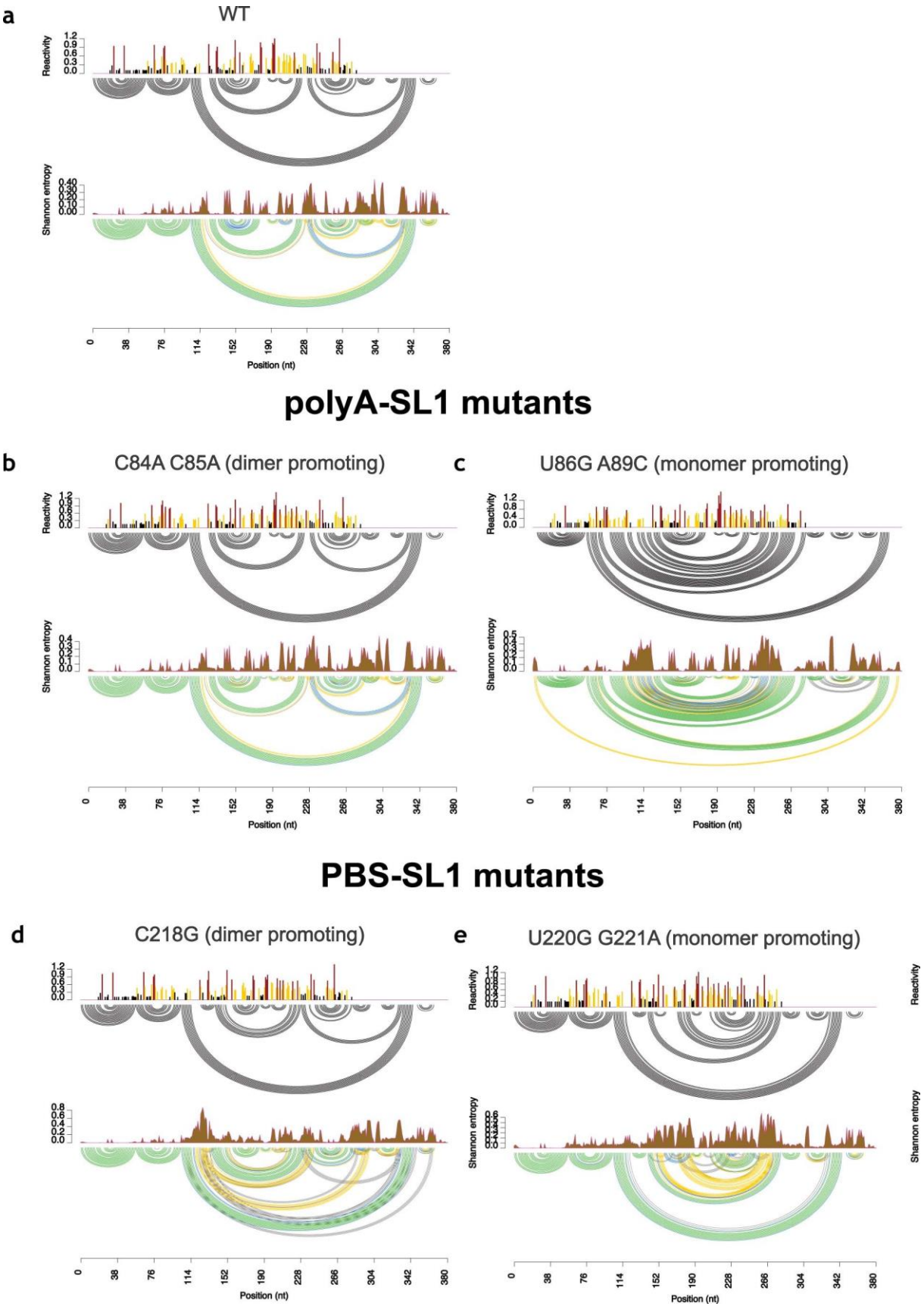

a
WT
Reactivity
Shannon entropy
Position (nt)
polyA-SL1 mutants
b
C84A C85A (dimer promoting)
c
U86G A89C (monomer promoting)
PBS-SL1 mutants
d
C218G (dimer promoting)
e
U220G G221A (monomer promoting)

Figure 3.17| In cell DMS reactivities and secondary structure predictions for dimer and monomer promoting mutants. In cell DMS reactivities and secondary structure predictions obtained for dimer and monomer promoting mutants targeting the polyA-SL1 and PBS-SL1 interactions. DMS reactivities for each nucleotide position are show in the upper barchart. Shannon entropies are shown in lower chart. Upper arc plots show consensus structure. Lower arc plots show base pairing probabilities (green = 70-100%; blue=40-70%; yellow=10-40%; grey=5-10%). **(a)** Wild-type HIV-1 **(b)** dimer promoting mutant C84A-C85A folds into the canonical 5'UTR structure **(c)** Monomer promoting mutant U86G-A89C contains the polyA-SL1 interaction. **(d)** Dimer promoting mutant C218G folds into the structure containing SL1 **(e)** Monomer promoting mutant U220G-G221A folds into a structure containing the PBS-SL1 interaction.

3.3.8 Dimerization regulation in HIV-1 strain, Mal

Recently, the structure of the 3G capped transcript for a different strain of HIV-1 (M group subtype A; HIV-1_{MAL}) was solved by NMR [36]. In contrast to our results on HIV-1 the NL43 strain, the identified a disruption of the polyA stem in 3G transcripts and the formation of a long-range interaction between SL1 and U5. Thus, our results obtained from NL43 agree that 3G transcripts are preferentially monomeric, yet disagree with some precise structural details, in particular the base-pairing partner of SL1. One explanation for these results is that there are different regulatory mechanisms in HIV-1 strains. Compared to the prototypic subtype B strain NL43, Mal contains a 23-nucleotide duplication in the same region in PBS that we found to be a regulator of dimerization. Furthermore, this duplication leads to structural differences in the initiation of reverse transcription[82][83].

To assess the difference of dimerization regulation in Mal and NL43, we applied FARS-seq on HIV-1_{MAL}. Using the same process as NL43, we generated the 1G or 3G RNA transcripts with or without Cap using randomly mutated HIV-1_{MAL} 5'UTR as templates, followed by dimer/monomer isolation, DMS probing, mutational profiling, mutation co-relation analysis, and ultimately, RNA structure remodelling.

Similar to NL43, native gel of HIV-1_{MAL} 5'UTR showed that the dimerization of capped RNAs and non-capped RNAs were similar, and 1G transcripts favor dimer formation compared with 3G transcript (**Figure 3.18a**). However, it seemed like that for Mal, no matter the buffer condition (low or high salt buffer), the monomer population was more dominant than the dimer population (**Figure 3.18a**); while for NL43, the dimer/monomer ratio of 1G/2G RNAs is higher in high salt buffer **(Figure 3.3a)**.

We next plotted the median $\log_2(K^{dimer})$ values at each nucleotide position for all transcript variants tested under the two buffer conditions (**Figure 3.18b**). As expected, the most significant peak localized to SL1 in all conditions. Interestingly, a double negative peak emerged within the 3' sequence on the bottom stem of SL1. Moreover, we identified additional sequences outside of SL1 involved in the dimerization regulation process: two peaks within the TAR region indicated that these regions negatively regulated dimerization, as well as the region surrounding U5 (110-116) and AUG (350-352), which indicated that the U5-AUG interaction was important for the dimerization process.

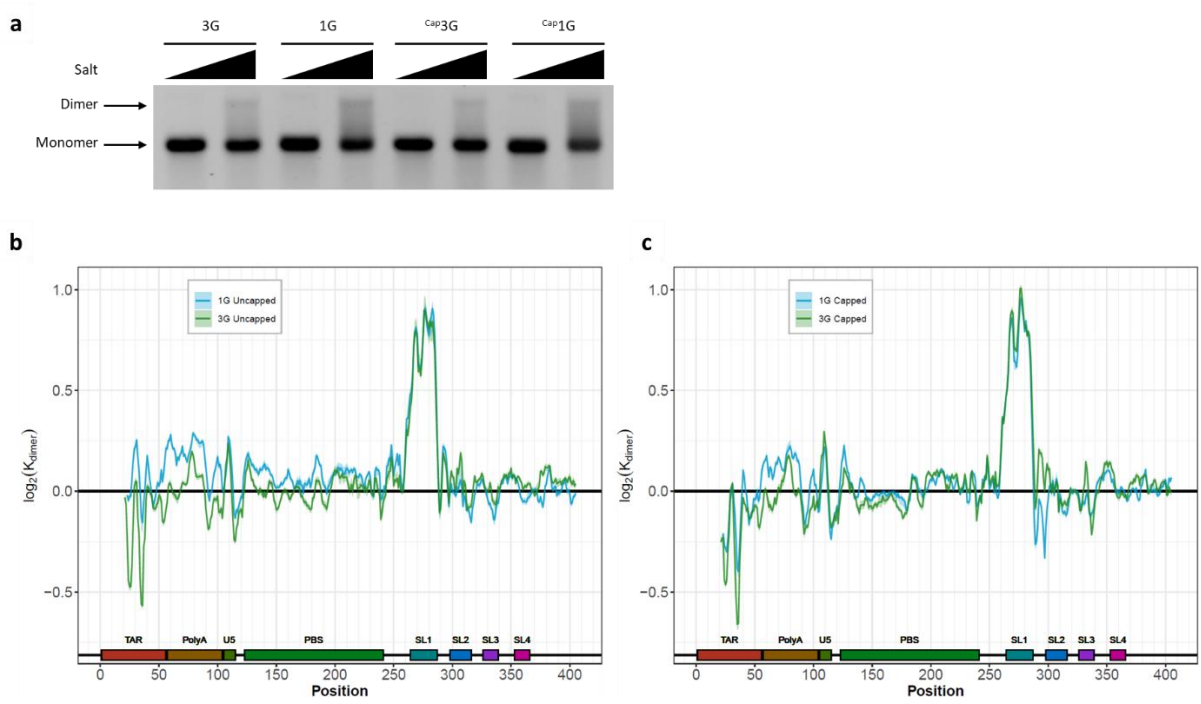


Figure 3.18| Functional and structural analysis of HIV-1_{MAL} 5'UTR RNA dimerization. (a) 1G and 3G capped and uncapped transcript variants migrate as distinct monomer and dimer bands on native agarose gels in both low and high salt buffers. **(b)(c)** median $\log_2(k^{dimer})$ values for each genome position for all transcript variants in high buffer.

We next analyzed the secondary structure of dimer and monomer population based on DMS reactivities **(Figure 3.19a, c)**. For the 1G dimer class, we could see that the RNA structure contained TAR, polyA (partial), PBS, SL1, SL2 and SL3 stem loops, as well as the U5-AUG interaction, which was similar to the dimer structure of NL43. The predicted monomer structure also presented the same TAR, polyA (partial), PBS, as well as the U5-AUG interaction, which is different from the published results that U5 interacted with SL1. In the monomer structure, the 3' portion of the SL1 stem was predicted to base-pair with the region 114-121 near U5, rather than PBS or polyA. These results indicate that HIV-1 can use alternative mechanisms to regulate dimerization. Thus, the metastable SL1 was still the key regulator for Mal dimerization, but because of the 23-nucleotide duplication in PBS, Mal used a different mechanism from NL43. Further 2D analysis and function validation is still on going.

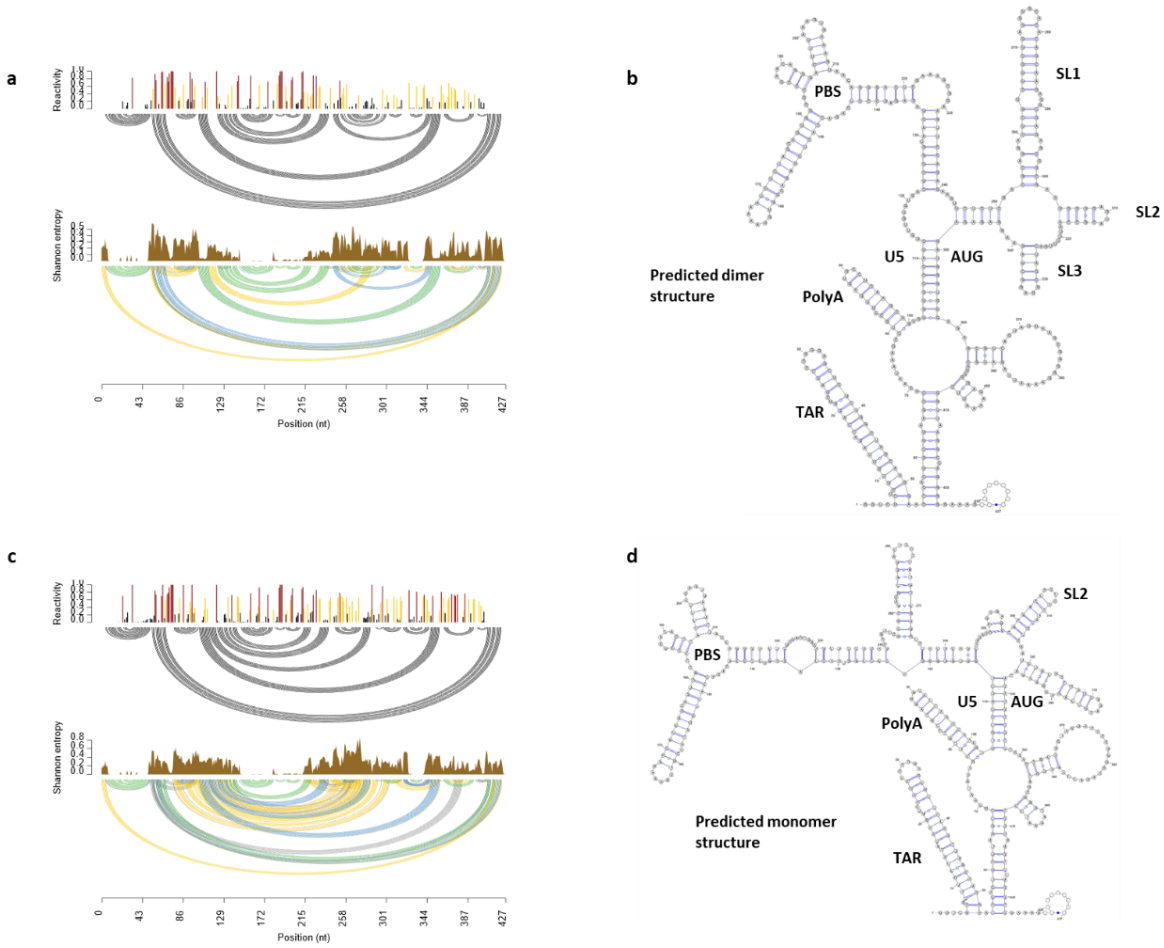


Figure 3.19| Predicted secondary structure model for 1G dimer and 3G monomer populations. DMS reactivity of dimer **(a)** and monomer**(c)**; Secondary structure model of dimer **(b)** and monomer class **(d)**. Models were obtained using DMS reactivities as soft constraints for in silico folding in the Vienna RNA structure package.

3.4 Discussion

Accumulating evidence emphasizes dimerization as a key step in HIV-1 life cycle that is regulated, at least in part, through the folding of the HIV-1 genomic RNA[19], [32]–[36], [49], [89]–[92]. Here, we resolved the structure of the monomeric and dimeric RNAs using a novel approach that integrates information from RNA structural probing with high-throughput functional profiling. This experimental strategy has significant advantages over other chemical probing methods that make ensemble measurements over all possible conformations of the RNA in solution. Such ensemble measurements, unless cautiously interpreted, can lead to false predictions when mapped to a single structure. We overcome this problem by physically isolating RNA structural conformations with respect to their function, akin to in gel SHAPE which was first developed to resolve structural differences between monomeric and dimeric species of the HIV-1 5'UTR[34]. Moreover, by performing chemical probing

on mutagenic libraries we obtain model-free information on RNA helices in the same way as “mutate and map”[86] or “M2-seq”[59]. Finally, and most importantly, our approach enables a deep understanding of how RNA structures relate to RNA function by uniquely coupling structural information with a functional read-out.

Taken together, our data recognizes a core dimerization domain of SL1 comprised of a 7 bp apical stem and 4 bp basal stem separated by an internal loop. This core dimerization domain is present in most structural models of SL1, but there is significant disagreement on whether SL1 is further extended[93]–[96]. In some structures, extensions to SL1 even lead to the complete disruption of SL2[32]. Here, we found no direct evidence that SL1 is in an extended form in dimeric RNA and consistently observe signals for SL2 as a short imperfect stem containing a bulged adenosine. Nevertheless, functional profiling provides strong evidence that mutually exclusive extended forms of SL1 can be readily generated, either directly through stabilizing mutations, or indirectly by destabilizing SL2. The fact that single point mutations could have such dramatic effects on dimerization provides evidence that the 5’UTR is dynamic and metastable. In the context of viral infection this is noteworthy because it provides a mechanism to regulate dimerization through the binding of viral or cellular factors to the genome (**Figure 3.15c**).

The metastable nature of SL1 was strikingly revealed in monomeric RNA. In contrast to SL3, which was present in both monomer and dimer structures, SL1 was destructured in monomeric RNA. Instead of a stem-loop, SL1 was reorganized into a short-range interaction with PBS and a long-distance interaction with polyA. These results are in agreement with the prevalent idea that RNA conformational switches regulate HIV-1 replication[50][97]. The dimer and monomer structural conformations we present here are reminiscent of the branched multiple hairpin (BMH) and long-distance interaction (LDI) models that were proposed as alternative structures that would regulate the dimerization, packaging, splicing and translation of the HIV-1 genome[19], [33], [90], [91]. The BMH exposes the TAR, polyA, PBS, SL1, SL2 and SL3 structures, and contains the U5-AUG interaction. The LDI model includes the interaction between polyA and SL1, but also includes additional rearrangements that we did not observe, such as an extension of SL3 and a disruption of SL2. Moreover, the LDI model does not include the novel PBS-SL1 interaction. Nevertheless, certain mutants designed to alter the LDI/BHM equilibrium are directly applicable to our structural model. In particular, mutations destabilizing the polyA stem inhibit dimerization and packaging[33], [90], whereas mutations disrupting the polyA-SL1 interaction enhanced dimerization[91]. These data are in agreement with our results showing that polyA-SL1 regulates not only dimerization, but also genome packaging. Significantly, recent work has identified the primary $Pr55^{Gag}$ binding site for HIV-1 as SL1[47], [48], [52], [98] with polyA providing an additional packaging signal in cells[48][99]. The fact that SL1 and polyA are completely disrupted in the monomer population provides a mechanistic explanation for the long-postulated link between dimerization and packaging.

Recently the structure of the 3G capped transcript was solved by NMR revealing the disruption of the polyA stem in 3G transcripts and the formation of a long-range interaction between SL1 and U5[36]. Thus, our results agree that 3G transcripts are preferentially monomeric, yet disagree with precise

structural details, in particular the base pairing partner of SL1. One way to reconcile these data is that the NMR structure was obtained with the Mal isolate, in contrast to the NL43 isolate used in the present study. Therefore, we applied FARS-seq on Mal strain too. Interestingly, our DMS probing data showed that the dimer structure of Mal did form an A-rich loop PBS and a U5-AUG interaction, but U5 did not interact with SL1 within monomer structure. Further analysis is still on going, we can still see that Mal use related, yet distinct, structural rearrangements to regulate dimerization, since the Mal isolate contains a 23-nucleotide duplication in the same region in PBS that we find as a regulator of dimerization in NL43. Nonetheless, both the polyA-SL1 and PBS-SL1 interactions are conserved amongst 800 curated sequences in the Los Alamos HIV-1 sequence database indicating regulation of dimerization by polyA and PBS is widespread (https://static-content.springer.com/esm/art%3A10.1038%2Fs41594-022-00746-2/MediaObjects/41594_2022_746_MOESM1_ESM.pdf).

Finally, we identified a novel interaction between PBS and SL1 that acts as negative regulator of dimerization, $Pr55^{Gag}$ binding and packaging. We demonstrated that this negative regulation can be counteracted through the binding of oligos to the primer binding site. Disruption of this negative regulation would mechanistically explain why tRNA annealing enhances dimerization[17], [100], and also opens up the possibility that primer binding to the PBS affects other steps of the HIV-1 life-cycle, such as translation, by altering the monomer / dimer equilibrium. It also reveals a general principle by which RNA structural changes induced by host factors can regulate key stages of the HIV-1 life cycle (**Figure 3.15d**).

Acknowledgements: We thank J.-C. Paillart and R. Marquet for critical feedback. We also thank the Helmholtz Association (grant no. VH-NG-1347 to R.S.) and the Bundesministerium für Bildung und Forschung (grant no. COMPLS-182 to R.S. and M.v.K.). A.S.G.-B. was supported with a fellowship from the Peter und Traudl Engelhorn Stiftung. N.C. received funding from the European Research Council (ERC) grant no. 948636. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions: R.P.S., A.S.G.-B. and L.Y. conceived the study. L.Y., A.S.G.-B., C.B., U.B.A., S.A. and M.O.-N. performed the experiments. R.P.S., M.v.K., P.B. and M.S. performed the analysis. A.K. and N.C. purified Pr55Gag and performed MST measurements. R.P.S. and L.Y. wrote the manuscript with contributions from the other authors.

Funding: Open access funding provided by Helmholtz-Zentrum für Infektionsforschung GmbH (HZI).

3.5 References

[1] D. C. Douek, M. Roederer, and R. A. Koup, “Emerging concepts in the immunopathogenesis of

AIDS," *Annu. Rev. Med.*, vol. 60, pp. 471–484, 2009, doi: 10.1146/annurev.med.60.041807.123549.

[2] R. A. Weiss, "How does HIV cause AIDS?," *Science (80-.).*, vol. 260, no. 5112, pp. 1273–1279, 1993, doi: 10.1126/science.8493571.

[3] A. L. Cunningham, H. Donaghy, A. N. Harman, M. Kim, and S. G. Turville, "Manipulation of dendritic cell function by viruses," *Curr. Opin. Microbiol.*, vol. 13, no. 4, pp. 524–529, 2010, doi: 10.1016/j.mib.2010.06.002.

[4] H. Garg, J. Mohl, and A. Joshi, "HIV-1 induced bystander apoptosis," *Viruses*, vol. 4, no. 11, pp. 3020–3043, 2012, doi: 10.3390/v4113020.

[5] G. Doitsh *et al.*, "Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection," *Nature*, vol. 505, no. 7484, pp. 509–514, 2014, doi: 10.1038/nature12940.

[6] J. M. Abramo *et al.*, "Individuality in music performance," *Assess. Eval. High. Educ.*, vol. 37, no. October, p. 435, 2012, doi: 10.1007/82.

[7] B. Chen, "HIV Capsid Assembly, Mechanism, and Structure," *Biochemistry*, vol. 55, no. 18, pp. 2539–2552, 2016, doi: 10.1021/acs.biochem.6b00159.

[8] B. Chen, "Molecular Mechanism of HIV-1 Entry," *Trends Microbiol.*, vol. 27, no. 10, pp. 878–891, 2019, doi: 10.1016/j.tim.2019.06.002.

[9] T. Xiao, Y. Cai, and B. Chen, "Hiv-1 entry and membrane fusion inhibitors," *Viruses*, vol. 13, no. 5, pp. 1–19, 2021, doi: 10.3390/v13050735.

[10] Z. Ambrose and C. Aiken, "HIV-1 uncoating: Connection to nuclear entry and regulation by host proteins," *Virology*, vol. 454–455, no. 1, pp. 371–379, 2014, doi: 10.1016/j.virol.2014.02.004.

[11] N. Jouvenet, S. Lainé, L. P. Vivares, and M. Mougel, "Cell biology of retroviral RNA packaging," *RNA Biol.*, vol. 8, no. 4, 2011, doi: 10.4161/rna.8.4.16030.

[12] V. D'Souza and M. F. Summers, "How retroviruses select their genomes," *Nat. Rev. Microbiol.*, vol. 3, no. 8, pp. 643–655, 2005, doi: 10.1038/nrmicro1210.

[13] N. Dubois, R. Marquet, J. C. Paillart, and S. Bernacchi, "Retroviral RNA dimerization: From structure to functions," *Front. Microbiol.*, vol. 9, no. MAR, pp. 1–19, 2018, doi: 10.3389/fmicb.2018.00527.

[14] J. C. Paillart, M. Shehu-Xhilaga, R. Marquet, and J. Mak, "Dimerization of retroviral RNA genomes: An inseparable pair," *Nat. Rev. Microbiol.*, vol. 2, no. 6, pp. 461–472, 2004, doi: 10.1038/nrmicro903.

[15] J. C. Paillart, R. Marquet, E. Skripkin, C. Ehresmann, and B. Ehresmann, "Dimerization of retroviral genomic RNAs: Structural and functional implications," *Biochimie*, vol. 78, no. 7, pp.

639–653, 1996, doi: 10.1016/S0300-9084(96)80010-1.

[16] M. Kuzembayeva, K. Dilley, L. Sardo, and W. S. Hu, “Life of psi: How full-length HIV-1 RNAs become packaged genomes in the viral particles,” *Virology*, vol. 454–455, no. 1, pp. 362–370, 2014, doi: 10.1016/j.virol.2014.01.019.

[17] B. S. Brigham, J. P. Kitzrow, J. P. C. Reyes, K. Musier-Forsyth, and J. B. Munro, “Intrinsic conformational dynamics of the HIV-1 genomic RNA 5’UTR,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 21, pp. 10372–10381, 2019, doi: 10.1073/pnas.1902271116.

[18] E. Skripkin *et al.*, “Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 11, pp. 4945–9, May 1994, doi: 10.1073/pnas.91.11.4945.

[19] M. Ooms, H. Huthoff, R. Russell, C. Liang, and B. Berkhout, “A Riboswitch Regulates RNA Dimerization and Packaging in Human Immunodeficiency Virus Type 1 Virions,” *J. Virol.*, vol. 78, no. 19, pp. 10814–10819, 2004, doi: 10.1128/jvi.78.19.10814-10819.2004.

[20] X. Heng *et al.*, “Identification of a minimal region of the HIV-1 5’-leader required for RNA dimerization, NC binding, and packaging,” *J. Mol. Biol.*, vol. 417, no. 3, pp. 224–239, 2012, doi: 10.1016/j.jmb.2012.01.033.

[21] N. Van Bel, A. T. Das, M. Cornelissen, T. E. M. Abbink, and B. Berkhout, “A short sequence motif in the 5’ leader of the HIV-1 genome modulates extended RNA dimer formation and virus replication,” *J. Biol. Chem.*, vol. 289, no. 51, pp. 35061–35074, 2014, doi: 10.1074/jbc.M114.621425.

[22] D. Muriaux, H. De Rocquigny, B. P. Roques, and J. Paoletti, “NCp7 activates HIV-1Lai RNA dimerization by converting a transient loop-loop complex into a stable dimer,” *J. Biol. Chem.*, vol. 271, no. 52, pp. 33686–33692, 1996, doi: 10.1074/jbc.271.52.33686.

[23] W. Fu, R. J. Gorelick, and A. Rein, “Characterization of human immunodeficiency virus type 1 dimeric RNA from wild-type and protease-defective virions,” *J. Virol.*, vol. 68, no. 8, pp. 5013–5018, 1994, doi: 10.1128/jvi.68.8.5013-5018.1994.

[24] M. Jalalirad and M. Laughrea, “Formation of immature and mature genomic RNA dimers in wild-type and protease-inactive HIV-1: Differential roles of the Gag polyprotein, nucleocapsid proteins NCp15, NCp9, NCp7, and the dimerization initiation site,” *Virology*, vol. 407, no. 2, pp. 225–236, 2010, doi: 10.1016/j.virol.2010.08.013.

[25] J.-I. Sakuragi, S. Ueda, A. Iwamoto, and T. Shioda, “Possible Role of Dimerization in Human Immunodeficiency Virus Type 1 Genome RNA Packaging,” *J. Virol.*, vol. 77, no. 7, pp. 4060–4069, 2003, doi: 10.1128/jvi.77.7.4060-4069.2003.

[26] E. Skripkin, J. C. Paillart, R. Marquet, B. Ehresmann, and C. Ehresmann, “Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 11, pp. 4945–4949, 1994, doi: 10.1073/pnas.91.11.4945.

[27] J. C. Paillart, R. Marquet, E. Skripkin, B. Ehresmann, and C. Ehresmann, “Mutational analysis of the bipartite dimer linkage structure of human immunodeficiency virus type 1 genomic RNA.,” *J. Biol. Chem.*, vol. 269, no. 44, pp. 27486–93, Nov. 1994.

[28] R. S. Russell, J. Hu, M. Laughrea, M. a. Wainberg, and C. Liang, “Deficient dimerization of human immunodeficiency virus type 1 RNA caused by mutations of the u5 RNA sequences.,” *Virology*, vol. 303, no. 1, pp. 152–63, Nov. 2002, doi: 10.1006/viro.2002.1592.

[29] N. Shen, L. Jetté, M. A. Wainberg, and M. Laughrea, “Role of stem B, loop B, and nucleotides next to the primer binding site and the kissing-loop domain in human immunodeficiency virus type 1 replication and genomic-RNA dimerization.,” *J. Virol.*, vol. 75, no. 21, pp. 10543–9, Nov. 2001, doi: 10.1128/JVI.75.21.10543-10549.2001.

[30] R. Marquet, J. christophe Paillart, E. Skripkin, C. Ehresmann, and B. Ehresmann, “Dimerization of human immunodeficiency virus type 1 RNA involves sequences located upstream of the splice donor site,” *Nucleic Acids Res.*, vol. 22, no. 2, pp. 145–151, 1994, doi: 10.1093/nar/22.2.145.

[31] R. Song, J. Kafaie, and M. Laughrea, “Role of the 5′ TAR stem-loop and the U5-AUG duplex in dimerization of HIV-1 genomic RNA,” *Biochemistry*, vol. 47, no. 10, pp. 3283–3293, 2008, doi: 10.1021/bi7023173.

[32] S. C. Keane *et al.*, “RNA structure. Structure of the HIV-1 RNA packaging signal.,” *Science*, vol. 348, no. 6237, pp. 917–21, May 2015, doi: 10.1126/science.aaa9266.

[33] T. E. M. Abbink and B. Berkhout, “A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon.,” *J. Biol. Chem.*, vol. 278, no. 13, pp. 11601–11, Mar. 2003, doi: 10.1074/jbc.M210291200.

[34] J. C. Kenyon, L. J. Prestwood, S. F. J. Le Grice, and A. M. L. Lever, “In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader,” *Nucleic Acids Res.*, vol. 41, no. 18, pp. 1–11, 2013, doi: 10.1093/nar/gkt690.

[35] S. Kharytonchyk *et al.*, “Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 47, pp. 13378–13383, 2016, doi: 10.1073/pnas.1616627113.

[36] J. D. Brown *et al.*, “Structural basis for transcriptional start site control of HIV-1 RNA fate.,” *Science*, vol. 368, no. 6489, pp. 413–417, 2020, doi: 10.1126/science.aaz7959.

[37] C. M. Obayashi, Y. Shinohara, T. Masuda, and G. Kawai, “Influence of the 5′-terminal sequences on the 5′-UTR structure of HIV-1 genomic RNA,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-90427-9.

[38] K. Lu, X. Heng, and M. F. Summers, “Structural determinants and mechanism of HIV-1 genome packaging.,” *J. Mol. Biol.*, vol. 410, no. 4, pp. 609–33, Jul. 2011, doi: 10.1016/j.jmb.2011.04.029.

[39] R. S. Russell *et al.*, “Sequences downstream of the 5’ splice donor site are required for both packaging and dimerization of human immunodeficiency virus type 1 RNA.,” *J. Virol.*, vol. 77, no. 1, pp. 84–96, Jan. 2003.

[40] M. D. Moore, O. A. Nikolaitchik, J. Chen, M. L. Hammarskjöld, D. Rekosh, and W. S. Hu, “Probing the HIV-1 genomic RNA trafficking pathway and dimerization by genetic recombination and single virion analyses,” *PLoS Pathog.*, vol. 5, no. 10, 2009, doi: 10.1371/journal.ppat.1000627.

[41] R. S. Russell, C. Liang, and M. A. Wainberg, “Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably?,” *Retrovirology*, vol. 1, no. Mlv, pp. 1–14, 2004, doi: 10.1186/1742-4690-1-23.

[42] L. Ye *et al.*, “RNA Structures and Their Role in Selective Genome Packaging,” *Viruses*, vol. 13, no. 9, 2021, doi: 10.3390/v13091788.

[43] A. Lever, H. Gottlinger, W. Haseltine, and J. Sodroski, “Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions,” *J. Virol.*, vol. 63, no. 9, pp. 4085–4087, Sep. 1989, doi: 10.1128/jvi.63.9.4085-4087.1989.

[44] A. Aldovini and R. A. Young, “Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus,” *J. Virol.*, vol. 64, no. 5, pp. 1920–1926, 1990, doi: 10.1128/jvi.64.5.1920-1926.1990.

[45] F. Clavel and J. M. Orenstein, “A mutant of human immunodeficiency virus with reduced RNA packaging and abnormal particle morphology,” *J. Virol.*, vol. 64, no. 10, pp. 5230–5234, 1990, doi: 10.1128/jvi.64.10.5230-5234.1990.

[46] A. Rein, “RNA Packaging in HIV,” *Trends Microbiol.*, vol. 27, no. 8, pp. 715–723, 2019, doi: 10.1016/j.tim.2019.04.003.

[47] E. W. A. El-Wahab *et al.*, “Specific recognition of the HIV-1 genomic RNA by the Gag precursor,” *Nat. Commun.*, vol. 5, 2014, doi: 10.1038/ncomms5304.

[48] R. P. Smyth *et al.*, “In cell mutational interference mapping experiment (in cell MIME) identifies the 5’ polyadenylation signal as a dual regulator of HIV-1 genomic RNA production and packaging,” *Nucleic Acids Res.*, vol. 46, no. 9, pp. 1–16, 2018, doi: 10.1093/nar/gky152.

[49] K. Lu *et al.*, “NMR detection of structures in the HIV-1 5’-leader RNA that regulate genome packaging.,” *Science*, vol. 334, no. 6053, pp. 242–5, Oct. 2011, doi: 10.1126/science.1210460.

[50] E. Mailler, S. Bernacchi, R. Marquet, J. C. Paillart, V. Vivet-Boudou, and R. P. Smyth, “The life-cycle of the HIV-1 gag–RNA complex,” *Viruses*, vol. 8, no. 9, pp. 1–19, 2016, doi: 10.3390/v8090248.

[51] J. C. Kenyon, L. J. Prestwood, and A. M. L. Lever, “A novel combined RNA-protein interaction analysis distinguishes HIV-1 Gag protein binding sites from structural change in the viral RNA leader,” *Sci. Rep.*, vol. 5, no. August, pp. 1–11, 2015, doi: 10.1038/srep14369.

[52] R. P. Smyth *et al.*, “Mutational interference mapping experiment (MIME) for studying RNA structure and function,” *Nat. Methods*, vol. 12, no. 9, pp. 866–872, 2015, doi: 10.1038/nmeth.3490.

[53] E. Mailler, J. C. Paillart, R. Marquet, R. P. Smyth, and V. Vivet-Boudou, “The evolution of RNA structural probing methods: From gels to next-generation sequencing,” *Wiley Interdiscip. Rev. RNA*, vol. 10, no. 2, pp. 1–20, 2019, doi: 10.1002/wrna.1518.

[54] X. W. Wang, C. X. Liu, L. L. Chen, and Q. C. Zhang, “RNA structure probing uncovers RNA structure-dependent biological functions,” *Nat. Chem. Biol.*, vol. 17, no. 7, pp. 755–766, 2021, doi: 10.1038/s41589-021-00805-7.

[55] P. C. Bevilacqua and S. M. Assmann, “Technique development for probing RNA structure in vivo and genome-wide,” *Cold Spring Harb. Perspect. Biol.*, vol. 10, no. 10, 2018, doi: 10.1101/cshperspect.a032250.

[56] M. Zubradt, P. Gupta, S. Persad, A. M. Lambowitz, J. S. Weissman, and S. Rouskin, “DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo,” *Nat. Methods*, vol. 14, no. 1, pp. 75–82, 2016, doi: 10.1038/nmeth.4057.

[57] D. Mitchell, S. M. Assmann, and P. C. Bevilacqua, “Probing RNA structure in vivo,” *Curr. Opin. Struct. Biol.*, vol. 59, no. 814, pp. 151–158, 2019, doi: 10.1016/j.sbi.2019.07.008.

[58] D. Incarnato, E. Morandi, L. M. Simon, and S. Oliviero, “RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications,” *Nucleic Acids Res.*, vol. 46, no. 16, p. e97, Sep. 2018, doi: 10.1093/nar/gky486.

[59] C. Y. Cheng, W. Kladwang, J. D. Yesselman, and R. Das, “RNA structure inference through chemical mapping after accidental or intentional mutations,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 37, pp. 9876–9881, 2017, doi: 10.1073/pnas.1619897114.

[60] L. Ye *et al.*, “Short- and long-range interactions in the HIV-1 5’ UTR regulate genome dimerization and packaging,” *Nat. Struct. Mol. Biol.*, vol. 29, no. 4, pp. 306–319, 2022, doi: 10.1038/s41594-022-00746-2.

[61] J. S. Gibbs, D. A. Regier, and R. C. Desrosiers, “Construction and in vitro properties of HIV-1 mutants with deletions in ‘nonessential’ genes.,” *AIDS Res. Hum. Retroviruses*, vol. 10, no. 4, pp. 343–50, Apr. 1994, doi: 10.1089/aid.1994.10.343.

[62] W. J. McKinstry *et al.*, “Expression and purification of soluble recombinant full length HIV-1 Pr55Gag protein in Escherichia coli,” *Protein Expr. Purif.*, vol. 100, pp. 10–18, 2014, doi: 10.1016/j.pep.2014.04.013.

[63] M. R. Smith, R. P. Smyth, R. Marquet, and M. Von Kleist, “MIMEAnTo: Profiling functional RNA in mutational interference mapping experiments,” *Bioinformatics*, vol. 32, no. 21, pp. 3369–3370, 2016, doi: 10.1093/bioinformatics/btw479.

[64] S. Busan and K. M. Weeks, “Accurate detection of chemical modifications in RNA by

mutational profiling (MaP) with ShapeMapper 2.," *RNA*, Nov. 2017, doi: 10.1261/rna.061945.117.

[65] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman, "Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo," *Nature*, vol. 505, no. 7485, pp. 701–705, Jan. 2014, doi: 10.1038/nature12894.

[66] D. Incarnato, F. Neri, F. Anselmi, and S. Oliviero, "RNA structure framework: Automated transcriptome-wide reconstruction of RNA secondary structures from high-throughput structure probing data," *Bioinformatics*, vol. 32, no. 3, pp. 459–461, 2016, doi: 10.1093/bioinformatics/btv571.

[67] R. Lorenz *et al.*, "ViennaRNA Package 2.0," *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 26, Nov. 2011, doi: 10.1186/1748-7188-6-26.

[68] M. J. Smola, J. M. Calabrese, and K. M. Weeks, "Detection of RNA-Protein Interactions in Living Cells with SHAPE," *Biochemistry*, vol. 54, no. 46, pp. 6867–6875, 2015, doi: 10.1021/acs.biochem.5b00977.

[69] G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet, "Alignments of RNA structures," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 7, no. 2, pp. 309–322, 2010, doi: 10.1109/TCBB.2008.28.

[70] P. Cordero, J. B. Lucks, and R. Das, "An RNA mapping data base for curating RNA structure mapping experiments," *Bioinformatics*, vol. 28, no. 22, pp. 3006–3008, 2012, doi: 10.1093/bioinformatics/bts554.

[71] J. I. Sakuragi and a T. Panganiban, "Human immunodeficiency virus type 1 RNA outside the primary encapsidation and dimer linkage region affects RNA dimer stability in vivo.," *J. Virol.*, vol. 71, no. 4, pp. 3250–4, Apr. 1997.

[72] K. L. Jones, S. Sonza, and J. Mak, "Primary T-lymphocytes rescue the replication of HIV-1 DIS RNA mutants in part by facilitating reverse transcription.," *Nucleic Acids Res.*, vol. 36, no. 5, pp. 1578–88, 2008, doi: 10.1093/nar/gkm1149.

[73] J. L. Clever and T. G. Parslow, "Mutant human immunodeficiency virus type 1 genomes with defects in RNA dimerization or encapsidation.," *J. Virol.*, vol. 71, no. 5, pp. 3407–14, May 1997.

[74] P. Cordero, W. Kladwang, C. C. Vanlang, and R. Das, "Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference," *Biochemistry*, vol. 51, no. 36, pp. 7037–7039, 2012, doi: 10.1021/bi3008802.

[75] C. Pop *et al.*, "Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation.," *Mol. Syst. Biol.*, vol. 10, p. 770, 2014, doi: 10.15252/msb.20145524.

[76] S. Tian and R. Das, "RNA structure through multidimensional chemical mapping.," *Q. Rev. Biophys.*, vol. 49, no. 37, p. e7, Sep. 2016, doi: 10.1017/S0033583516000020.

[77] J. C. Paillart *et al.*, "A dual role of the putative RNA dimerization initiation site of human immunodeficiency virus type 1 in genomic RNA packaging and proviral DNA synthesis.," *J. Virol.*, vol. 70, no. 12, pp. 8348–54, Dec. 1996.

[78] C. Gavazzi *et al.*, "A functional sequence-specific interaction between influenza A virus genomic RNA segments," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 41, pp. 16604–16609, 2013, doi: 10.1073/pnas.1314419110.

[79] C. Gavazzi *et al.*, "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 Res.*, vol. 41, no. 2, pp. 1241–1254, 2013, doi: 10.1093/nar/gks1181.

[80] T. Fajardo, P. Y. Sung, C. C. Celma, and P. Roy, "Rotavirus genomic RNA complex forms via specific RNA–RNA interactions: Disruption of RNA complex inhibits virus infectivity," *Viruses*, vol. 9, no. 7, pp. 1–15, 2017, doi: 10.3390/v9070167.

[81] T. Masuda *et al.*, "Fate of HIV-1 cDNA intermediates during reverse transcription is dictated by transcription initiation site of virus genomic RNA," *Sci. Rep.*, vol. 5, no. December, pp. 1–15, 2015, doi: 10.1038/srep17680.

[82] V. Goldschmidt *et al.*, "Structural variability of the initiation complex of HIV-1 reverse transcription," *J. Biol. Chem.*, vol. 279, no. 34, pp. 35923–35931, 2004, doi: 10.1074/jbc.M404473200.

[83] V. Goldschmidt, M. Rigourd, C. Ehresmann, S. F. J. Le Grice, B. Ehresmann, and R. Marquet, "Direct and indirect contributions of RNA secondary structure elements to the initiation of HIV-1 reverse transcription," *J. Biol. Chem.*, vol. 277, no. 45, pp. 43233–43242, 2002, doi: 10.1074/jbc.M205295200.

[84] O. A. Nikolaitchik, X. Somoulay, J. M. O. Rawson, J. A. Yoo, V. K. Pathak, and W.-S. Hu, "Unpaired Guanosines in the 5' Untranslated Region of HIV-1 RNA Act Synergistically To Mediate Genome Packaging.," *J. Virol.*, vol. 94, no. 21, 2020, doi: 10.1128/JVI.00439-20.

[85] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks, "Accurate SHAPE-directed RNA structure determination.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 1, pp. 97–102, Jan. 2009, doi: 10.1073/pnas.0806929106.

[86] W. Kladwang, C. C. VanLang, P. Cordero, and R. Das, "A two-dimensional mutate-and-map strategy for non-coding RNA structure.," *Nat. Chem.*, vol. 3, no. 12, pp. 954–62, Oct. 2011, doi: 10.1038/nchem.1176.

[87] J. L. Clever, M. L. Wong, and T. G. Parslow, "Requirements for kissing-loop-mediated dimerization of human immunodeficiency virus RNA.," *J. Virol.*, vol. 70, pp. 5902–5908, 1996.

[88] H. Huthoff, K. Bugala, J. Barciszewski, and B. Berkhout, "On the importance of the primer activation signal for initiation of tRNA(lys3)-primed reverse transcription of the HIV-1 RNA genome.," *Nucleic Acids Res.*, vol. 31, no. 17, pp. 5186–94, Sep. 2003, doi: 10.1093/nar/gkg714.

[89] W. Kasprzak, E. Bindewald, and B. a Shapiro, “Structural polymorphism of the HIV-1 leader region explored by computational methods.,” *Nucleic Acids Res.*, vol. 33, no. 22, pp. 7151–63, Jan. 2005, doi: 10.1093/nar/gki1015.

[90] H. Huthoff and B. Berkhout, “Two alternating structures of the HIV-1 leader RNA.,” *RNA*, vol. 7, no. 1, pp. 143–57, Jan. 2001, doi: 10.1017/s1355838201001881.

[91] T. E. M. Abbink, M. Ooms, P. C. J. Haasnoot, and B. Berkhout, “The HIV-1 Leader RNA Conformational Switch Regulates RNA Dimerization but Does Not Regulate mRNA Translation †,” *Biochemistry*, vol. 44, no. 25, pp. 9058–9066, Jun. 2005, doi: 10.1021/bi0502588.

[92] M. Ooms, K. Verhoef, E. Southern, H. Huthoff, and B. Berkhout, “Probing alternative foldings of the HIV-1 leader RNA by antisense oligonucleotide scanning arrays.,” *Nucleic Acids Res.*, vol. 32, no. 2, pp. 819–27, 2004, doi: 10.1093/nar/gkh206.

[93] K. a Wilkinson *et al.*, “High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states.,” *PLoS Biol.*, vol. 6, no. 4, p. e96, Apr. 2008, doi: 10.1371/journal.pbio.0060096.

[94] J. M. Watts *et al.*, “Architecture and secondary structure of an entire HIV-1 RNA genome,” *Nature*, vol. 460, no. 7256, pp. 711–716, 2009, doi: 10.1038/nature08237.

[95] N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson, and K. M. Weeks, “RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP).,” *Nat. Methods*, Jul. 2014, doi: 10.1038/nmeth.3029.

[96] J. C. Paillart, M. Dettenhofer, X. F. Yu, C. Ehresmann, B. Ehresmann, and R. Marquet, “First snapshots of the HIV-1 RNA structure in infected cells and in virions,” *J. Biol. Chem.*, vol. 279, pp. 48397–48403, 2004, doi: 10.1074/jbc.M408294200.

[97] L. Balvay, M. L. Lastra, B. Sargueil, J. L. Darlix, and T. Ohlmann, “Translational control of retroviruses,” *Nat. Rev. Microbiol.*, vol. 5, no. 2, pp. 128–140, 2007, doi: 10.1038/nrmicro1599.

[98] S. Bernacchi *et al.*, “HIV-1 Pr55 Gag binds genomic and spliced RNAs with different affinity and stoichiometry,” *RNA Biol.*, vol. 14, no. 1, pp. 90–103, Jan. 2017, doi: 10.1080/15476286.2016.1256533.

[99] L. Houzet *et al.*, “HIV controls the selective packaging of genomic, spliced viral and cellular RNAs into virions through different mechanisms.,” *Nucleic Acids Res.*, vol. 35, no. 8, pp. 2695–704, Jan. 2007, doi: 10.1093/nar/gkm153.

[100] E. Seif, M. Niu, and L. Kleiman, “Annealing to sequences within the primer binding site loop promotes an HIV-1 RNA conformation favoring RNA dimerization and packaging.,” *RNA*, vol. 19, no. 10, pp. 1384–93, Oct. 2013, doi: 10.1261/rna.038497.113.

Chapter 4 Defining the architecture of the influenza RNA genome by RNA-RNA-seq

This project is ongoing and all data is unpublished.

Contributions: Redmond Smyth and Liqing Ye designed the project, Liqing Ye performed all experiments unless otherwise stated, Uddhav Ambi contributed to virus propagation; Liqing Ye and Redmond Smyth performed the sequencing data analysis.

Abstract

Influenza A viruses are responsible for recurrent epidemics and occasional pandemics, and are a major burden on public health worldwide. Their segmented genome, composed of eight negative-sense viral RNAs (vRNAs), complicates virus assembly but offers evolutionary advantages by enabling reassortment. Current evidence suggests that influenza vRNAs are organized during assembly into a supramolecular complex. However, its molecular details are poorly understood. Our goal is to define the quaternary RNA architecture of the genome in virions by identifying sites of interaction between the eight vRNAs. To this end, we developed RNA-RNA-seq, which can measure direct (RNA-RNA) and indirect (protein-mediated) interactions without being limited by specific protein or RNA baits. I optimized each step of the protocol and proved that RNA-RNA seq worked efficiently on model substrate, like HIV DIS RNA, purified ribosome, as well as influenza A virus infected cells.

4.1 Introduction

4.1.1 Overview of Influenza viruses

Influenza viruses belong to the Orthomyxoviridae family, a family of enveloped viruses composed of negative-sense (-), single-stranded, segmented viral (v)RNA genomes. There are four types of influenza viruses: influenza A, B, C and D[1]–[9]. Influenza virus infects about 5-10% adults and 20-30% children, and is responsible for about 250,000 to 500,000 deaths each year according to the World Health Organization (https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)). Both Influenza virus types A and B are common causes of acute respiratory illnesses, while influenza A viruses are the main cause for large epidemics with high mortality. In 1918, the Spanish flu (A/H1N1) killed about 50 million people[10]. Later in 1957, 1.1 million people died from the Asian flu (A/H2N2), and the 1968 flu (A/H2N2) pandemic resulted in an estimated 1 to 4 million deaths[11], [12] (https://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html).

Influenza A viruses are classified based on their glycoproteins hemagglutinin (HA) and neuraminidase (NA). There are 18 HA subtypes and 11 NA subtypes have been identified until now (https://www.cdc.gov/flu/avianflu/influenza-a-virus-subtypes.htm). On the other hand, influenza B has only one type of HA and NA, influenza C and D have only one surface glycoprotein, hemagglutinin-esterase-fusion (HEF), which has similar functions to HA and NA of influenza A and B[15], [16]. The natural reservoir for Influenza A viruses is wild aquatic birds, in which they often cause enteric infections with no apparent disease. In a range of mammals, such as humans, horses, canines, pigs, influenza viruses can also infect and cause mild or severe respiratory diseases. The symptoms for infected humans include headache, fever, muscle and joint pain, sore throat, dry cough, runny nose and weakness[1], [13], [14].

4.1.2 Influenza virus structure

Influenza viruses have lipid membranes derived from the host cell, embedded with surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA). The matrix protein (M1) lies just beneath the envelope, and the core of the virus particle is a large complex consisting of segmented negative genomic RNAs (eight segments for the influenza A and B viruses, seven for the influenza C and D viruses). Each segment encodes at least one essential protein[7]–[9], [13], [17] (**Figure 4.1a, Table 4.1**): PB1, PB2, PA, HA, NP, NA, M, NS. Consequently, every infective virion must contain 8 segments. Consequently, every infective virion must contain 8 segments. Every segment is packaged with NP protein and polymerase into RNP (**Figure 4.1b**).

Each genome segment is bound to multiple copies of nucleoprotein (NP) and single copy of the heterotrimeric polymerase complex (PA, PB1 and PB2) [17], [18](**Figure 4.1**). The overall composition of viral particles is about 1% RNA, 5-8% carbohydrate, 20% lipid, and approximately 70% protein.

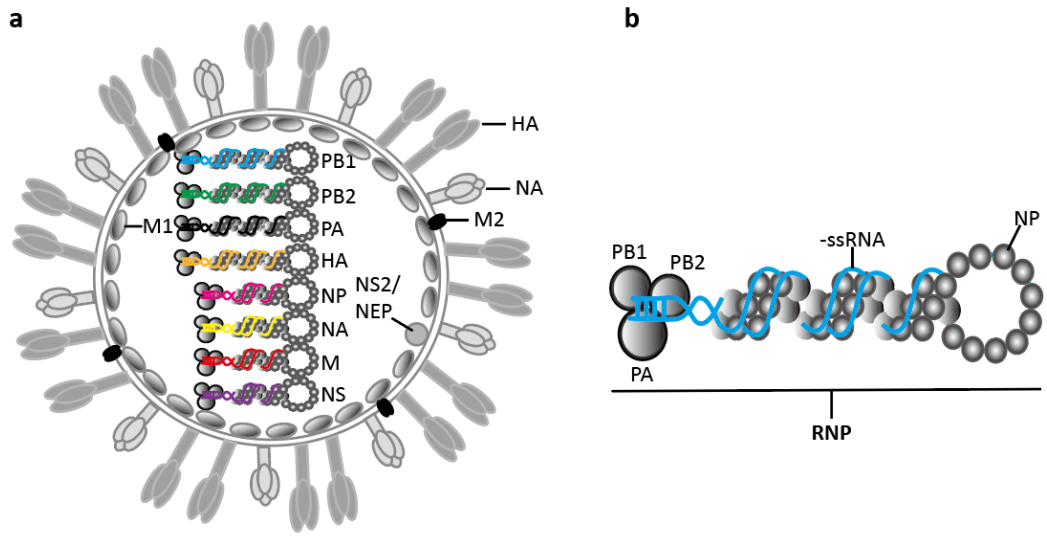


Figure 4.1| Influenza A virus structure and genome. (a) Influenza virus has a segmented genome comprising 8 negative-stranded RNAs, and each segment encodes an essential protein; **(b)** Influenza virus genome is packaged as ribonucleoprotein (RNP), in which the negative single strand RNA is bound by a heterotrimeric polymerase complex and nucleocapsid protein (NP).

Table 4.1 Eight segments of influenza A virus and the encoding proteins and function [19]

VRNA	LENGTH (NT)	PROTEIN	FUNCTIONS
PB2	2341	PB2	Cellular mRNA cap recognition and binding
PB1	2341	PB1	RNA-dependent RNA polymerase
PA	2233	PA	Endonuclease
HA	1778	HA	Surface glycoprotein, cellular receptor binding and viral endosome fusion
NP	1565	NP	Nucleo protein, vRNA binding, vRNP nuclear export and vRNA replication
NA	1413	NA	Neuraminidase (cleavage between HA and sialic acid)
M	1027	M1	Matrix protein, vRNP nuclear export, and virus budding
		M2	Ionic channel
NS	890	NS1	Inhibition of immune response
		NS2/NEP	vRNP nuclear export

4.1.3 Influenza A virus life cycle

The first step for influenza virus infection is binding of HA molecules on the viral envelope to sialic acids (SAs) on the surface of cells. Typically, avian HAs prefer to bind with α-2,3-linked SAs that have a "linear" structure, while human adapted HAs have higher specificity for α-2,6 linkage SAs. HA-mediated binding to the receptor triggers endocytosis of the virion[8]. The low pH in the endosome activates the M2 ion channel and causes a large conformational change in HA that triggers the fusion peptide to fuse viral and endosomal membranes, leading to vRNPs release into the cytoplasm. Subsequently, vRNPs traffic to the nucleus by hijacking the host cell machinery and transport pathways[18]. Inside the nucleus, the heterotrimeric viral RNA-dependent RNA polymerase carries out transcription and replication of the vRNAs to produce messenger RNA (mRNA), complementary RNA (cRNA) and vRNA[20][21]. mRNAs are primed by a "cap snatching" mechanism, which provides the host like 5'methlyated cap structure[13], [22], [23]. mRNAs are then transported to the cytoplasm to be translated into proteins[24]. The newly translated NP, PB1, PB2, PA are transported back to the nucleus for RNPs assembly. HA, NA, M2 are co-translationally directed to the endoplasmic reticulum (ER) and translocated through Golgi to the cytoplasmic membrane for virus assembly and release. Replication of influenza genome is carried out in two steps: cRNA synthesis and vRNA transcription. cRNA synthesis starts with the *de novo* synthesis of a di-nucleotide complementary to the 3' end of vRNA[13][25]. This di-nucleotide functions as transcription primer for elongation. vRNAs are produced using a similar strategy, except they are transcribed using cRNAs as template. Newly transcribed vRNAs then assemble in the nucleus with translated nucleoprotein (NP) and polymerase subunit (PB1, PB2, and PA) into vRNPs. Finally, vRNPs are exported to the cytoplasm by M1, NEP proteins, CRM1 (also known as exportin-1). vRNPs are then thought to associate in the cytoplasm into sub-bundles on Rab11+ vesicles where they are transferred to the plasma membrane for assembly and budding[13], [21], [26].

4.1.4 Influenza A virus genome packaging

Historically, there were two different packaging models for influenza [17], [27], [28] (**Figure 4.2a**). The random packaging model hypothesizes that influenza virus packages its genomic segments randomly, and as a consequence only a few viruses would be infectious (**Figure 4.2b**). The selective model hypothesizes that each segment is selectively packaged into virion, which is much more efficient than the random model, but would presumably require a complex assembly mechanism to allow the virus to discriminate between different segments. Current evidence suggests that influenza viruses use the selective model to assemble the segmented genome into viral particles. In support of this view, Cryo-EM and cryo-tomography imaging showed most virus particle contained 8 segments, organized into a characteristic 7+1 structure [29]–[31] (**Figure 4.2c**, **4.2d**, **4.2e**). Each vRNA is packaged as a rod like vRNP, which contains polymerase and NP. But it is still unclear how these 8 segments reassort and assemble. From imaging studies, we can observe connections between segments (**arrows shown in Figure 4.2d**) [30], but its molecular details underlying these connections remain unknown. These putative connections might be direct RNA-RNA interactions or alternatively mediated by protein-protein or protein-RNA interaction.

Our goal is to define the quaternary RNA architecture of influenza genomes in cells and virions, by identifying sites of direct and indirect RNA-RNA interactions between the 8 vRNPs. To this end, I have developed a proximity ligation and next generation sequencing strategy to define the global vRNA organization (inter-molecular RNA interactions) and local RNA structure (intra-molecular RNA interactions) of an entire influenza A virus particle.

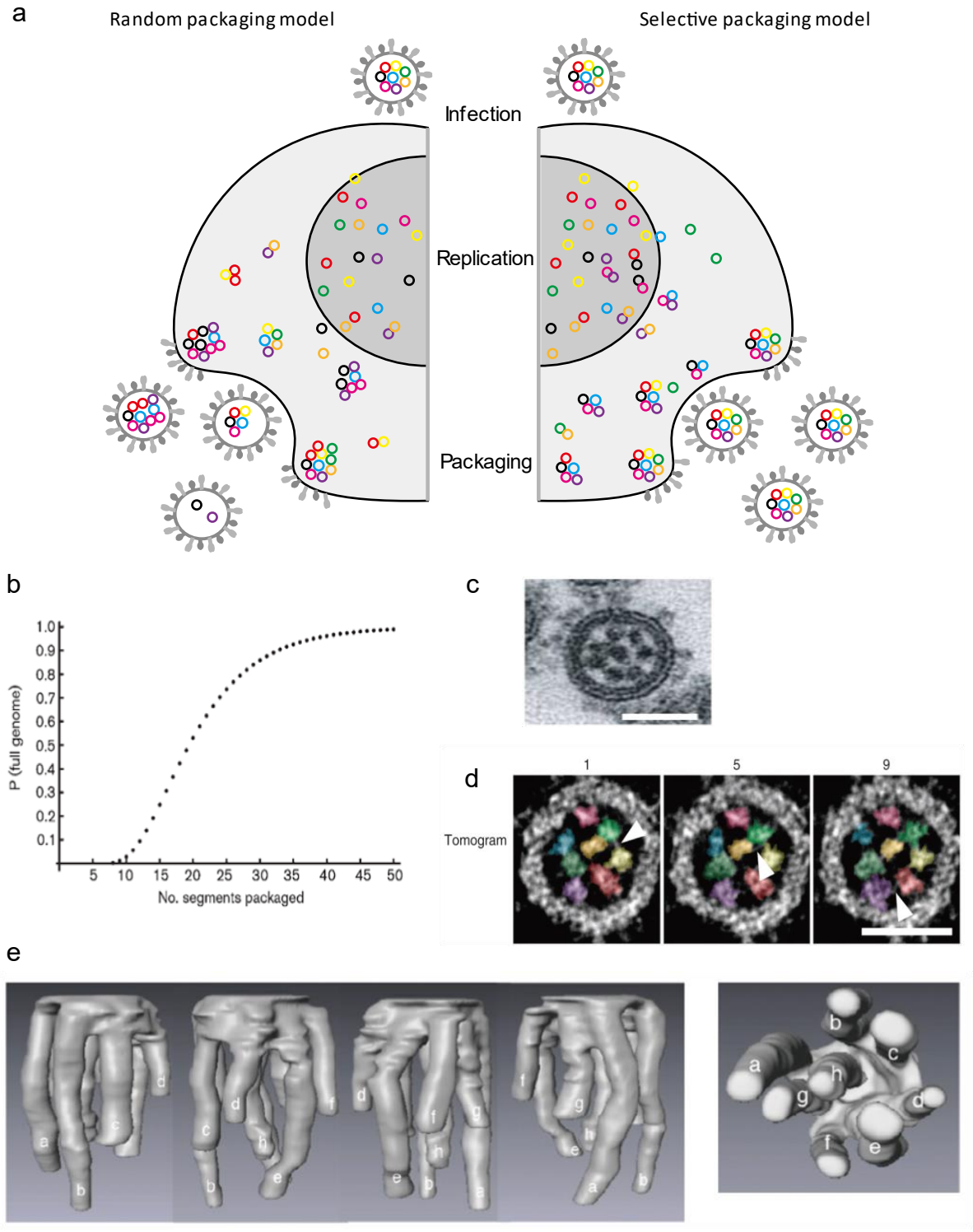

a
Random packaging model
Selective packaging model
Infection
Replication
Packaging
b
P (full genome)
No. segments packaged
c
d
Tomogram
1
5
9
e

Figure 4.2| Influenza A virus genome packaging. (a) random packaging model and selective packaging model[27]; **(b)** A simple mathematical model of random genome packaging. The probabilities of obtaining one copy of each segment in a single virus particle are plotted for random selection with increasing numbers of RNPs per virion [using standard probability theory (Enami et al., 1991)][17][32]; **(c)** Electron-dense dots of influenza A virus[30]; **(d)** Tomograms[31] to show that the 8 segments of influenza virus are organized into 7+1 structure; **(e)** 3D surface rendering of the vRNPs in budding H3N2 influenza A virions from electron tomography image[29].

4.1.4 High throughput sequencing-based RNA-RNA interaction methods

RNA-RNA interactions can be mediated either by direct base pairing or indirectly through RNA binding proteins (RBPs). There are several published methodologies for RNA-RNA interaction identification[33]–[43], and all have a similar workflow. In brief, duplexes of RNA are crosslinked using a variety of reagents, fragmented, and then ligated and sequenced (**Figure 4.3**). The first approach developed for such purpose is called 'cross-linking, ligation and sequencing of hybrids' (CLASH)[44] and has been successfully applied to identify *in vivo* targets of small nucleolar RNAs (snoRNAs)[44], microRNA[45] and Piwi-interacting RNA[46]. UV light can be used to crosslink RNA-RNA duplexes and RBP-RNA interactions. In CLASH, RNAs are crosslinked to proteins using UV irradiation followed by purification, fragmentation, proximity ligation, sequencing and bioinformatics analysis. An interaction between two RNAs is revealed as chimeric reads mapped to two different transcripts. Similar to CLASH, RNA hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation (hiCLIP)[33][34] can also identify RNA duplexes bound by a specific RBP. The workflow includes *in vivo* UV crosslinking, partial RNA digestion, RNA duplex purification by immunoprecipitation, adaptor ligation, proximity ligation, reverse transcription,library amplification, and sequencing. hiCLIP uses two adaptors during ligation, which demarks the boundary of interacted RNAs, thus achieving higher specificity. RNA antisense purification (RAP)[47] and RNA interactome analysis followed by deep sequencing (RIA-seq)[48] are similar methods, and they are designed to identify the interaction partners for a specific target RNA. In RAP/RIA-seq, the UV crosslinked interacted RNAs are pulled down by biotinylated anti-sense DNA probes (50-120nt for RAP and around 20nt for RIA), which are designed to target the full-length of the investigated RNA. After RNA pulldown, the enriched interacted RNAs are eluted, reverse transcribed into cDNAs, and sequenced.

However, these methods can only study interactions between specific RBP-bound RNAs or for one target RNA, which limits the exploration of unknown RNA-RNA interactions.

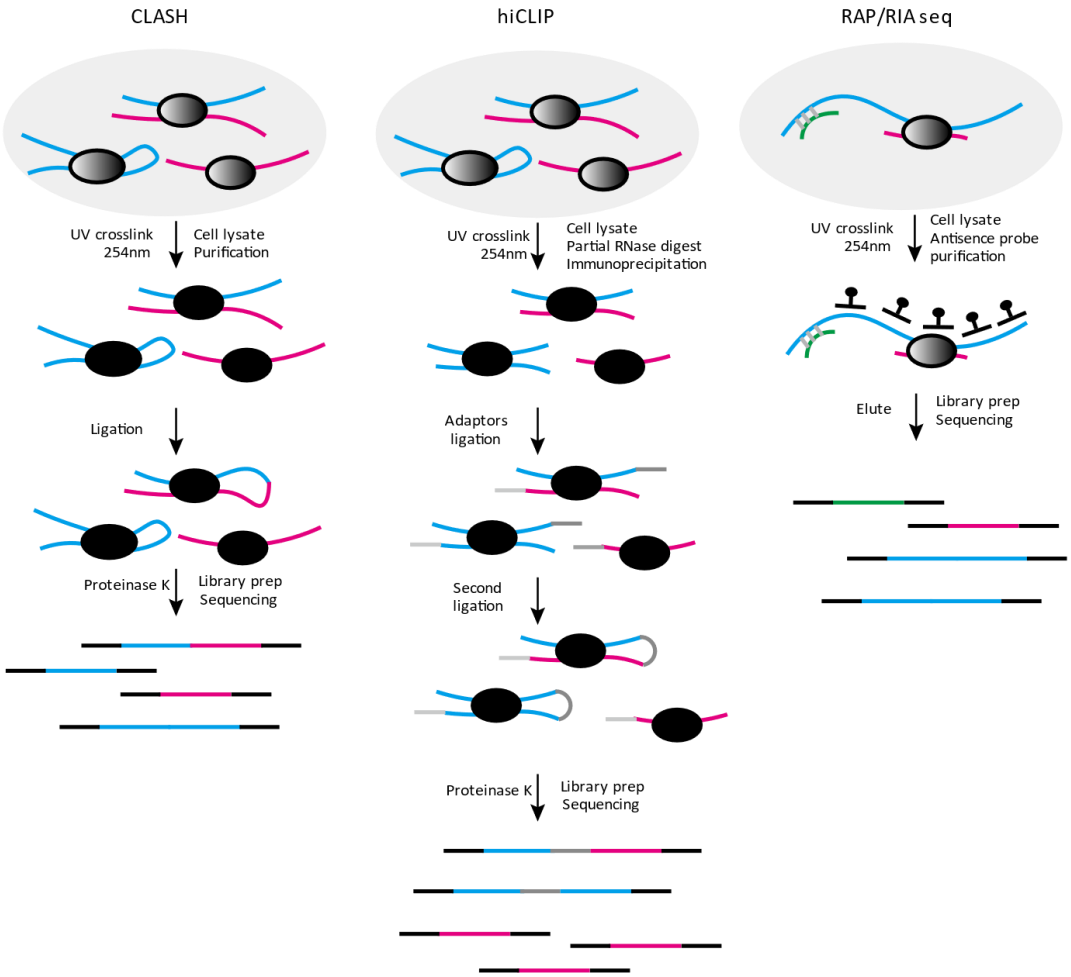


Figure 4.3| High throughput sequencing-based methods to study specific interactions between specific types of RNAs or for one target RNA. In CLASH, RNAs are crosslinked to proteins under UV irradiation followed by protein pull down purification, fragmentation, proximity ligation, proteinase K to digest the crosslinked protein, sequencing and bioinformatics analysis; The workflow of hiCLIP is similar to CLASH, but it includes adaptors ligation before proximity ligation; in RAP/RIA seq, the UV crosslinked RNAs are pulled down and purified by antisense DNA probe, then the interacted RNAs are eluted and sequenced.

To study RNA-RNA interactions transcriptome-wide, more sophisticated methods have been developed that have the potential to cover all RNAs in a cell, such as PARIS, SPLASH, LIGR-seq, COMRADES, MARIO and RIC-seq (**Figure 4.4**). These methods are structurally similar, including the main steps: crosslinking, enrichment of crosslinked interacted RNAs, proximity ligation, sequencing, and bioinformatics analysis.

Psoralen Analysis of RNA Interactions and Structures (PARIS)[36], [37] employs a psoralen derivative 4'-aminomethyltrioxsalen (AMT) as the nucleic acid crosslinker for RNA base pairs. RNA fragments

are purified with partial RNase and complete proteinase digestion. Crosslinked duplexes are then enriched through two-dimension electrophoresis. The enriched and purified RNA duplexes are proximity ligated, reverse crosslinked and sequenced. Similar to PARIS, Sequencing of Psoralen crosslinked, Ligated, And Selected Hybrids (SPLASH)[42], [49]–[51] and Ligation of Interacting RNA followed by high-throughput sequencing (LIGR-seq)[43] also uses AMT crosslinking. SPLASH crosslinks RNA-RNA duplexes using a biotinylated psoralen and enriches the crosslinked duplex by biotin selection. LIGR-seq applies a circularization step and enriches the crosslinked RNA duplex by RNase R, which has a 3'-->5' exoribonuclease activity to digest the un-circularized RNA. Crosslinking of matched RNAs and deep sequencing (COMRADES)[35], [52] uses psoralen-triethylene glycol azide to crosslink interacted RNAs and following with biotinylated DNA probes to pulldown the target RNA. After fragmentation, crosslinked RNAs are labelled with biotin by click chemistry, which allows a second streptavidin-based affinity selection of cross-linked regions. Then the enriched duplexes are proximity ligated and reverse crosslinked, and sequenced.

Besides the methods which detect direct RNA-RNA interaction, Mapping RNA interactome *in vivo* (MARIO)[41] is designed to study the RNA-RNA interactions mediated by protein. RNA *In situ* Conformation sequencing (RIC-seq) can detect both direct RNA-RNA interactions, and indirect interactions through proximity [39], [53], [54]. MARIO applies UV-C (254 nm), formaldehyde and EthylGlycol bis to crosslink RNA-protein and protein-protein interactions, followed by pulldown the crosslinked complex with the biotinylated protein, labelling with a biotinylated adaptor, proximity ligation, and finally high throughput sequencing. Similarly, RIC-seq uses formaldehyde to crosslink the RNA-RNA interactions *in situ*, and enriches chimeric reads using a biotinylated cytidine (bis) phosphate (pCp–biotin), after that proximity ligation join the interacting RNAs, then the duplex is purified and sequenced (**Figure 4.4**).

The biggest advantage of these high throughput sequencing-based methods is that they can identify unknown RNA-RNA interactions at transcriptome level, which can be mapped with high resolution. But none of them can perfectly match our goal. For example, the PARIS protocol uses 2D gels, which needs a lot of material, while the amount obtainable from virus RNA is low. SPLASH and LIGR-seq use AMT, or psoralen, for crosslinking, which specifically crosslinks RNA-RNA interactions. But it cannot yet be excluded that influenza virus assembly is mediated by protein-based interactions. For the same reason, we cannot use MARIO for our project, because it is used to detect only interactions mediated by protein. RIC-seq was originally designed for cell samples. Later adaptations for virus samples, also required large amounts of starting material.

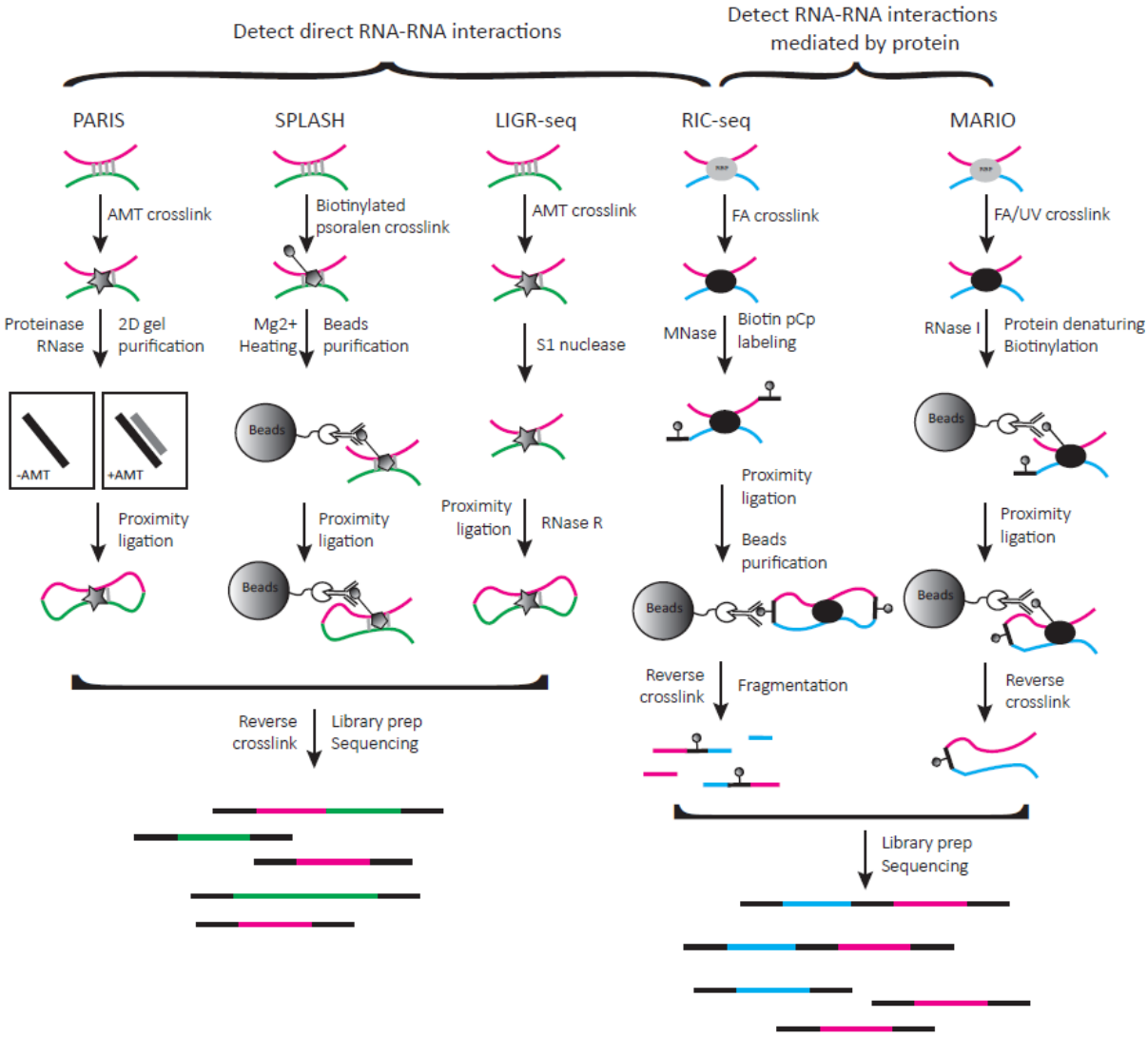


Figure 4.4| High throughput sequencing-based methods to study transcriptome-wide RNA-RNA interactions. There are two types of methods: to detect direct RNA-RNA interactions (PARIS, SPLASH, LIGR-seq, RIC-seq) and to detect the interactions mediated by proteins (MARIO, RIC-seq). In the direct RNA-RNA interaction methods, the RNAs are crosslinked by psoralen or its derived chemical AMT, following by fragmentation, then the crosslinked RNAs are purified, proximity ligated, reverse crosslinked and sequenced. The difference is that PARIS applies 2D-gel electrophoresis to purify crosslinked RNAs, while SPLASH uses magnetic beads to pull down the crosslinked RNA and LIGR-seq takes the advantage of RNase R to enrich crosslinked and ligated RNAs. In MARIO, RNAs are crosslinked by formaldehyde, following by fragmentation, beads purification, proximity ligation, proteinase K to reverse crosslink, and sequencing; the workflow of RIC-seq is similar to MARIO, but the protocol is performed *in situ* until the proximity ligation and beads purification, so it also detects direct RNA-RNA interactions.

Table 4.2 Comparison of high throughput sequencing-based methods to study RNA-RNA interactions

METHODS	ADVANTAGES	LIMITATIONS
CLASH	✓ High throughput ✓ Stringent purification conditions remove non-physiological interactions	○ Requires prior knowledge of an RNA-binding protein ○ Requires an antibody ○ Only detects interactions mediated by protein
HICLIP	✓ High throughput ✓ Adaptors incorporation demark the interacted RNAs boundary	○ Requires prior knowledge of an RNA-binding protein. ○ Requires an antibody ○ Only detects interactions mediated by protein
RIA/RAP SEQ	✓ High throughput ✓ Stringent purification conditions	○ Only identifies RNA interactions with specific RNA
PARIS	✓ High throughput ✓ Many-to-many mapping	○ 4'-Aminomethyl trioxsalen (AMT) preferentially crosslinks pyrimidine bases and may introduce bias ○ Low crosslinking efficiency ○ Only detects RNA-RNA direct interactions ○ Gel separation requires a lot of start material
SPLASH	✓ High throughput ✓ Improves signal-to-noise ratio by leveraging biotinylated psoralen ✓ Many-to-many mapping ✓ Biotin pull down	○ Psoralen preferentially crosslinks pyrimidine bases and may introduce bias ○ Low crosslinking efficiency ○ Only detects RNA-RNA direct interactions

LIGR-SEQ	✓ High throughput ✓ Many-to-many mapping	○ AMT preferentially crosslinks pyrimidine bases and may introduce bias ○ Low crosslinking efficiency ○ Only detects RNA-RNA direct interactions
MARIO	✓ High throughput ✓ Many-to-many mapping ✓ Biotin pull down ✓ Incorporation of an adaptor between two RNA molecules increases ligation efficiency and improves accuracy in sequence mapping	○ Only identifies the RNA-RNA interactions associated with proteins
RIC-SEQ	✓ High throughput ✓ Many-to-many mapping ✓ Biotin pull down ✓ Biotin pCp incorporation to demark interacted RNA boundary ✓ *In situ* ligation	○ Only identifies the RNA-RNA interactions associated with any proteins ○ Need large amount of start material

4.1.5 RNA-RNA seq

Taking into account the disadvantages and advantages of the current methods to study RNA based interactomes (**Table 4.2**), we developed our own workflow called RNA-RNA-seq (**Figure 4.5**). This novel protocol can measure direct (RNA-RNA) and indirect (protein-mediated) interactions without being limited by specific protein or RNA baits by applying different crosslink reagents. We use on-bead ligation to avoid difficult 2D-gel isolation that is not suitable for small quantities of RNA isolated from viral samples, and we use specific adaptors to clearly demark the boundary between interacting RNAs. We also enrich RNA duplexes for sequencing to avoid wasting sequencing depth.

Our protocol starts with crosslinking. We use AMT and SHARC to crosslink direct RNA-RNA interactions. AMT (4'-aminomethyltrioxsalen hydrochloride) is a psoralen derived, cell-permeable and reversible photo-cross-linker that specifically crosslinks RNA-RNA interactions under long wavelength UV irradiation, which can then be reversed under short wavelength UV irradiation. However, psoralen reagents have low efficiency and crosslinks are biased toward RNA-RNA duplexes containing opposing uridine bases. SHARC[55] (Spatial 2'-Hydroxyl Acylation Reversible Crosslinking) is a 2'-hydroxyl acylation based crosslinker that is easy to prepare, and has high crosslink efficiency for RNA-RNA interactions. The resulting crosslink can be reversed under mild alkaline conditions without RNA damage. On the other hand, we use formaldehyde to crosslink indirect RNA-RNA interactions. Formaldehyde crosslinks RNA-protein and protein-protein interactions, thus, capturing indirect RNA-RNA interactions that occur through protein intermediates as well as direct interactions that are flanked or caged by proteins.

Following crosslinking, we perform RNA fragmentation. Depending on the fragmentation method, 5' and 3' RNA ends may not be available for subsequent ligation. We therefore performed end repair to make the RNA ends suitable for adaptor ligation, and to prevent fragmented RNA from undergoing unwanted ligation (e.g. fragmented RNA ligated together or RNA self-ligation) on beads or in solution. Our protocol uses two adaptors for ligation. First, a biotinylated adaptor for enrichment of RNA on streptavidin magnetic beads. Second, an Illumina adaptor-3'5'P adaptor, for reverse transcription and library preparation. Both adaptors contain 3' and 5' phosphate, therefore, adaptor self-ligation can be avoided. First, ligation is performed with T4 RNA ligase 1, which catalyzes the ligation of a 5' phosphate-terminated adaptor donor to 3' hydroxyl-terminated RNA ends acceptor through the formation of a 3' → 5'phosphodiester bond with hydrolysis of ATP to AMP and PPi[56]. Then, streptavidin magnetic beads are used for biotin selection, ensuring that ligation products containing at least one biotin adaptor can bind on beads. The proximity ligation is performed on beads using an unusual RNA ligase, known as RtCB, to join the 3' phosphate end of adaptors to 5' hydroxyl of RNA ends to form a circularized ligation product[57], [58]. In principle, RtCB ligase is simpler than T4 RNA ligase used in other methods, as this avoids an additional end repair step thanks to the 3' phosphate group present in our adaptors. Next, the crosslink is reversed and stringent washes are applied to remove unwanted side reactions. In the end, only ligation products containing biotin adaptor are left. But only the expected ligation products containing one of each of the two adaptors can be reverse transcribed and amplified.

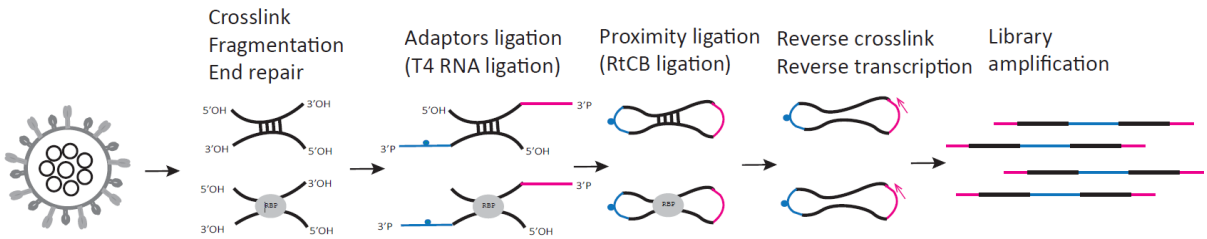


Figure 4.5| RNA-RNA seq workflow. First, crosslink RNAs with different reagents, AMT/SHARC reagent for direct RNA-RNA interactions, and formaldehyde for interactions mediated by protein, following by fragmentation and end repairing; then ligate two adaptors with T4 RNA ligase, and RtCB ligase for proximity ligation; finally reverse crosslink and prepare the library for sequencing.

4.2 Materials and methods

Table 4.3 Model subtract RNA oligos, adaptors and primers sequences

Name	*Sequence*
HIV-dimerizing	HO5'rArGrGrArCrUrCrGrGrCrUrUrGrCrUrGrArArGrCrGrCrGrCrArCrGrGrCrArArGrArGrGrCrGrArGrGrGrC3'OH
HIV-Non-dimerizing	HO5'rArGrGrArCrUrCrGrGrCrUrUrGrCrUrGrArArGrCrGrCrarCrArCrGrGrCrArArGrArGrGrCrGrArGrGrGrC3'OH
RNA-B	HO5'rCrUrGrGrCrUrCrCrGrcrUrArUrCrArCrGrCrUrU3'OH
MS2-Linker	/5Phos/rArArArArArArArArArcrgrarcrgrarcrgrarurcrarcrgrcrgrurcrgrArArArArArArArArArA/3Phos/
RNA adaptor	/5Phos/rArArArArArCrUrCrGrGrCrGrGrArUrArGrArArArArArArArArArA/3Phos/
3'5'P adaptor	/5Phos/rUrGrGACACCGACGATTACCCT/iSp18/GCAATGAAGTrCrGrCrArGrGrGrUrUrG/3Phos/
Biotin adaptor	/5Phos/rArArACAACAAAcgacgacga/iBiodT/cacgcgtcgrCrArArCrArArCrArArA/3Phos/
RRseq RT/PCR-Rv	AGGGTAATCGTCGGTGTCCA
RRseq PCR-Fw	GCAATGAAGTCGCAGGGTTG
RNA_A	/5Phos/rNrNrNrNrNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGaaaaaaaaaa/3Bio/
DNA_B	ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Plasmids

The sequences of PR8 were obtained from pHW2000 vector based 8 plasmids contain 8 segments of PR8 [59], [60]

Table 4.4 Plasmids name and sequences

Segment name	*sequence*
PB2	https://www.ncbi.nlm.nih.gov/nuccore/NC_002023.1?from=28&to=2307&report=genbank
PB1	https://www.ncbi.nlm.nih.gov/nuccore/NC_002021.1?from=25&to=2298&report=genbank
PA	https://www.ncbi.nlm.nih.gov/nuccore/NC_002022.1?from=25&to=2175&report=genbank
HA	https://www.ncbi.nlm.nih.gov/nuccore/NC_002017.1?from=33&to=1733&report=genbank
NP	https://www.ncbi.nlm.nih.gov/nuccore/NC_002019.1?from=46&to=1542&report=genbank
NA	https://www.ncbi.nlm.nih.gov/nuccore/NC_002018.1?from=21&to=1385&report=genbank
M	https://www.ncbi.nlm.nih.gov/nuccore/NC_002016.1?from=26&to=784&report=genbank
NS	https://www.ncbi.nlm.nih.gov/nuccore/NC_002020.1?from=27&to=719&report=genbank

RNA-RNA seq protocol

Crosslinking

***AMT*:** AMT (Biomol) was added to final concentration of 0.25 mg/ml and incubated on ice for 10 min, then the sample was irradiated 3 times at 365 nm UV for 10 min. Crosslinking was performed at a distance of 15 cm from UV light source on ice with regular mixing.

***SHARC*:** SHARC reagents were made by dissolving 1-part SHARC reactant (2,6-Pyridindicarbonsäure) in 200 µl anhydrous DMSO (Sigma, 276855) and 2 parts 1,1'-Oxalyldiimidazol (CDI, Sigma, 115533) in 250 µl DMSO. Dissolved SHARC reactant was pipetted into the tube containing CDI. After briefly vortex and spinning down, a needle was inserted into the top of the 1.5 mL centrifuge tube to allow the CO_2 product to escape. Mixed solutions were left at room temperature to react for 60 min before crosslinking. The model RNA or biological samples incubated with SHARC reagent at final concentration 25 mM (cell or virus) or indicated (test) at room temperature on rotor for 60 min.

***Formaldehyde*:** Cells or RNA-Protein complexes were incubated with the indicated concentration of formaldehyde at room temperature for 10 min. The reaction was quenched by adding Tris-HCl buffer (pH 7.5) or glycine to final concentration at 250 mM.

Fragmentation

***Mg^{2+}*:** 2 µg/0.2 µg of a pool of *in vitro* transcribed RNAs from all 8 segments of the PR8 influenza strain were heated to 94°C for 3 or 5 min in RNA fragmentation buffer with 20 mM Tris-Acetate (pH 8.1), 50 mM K-Acetate, 15 mM Mg-Acetate. Fragmentation was stopped by transferring the reaction to ice and adding EDTA to final concentration at 50 mM. Mg^{2+} fragmentation creates RNA fragments with 5´ hydroxyl and 3´ phosphate termini.

***Sonication (Bioruptor)*:** 2 µg/0.2 µg of a pool of *in vitro* transcribed RNAs from all 8 segments of the PR8 influenza strain were mixed in folding buffer (100 mM Hepes-NaOH (pH8.0), 100 mM NaCl, 10 mM $MgCl_2$) or TE buffer (10 mM Tris-HCL (pH 7.5), 1 mM EDTA). Samples were sonicated with the indicated sonication time and intensity.

***RNase A*:** 2 µg PR8 of a pool of *in vitro* transcribed RNAs from all 8 segments of the PR8 influenza strain were mixed with RNase A (Promega) at 1/10/100 µg/ml in 1X T4 RNA ligation buffer (50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM DTT) and incubated at room temperature for 10/30 min, the reaction was stopped by Trizol (Sigma) extraction according to the manufacturer's instructions.

***S1 nuclease*:** 2 µg of a pool of *in vitro* transcribed RNAs from all 8 segments of the PR8 influenza strain were mixed with 2 µl of 1:100 diluted S1 nuclease (Promega) in 1X S1 nuclease buffer (50 mM sodium acetate (pH 4.5), 280 mM NaCl, 4.5 mM $ZuSO_4$) and incubated at room temperature for 10/30 min. The reaction was stopped by Trizol (Sigma) extraction according to the manufacturer's instructions.

***RNase If*:** 2 µg PR8 of a pool of *in vitro* transcribed RNAs from all 8 segments of the PR8 influenza strain were mixed with 12.5 U of RNase If (NEB) in 1X T4 RNA ligation buffer (50 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, 1 mM DTT (pH 7.5) and incubated at 37°C for 10/20/30 min. The reaction was stopped by adding Ribonucleoside Vanadyl Complex (RVC) at 1 mM final concentration, followed by Trizol (Simga) extraction according to the manufacturer's instructions.

***Pb^{2+}*:** 1 µg of purified human ribosome mix with indicated $PboAC_2$ in 1X ribosome resuspension buffer (20 mM HEPES-KOH (pH7.4), 120 mM K-Acetate, 2 mM Mg-Acetate) and incubate at 37°C for 5 min. EDTA was added to stop the reaction at final concentration 50 mM EDTA, and following Trizol (Sigma) extraction according to the manufacturer's instructions.

End repair

The fragmented RNA was incubated with 20 U T4 Polynucleotide Kinase and 2 U Shrimp Alkaline Phosphatase (rSAP) in 1X T4 Polynucleotide Kinase Reaction Buffer (70 mM Tris-HCl (pH 7.6), 10 mM

$MgCl_2$, 5 mM DTT) for 1h at 37°C. To inactivate the rSAP, incubated the samples at 80 °C for 4 min, mix once during the inactivation.

T4 RNA ligation

in vitro: 100 pmol of each adaptor and 50 pmol of synthesized HIV-d or HIV-nd RNA were mixed with 1X T4 RNA ligase buffer (50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM DTT), 1mM of ATP (NEB), 12.5% of PEG8000 (NEB), 10 U of RNasin (Molox), 10 U of T4 RNA Ligase 1 (NEB, M0204) in a total volume of 25 µl. Samples were mixed and incubated at room temperature for 2h or overnight.

in vivo: 4-fold excess of adaptor were added to fragmented RNA or formaldehyde crosslinked cells together with 1X T4 RNA ligase buffer (50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM DTT), 1 mM of ATP (NEB), 12.5% of PEG8000 (NEB), 10 U of RNasin (Molox), 10 U of T4 RNA Ligase 1 (NEB, M0204) in a total volume of 25 µl. Samples were mixed and incubated at room temperature overnight.

Biotin selection

MyOne Streptavidin C1 Beads (Invitrogen) were made RNase-free by washing twice in Solution A (DEPC-treated 0.1 M NaOH, DEPC-treated 0.05 M NaCl) for 2 min using a volume of Solution A equal to, or larger than the initial volume of beads originally taken from the vial. Then washed the beads once in Solution B (DEPC-treated 0.1 M NaCl) using a volume equal to the volume used for Solution A. Resuspend the beads in Solution B. Resuspended washed Dynabeads™ magnetic beads in 2X B&W Buffer (10 mM Tris-HCl (pH 7.5) 1 mM EDTA 2 M NaCl) to a final concentration of 5 µg/µL (twice original volume). Added an equal volume of biotinylated DNA or RNA (in distilled water) following with 15 min incubation at room temperature using gentle rotation. Separated the biotinylated DNA or RNA coated beads with a magnet for 2-3 min. Finally washed the coated beads 3 times with 1X B&W Buffer and 1 time with 1X RtCB ligase buffer (50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT (pH 8.3 @ 25°C)).

RtCB ligation

The magnetic beads coupled T4 RNA ligation product mixed with 1X RtCB ligase buffer (50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT (pH 8.3 @ 25°C)), 0.1 mM GTP, 1 mM $MnCl_2$, 40 U of RNasin (Molox), 15% of PEG8000 (NEB), 8 µl of homemade RtCB ligase, made volume with H_2O to 200 µl, mix well, incubated on shaker at 37 °C for 1 h.

RNA elution from streptavidin magnetic beads

Added 100 µl elution buffer (10 mM Tris-HCl (pH 7.5), 1 mM EDTA) with beads, heated at 90°C for 1 min, applied magnet, collected supernatant as elution, repeat for 3 times.

Reverse crosslinking

***AMT*:** the sample was irradiated for 10 min at 254 nm, 15 cm from the UV bulb on ice. After reverse crosslinking, RNA was purified with three volumes of ethanol, 1/10 volume of 3M NaOAC (pH 5.2), and 1µl of GlycoBlue.

***SHARC*:** 5x decrosslinking buffer (500 mM Boric acid, pH 11) was added to the eluted proximity ligated RNA, and nuclease free water was added to bring decrosslinking buffer to 1x. Samples were incubated for 2 h at 45°C. After reverse crosslinking, RNA was purified with three volume of ethanol, 1/10 volume of 3M NaOAC (pH 5.2), and 1µl of GlycoBlue.

***Formaldehyde*:** the crosslinked sample was incubated with proteinase K (NEB) (final concentration 0.03 U/µl) in the buffer contain 20 mM Tris-HCl pH 7.4, 10 mM EDTA, 2% N-lauroylsarcosine, 2.5 mM TCEP, and incubated at 65°C for 1.5 h. After proteinase K treatment, RNA was purified by phenol chloroform extraction and ethanol precipitation.

RNA purification

An equal volume of phenol chloroform was added to the sample, vortexed for 1 min and centrifuged at 17000x*g* for 2 min. The upper aqueous phase was collected into fresh tube. An equal volume of chloroform was then aqueous phase, vortexed for 1 min and centrifuge again at 17000x*g* for 2 min. The upper aqueous phase was collected into a fresh tube and 1/10 volume of 3M NaOAC (pH 5.2) was added along with 1µl of Glycoblue (Invitrogen) and 3 volumes of 100% ethanol. The sample was stored at -80^{o}C for 30 min to precipitate the RNA. The RNA was then pelleted by centrifugation for 30 min at 17000x*g* at 4^{o}C. The pellets were washed twice with 75% of ethanol and centrifuged for 5 min at 4^{o}C at 17000x*g.* Ethanol was decanted and the pellet was air dried and resuspend in RNase free H_2O.

Stringent wash

200 µl of stringent wash buffer (x2 SSC, 70% formamide, 1 mM EDTA) was added to the magnetic beads coupled with RNA and incubated at 40^{o}C for 5 min. Beads were pelleted with a magnet stand. Stringent washing was carried out three times.

Reverse transcription

Purified RNA was mixed with 0.5 µM of reverse transcription primer, 0.5 mM dNTPs and incubated at 65^{o}C for 5 min before chilling on ice for 2 min. 100 U SuperScript IV reverse transcriptase (Invitrogen), 50 mM Tris-HCl (pH 8.3), 4 mM $MgCl_2$, 10 mM DTT, 50 mM KCl was then added to the RNA in a total volume of 20 µl. The sample was then incubated at 52^{o}C for 60 min. Reverse transcriptase was inactivated by incubating at 80^{o}C for 10 min.

PCR

PCR amplifications were performed on 10-fold diluted reverse transcribed cDNAs with 250 nM forward primer and reverse primer, 200 µM dNTPs, 1X Q5 reaction buffer, 0.01 U/µl Q5 polymerase (NEB) in 50 µl using the PCR cycling conditions: 98°C for 30 s, followed by 10-34 cycles of 98°C for 10 s, 55°C for 30 s, and 72°C for 30 s. Products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer.

***In vitro* transcription**

Plasmids containing sequences of the influenza 8 segments were linearized with BstEII (NEB) in 1X Cutsmart buffer (NEB) for 3 h at 37°C. Linearized plasmids are loaded on 1% agarose gel in 1X TAE buffer with unlinearized plasmids, followed by phenol chloroform extraction and ethanol precipitation. The purified and linearized plasmids were used as templates for RNA *in vitro* transcription with a homemade T7 RNA polymerase. Reactions contained 1X reaction buffer (40 mM Tris-HCl (pH 7.5), 18 mM $MgCl_2$, 10 mM DDT, 1 mM Spermidine), 5 mM NTPs, 40 U RNasin (Molox), DNA template, 0.05 U of Pyrophosphatase (NEB) and 5 µl of homemade T7 RNA polymerase. The reaction was incubated at 37°C for 3 h, followed by DNase I treatment for 30 min at 37 °C. RNAs were gel purified after electrophoresis on 1% agarose gels in 1X TAE buffer using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) with NTC buffer (Macherey-Nagel).

Cell culture

HEK 293T cells and MDCK cells were cultured in DMEM media containing 10% FBS and 100 U/ml of Penicillin-Streptomycin in a 37°C incubator with 5% CO_2.

Ribosome purification

A confluent 10-cm^2 dish of HEK293T cells were harvested by scraping. Cells were collected by centrifugation at 1000x*g*, followed by washing with ice-cold PBS buffer wash by centrifugation at 1000x*g*. Cell pellets were resuspended in 150µl lysis buffer (10 mM HEPES-KOH (pH 7.5), 100 mM KCl, 5 mM $MgCl_2$, 2 mM DTT, 0.16 U/µl RNasin, 1X complete protease inhibitor, 1 mM PMSF) and sheared with a 26-gauge needle 20 times. Lysates were centrifuged at 17000x*g* for 15 min at 4°C, and soluble supernatant transferred onto a sucrose-cushion (20 mM HEPES-KOH (pH7.4), 150 mM K-Acetate, 2 mM Mg-Acetate, 1 mM PMSF, 25% sucrose, 1X protease inhibitor cocktail). Sucrose cushioned sampled were then centrifuge 25 min in a MLA130/TLA100-2 rotor at 95000 rpm at 4°C. The supernatant was discarded and the ribosome pellet resuspended in 150 µl of lysis buffer. A 5%-45% sucrose gradient in lysis buffer was prepared and equilibrated at 4°C for 30 min. Subsequently, 150 µl of the ribosomal solution was added to the top without perturbing the gradients, and centrifuged at 38000 rpm for 2 h at 4°C with the rotor SW40. Fractions were collected the fraction containing pure 80S ribosome were pooled based on the A260 profile. Pure 80S ribosomes were pelleted by centrifugation for 25 min at 95,000 rpm at 4°C in a MLA130 rotor. Finally, pellets were resuspended in ribosome resuspension buffer (20 mM HEPES-KOH (pH7.4), 120 mM K-Acetate, 2 mM Mg-Acetate).

Virus propagation

***In MDCK cells*:** A 15-cm^2 dish of MDCK cells at 80-90% confluency containing approximately $22x10^6$ cells was washed with 5 mL of 1xPBS buffer for 3 times. 0.001 MOI of virus inoculum was added in 5 ml PBS buffer Ca^{2+}/Mg^{2+} with 0.2% BSA. Plates were incubated at 37°C for 60 min with front-to-back and side-to-side rocking every 15 mins. Infected MDCKs were washed 3 times with sterile PBS buffer. 20 ml of opti-MEM with Pen/Strep and 1.0 µg/mL trypsin-TPCK were added and dishes were incubated in a 37°C incubator with 5% CO_2 until only ~10% of the cells were still attached (about 36-48 h for PR8).

In specific pathogen-free (SPF) embryo eggs: 10-days-old eggs were ordered from Charles River. Upon arrival eggs were placed in a humidified egg incubator with an automatic egg turner to rotate eggs regularly. Eggs were candled to eliminate any dead embryos, and the air space of the live eggs were washed with 70% ethanol and marked with a permanent marker. Virus inoculum was diluted to 100 PFU/100 µl with PBS buffer. Eggs were placed into a biosafety hood (BSL-2), and a sterile 18G needle was used to punch a small hole in the shell over the air sac of each egg. 200 µl of virus inoculum was then carefully injected into the allantoic cavity using a 22G needle at a 45° angle. The small hole was sealed with a drop of glue from a glue gun, and the eggs were placed back into the egg incubator with the air space pointed upward. The infected eggs were then incubated without turning at 37°C and ~60% humidity for 48 h. After the incubation period, the eggs were chilled at 4 °C overnight to kill the embryo and constrict the blood vessels to reduce the risk of contaminating the infected allantoic fluid with blood. After chilling, the eggs were transferred to a biosafety hood. Eggs were placed into a holder with the air sac facing up, and the surface cleaned with 70% ethanol. Sterile scissors were used to remove the eggshell above the air sac, while being careful not to destroy the chorioallantoic membrane. The chorioallantoic membrane was then opened with sterile blunt forceps, and without rupturing the yolk, the embryo and the yolk sac were moved aside with a small spatula. The allantoic fluid was then carefully collected with a 1 ml pipette and pooled into a 50 ml plastic conical tube on ice. One egg yields 5-10 ml of a slightly yellowish fluid. The virus-containing allantoic fluid was centrifuged at 1,000xg for 10 min at 4°C to pellet debris. The clear fluid was transferred into a new tube and snap-frozen in liquid nitrogen, before transfer to a -80 °C freezer for long-term storage.

Plaque assay

MDCK cells were seeded into 12-well plate at $1x10^6$ cell per well in 1 mL of DMEM growth medium 24 h prior to plaque assay to assure a confluent monolayer. 100µl of virus sample was diluted with 900 µl of PBS buffer (Mg^{2+}/Ca^{2+}) containing 0.2% of BSA, from 10^{-1} to 10^{-9} at 10-fold serial dilution. The MDCK monolayer was washed 3 times with PBS buffer. 100 µl of each serially diluted viral sample was added to each well and mixed well by rocking the plates side-to-side and forward-to-back every 15 minutes to distribute the virus inoculum evenly over the cells. Samples were incubated at 37°C with 5% CO_2 for 60 min. During the incubation, the agar overlayer was prepared. For one 12-well plate, 6.25 ml of 2x MEM, 2.25 ml of sterile water, 125 µl of 1% DEAE Dextran, 187.5 µl of 7.5% $NaHCO_3$, 12.5 µl of 1 mg/mL trypsin-TPCK were mixed in a sterile 50 mL conical tube and incubated in a 37°C water bath. 2% agarose to liquefied in a microwave and left in in a 56°C water bath. After the

60 min incubation, the virus inoculum was removed by aspiration. 3.75 ml of 2% Oxoid agar was mixed with the overlay components and left to cool until 1mL could be added to the cells. The 12-well plates were then left to rest in the biosafety cabinet for approximately 10 min to allow the overlay to solidify, before transfer into the 37^{o}C incubator. For PR8, plaques become visible by eye after 48 h. To fix cells and neutralize infectious virus, 1 mL of 5% formaldehyde (in PBS buffer) was added to each well followed by incubation plates for approximately 1 h (or overnight). The agar overlay was removed from the wells with a spatula, and 500 µl of Crystal Violet Staining solution added to each well and incubate at room temperature for 30 min. The plates were then rinsed with tap water and the titer calculate based on the plaque number and dilution factor.

Virus NHS beads binding

Magnetic beads were equilibrated to room temperature. The appropriate volume of beads were the placed into a 1.5 mL microcentrifuge tube in a magnetic stand for 2 min and the supernatant discarded. Bead pelleted were activated with 1mL 1mM ice cold HCl by gentle vortexing for 15 seconds followed by 2 min a magnetic stand so that the activation solution could be discarded. Virus diluted in 100 mM HEPES (0.2% BSA) was then added to the activated NHS beads in a minimum volume of 500 µl and mixed for 1.5 h at room temperature on rotor. The tube was then placed into a magnetic stand for 2 min to collect the beads and discard the supernatant. Beads were washed beads with 500 µl 10 mM Tris pH 7-8 (0.2% BSA) for 3 times.

Benzonase treatment

Beads were resuspended in in 250 µl of 1X benzonase buffer (50 mM Tris-HCl (pH 8), 1 mM $MgCl_2$, 0.2% BSA) with 1 µl of benzonase and incubated at 37°C for 1 h on a shaker.

qPCR

Each sample was mixed with 5 µl of 10-fold diluted cDNA together with 7.5 µl of X2 Power SYBR and 350 nM of forward and reverse primer in 15µl total volume using the cycling conditions: 95°C for 30 s, followed by 39 cycles of 95°C for 15 s, 55°C for 15 s, and 68°C for 30 s with melting curve and appropriate standards.

Library preparation

Two-read sequencing library

There were two step PCR to amplify the two-read sequencing library: 1) library amplification, 20 µl of 10-fold diluted reverse transcribed cDNAs as template with 250 nM forward primer (ACACTCTTTCCGCAATGAAGTCGCAGGGTTG) and reverse primer (TCGGAGATGTAGGGTAATCGTCGGTGTCCA), 200 µM dNTPs, 1X Q5 reaction buffer, 0.01 U/µl Q5 polymerase (NEB) in 200 µl reaction using the PCR cycling conditions: 98°C for 30 s, followed by 10-20 cycles of 98°C for 10 s, 55°C for 30 s, and 72°C for 30 s. Products were visualized by

electrophoresis on 1% agarose gels in 1X TAE buffer and isolated, column purified with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel); 2) index PCR, half of the purified PCR product was used as template with with 250 nM forward primer (AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCGCAATGAAGTCGCAGGGTTG) and reverse primer (CAAGCAGAAGACGGCATACGAGAT(*index7*)TCGGAGATGTAGGGTAATC), 200 µM dNTPs, 1X Q5 reaction buffer, 0.01 U/µl Q5 polymerase (NEB) in 50 µl using the PCR cycling conditions: 98°C for 30 s, followed by 5 cycles of 98°C for 10 s, 55°C for 30 s, and 72°C for 30 s. Products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer and isolated, column purified with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). The final libraries were quantified by NEBNext® Library Quant Kit for Illumina (E7630L). Paired end PE150 sequencing was carried out on an Miniseq instrument (Illumina) according to the manufacturer's instructions.

Four-read sequencing library

20 µl of 10-fold diluted reverse transcribed cDNAs as template with 250 nM forward primer (AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCGCAATGAAGTCGCAGGGTTG) and reverse primer (CAAGCAGAAGACGGCATACGAGATTCGGAGAGAG(*index7*)AGGGTAATCGTCGGTGTCCA), 200 µM dNTPs, 1X Q5 reaction buffer, 0.01 U/µl Q5 polymerase (NEB) in 200 µl reaction using the PCR cycling conditions: 98°C for 30 s, followed by 10-20 cycles of 98°C for 10 s, 55°C for 30 s, and 72°C for 30 s. Products were visualized by electrophoresis on 1% agarose gels in 1X TAE buffer and isolated, column purified with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). The final libraries were quantified by NEBNext® Library Quant Kit for Illumina (E7630L) following the manufacturer's instruction.

4.3 Results

4.3.1 Model substrate

To establish the RNA-RNA-seq protocol, a short sequence (42nts) from the HIV-1 5'UTR corresponding to the dimerization initiation site (DIS) was used as a model substrate [61]–[65] (**Figure 4.6**). This short sequence undergoes spontaneous dimerization *in vitro* (HIV-Dimerizing, HIV-d), and dimerization can be prevented with a point mutation within the 6-nucleotide palindromic loop sequence (HIV-Non-dimerizing, HIV-nd) (**Figure 4.6a**). We confirmed the expected behaviour of the HIV-d and HIV-nd on a native agarose gel.

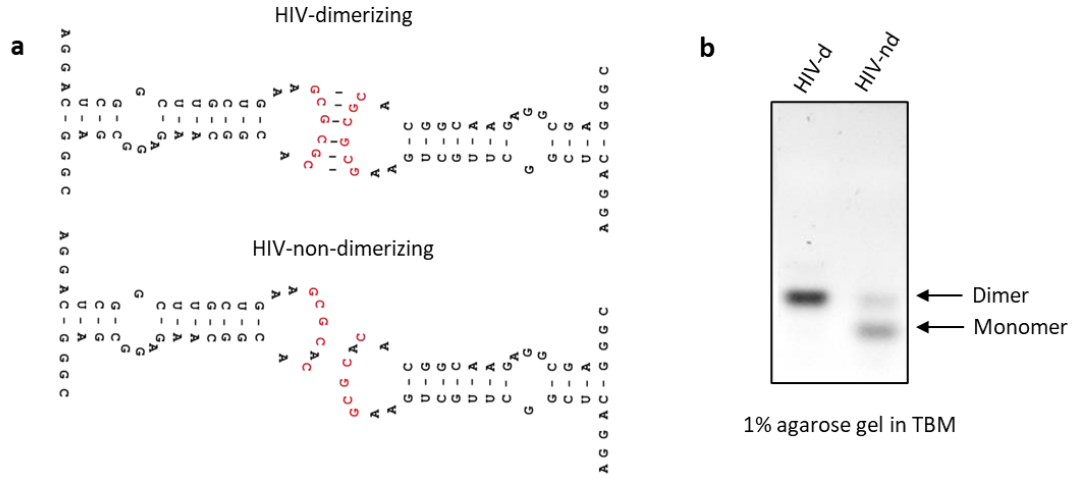


Figure 4.6| DIS from HIV-1 as model RNA-RNA interaction. The diagram **(a)** and native agarose gel **(b)** to show the HIV dimer and non-dimer form: HIV-Dimerizing, HIV-d; HIV-Non-dimerizing, HIV-nd.

4.3.2 Ligation test

Ligation is an essential step in proximity ligation techniques, including RNA-RNA seq. Our protocol uses two different ligases for different purposes: T4 RNA ligase 1 for the ligation of the first set of adaptors to the RNA, and RtCB ligase for the proximity ligation of RNA duplexes (**Figure 4.5**).

To determine the ligation efficiency of T4 RNA ligase, we performed test ligations between the adaptor oligos containing a 5' phosphate and 3' phosphate block and HIV-d and HIV-nd oligos containing 5' -hydroxyl and a 3'-hydroxyl. This set up ensures that only a single ligation takes place and avoids the formation of concatemers that would complicate analysis. To carry out the experiment, 50 pmol of oligo with 3' and 5' -hydroxyl and 100 pmol of adaptors with 3' and 5' phosphate were denatured at 90°C for 2 min, chilled on ice for 2 min, then added the mixture with 1X T4 RNA ligase buffer (50 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, 1 mM DTT (pH 7.5 @ 25°C)), 1mM of ATP (NEB), 12.5% of PEG8000 (NEB), 10 U of RNasin (Molox), 10 U of T4 RNA Ligase 1 (NEB, M0204) in a total volume of H_2O to 25 µL. Samples were mixed, and incubated at room temperature for 2 h. The ligation products were purified by ethanol precipitation and then loaded on 7.5% of denaturing polyacrylamide urea gel. The result showed that T4 RNA ligation was carried out efficiently (**Figure 4.7a**).

To determine RtCB ligation efficiency on adaptors containing a 3'-phosphate, we performed a test ligation between an adaptor containing a 3′-phosphate and 5′-phosphate block and an RNA containing 5-hydroxyl and 3′-hydroxyl. Like the previous T4 RNA ligase test, in this setup RtCB can only carry out a single intermolecular ligation. 50 pmol of oligo with 3' and 5' -hydroxyl and 100 pmol

of adaptors with 3’ and 5’ phosphate was denatured at 90°C for 2 min, chilled on ice for 2 min, then added the mix with 1X RtCB ligase buffer (50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT (pH 8.3 @ 25°C)), 0.1 mM GTP, 1 mM $MnCl_2$, 40 U of RNasin (Molox), 15% of PEG8000 (NEB), 2 µL of homemade RtCB ligase, in a total volume of 25 µL. Samples were mixed well, and incubate at 37°C for 1 h. The ligation product was determined by 7.5% denaturing polyacrylamide urea gel (**Figure 4.7b**). The result showed that all of the RNA was ligated with the 3’5’P adaptor.

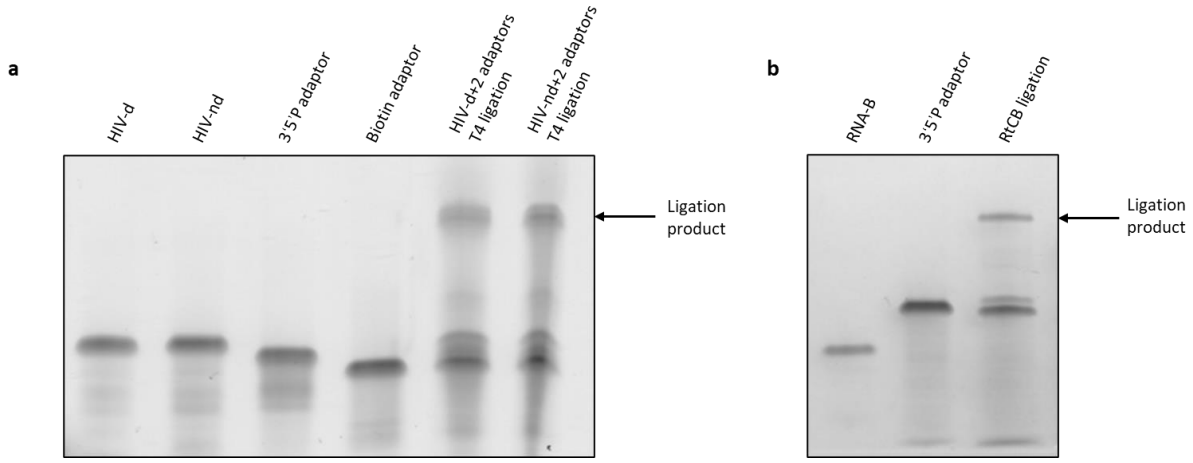


Figure 4.7| Denaturing gel to show the ligation product. (a) T4 RNA ligation, HIV-d, HIV-nd, 3’5’P adaptor and biotin adaptor were denatured and loaded on lane 1-4, lane 5 and lane 6 are HIV-d or HIV-nd mixed with two adaptors and ligated with T4 RNA ligase 1 product, respectively; **(b)** RtCB ligation, RNA-B is another model substrate RNA has 3’ and 5’ -hydroxyl (lane 1), and 3’5’P adaptor (lane 2), and RtCB ligation product from RNA-B and 3’5’P adaptor.

4.3.3 Ligation, selection, reverse transcription and library amplification worked on model substrate RNA without crosslink

From the ligation test, we could observe the expected ligation products. Next, we tested whether two sequential ligations with T4 RNA ligase and RtCB ligase could produce a ligation product that can be reverse transcribed and amplified into a sequencing library. The test protocol included the following steps: T4 RNA ligation, streptavidin magnet beads binding (Biotin selection), RtCB ligation, washing, elution of RNA from beads, RNA purification, reverse transcription, and PCR amplification.

T4 RNA ligation and RtCB ligation test results were shown above (**Figure 4.7a, 4.7b**). For streptavidin magnetic beads (S1420S) binding, , I optimized the binding and elution condition with a biotinylated oligo following the manufacturer’s recommended protocol (NEB). The optimized conditions for

binding were 100 µl of beads per 100 pmol biotinylated oligo, incubation at room temperature in 1X T4 RNA ligase buffer for 15 min with regular mixing. The optimized conditions for elution were adding 100 µl of Elution Buffer [10 mM Tris-HCl (pH 7.5), 1 mM EDTA] at 90°C for 1 min, repeated for three times.

To test whether the protocol faithfully captured RNA duplexes, and not artificial interactions between single stranded RNAs in solution, I carried out the protocol with HIV-d, and HIV-nd as a negative control. 50 pmol of HIV-d or HIV-nd and 100 pmol of each two adaptors were ligated using T4 RNA ligase I at room temperature for 2 h. 100 µl of streptavidin magnetic beads were added in 1X T4 RNA ligase buffer and incubated at room temperature for 15 min with regular mixing. Beads were washed with 1xT4 RNA ligase buffer for three times to remove unligated substrate and ligation products without biotin adaptor. To perform proximity ligation, 8 µl of homemade RtCB ligase was added in 1x RtCB ligation buffer (50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT (pH 8.3 @ 25°C)), 0.1 mM GTP, 1 mM $MnCl_2$, 40 U of RNasin (Molox), 15% of PEG8000 (NEB), with a final volume of 100 µl, and incubated on shaker at 700 rpm/37^{o}C for 1 h. Samples were applied to the magnetic stand, and supernatant collected as flow-through for analytical analysis. Beads were washed with elution buffer three times. Finally, ligation products were eluted from the beads with elution buffer. Both flow through and eluted RtCB ligation products were precipitated by ethanol precipitation and resuspended in 10 µl of H_2O. The RNA was reverse transcribed, and PCR amplified with specific primer annealing to the 3’5’P adaptor.

A successful protocol would result in a product of 144 bp for HIV-d and no band for HIV-nd. Unexpectedly, for the HIV-d sample, were observed a band at approximately 100 bp, , and for the HIV-nd samples, we observed an additional product between 100-200bp (**Figure 4.8a**). We hypothesized that the washing was not stringent enough to remove an unwanted side product that only contained 3’5’P adaptor but was captured on streptavidin beads through a non-covalent interaction with a RNA containing a biotinylated adaptor (**Figure 4.8d, S3-1**). We therefore decided to increase the stringency of the wash buffer. Using the wash buffer obtained from the PARIS protocol[36], [37], [66], I repeated the protocol, this time washing beads with bead wash buffer [100 mM Tris pH 7.0, 4 M NaCl, 10 mM EDTA, and 0.2 v/v % Tween 20] six times at 37^{o}C/50^{o}C/ 60^{o}C. The results were encouraging as the expected product was now present in the HIV-d samples, and increased in abundance with increasing wash temperatures (**Figure 4.8b**). Finally, I increased wash temperature to 80^{o}C which gave only a single product in the HIV-d samples and no product in the HIV-nd samples (**Figure 4.8c**). Sequencing of this product showed that it contained the expected sequence with inconsequential small deletions at the HIV-d – adaptor junctions (**Figure 4.8d**).

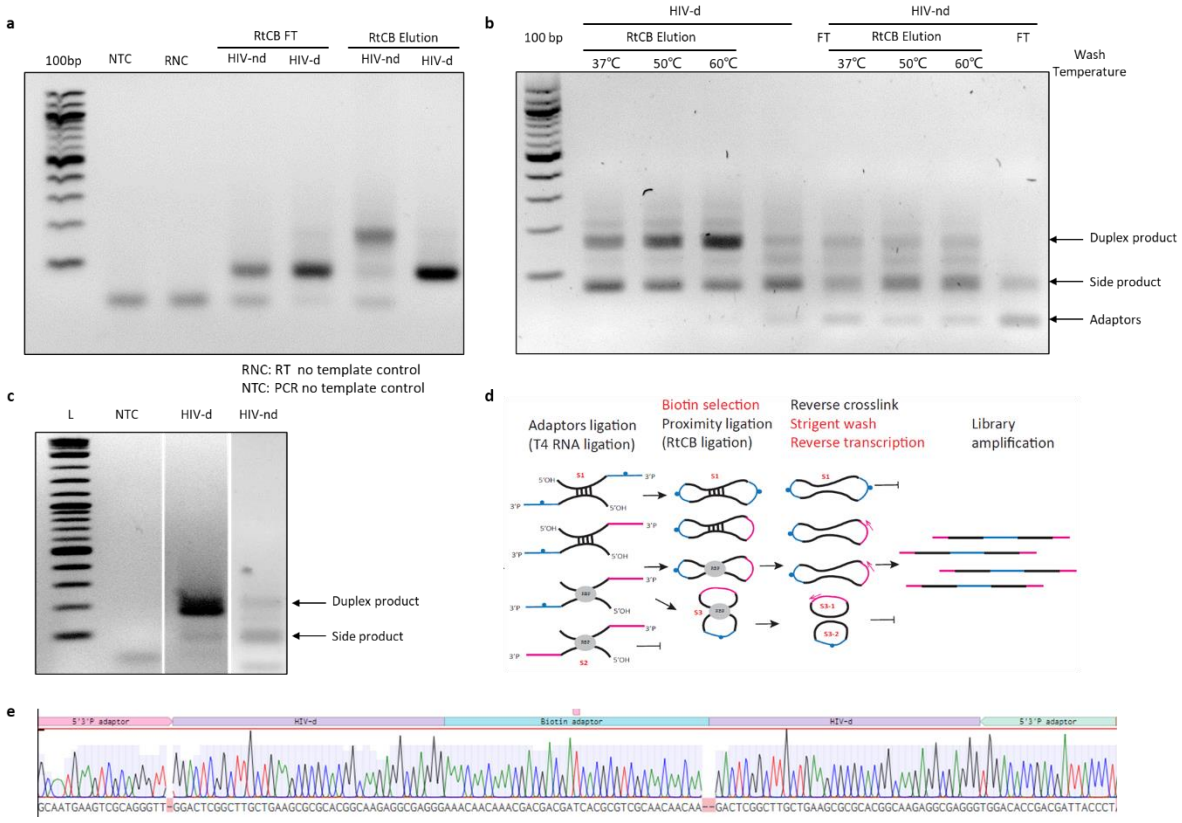


Figure 4.8| RNA-RNA seq on model substrate. (a) Library after RNA-RNA seq on HIV-d and HIV-nd RNA without stringent wash; **(b)** Library after RNA-RNA seq on HIV-d and HIV-nd with different stringency wash; **(c)** Library after RNA-RNA seq on HIV-d with stringent wash, the arrows indicate the expected (duplex) product and side product;**(d)** diagram to show the side products in RNA-RNA seq protocol (side product 1-S1, side product 2-S2, side product 3-S3, etc.); (**e**) duplex product sequence of **(c)**.

4.3.4 Crosslink and reverse crosslink optimization

Crosslinking is an essential step in RNA-RNA-seq. In our protocol, crosslinking reagents can be varied to capture different types of interactions. We applied formaldehyde crosslinking to identify the interactions mediated by protein and psoralen derived AMT and SHAPE based SHARC reagents to identify direct RNA-RNA interactions. Before processing the biological samples, I first tested and optimized crosslinking conditions.

For optimizing formaldehyde crosslinking, I used a short RNA sequence from the MS2 bacteriophage (**Figure 4.9a**), which can form a specific stem loop[67]; and MS2 coat protein (MCP-MBP)[68] which specifically binds to the MS2 stem loop. First, 50 pmol of MS2 RNA was folded in 1x T4 RNA ligase buffer at 37°C for 15 min. Then 200 pmol of MCP-MBP protein was added, and samples were incubated on ice for 15 min. Formaldehyde was added to final concentration at 0%/2%/4%, and

samples were incubated at room temperature for 10 min, and quenched with glycine at final concentration of 250 mM. The samples were then purified by Trizol extraction. Crosslinked RNA partitions to the organic phase along with the proteins. Non-crosslinked RNA was extracted from the aqueous phase and precipitated by isopropanol precipitation. By running the non-crosslinked RNA on a 1% agarose gel, we could estimate crosslinking efficiency as a loss of RNA-protein from the aqueous phase into the organic phase. The results showed that 4% of formaldehyde can crosslink more than 95% of RNA (**Figure 4.9b**).

After proximity ligation, we need to wash away ligation side products and reverse transcribe the ligated duplex to make the sequencing library. Consequently, reverse crosslinking must be carried out before stringent washing. Following a published reverse crosslinking protocol[69], the crosslinked samples were incubated at 70^{o}C for 10 min, and subsequently purified by Trizol extraction. The free RNA is precipitated by isopropanol and determined on a 1% agarose gel. Unfortunately, the results showed the reverse crosslink was inefficient by heat (**Figure 4.9b**). We next tried to reverse the formaldehyde crosslink by proteinase K digestion. Here, a 4% crosslinked MS2-MCP-MBP complex was incubated in 1% SDS, 1 mM EDTA, 20 mM Tris-HCl pH7.5 with proteinase K (NEB) (final concentration 0.012 U/µl) at 42°C for 20 min or 30 min. About 20% of crosslinked RNA was reversed by proteinase K, which still could be improved (**Figure 4.9c**).

As an additional crosslink and reverse crosslink test model, we used cultured HEK293T cells as it is more similar to most biological samples. Here, after 0.5% formaldehyde crosslinking, I tried four conditions of proteinase K digestion to reverse the crosslinks :

1#: 20 mM Tris-HCl pH7.5, 1% SDS, 1 mM EDTA with proteinase K (NEB) (final concentration 0.015U/µl) at 55 °C for 60min;

2#: 100 mM Tris-HCl pH 7.4, 150 mM NaCl, 12.5 mM EDTA, 0.5% (w/v) SDS with proteinase K (NEB) (final concentration 0.12U/µl) at 55°C for 60 min;

3#: 16 mM Tris-HCl pH 7.4, 8 mM EDTA, 1.6% N-lauroylsarcosine, 2 mM TCEP, 250mM NaCl with proteinase K (NEB) (final concentration 0.03U/µl) at 65°C for 1.5 h;

4#: 38 mM Tris-HCl pH 7.4, 411 mM NaCl, 7.8mM EDTA, 0.09% (w/v) SDS with proteinase K (NEB) (final concentration 0.02U/µl) at 65°C for 2 h

The results showed that 0.5% formaldehyde was sufficient to crosslink the cellular RNA; and condition 3# is the most efficient for proteinase K digestion to reverse formaldehyde crosslinks (**Figure 4.9d**). Therefore, in the later assays condition 3# is implemented.

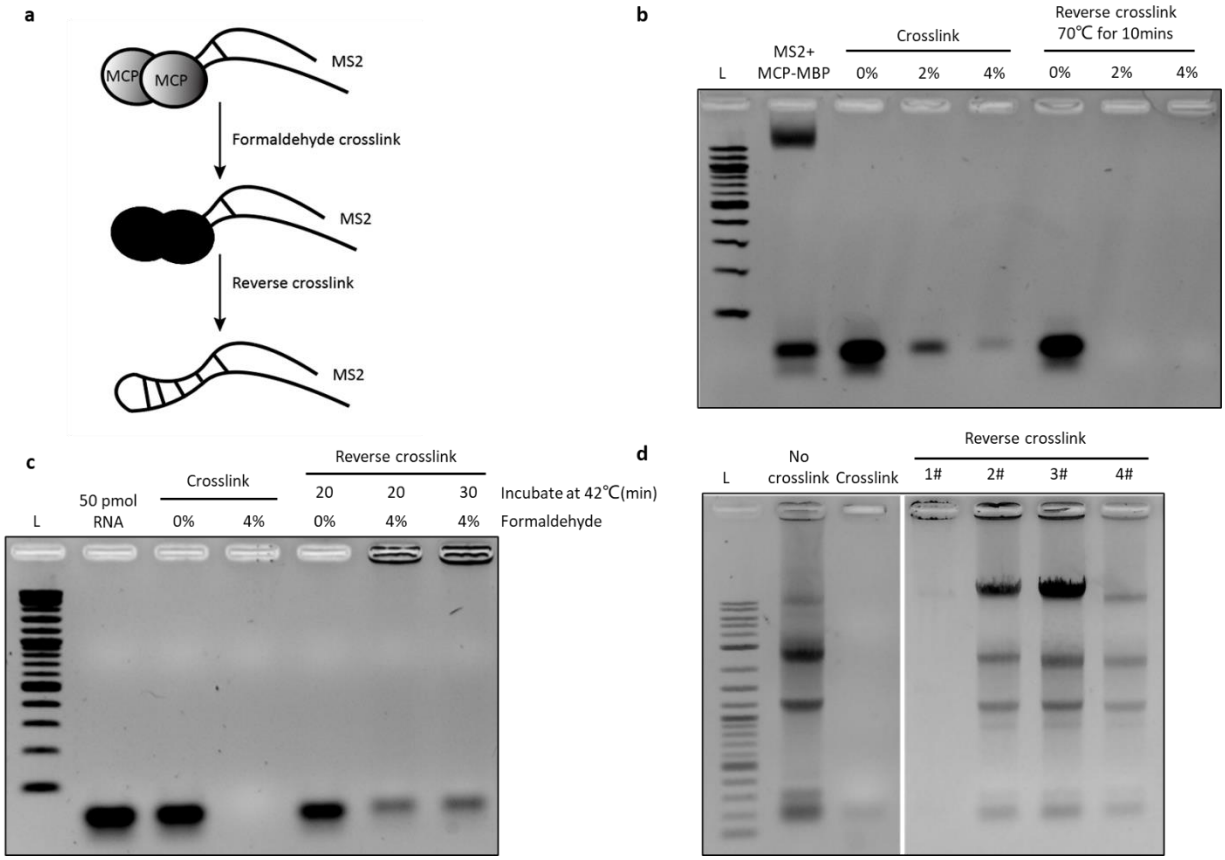


Figure 4.9| Formaldehyde crosslink and reverse crosslink optimization. (a) MS2 RNA and MCP-MBP protein are used as a model substrate for formaldehyde crosslink and reverse crosslink test; **(b, c)** Formaldehyde crosslink and reverse crosslink test; **(d)** Crosslink test on HEK293T cells, and proteinase K to reverse crosslink under different condition.

AMT (4'-aminomethyltrioxsalen hydrochloride) is a psoralen derived crosslinker for direct RNA-RNA interaction crosslinking. AMT specifically crosslinks opposing uridine bases in AU base pairing under long wavelength UV irradiation. The resulting duplex can be reverse crosslinked under short wave length UV irradiation[51], [66], [70].

To test crosslinking, eight *in vitro* transcribed influenza (PR8) RNAs and the HIV-d short RNA were used as model substrates (**Figure 4.10a**). We tested crosslinking with different concentrations of AMT under 365nm UV for 800mJ or (10+10+10) min. With increasing AMT concentration under 365nm UV for (10+10+10) min, crosslinking could be observed as an upwards gel mobility shift (**Figure 4.10c**). As a negative model substrate, no crosslink product could be observed on the HIV-d short RNA (**Figure 4.10b**), since the kiss loop interaction is mediated by GC based pairing.

To test the reverse crosslink, the AMT crosslinked PR8 RNAs from *in vitro* transcription (IVT) were irradiated with 254nm UV for 2/5/10 min on ice, following by ethanol precipitation and gel

electrophoresis. The results showed that the AMT crosslinking products decreased significantly after 2 min (**Figure 4.10d**). However, RNAs started to degrade after 10 min (**Figure 4.10d**). Therefore, 254nm UV irradiation for 5 min was used for reverse AMT crosslinks in later experiments.

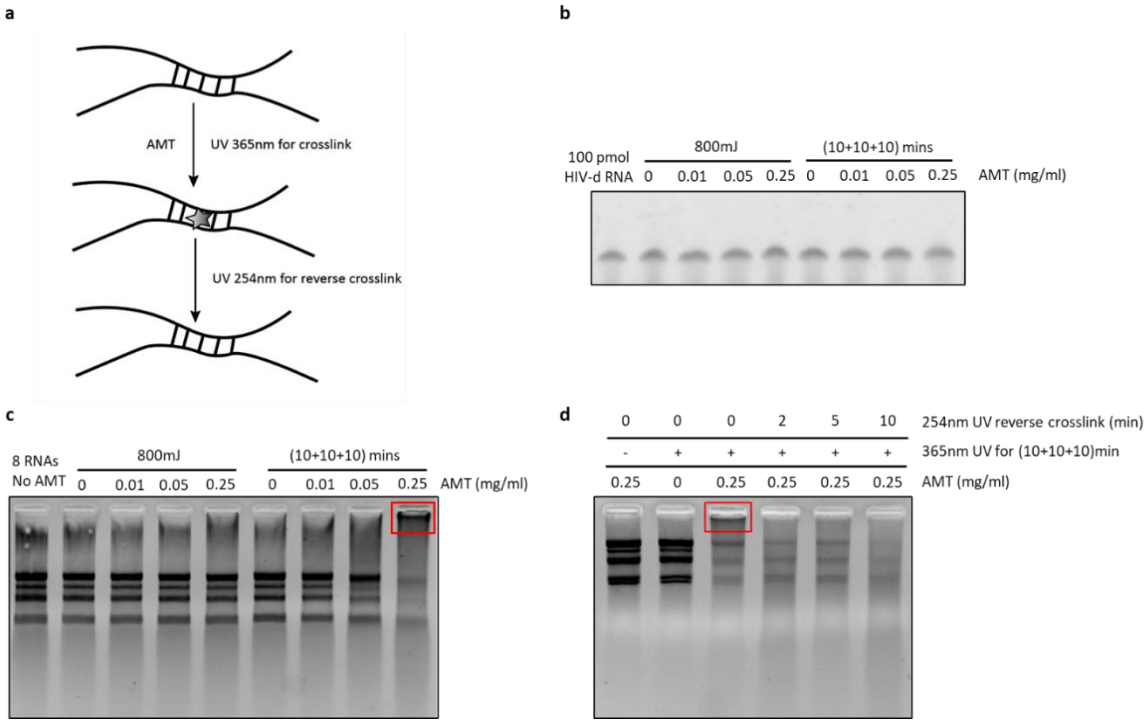


Figure 4.10| AMT crosslink optimization. (a) diagram to show AMT crosslink and reverse crosslink; **(b)** AMT crosslink test on HIV DIS dimerization short RNA; **(c)** AMT crosslink test on *in vitro* transcribed PR8 RNAs, the red frame highlights the crosslink product; **(d)** reverse AMT crosslink test on IVT PR8 RNAs, the crosslink product disappeared after 254nm UV irradiation.

Spatial 2'-Hydroxyl Acylation Reversible Crosslinking (SHARC) is another recently developed RNA-RNA crosslinking reagent [55]. Unlike AMT, which has a preference for AU base-pair, SHARC was reported to have less bias, with the advantage of easy preparation, high efficiency, and ease of crosslink reversal.

To prepare the SHARC reagent, we mixed one part of 2,6-Pyridinedicarboxylic acid and 2 parts of 1,1'-carbonyldiimidazole (CDI) in DMSO for 1h (**Figure 4.11a**). To test crosslinking, HIV-d RNA was used as a model substrate (**Figure 4.11b**), with the results showing that SHARC could crosslink HIV-d efficiently, as more that 50% of the substrate shifted up when 100mM SHARC reagent was added (**Figure 4.11c**). To reverse crosslinking, decrosslinking buffer was added into crosslinked samples (final concentration at 100 mM Boric acid, pH 11), followed by incubation at 45°C for 2 h. The reverse crosslink efficiency was fairly high, as at least 90% of the crosslinked model substrate was reversed (**Figure 4.11c**).

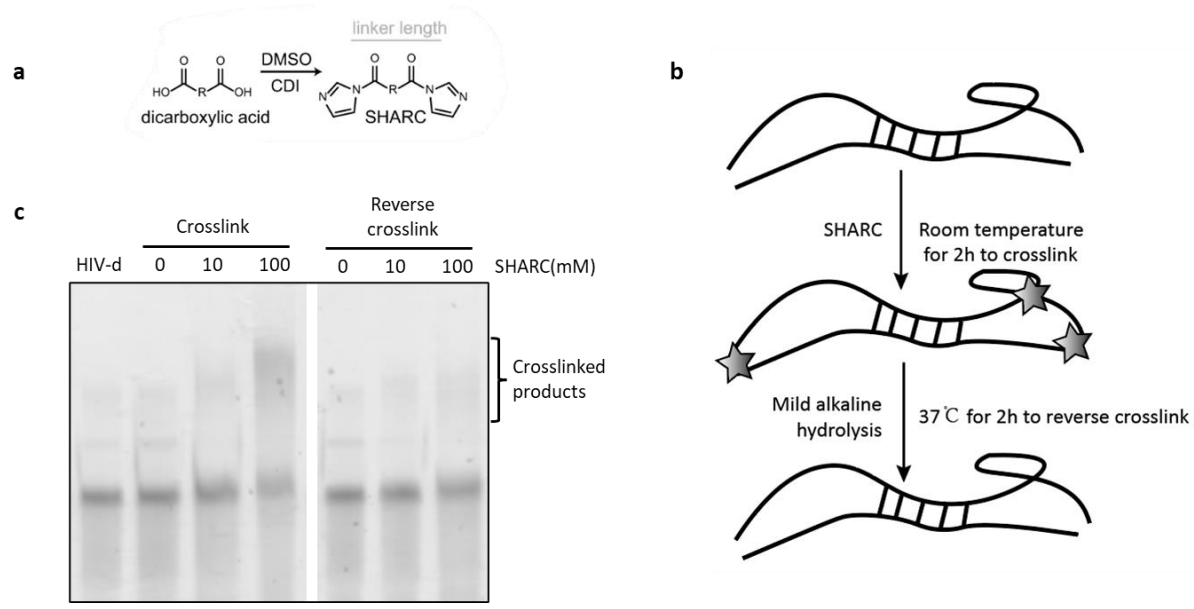


Figure 4.11| SHARC crosslink optimization. (a) Dicarboxylic acid reacts with 1,1'-carbonyldiimidazole (CDI) in DMSO to get SHARC reagent[55]; **(b)** diagram to show SHARC crosslink and reverse crosslink; **(c)** Urea-PAGE to show the SHARC crosslink and reverse crosslink on HIV-d RNA.

4.3.5 Fragmentation optimization

Following crosslinking, the next step in RNA-RNA-seq is RNA fragmentation. To accurately determine RNA-RNA interaction sites, and RNA duplexes must not be too long, but should have sufficient length to align to a reference sequence. The ideal RNA fragment is in the range of 50-200 nts. Here, I tested several fragmentation methods: nucleases digestion, heating in the presence of Mg^{2+}, sonication (Bioruptor), and Lead (II) fragmentation.

RNase A is a stable, sufficient endonuclease from mammalian pancreas, which cleaves single-stranded RNA and forms 3' -phosphate and 5' -hydroxyl ends. 2 µg of IVT PR8 RNAs were incubated with RNase A (final concentration 0.4 µg/µl) for 10 min or 30 min. The reaction was stopped by adding EDTA to 50 mM and heating at 70°C for 10 min followed by Trizol extraction. Under these conditions, RNA degradation was complete (**Figure 4.12a**).

S1 nuclease is an endonuclease that cleaves single-stranded RNA but forms 5' -phosphate and 3' -hydroxyl ends. Nuclease digestion was carried out on IVT transcribe substrates as indicated above, with the inclusion of 2µl of 1:100 diluted S1 nuclease (Promega). The result showed that S1 nuclease was similar to RNase A, with complete digestion of the substrate (**Figure 4.12a**).

RNase If is a recombinant protein fusion of RNase I which cleaves both single-stranded and double-stranded RNA but prefers single-stranded, and leaves a 5´-hydroxyl and 3´-phosphate. In addition, it

can be inhibited by SuperRNasin. The result showed that it could digest RNA very efficiently (**Figure 4.12b**). However, similar to RNase A and S1 nuclease, it was difficult to stop the reaction.

Magnesium can be used to fragment RNA as a divalent cation at elevated temperatures. PR8 IVT RNAs were heated to 94^{o}C for 3 or 5 min in 20 mM Tris-Acetate, pH 8.1, 50 mM K-Acetate, 15 mM Mg-Acetate. The reaction was stopped by transferring the RNA samples on ice. The results showed that the fragmentation was sufficient, and also that it was easy to stop the reaction at the desired fragmentation size (**Figure 4.12c**).

Sonication is also a common method to fragment DNA or RNA. Here, we tested a Bioruptor ultrasonication device to fragment IVT PR8 RNAs. The results indicated that it was not as efficient as the previous fragmentation methods, and the efficiency depended on the buffer component (**Figure 4.12d**).

Lead(II) is known to cleave RNA within single-stranded regions at mild temperature, whereas cleavages in double-stranded regions are weaker or absent. Thus, it can be used as a method to probe RNA structure[71][72]. I used purified human ribosome as substrate to test the $PbOAC_2$ fragmentation manner at 37^{o}C. The results showed that at 10mM, $PbOAC_2$ could fragment ribosome RNA effectively (**Figure 4.12e**).

Summarizing these data, we could see that RNase A, S1 nuclease and RNaseIf nucleases can efficiently cleave RNA, but they were difficult to control. Magnesium fragmentation was efficient and easy to control, but the high temperature was not good for samples with protein and might denature RNA to create artificial interactions. Sonication was very easy to control, but was not as efficient as RNases and the efficiency depends on the buffer condition. Finally, Lead(II) fragmentation was efficient under mild conditions, was easy to stop by EDTA, and in theory also acts as an RNA structural probe by cleaving specifically at single stranded RNA. Given the above, sonication or Lead(II) were the fragmentation methods that we chose toapply in the biological samples during the RNA-RNA seq protocol.

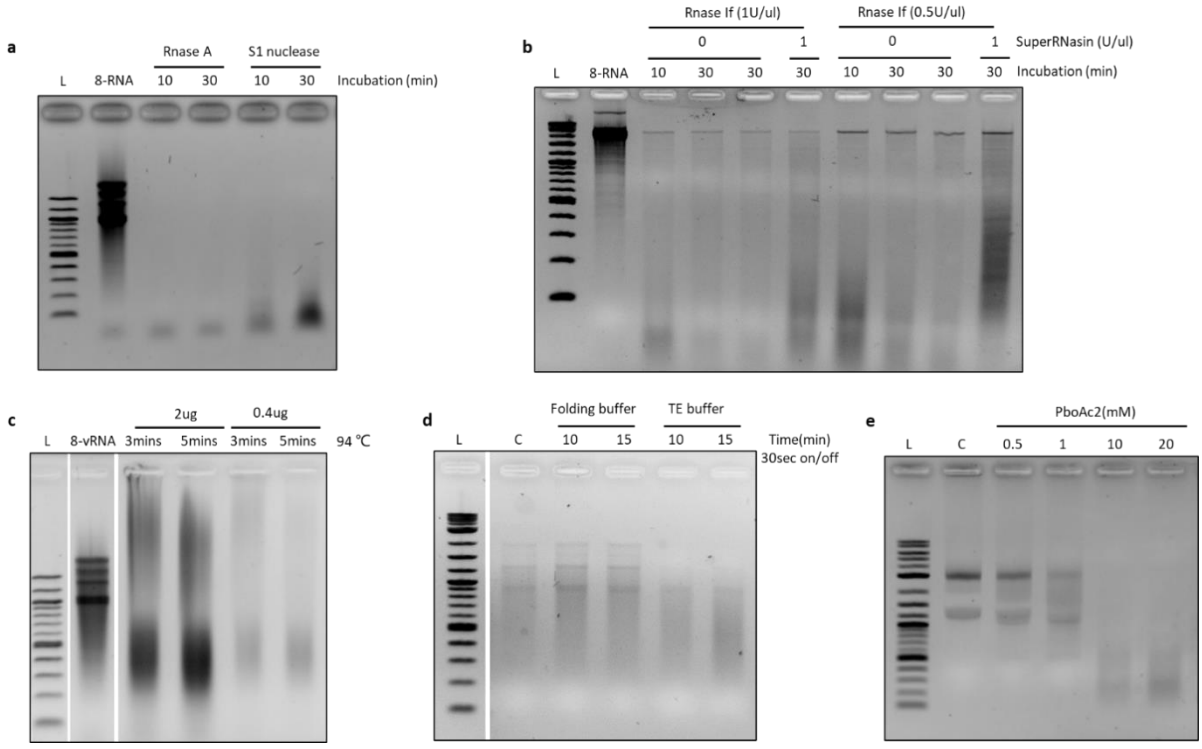


Figure 4.12| RNA fragmentation test. (a) RNaseA and S1 nuclease fragmentation test on *in vitro* transcribed PR8 RNAs; **(b)** RNaseIf fragmentation and SuperRNasin inhibition on *in vitro* transcribed PR8 RNAs; **(c)** Magnesium fragmentation; **(d)** sonication (Bioruptor) fragmentation; **(e)** Lead(II) fragmentation.

4.3.6 Run whole RNA-RNA seq protocol on ribosome

Ribosomal RNA is the primary component of ribosomes, which is the most abundant cellular RNAs. In addition, there are cryo-electron microscopy reconstructions and X-ray crystal structures of eukaryotic ribosomes [73]–[75]. Thus, the ribosome is an excellent structural model to validate the RNA-RNA seq protocol.

After optimization of crosslinking and reverse crosslinking conditions, fragmentation, adaptor ligation (T4 RNA ligation), proximity ligation (RtCB ligation), stringent wash, and library preparation on model substrates, I validated the RNA-RNA seq protocol on the purified eukaryotic ribosome.

The ribosome was purified from HEK293T cells by sucrose gradient. Here, I had six samples in total:

1. No crosslink ribosome with low stringency wash buffer;
2. AMT crosslink ribosome with low stringency wash buffer;
3. Formaldehyde crosslink ribosome with low stringency wash buffer;

4. No crosslink ribosome with high stringency wash buffer;

5. AMT crosslink ribosome with high stringency wash buffer;

6. FA crosslink ribosome with high stringency wash buffer;

4 µg of purified ribosome was prepared following the crosslinking and fragmentation protocols determined in the optimization test. RNA fragments were dephosphorylated by Shrimp Alkaline Phosphatase (rSAP) at 37°C for 40 min to obtain end repaired RNA with 3'-hydroxyl and 5'-hydroxyl. rSAP was then inactivated by heating to 80°C for 4 min. The no crosslink and AMT crosslinked samples were purified by RNA Clean & Concentrator™-5 Zymo-Spin™ IC Columns, while formaldehyde crosslinked samples were precipitated by ethanol. Next, RNAs were dual ligated with excess biotin and 3'5'-P adaptors, then purified by biotin-streptavidin selection to remove unligated products and ligation products without biotin adaptor. On bead proximity ligation was then performed using optimized RtCB ligation conditions. Following proximity ligation, crosslinking was reversed and washing performed with low stringent wash buffer ([x2 SSC, 1mM EDTA], 5 min @ 40°C, 3 times) or high stringent wash buffer ([100 mM Tris pH 7.0, 4 M NaCl, 10 mM EDTA, and 0.2 v/v % Tween 20]: for six times at 80°C). Lastly, RNA was reverse transcribed with a primer designed to anneal specifically to the 3'5'P adaptor, so that only the expected duplex ligation products with both adaptors can be amplified for sequencing.

The gel results showed that we were able to successfully obtain a sequencing library using the RNA-RNA seq protocol (**Figure 4.13a**). Unexpectedly, we obtained sequencing libraries not only from AMT crosslinking samples, but also from the no crosslink control samples, which indicated that washing steps were unable to completely remove non-crosslinked interactions. In other words, either these interactions are very strong, or there were strong artificial interactions that had been generated during the protocol. Unfortunately, the libraries from formaldehyde crosslink sample did not look as good as other samples. Nevertheless, we continued the sequencing on all samples that generated a sequencing library.

We successfully obtained sequencing data from no crosslinked, AMT crosslinked samples, and as expected from the poor quality of the sequencing library, very few reads from the formaldehyde sample. From these data, we plotted the interactions in 28S as heat map matrix (**Figure 4.13b**), and we could see that our sequencing data cover most regions of the 28S ribosomal RNA. By calculating an interaction strength based on the read counts, we observed some structured domains such as those between 0-100nt, 600-800nt, 1200-1400nt. By comparing to the interaction matrix from the crystal structure (**Figure 4.13b**), we observed that a lot of our interactions were consistent with the crystal structure, which indicated that the interactions we detected by RNA-RNA seq were real. Interestingly, flexible structural regions missing in the crystal structure were captured by RNA-RNA-seq indicating that it can provide complementary information.

However, the sequencing data also revealed that the protocol still needs to be optimized. We found many reads without sequences derived from RNA duplexes, rather they are the result of direct

ligation between the two adaptors. We suspect that over time the blocking 3' phosphate on the adaptors is degraded resulting in spurious adaptor-adaptor ligation. To solve this problem, we purified the two adaptors by HPLC.

Another adaptation to improve RNA-RNA-seq is the use of limited concentrations of one of the two adaptors in a two-step ligation strategy. First, T4 RNA ligation was performed using a limited concentration of the biotin adaptor followed by the streptavidin bead selection. Next, a second T4 RNA ligation was performed using high concentrations of the 3'5'P adaptor. In theory, this strategy ensures that the majority of RNA duplexes contain one of each adaptor. However, this strategy did not improve sequencing library quality, so it was not pursued further (**Figure 4.13d**).

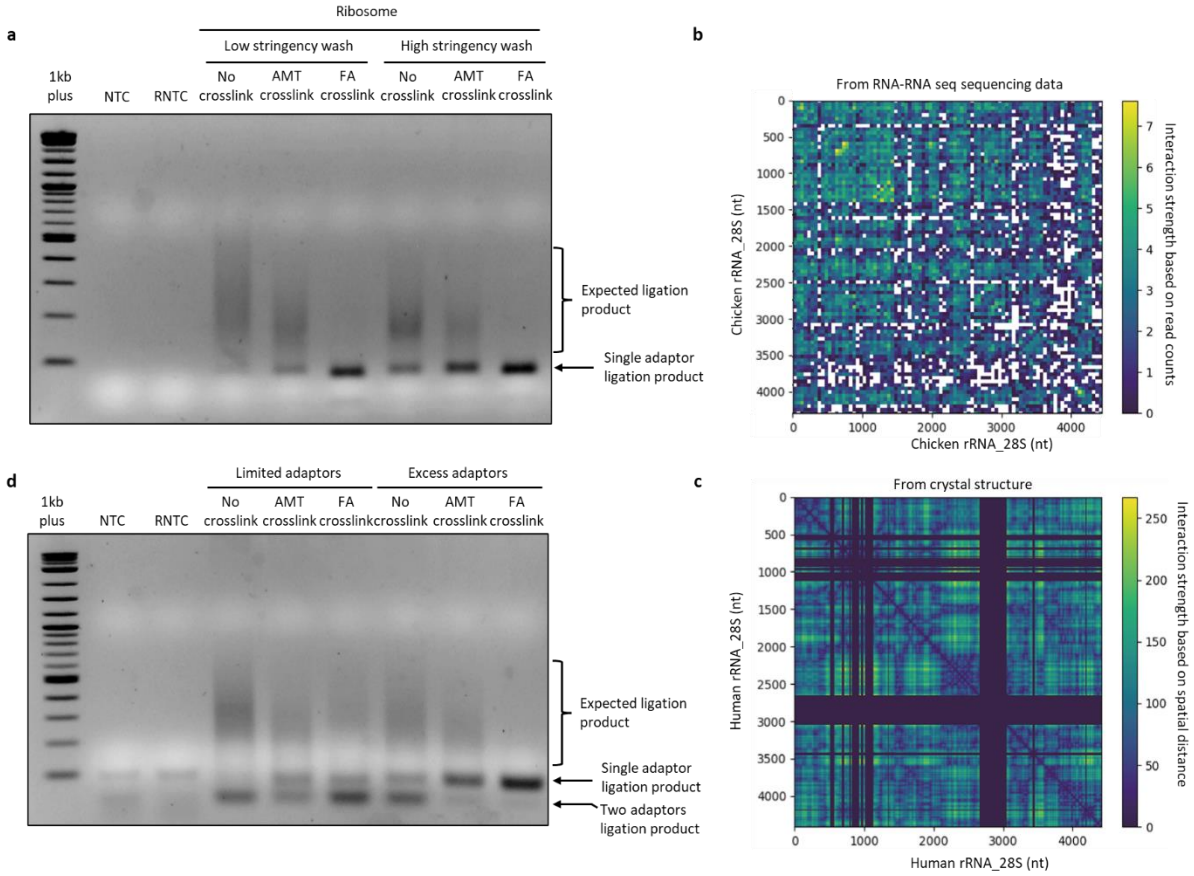


Figure 4.13| RNA-RNA seq on purified ribosome. (a) Agarose gel to show libraries after RNA-RNA seq; **(b)** RNA interactions in no crosslink ribosome (Chunk size: 50, threshold: 10); **(c)** Ribosome RNA interactions from crystal structure. The dark region indicates the structure information is missing in the crystal structure of the human ribosome **(d)** agarose gel to show the libraries after RNA-RNA seq on purified ribosome with limited or excess adaptors.

4.3.7 Stringent wash optimization

From the results above, we could see that we can generate library and detect RNA-RNA interactions from no crosslinked samples. This was unexpected because stringent washing should remove all non-crosslinked interactions, suggesting that washing for six times at 80^{o}C with Bead Wash Buffer (100 mM Tris pH 7.0, 4 M NaCl, 10 mM EDTA, and 0.2 v/v % Tween 20) was not stringent enough to remove all non-crosslinked interaction. Optimal wash conditions should remove all side products and unspecific interactions whilst keeping the streptavidin-biotin interaction intact. This is not so straightforward because conditions that break RNA-RNA interactions may also break the biotin-streptavidin interaction. Here, I used a biotinylated RNA oligo annealed to a reverse complementary DNA oligo as a model to test the stringency of different wash buffer (**Figure 4.14a**).

Wash buffers and wash condition:

wash 1: 4M Urea, 2% SDS, 150 mM NaCl, 1mM EDTA, 5 min @ 60°C, repeat 3 times

wash 2: 4M Urea, 2% SDS, 150 mM NaCl, 1mM EDTA, 5 min @ 50°C, repeat 3 times

wash 3: 4M Urea, 2% SDS, 150 mM NaCl, 1mM EDTA, 5 min @ 40°C, repeat 3 times

wash 4: x2 SSC, 70% formamide, 1mM EDTA, 5 min @ 40°C, repeat 3 times

wash 5: x2 SSC, 70% formamide, 1mM EDTA, 5 min @ room temperature, repeat 3 times

wash 6: x2 SSC, 60% formamide, 1mM EDTA, 5 min @ 40°C, repeat 3 times

wash 7: x2 SSC, 60% formamide, 1mM EDTA, 5 min @ room temperature, repeat 3 times

First, equimolar amounts of the biotinylated RNA oligo (RNA-A) and the reverse complementary DNA oligo (DNA-B) were hybridized by denaturing at 90^{o}C for 2 min, chilling on ice, and annealing in 1x T4 RNA ligase buffer at 37^{o}C for 30 min to form the interaction. Next, the annealed duplex was coupled to streptavidin magnetic beads, as previously described. Flow-through was collected and the beads were then washed with the listed wash buffer under indicated condition. Lastly, RNA was eluted from the streptavidin magnetic beads and purified by ethanol precipitation and loaded on 10% Urea-PAGE.

From the gel we could see, the that the flow-through was clean, which indicated successful coupling of the hybrid to the streptavidin magnetic beads. Surprisingly, many of these wash conditions were unable to completely remove DNA-B from the RNA oligo coupled to beads. Wash 2, wash 3, wash 5, and wash 7 were particularly inefficient, as most DNA-B remained annealed with RNA oligo. Wash 1 and wash 6 washed were able to partially remove DNA-B, but also disrupt the interaction between streptavidin-biotin. Wash 4 was able to remove all DNA-B oligo, although some loss of the RNA-A from the beads was still observed. It seems that it is possible to break a strong DNA-RNA hybrid by

stringent washing, but that this also results in the weakening of the biotin-streptavidin interaction. Thus, there is a trade-off in potential specificity of the protocol and the ability to capture material for analysis.

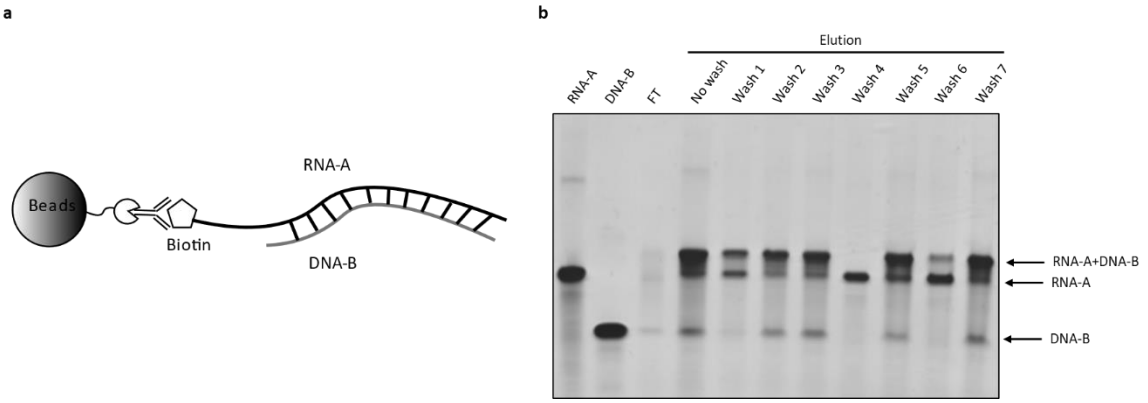


Figure 4.14| Stringent wash test. (a) wash test model; **(b)** Urea-PAGE to show the wash efficiency. The first two lanes showed the oligos we used, then flow-through (FT). Comparing with no wash, wash 2, wash 3, wash 5, and wash 7 were not efficient enough to wash away DNA-B; wash 1 and wash 6 washed were able to partially remove DNA-B, but also disrupt the interaction between streptavidin-biotin; wash 4 was able to remove all DNA-B oligo.

4.3.8 RNA-RNA seq optimization on formaldehyde crosslink samples

One possibility for why the RNA-RNA seq protocol didn't work on formaldehyde crosslink samples is that the fragmentation was not sufficient, resulting in insufficient RNA substate for ligation.To test whether Lead(II) fragmentation is impaired in the formaldehyde crosslinked sample, we performed a fragmentation test after crosslinking with or without quenching with glycine[76]. We performed this fragmentation test on formaldehyde crosslink samples with glycine at different pHs. The gel results showed that glycine decreased fragmentation efficiency of Lead(II), especially at lower pH (**Figure 4.15a**). Formaldehyde crosslinking itself also decreased fragmentation efficiency. To increase the fragmentation of the formaldehyde crosslinked samples, we combined bioruptor sonication and Lead(II) fragmentation. The results showed the 0.5% formaldehyde crosslinked polysome can be fragmented efficiently by 20 mM of PboAC2 incubated at 37^{o}C for 20 min with 10 min bioruptor sonication (**Figure 4.15b**).

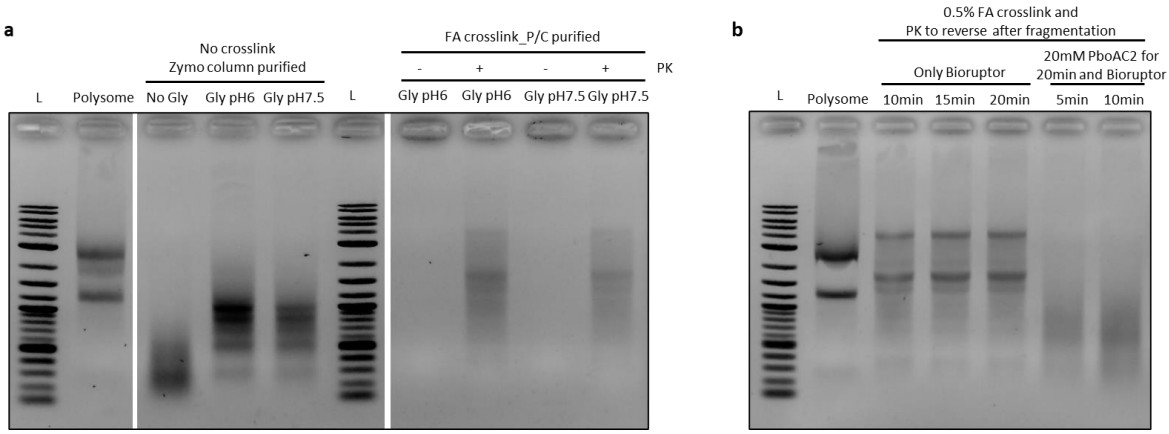


Figure 4.15| Fragmentation optimization on formaldehyde crosslink samples. (a) Lead(II) fragmentation efficiency is diminished by glycine, low pH, and formaldehyde crosslink; **(b)** Fragmentation with higher concentration PboAC2 and bioruptor.

4.3.9 RNA-RNA seq optimization on cell samples

After the optimization of stringent wash and fragmentation on formaldehyde crosslink samples we implemented the optimized RNA-RNA seq protocol on virus infected cells. This sample has the additional advantage that the protocol can be simultaneously validated on the human ribosome structure.

There are eight samples in total and the experiment is carried out in replicate:

1. PR8 infected MDCK at 1MOI and harvest at 12hpi; No crosslink;
2. PR8 infected MDCK at 1MOI and harvest at 12hpi; AMT crosslink;
3. PR8 infected MDCK at 1MOI and harvest at 12hpi; SHARC crosslink;
4. PR8 infected MDCK at 1MOI and harvest at 12hpi; Formaldehyde crosslink;
5. NL43 transfected HEK293T and harvest at 24h; No crosslink;
6. NL43 transfected HEK293T and harvest at 24h; AMT crosslink;
7. NL43 transfected HEK293T and harvest at 24h; SHARC crosslink;
8. NL43 transfected HEK293T and harvest at 24h; Formaldehyde crosslink;

The infected or transfected cells were harvest as indicated, using the following conditions: no crosslink, AMT crosslink (0.25 mg/ml, 365nm UV irradiate for 10+10+10 min), and formaldehyde crosslink (0.5% Formaldehyde). Next, RNAs from the no crosslink, AMT crosslink and SHARC crosslink cells were purified by Trizol extraction and Turbo DNase treatment according to the manufacturer's instructions. Purified RNA was fragmented by Lead(II) and end repaired with rSAP and T4 PNK at 37°C for 40 min following by Zymo column purification. Formaldehyde crosslink cells were permeabilized and fragmented as the optimized protocol above. End repair was performed with rSAP and T4 PNK at 37°C for 40 min. Cells were washed with cold PBS buffer and centrifuged at 1200*x*g for 10 min at 4°C between each step. All samples were ligated to excess biotin adaptor and 3'5'P adaptor using T4 ligase, then bound to streptavidin beads to remove ligation products without a biotin adaptor. Subsequently, RtCB ligase was used for proximity ligation on beads for no crosslink, AMT crosslink and SHARC crosslink samples. *In situ* ligation was performed on formaldehyde crosslink samples. After, crosslinks were reversed using the optimized protocols: for AMT crosslinked samples, 254nm UV irradiation for 10 min on ice; for SHARC crosslinked samples, 5x decrosslinking buffer (500mM Boric acid, pH 11) was added to 1x and incubated for 2 hours at 45°C; formaldehyde crosslinked samples were incubated with proteinase K (NEB) (final concentration 0.03U/µl) in the buffer contain 20 mM Tris-HCl pH 7.4, 10 mM EDTA, 2% N-lauroylsarcosine, 2.5 mM TCEP, and incubate at 65°C for 1.5 h. Subsequently, all purified reversed crosslinked and no crosslinked RNA were subject to high stringent wash ([x2 SSC, 70% formamide, 1mM EDTA], 5 min @ 40°C, repeat 3 times) on beads to wash away non-specific products. Lastly, the RNA was reverse transcribed with a primer that specifically binds to 3'5'P adaptor. This step enriches for the expected duplex ligation product.

The agarose gel showed that I was able to obtain a sequencing library from all samples (Figure 4.16a), especially the libraries from no crosslink, AMT crosslink, SHARC crosslink samples. The library size indicated that the inserts were between 20-200bp, as desired. For the formaldehyde crosslink samples, the libraries were of poor quality, and had a much smaller predicted insert size.

We sequenced all libraries, and the results looked very promising. Using the more stringent wash, we managed to detect more interactions on the crosslink sample compared to the no crosslink sample. For instance, we were able to see a known interaction between 28S rRNA and 5.8S rRNA within the large subunit of ribosome. 342 GCACGAGACC 351 (28S) base-paired with 24 GGCUCGUGC 32 (5.8S) showed up in no crosslink sample interaction matrix, but was more intensely captured with AMT crosslinking. Similarly, an interaction between 2 GCGACCUCAGAUCAGACG 19 (28S) and 138 CGCCUGUCUGAGCGUCGCC 155(5.8S) showed up in AMT crosslink sample, which was stronger again in SHARC crosslink sample. These results indicate that the stringent wash cannot wash away all non-crosslinked interaction. Nevertheless, these results strongly argue that the interactions from no crosslink sample were valid, and likely represent very strong RNA-RNA interactions that cannot be broken during stringent washing (**Figure 4.16b, 4.16c**). As expected from the gel of the sequencing library, we didn't obtain high quality information from formaldehyde crosslink samples.

Unfortunately, we were unable to recover many viral sequences from this pilot experiment, due to the high abundance of ribosomal RNA in the samples. Nevertheless, the quality of the ribosome

interaction data indicates that the protocol is able to faithfully capture native and biologically relevant RNA-RNA interactions. To perform RNA-RNA interaction analysis on viral RNA, we will need to sequence cellular samples deeper, perform ribosomal RNA depletion before sequencing, or perform experiments on large quantities of purified virus.

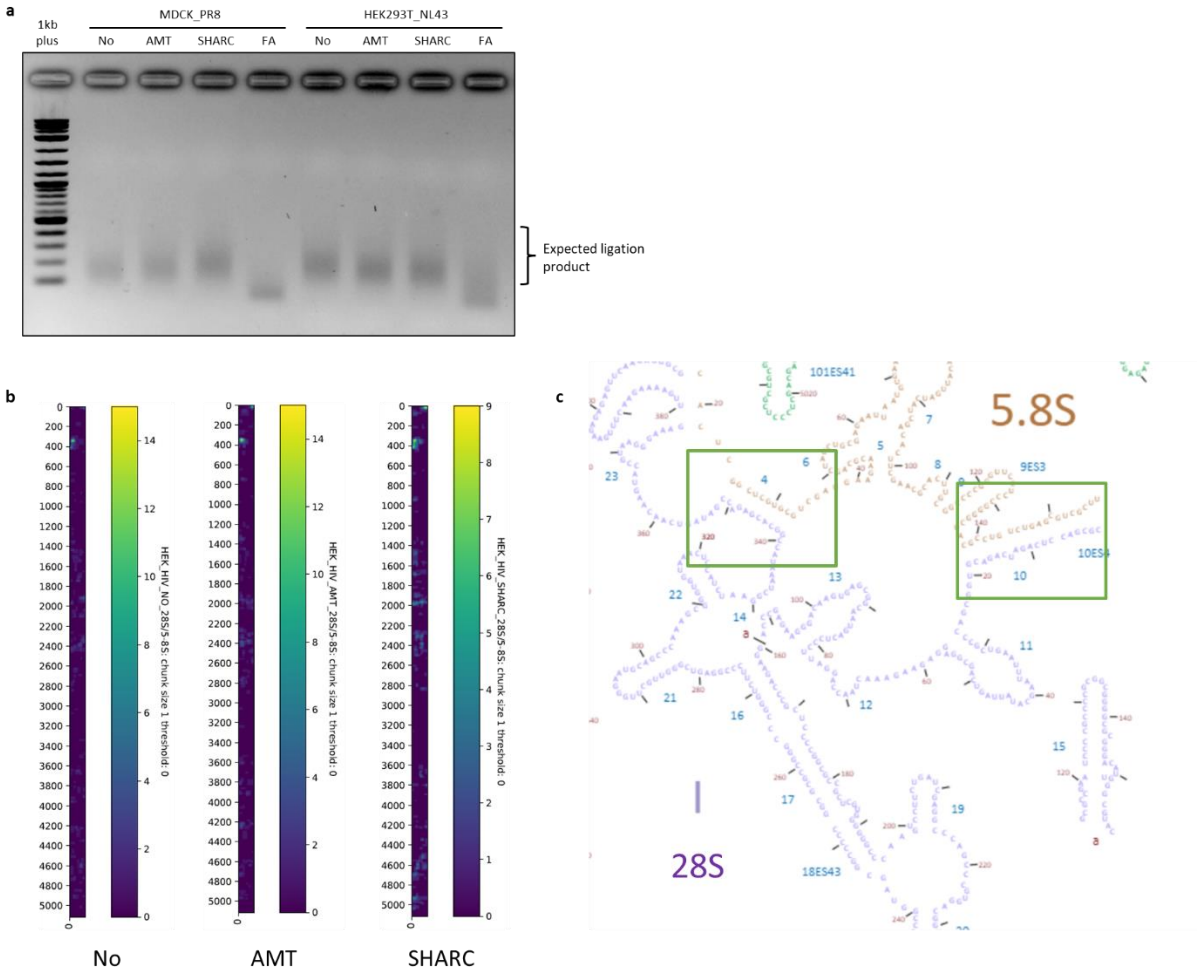


Figure 4.16| RNA-RNA seq on infected or transfected cells. (a) Agarose gel to show the libraries after RNA-RNA seq on PR8 infected MDCK cells and HIV(NL43) transfected HEK293T cells; **(b)** interaction matrix to show the interactions between 28S and 5.8S rRNA in no crosslink, AMT crosslink and SHARC crosslink samples; **(c)** human ribosome secondary structure (Adapt from http://apollo.chemistry.gatech.edu/RiboVision/#HS_LSU_3D)

4.3.10 RNA-RNA seq on purified influenza virus

After the optimization of RNA-RNA seq on a model substrate, the human ribosome, and cellular lysates, I decided to apply the optimized protocol on virus. Our goal included identifying the intra-molecular RNA structure within each segment and the inter-molecular interactions between the

eight segments of the virus genome in high resolution. In principle, we will be able to define the nature of the interactions (RNA-RNA interactions vs RNA-protein-RNA interactions) by comparing the interactome between formaldehyde crosslink sample and AMT crosslink sample. Ultimately, we would aim to obtain a high-resolution view of influenza genome architecture and find out how these interactions impact the influenza replication and evolution. For the first attempt on virus, I planned to apply the protocol on PR8, which is a laboratory adapted strain and easy to propagate. To test if the protocol works well on virus, we prepared the PR8 from specific pathogen free (SPF) chicken embryo following the published protocol[77]. The titre was 1.4*10^7 PFU/ml. 10 ml of virus were pelleted through 20% sucrose cushion at 17000*xg* for 3 h and then resuspended with 20 µl of TE buffer (10 mM Tris, 1 mM EDTA, pH8.0).

There were three samples in total: no crosslink/ AMT crosslink/ formaldehyde crosslink. The crosslink and fragmentation protocols followed the parameters in the optimization test. Then, the virions were permeabilized with 1% NP40. RNA fragments were dephosphorylated by Shrimp Alkaline Phosphatase (rSAP) at 37°C for 40 min to obtain RNA with 3' -hydroxyl and 5' -hydroxyl end, followed by rSAP inactivation by heating to 80°C for 4 min. The no crosslink and AMT crosslink virus samples were purified by phenol chloroform and precipitated by ethanol, while formaldehyde crosslink viruses were only precipitated by ethanol. Next, RNAs were ligated with biotin adaptor and 3'5'P adaptor, then captured by streptavidin selection to remove ligation products that didn't have a biotin adaptor. Subsequently, RtCB ligase was used for proximity ligation on beads. After crosslink reversal we performed stringent washing to remove unwanted side products. Lastly, we reverse transcribed the RNA with specific primer that binds to the 3'5'P adaptor. Only the expected duplex ligation product would be selected and amplified for sequencing in the end.

The gel results showed that the RNA-RNA seq protocol worked as expected (**Figure 4.17a**). Similar to the assays on previous biological samples, we obtained sequencing libraries from all samples. Importantly, we recovered sequences from all eight segments of influenza virus. Based on read number, we plotted the interaction network of PR8 genome. We found interactions between every segment pair indicating a complex network that potentially differs between viruses. However, the interactions between PB1 and PB2 were detected most frequently. Using an unbiased network analysis, we saw that PB2 was organised in the centre of the network. As PB2 is one of the longest segments, these results are consistent with the published electron tomography of influenza A virions[29] (**Figure 4.17b**). Intriguingly, we found several positions within the network which were able to interact with more than one RNA. This strongly suggests heterogeneity and/or flexibility in influenza assembly mediated by RNA-RNA interactions, as recently reported[78][79].

Focusing on RNA structure of individual segments, we found intra-molecular interactions, especially between the 3' and 5' termini of the vRNAs. This is consistent with our knowledge about the influenza genome that each segment has conserved region U12 and U13, and these conserved sequences are partially complementary and anneal to form a hairpin structure essential for transcription and replication[19]. Interestingly, even though I pelleted the virus through sucrose cushion, we were able to detect interactions with ribosomal RNA. This could either indicate ribosome

contamination, or alternatively that influenza viruses can package ribosomal RNA, as previously suggested[80]. Despite a highly optimized protocol, we still obtained large number of sequencing reads without biotin adaptor, and for the formaldehyde sample, there was only a few reads after adaptor trimming. Formaldehyde may overcrosslink the samples leading to very short insert sizes.

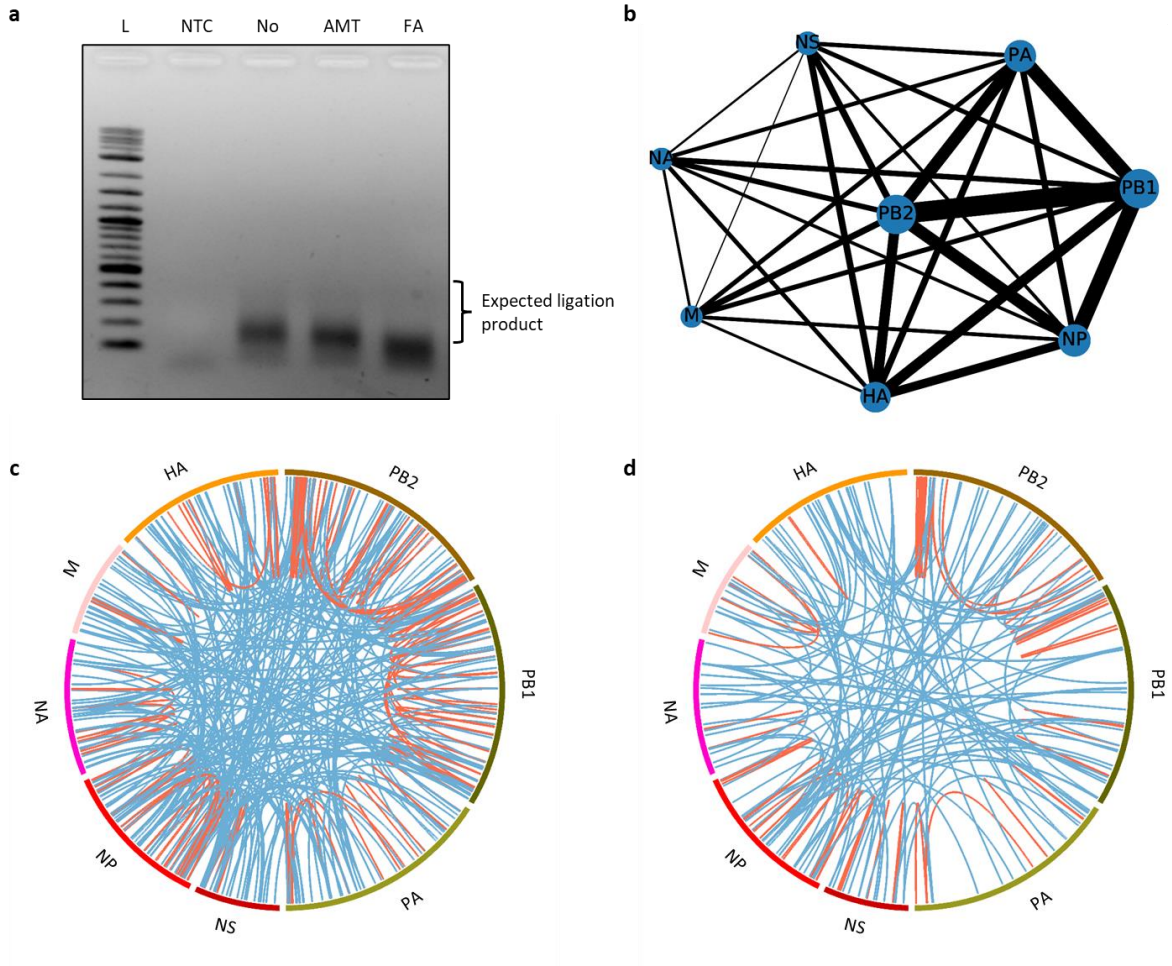


Figure 4.17| RNA-RNA seq on PR8 virus. (a) Agarose gel to show libraries after the whole protocol; **(b)** Inter-molecular interactions between eight segments; **(c)** Inter and intra-molecular interactions found in no crosslink and **(d)** AMT crosslink samples, red lines represent intra-molecular interactions, and blue lines represent inter-molecular interactions;

4.3.11 Virus purification

RNA-RNA-seq was successful on cellular and virus samples, but we didn't obtain sufficient data from the virus genome because of ribosomal RNA was too abundant. In an attempt to reduce this potential contamination, I tested several virus purification protocols.

I first determined the efficiency of pelleting virus through sucrose cushion, which is a common method to purify biological samples [81]–[83], and also the method that I used in the previous RNA-RNA seq run on virus. The result showed that pelleting through sucrose cushion can remove more than half of the ribosomal RNA, but the overall recovery of virus was only about 12% (**Figure 4.18a**). We speculate that the reason for the low recovery was that the pellet resuspension was not sufficient. I also tried to purify virus by Optiprep gradient[84], which is a powerful technique for fractionating particles with different sizes. However, I lost more than 90% of the virus. Second, I tested coupling virus to NHS-Activated Magnetic Beads, which forms covalent conjugation between the magnetic beads and amine groups on the virus surface proteins (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0011758_Pierce_NHSActiv_Mag_Bead_UG.pdf). Here I used 20 µl of beads to couple 50 µl of PR8 virus (4*10^8 PFU/ml). The qPCR result showed more than 90% of virus coupled to beads (**Figure 4.18b**), and after benzonase treatment[85] the majority of free ribosomal RNA could be removed. Next, I tested the binding capacity of NHS beads, since 50 µl of virus was not enough for an RNA-RNA seq run. With fixed amount of NHS beads, I added 50 µl/150 µl/500 µl/1000 µl of PR8 virus, and quantified the bound and unbound virus. The qPCR results (**Figure 4.18c**) showed 20 µl of NHS beads can bind up to 100 µl of virus at 4*10^8 PFU/ml (2-fold to 50 µl of input virus).

To sum up, both pelleting virus through sucrose cushion and NHS beads coupling can be used to highly purify virus from contaminating ribosomal RNA. Even though the virus recovery of sucrose cushion pelleting is not high, it can be used for large volume of virus. NHS beads coupling is more efficient, but the binding capacity is not high enough for large volumes of virus.

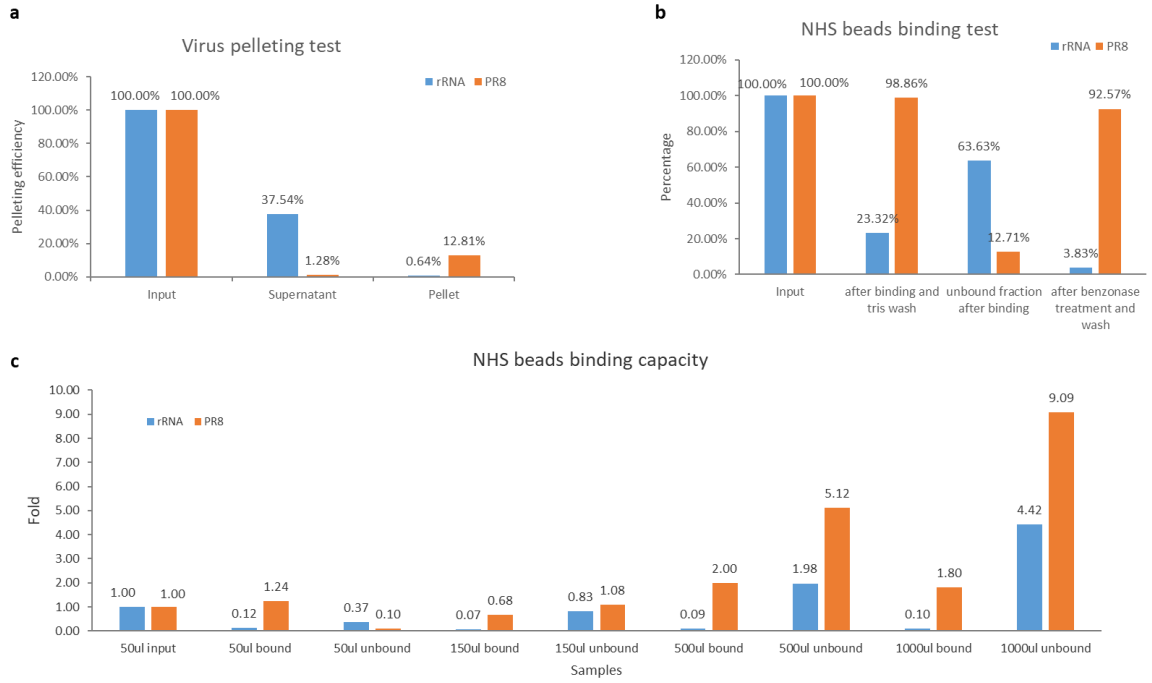


Figure 4.18| Virus purification test. (a) virus pelleting and purification efficiency through opti-prep cushion; **(b)** NHS beads binding efficiency; **(c)** NHS beads binding capacity.

4.3.12 Four-read sequencing for RNA-RNA seq library

As previously mentioned, our RNA-RNA-seq sequencing libraries contain many reads without a biotin adaptor. Presumably, these are side ligation products, or RNA molecules non-specifical bound to the streptavidin beads, that were not removed by stringent washing. To enrich the expected duplex ligation product and avoid wasting sequencing capacity, we developed a custom sequencing protocol: four-read sequencing (**Figure 4.19a**). In standard pair-end sequencing, sequencing follows the order: Read1, index 7, index 5, and Read 2, where read 1 and read2 are derived from the illumina adaptors attached to the ends of the interacting RNA molecules. In the four-read sequencing includes read X and read Y instead of two index reads. Importantly, read X and Y begin on the biotin adaptor itself. Using this custom sequencing strategy, only the expected duplex ligation product with biotin adaptor would be sequenced. After the clustering, libraries will be sequenced on the following order: Read X, Read 1, Read 2, Read Y.

We sequenced libraries from RNA-RNA seq on purified ribosome and PR8/HXB2 mixed samples (**Figure 4.19b**). As expected, all sequencing reads contained the biotin adaptor, which means the four-read sequencing strategy worked (**Table 4.5**). However, due to loading issues on this pilot run, we did not get enough reads as we expected. This issue occurred because ligation products without biotin adaptor were quantified by the library quantification kit, even though our custom protocol did

not sequence them. Thus, our sequencing run was underloaded compared to typical parameters recommended by Illumina. One solution is to quantify the library based using qPCR primers on the biotin adaptor. This will ensure correct loading of the Illumina flowcell. In this experiment, we obtained many ribosome reads, more than 97%, even the virus was purified through sucrose cushion.

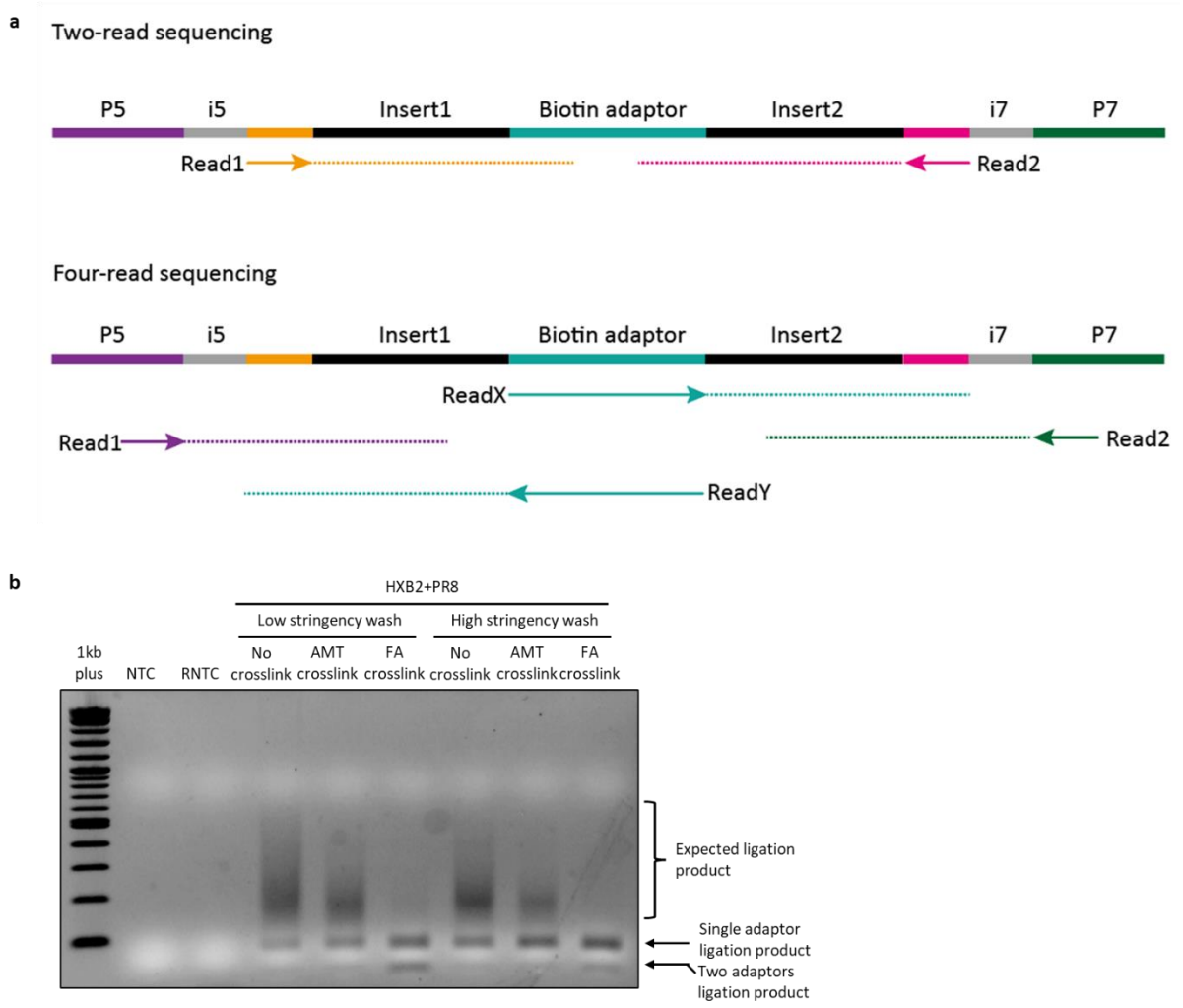


Figure 4.19| Two-read sequencing and four-read sequencing strategy. (a) diagram to show two-read sequencing and four-read sequencing; **(b)** agarose gel to show the library of RNA-RNA seq on mixed PR8 and HXB2 that purified by sucrose cushion pelleting.

Table 4.5 Ratio of different interactions from four-read sequencing

		Reads				Ratio (%)		
Crosslink	Wash stringency	HIV-PR8 inter	HIV-HIV	PR8-PR8	Total	Ribosome	HIV-HIV	PR8-PR8
NC	low	2197	5718	301	688048	98.81	2.60	0.14
AMT	low	392	2166	171	287019	99.05	5.53	0.44
FA	low	671	2098	96	133277	97.85	3.13	0.14
NC	high	50	369	224	605992	99.89	7.38	4.48
AMT	high	32	2999	208	239455	98.65	93.72	6.50
FA	high	6	76	48	16835	99.23	12.67	8.00

4.4 Conclusion and discussion

Influenza viruses contain eight segments which form a '7+1' supramolecular complex [29]–[31], but how influenza assembles these eight segments into viral particles and the details of the interactions are still not well understood. To sum up, I developed a high throughput sequencing and proximity ligation-based method, RNA-RNA seq, to study the interactions between the eight genomic segments of influenza virus. In developing this protocol, we aimed to maintain the advantages of the published high throughput sequencing and proximity ligation-based methods without their disadvantages. In principle, our RNA-RNA seq can measure direct (RNA-RNA) and indirect (protein-mediated) interactions without being limited by specific protein or RNA baits. Furthermore, two adaptors are used to demark the boundary of the interacted RNA and enrich the expected duplex, thus simplifying the analysis.

I validated the RNA-RNA-seq protocol on model substrates *in vitro* (HIV-d and HIV-nd), optimized the AMT, SHARC, formaldehyde crosslinking and reverse crosslinking conditions, RNA fragmentation methods including RNase A, S1 nuclease, RNase If, Magnesium and heating, sonication, and Lead(II),

T4 RNA ligation for adaptors ligation, and RtCB ligation for proximity ligation. With these optimized conditions, I was able to obtain interaction data from biological samples including ribosome, virus infected or transfected cells, and PR8 and HXB2 virions. Using the ribosome as a biological model to validate RNA-RNA-seq, we were able to identify known intermolecular interactions between the 18S and 5.8S rRNA. Importantly, we also obtained information that is missing from ribosome crystal and cryo-EM structures. We postulate that RNA-RNA-seq could be used to model the structure of rRNA regions that are very flexible, and thus invisible using biophysical structural approaches. Because the initial sequencing results showed there were a lot of reads without biotin adaptor, I optimized the stringent wash to enrich the expected duplex ligation product on model substrate. In addition, I also observed that the protocol was less efficient on formaldehyde crosslink samples compared with no crosslink and AMT crosslink samples. The test results suggested that the reason might be the fragmentation was not sufficient. With the optimization of stringent wash and crosslink and fragmentation on formaldehyde crosslink sample, I finally obtained good quality sequencing libraries from PR8 infected MDCK cells and NL43 transfected HEK293T cells.

So far attempts to perform RNA-RNA-seq on virus samples has been problematic because it is very difficult to generate enough start material for viral samples. I observed that in the absence of sufficient start material, I obtained poor sequencing libraries with many side products from adaptor self-ligation. PCR bias for short products may exacerbate this problem, but it may also reflect poor purity of the adaptors. In addition, the purification of virus samples is also very important because there were lots of ribosomes in influenza virus that was propagated from MDCK cell and SPF chicken embryo. Most likely, ribosomal rRNA is released from dying infected cells. Ultracentrifugation using a sucrose gradient to purify virus led to loss of much of the virus. Sucrose cushion pelleting is suitable for large amount of material and but results in low purity. I also tested the NHS beads coupling, which was very efficient at removing ribosomal RNA but not realistic for large volumes of material. An alternative solution is depleting rRNA before starting the RNA-RNA-seq protocol by commercial kits or depleting rRNA before library prep by Cas9-based depletion: Depletion of Abundant Sequences by Hybridization, DASH[86].

During the optimization of RNA-RNA seq protocol, there were several similar projects published [87][79]. Dadonaite et al. applied SPLASH on WSN (H1N1), PR8 (H1N1), Udorn (H3N2), and a reassorted PR8:Udorn virus (PB1+NA) (H1N2). They showed that both intra- and intersegment interactions were influenza strain dependent. They were able to validate some interaction hot spots between two segments and showed that they were essential for reassortment. Le Sage et al. used UV light and psoralen to crosslink WSN (H1N1) and their results were different from Dadonaite et al [87]. They observed that a single region could interact with several different segments, and that mutation of these hotspots would result in interaction network rearrangement. They claimed RNA-RNA interactions between influenza virus genome segments were redundant and flexible. From my point of view, the interactions between eight segments of influenza A virus are flexible in some way, something which may be evolutionary advantageous for a virus that commonly evolves by reassortment. Intriguingly, the interaction networks of WSN (H1N1) in the two publications are very different, which indicated that there may be methodical biases that are not understood.

More recently, a similar method, RIC-seq[39], [54] has been developed for *in situ* detection of RNA proximity interactions. RIC-seq applies formaldehyde to crosslink the RNA-RNA interactions, and similar to RNA-RNA seq, it ligates the duplex with biotin-pCp. In the end the proximity ligation products are re-fragmented and following library preparation. I attempted RIC-seq on our formaldehyde crosslink sample, but was unable to obtain a good quality library for sequencing. Similar to RNA-RNA-seq, RIC-seq requires a very large amount of starting material.

Overall, whilst there are still some difficulties to overcome, RNA-RNA seq still has the advantage that we can choose different crosslink reagent to limit the bias of a specific reagent; and with two specific adaptors, we can easily demark the boundary of the duplex, enrich the expected ligation duplex by biotin selection, specific primer to reverse transcribe only duplexed RNA.

In the near future, I will apply the optimized RNA-RNA seq protocol on large amounts of highly purified virus.

4.5 References

[1] M. Javanian, M. Barary, S. Ghebrehewet, V. Koppolu, V. K. R. Vasigala, and S. Ebrahimpour, "A brief review of influenza virus infection," *J. Med. Virol.*, vol. 93, no. 8, pp. 4638–4646, 2021, doi: 10.1002/jmv.26990.

[2] V. N. Petrova and C. A. Russell, "The evolution of seasonal influenza viruses," *Nat. Rev. Microbiol.*, vol. 16, no. 1, pp. 47–60, 2018, doi: 10.1038/nrmicro.2017.118.

[3] J. K. Taubenberger and J. C. Kash, "Influenza virus evolution, host adaptation, and pandemic formation," *Cell Host Microbe*, vol. 7, no. 6, pp. 440–451, 2010, doi: 10.1016/j.chom.2010.05.009.

[4] S. Pleschka, "Overview of Influenza Viruses," in *Assessment & Evaluation in Higher Education*, vol. 37, no. October, 2012, pp. 1–20.

[5] E. C. Hutchinson, "Influenza Virus," *Trends Microbiol.*, vol. 26, no. 9, pp. 809–810, 2018, doi: 10.1016/j.tim.2018.05.013.

[6] C. Paules and K. Subbarao, "Influenza," *Lancet*, vol. 390, no. 10095, pp. 697–708, 2017, doi: 10.1016/S0140-6736(17)30129-0.

[7] F. G. Genus, "Orthomyxoviridae," in *Virus Taxonomy*, no. Table 1, Elsevier, 2012, pp. 749–761.

[8] S. Payne, "Family Orthomyxoviridae," in *Viruses*, Elsevier, 2017, pp. 197–208.

[9] E. M. Abdelwhab and A. S. Abdel-Moneim, "Orthomyxoviruses," in *Recent Advances in Animal Virology*, Singapore: Springer Singapore, 2019, pp. 351–378.

[10] A. Trilla, G. Trilla, and C. Daer, “The 1918 ‘Spanish Flu’ in Spain,” *Clin. Infect. Dis.*, vol. 47, no. 5, pp. 668–673, 2008, doi: 10.1086/590567.

[11] L. Akin and M. G. Gözel, “Understanding dynamics of pandemics,” *Turkish J. Med. Sci.*, vol. 50, no. SI-1, pp. 515–519, 2020, doi: 10.3906/sag-2004-133.

[12] J. K. Taubenberger, J. C. Kash, and D. M. Morens, “The 1918 influenza pandemic: 100 years of questions answered and unanswered,” *Sci. Transl. Med.*, vol. 11, no. 502, 2019, doi: 10.1126/scitranslmed.aau5485.

[13] D. Dou, R. Revol, H. Östbye, H. Wang, and R. Daniels, “Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement.,” *Front. Immunol.*, vol. 9, no. JUL, p. 1581, 2018, doi: 10.3389/fimmu.2018.01581.

[14] C. Peteranderl, S. Herold, and C. Schmoldt, “Human In fl uenza Virus Infections,” *Int. J. Mol. Sci.*, 2017.

[15] Q. Gao, E. W. A. Brydon, and P. Palese, “A Seven-Segmented Influenza A Virus Expressing the Influenza C Virus Glycoprotein HEF,” *J. Virol.*, vol. 82, no. 13, pp. 6419–6426, 2008, doi: 10.1128/jvi.00514-08.

[16] S. Su, X. Fu, G. Li, F. Kerlin, and M. Veit, “Novel Influenza D virus: Epidemiology, pathology, evolution and biological characteristics,” *Virulence*, vol. 8, no. 8, pp. 1580–1591, 2017, doi: 10.1080/21505594.2017.1365216.

[17] E. C. Hutchinson, J. C. von Kirchbach, J. R. Gog, and P. Digard, “Genome packaging in influenza A virus,” *J. Gen. Virol.*, vol. 91, no. 2, pp. 313–328, 2010, doi: 10.1099/vir.0.017608-0.

[18] A. J. Eisfeld, G. Neumann, and Y. Kawaoka, “At the centre: Influenza A virus ribonucleoproteins,” *Nat. Rev. Microbiol.*, vol. 13, no. 1, pp. 28–41, 2015, doi: 10.1038/nrmicro3367.

[19] M. Dettenhofer and X. F. Yu, “Highly purified human immunodeficiency virus type 1 reveals a virtual absence of Vif in virions.,” *J. Virol.*, vol. 73, no. 2, pp. 1460–7, Feb. 1999, doi: 10.1128/JVI.73.2.1460-1467.1999.

[20] A. Pflug, M. Lukarska, P. Resa-Infante, S. Reich, and S. Cusack, “Structural insights into RNA synthesis by the influenza virus transcription-replication machine,” *Virus Res.*, vol. 234, pp. 103–117, 2017, doi: 10.1016/j.virusres.2017.01.013.

[21] T. Samji, “Influenza A: Understanding the viral life cycle,” *Yale J. Biol. Med.*, vol. 82, no. 4, pp. 153–159, 2009.

[22] G. G. Brownlee, E. Fodor, D. C. Pritlove, K. G. Gould, and J. J. Dalluge, “Solid phase synthesis of 5’-diphosphorylated oligoribonucleotides and their conversion to capped m7 Gppp-oligoribonucleotides for U.S. as primers for influenza A virus RNA polymerase in vitro,” *Nucleic Acids Res.*, vol. 23, no. 14, pp. 2641–2647, 1995, doi: 10.1093/nar/23.14.2641.

[23] C. De Vlugt, D. Sikora, and M. Pelchat, “Insight into influenza: A virus cap-snatching,” *Viruses*, vol. 10, no. 11, 2018, doi: 10.3390/v10110641.

[24] S. J. Plotch, M. Bouloy, I. Ulmanen, and R. M. Krug, “A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription,” *Cell*, vol. 23, no. 3, pp. 847–858, 1981, doi: 10.1016/0092-8674(81)90449-9.

[25] T. Deng, F. T. Vreede, and G. G. Brownlee, “Different De Novo Initiation Strategies Are Used by Influenza Virus RNA Polymerase on Its cRNA and Viral RNA Promoters during Viral RNA Replication,” *J. Virol.*, vol. 80, no. 5, pp. 2337–2348, 2006, doi: 10.1128/jvi.80.5.2337-2348.2006.

[26] E. C. Hutchinson and E. Fodor, “Transport of the influenza virus genome from nucleus to nucleus,” *Viruses*, vol. 5, no. 10, pp. 2424–2446, 2013, doi: 10.3390/v5102424.

[27] T. Noda and Y. Kawaoka, “Packaging of influenza virus genome: Robustness of selection,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 23, pp. 8797–8798, 2012, doi: 10.1073/pnas.1206736109.

[28] X. Li, M. Gu, Q. Zheng, R. Gao, and X. Liu, “Packaging signal of influenza A virus,” *Virol. J.*, vol. 18, no. 1, pp. 1–10, 2021, doi: 10.1186/s12985-021-01504-4.

[29] E. Fournier *et al.*, “A supramolecular assembly formed by influenza A virus genomic RNA segments,” *Nucleic Acids Res.*, vol. 40, no. 5, pp. 2197–2209, 2012, doi: 10.1093/nar/gkr985.

[30] T. Noda *et al.*, “Three-dimensional analysis of ribonucleoprotein complexes in influenza A virus,” *Nat. Commun.*, vol. 3, 2012, doi: 10.1038/ncomms1647.

[31] T. Noda *et al.*, “Architecture of ribonucleoprotein complexes in influenza A virus particles,” *Nature*, vol. 439, no. 7075, pp. 490–492, 2006, doi: 10.1038/nature04378.

[32] M. Enami, G. Sharma, C. Benham, and P. Palese, “An influenza virus containing nine different RNA segments,” *Virology*, vol. 185, no. 1, pp. 291–298, 1991, doi: 10.1016/0042-6822(91)90776-8.

[33] Y. Sugimoto *et al.*, “HiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1,” *Nature*, vol. 519, no. 7544, pp. 491–494, 2015, doi: 10.1038/nature14280.

[34] Y. Sugimoto, A. M. Chakrabarti, N. M. Luscombe, and J. Ule, “Using hiCLIP to identify RNA duplexes that interact with a specific RNA-binding protein,” *Nat. Protoc.*, vol. 12, no. 3, pp. 611–637, 2017, doi: 10.1038/nprot.2016.188.

[35] O. Ziv *et al.*, “COMRADES determines in vivo RNA structures and interactions,” *Nat. Methods*, vol. 15, no. 10, pp. 785–788, Oct. 2018, doi: 10.1038/s41592-018-0121-0.

[36] Z. Lu *et al.*, “RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure.,” *Cell*, vol. 165, no. 5, pp. 1267–1279, May 2016, doi: 10.1016/j.cell.2016.04.028.

[37] M. Zhang *et al.*, “Optimized photochemistry enables efficient analysis of dynamic RNA structuromes and interactomes in genetic and infectious diseases,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–14, 2021, doi: 10.1038/s41467-021-22552-y.

[38] J. Gong, Y. Ju, D. Shao, and Q. C. Zhang, “Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale,” *Quant. Biol.*, vol. 6, no. 3, pp. 239–252, Sep. 2018, doi: 10.1007/s40484-018-0146-5.

[39] Z. Cai *et al.*, “RIC-seq for global in situ profiling of RNA – RNA spatial interactions,” *Nature*, no. September 2019, 2020, doi: 10.1038/s41586-020-2249-1.

[40] A. Helwak and D. Tollervey, “Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH),” *Nat. Protoc.*, vol. 9, no. 3, pp. 711–728, Mar. 2014, doi: 10.1038/nprot.2014.043.

[41] T. C. Nguyen *et al.*, “Mapping RNA–RNA interactome and RNA structure in vivo by MARIO,” *Nat. Commun.*, vol. 7, no. 1, p. 12023, Nov. 2016, doi: 10.1038/ncomms12023.

[42] J. G. A. Aw *et al.*, “In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation,” *Mol. Cell*, vol. 62, no. 4, pp. 603–617, May 2016, doi: 10.1016/j.molcel.2016.04.028.

[43] E. Sharma, T. Sterne-Weiler, D. O’Hanlon, and B. J. Blencowe, “Global Mapping of Human RNA-RNA Interactions,” *Mol. Cell*, vol. 62, no. 4, pp. 618–626, May 2016, doi: 10.1016/j.molcel.2016.04.030.

[44] G. Kudla, S. Granneman, D. Hahn, J. D. Beggs, and D. Tollervey, “Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 24, pp. 10010–5, Jun. 2011, doi: 10.1073/pnas.1017386108.

[45] A. Helwak and D. Tollervey, “Identification of miRNA-Target RNA Interactions Using CLASH,” vol. 1358, E. Dassi, Ed. New York, NY: Springer New York, 2016, pp. 229–251.

[46] E. Z. Shen *et al.*, “Identification of piRNA Binding Sites Reveals the Argonaute Regulatory Landscape of the C. elegans Germline,” *Cell*, vol. 172, no. 5, pp. 937-951.e18, 2018, doi: 10.1016/j.cell.2018.02.002.

[47] J. M. Engreitz *et al.*, “RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites,” *Cell*, vol. 159, no. 1, pp. 188–199, 2014, doi: 10.1016/j.cell.2014.08.018.

[48] M. Kretz *et al.*, “Control of somatic tissue differentiation by the long non-coding RNA TINCR,” *Nature*, vol. 493, no. 7431, pp. 231–235, 2013, doi: 10.1038/nature11661.

[49] J. G. A. Aw, Y. Shen, N. Nagarajan, and Y. Wan, “Mapping RNA-RNA interactions globally using biotinylated psoralen,” *J. Vis. Exp.*, vol. 2017, no. 123, pp. 1–10, 2017, doi: 10.3791/55255.

[50] Y. Zhang *et al.*, “In vivo structure and dynamics of the SARS-CoV-2 RNA genome,” *Nat.*

Commun., vol. 12, no. 1, pp. 1–12, 2021, doi: 10.1038/s41467-021-25999-1.

[51] Z. Lu *et al.*, "RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure," *Cell*, vol. 165, no. 5, pp. 1267–1279, May 2016, doi: 10.1016/j.cell.2016.04.028.

[52] O. Ziv *et al.*, "The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2," *Mol. Cell*, vol. 80, no. 6, pp. 1067-1077.e5, 2020, doi: 10.1016/j.molcel.2020.11.004.

[53] C. Cao *et al.*, "The architecture of the SARS-CoV-2 RNA genome inside virion," *Nat. Commun.*, vol. 12, no. 1, pp. 1–14, 2021, doi: 10.1038/s41467-021-22785-x.

[54] C. Cao *et al.*, "Global in situ profiling of RNA-RNA spatial interactions with RIC-seq," *Nat. Protoc.*, vol. 16, no. 6, pp. 2916–2946, 2021, doi: 10.1038/s41596-021-00524-2.

[55] R. Van Damme *et al.*, "Chemical reversible crosslinking enables measurement of RNA 3D distances and alternative conformations in cells," *Nat. Commun.*, vol. 13, no. 1, pp. 1–13, 2022, doi: 10.1038/s41467-022-28602-3.

[56] T. E. England, R. I. Gumport, and O. C. Uhlenbeck, "Dinucleoside pyrophosphates are substrates for T4-induced RNA ligase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 11, pp. 4839–4842, 1977, doi: 10.1073/pnas.74.11.4839.

[57] N. Tanaka, B. Meineke, and S. Shuman, "RtcB, a novel RNA ligase, can catalyze tRNA splicing and HAC1 mRNA splicing in vivo," *J. Biol. Chem.*, vol. 286, no. 35, pp. 30253–30257, 2011, doi: 10.1074/jbc.C111.274597.

[58] S. E. Peach, K. York, and J. R. Hesselberth, "Global analysis of RNA cleavage by 5'-hydroxyl RNA sequencing," *Nucleic Acids Res.*, vol. 43, no. 17, pp. 1–13, 2015, doi: 10.1093/nar/gkv536.

[59] E. Hoffmann, G. Neumann, Y. Kawaoka, G. Hobom, and R. G. Webster, "A DNA transfection system for generation of influenza A virus from eight plasmids," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 11, pp. 6108–6113, 2000, doi: 10.1073/pnas.100133697.

[60] V. Czudai-Matwich, M. Schnare, and O. Pinkenburg, "A simple and fast system for cloning influenza A virus gene segments into pHW2000- and pCAGGS-based vectors," *Arch. Virol.*, vol. 158, no. 10, pp. 2049–2058, 2013, doi: 10.1007/s00705-013-1697-4.

[61] J. Sakuragi, S. Sakuragi, and T. Shioda, "Minimal Region Sufficient for Genome Dimerization in the Human Immunodeficiency Virus Type 1 Virion and Its Potential Roles in the Early Stages of Viral Replication," *J. Virol.*, vol. 81, no. 15, pp. 7985–7992, 2007, doi: 10.1128/jvi.00429-07.

[62] J. C. Paillart, R. Marquet, E. Skripkin, C. Ehresmann, and B. Ehresmann, "Dimerization of retroviral genomic RNAs: Structural and functional implications," *Biochimie*, vol. 78, no. 7, pp. 639–653, 1996, doi: 10.1016/S0300-9084(96)80010-1.

[63] J. C. Paillart *et al.*, "A dual role of the putative RNA dimerization initiation site of human immunodeficiency virus type 1 in genomic RNA packaging and proviral DNA synthesis.," *J. Virol.*, vol. 70, no. 12, pp. 8348–54, Dec. 1996.

[64] E. Skripkin *et al.*, “Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 11, pp. 4945–9, May 1994, doi: 10.1073/pnas.91.11.4945.

[65] M. K. Hill, M. Shehu-Xhilaga, S. M. Campbell, P. Poumbourios, S. M. Crowe, and J. Mak, “The Dimer Initiation Sequence Stem-Loop of Human Immunodeficiency Virus Type 1 Is Dispensable for Viral Replication in Peripheral Blood Mononuclear Cells,” *J. Virol.*, vol. 77, no. 15, pp. 8329–8335, 2003, doi: 10.1128/jvi.77.15.8329-8335.2003.

[66] Z. Lu, J. Gong, and Q. C. Zhang, “PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution,” vol. 1649, I. Gaspar, Ed. New York, NY: Springer New York, 2018, pp. 59–84.

[67] D. S. Peabody, “The RNA binding size of bacteriophage MS2 coat protein,” *EMBO J.*, vol. 12, no. 2, pp. 595–600, 1993, doi: 10.1002/j.1460-2075.1993.tb05691.x.

[68] Z. Zhou and R. Reed, “Purification of Functional RNA-Protein Complexes using MS2-MBP,” *Curr. Protoc. Mol. Biol.*, vol. 63, no. 1, pp. 27.3.1-27.3.7, 2003, doi: 10.1002/0471142727.mb2703s63.

[69] S. Niranjanakumari, E. Lasda, R. Brazas, and M. A. Garcia-Blanco, “Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo,” *Methods*, vol. 26, no. 2, pp. 182–190, 2002, doi: 10.1016/S1046-2023(02)00021-X.

[70] T. W. Nilsen, “Detecting RNA–RNA interactions using psoralen derivatives,” *Cold Spring Harb. Protoc.*, vol. 2014, no. 9, pp. 996–1000, 2014, doi: 10.1101/pdb.prot080861.

[71] M. Lindell, P. Romby, and E. G. H. Wagner, “Lead(II) as a probe for investigating RNA structure in vivo,” *Rna*, vol. 8, no. 4, pp. 534–541, 2002, doi: 10.1017/S1355838201020416.

[72] M. Lindell, M. Brännvall, E. G. H. Wagner, and L. A. Kirsebom, “Lead(II) cleavage analysis of RNase P RNA in vivo,” *Rna*, vol. 11, no. 9, pp. 1348–1354, 2005, doi: 10.1261/rna.2590605.

[73] D. N. Wilson *et al.*, “The Structure and Function of the Eukaryotic,” 2012, doi: 10.1101/cshperspect.a011536.

[74] P. B. Moore, “The PDB and the ribosome,” *J. Biol. Chem.*, vol. 296, pp. 1–11, 2021, doi: 10.1016/j.jbc.2021.100561.

[75] A. Jobe, Z. Liu, C. Gutierrez-Vargas, and J. Frank, “New insights into ribosome structure and function,” *Cold Spring Harb. Perspect. Biol.*, vol. 11, no. 1, pp. 1–18, 2019, doi: 10.1101/cshperspect.a032615.

[76] Y. Khayat, M. Cromer-Morin, and J.-P. Scharff, “Stability constants for lead(II) complexes of glycine, serine, aspartic acid and glycyl-l-leucine,” *J. Inorg. Nucl. Chem.*, vol. 41, no. 10, pp. 1496–1498, Jan. 1979, doi: 10.1016/0022-1902(79)80222-5.

[77] R. Brauer and P. Chen, “Influenza virus propagation in embryonated chicken eggs,” *J. Vis. Exp.*,

vol. 2015, no. 97, pp. 1–6, 2015, doi: 10.3791/52421.

[78] B. Dadonaite *et al.*, "The structure of the influenza A virus genome," *Nat. Microbiol.*, vol. 4, no. 11, pp. 1781–1789, 2019, doi: 10.1038/s41564-019-0513-7.

[79] V. Le Sage, J. P. Kanarek, D. J. Snyder, V. S. Cooper, S. S. Lakdawala, and N. Lee, "Mapping of Influenza Virus RNA-RNA Interactions Reveals a Flexible Network," *Cell Rep.*, vol. 31, no. 13, p. 107823, 2020, doi: 10.1016/j.celrep.2020.107823.

[80] T. Noda *et al.*, "Importance of the 1+7 configuration of ribonucleoprotein complexes for influenza A virus genome packaging," *Nat. Commun.*, vol. 9, no. 1, pp. 1–10, 2018, doi: 10.1038/s41467-017-02517-w.

[81] M. Li, N. Husic, Y. Lin, and B. J. Snider, "Production of Lentiviral Vectors for Transducing Cells from the Central Nervous System," *J. Vis. Exp.*, no. 63, pp. 2–7, 2012, doi: 10.3791/4031.

[82] A. Ali and M. J. Roossinck, "Rapid and efficient purification of Cowpea chlorotic mottle virus by sucrose cushion ultracentrifugation," *J. Virol. Methods*, vol. 141, no. 1, pp. 84–86, 2007, doi: 10.1016/j.jviromet.2006.11.038.

[83] M. C. Rivera, B. Maguire, and J. A. Lake, "Purification of 70S Ribosomes," *Cold Spring Harb. Protoc.*, vol. 2015, no. 3, pp. 300–302, 2015, doi: 10.1101/pdb.prot081356.

[84] M. Dettenhofer and X.-F. Yu, "Highly Purified Human Immunodeficiency Virus Type 1 Reveals a Virtual Absence of Vif in Virions," *J. Virol.*, vol. 73, no. 2, pp. 1460–1467, 1999, doi: 10.1128/jvi.73.2.1460-1467.1999.

[85] M. Olszewski and P. Filipkowski, "[Benzonase--possibility of practical application].," *Postepy Biochem.*, vol. 55, no. 1, pp. 21–24, 2009.

[86] G. Prezza, T. Heckel, S. Dietrich, C. Homberger, A. J. Westermann, and J. Vogel, "Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads," *Rna*, vol. 26, no. 8, pp. 1069–1078, 2020, doi: 10.1261/RNA.075945.120.

[87] B. Dadonaite *et al.*, "The structure of the influenza A virus genome," *Nat. Microbiol.*, vol. 4, no. 11, pp. 1781–1789, 2019, doi: 10.1038/s41564-019-0513-7.

Chapter 5 Summarizing discussion

RNA viruses are widespread and a lot of them can cause severe pandemics, like influenza, HIV, as well as SARS-CoV-2. RNA structure and RNA-RNA interactions play essential roles in all processes of RNA virus infection[1]–[15]. My PhD projects focuses on RNA secondary structure and RNA-RNA interactions on HIV-1 and influenza A virus genome packaging.

HIV-1 genome dimerization is assumed to be a requirement for binding to the viral structural protein $Pr55^{Gag}$ during genome packing and is a conserved characteristic of retroviral replication[16]–[22]. Many studies have shown that the dimerization is regulated by structure within the conserved HIV-1 5'UTR[23]–[26]. DIS within SL1 is the most essential motif for HIV-1 dimerization[26], other regions are reported to play a role in regulation of dimerization. For example, U5-AUG interaction promotes HIV-1 RNA dimerization[27]–[29], while when U5 base-pairing with SL1 suppress dimerization[12], [30], [31]. However, how HIV-1 regulates dimerization, and the RNA secondary structure features within HIV-1 5'UTR in dimer and monomer conformations are still under discussion. I developed FARS-seq to investigate how RNA-RNA interactions and secondary structure regulate HIV-1 genomic RNA dimerization and packaging. We showed HIV-1 5'UTR of NL4-3 can fold into two distinguish conformations: dimer form and monomer form. The dimer form presents the "classic" secondary structure which displays TAR, polyA, PBS, SL1, SL2, and SL3, while the monomer RNA folds into a structure that sequesters DIS and Gag binding site, which still displays TAR, PBS, SL2 and SL3 and without polyA, SL1 stem-loop. We demonstrated that the most essential domain SL1 for dimerization is metastable, and the interactions between polyA-SL1, PBS-SL1 negatively regulate HIV-1 dimerization and genome packaging. Furthermore, we suggest that the binding of host factors to genomic RNA, like tRNA, which function as reverse transcription primer for HIV-1, shifts the equilibrium of full-length RNA to dimer conformation and promotes genome packaging. I also implemented FARS-seq on Mal, another strain of HIV-1 and was reported to adapt different dimerization structure and regulation mechanism[30]. Our results on Mal showed that it did have a different profile compared to NL43. In general, Mal was much more monomeric, and the MIME data showed that TAR was a significant negative regulator for Mal dimerization, as well as U5 and 3' region of stem of SL1. In addition, the secondary structure based on DMS probing indicated that the SL1 was reorganized into another stem loop, in which the DIS was sequestered, and 3' region of SL1 base-paring with the region close to U5, rather than interacted with polyA and PBS. These results suggest that RNA structures are very dynamic and flexible. FARS-seq is a powerful approach and it has the advantage to connect the RNA structure to function. Compared with other structure probing methods based on high throughput sequencing, it overcomes the RNA secondary structure prediction limitation by mutation correlation analysis; and the mutational interference and function selection associates the RNA sequence and function. Here, we applied RNA dimerization as function selection to separate functional RNA populations for HIV-1 dimerization. Similarly, RNA secondary structure change, RNA translation, RNA splicing, protein or host factor binding can be also applied for

function selection. Therefore, FARS-seq can be also applied to other flexible RNAs efficiently, like IRES conformation change for protein translation; virus frameshifting motifs, which can result in multiple viral proteins without engaging longer coding sequence; or riboswitches, which display different RNA secondary structures to regulate gene expression depending on whether the ligand is bound. However, FARS-seq is limited by reverse transcription length and the Illumina sequencing length. Most critically, the detection of DMS modifications by mutations depend on SSII reverse transcription with Mn^{2+}, but SSII is only capable to reverse transcribe a limited length of the DMS modified RNA because reverse transcriptases tend to fall off the RNA upon hitting a modified base. Solutions to this problem include using ultra processive reverse transcriptases, like Marathon reverse transcriptase, or direct sequencing the modified RNA to skip the reverse transcription. Another limitation is the mutations correlation (2D) analysis requires mutations from the same sequencing read, while the Illumina sequencing reads usually is about 300bp, and maximum 600bp. Rapid development in nanopore technologies for sequencing single long RNA molecules give rise to substantial improvements in accuracy, read length and throughput, may overcome the limitations of FARS-seq.

I also worked on another project to develop a high throughput sequencing-based method, RNA-RNA seq to study the interactions between the eight negative single-stranded segments of influenza A virus genome. Virologists now agree that influenza A virus selectively packages its genome and there were evidences showing that the eight segments form a “7+1” supramolecular complex[32]–[35]. I optimized the protocol on model substrates and validated the method on purified ribosome, virus infected cell, and influenza A virus, and got promising interaction data. The project is still ongoing and it will be applied to other influenza A virus strains, as well as reassorted viruses, which will decipher the mechanism that influenza A virus genome packaging and help us to understand the evolution of influenza. An interesting feature of RNA-RNA-seq is that it identifies both the spatially proximal RNAs that do not physically interact, as well as directly interacting RNA. This feature will help us to build three-dimensional architectural models of influenza. We hope that combining these structural models with functional outcome of reassortment assays will allow us to better predict the emergence of reassortment, potentially pandemic, influenza viruses.

RNA virus genome packaging is a complicated and highly regulated process, and RNA secondary structures and RNA-RNA interactions play key roles in the process. The high-throughput sequencing and RNA structure determination technologies have revolutionized our ability to decipher the relationship of RNA sequence, structure and function, which will definitely reveal new principles in RNA virus life cycle. Identification of RNA structure motifs that regulate RNA virus genome packaging facilitates our knowledge of virus evolution, which may lead to new therapeutic prospects.

References

[1] C. Romero-López and A. Berzal-Herranz, “The role of the RNA-RNA interactome in the hepatitis C virus life cycle,” Int. J. Mol. Sci., vol. 21, no. 4, 2020, doi: 10.3390/ijms21041479.

[2] T. C. Nguyen, K. Zaleta-Rivera, X. Huang, X. Dai, and S. Zhong, “RNA, Action through Interactions,” Trends Genet., vol. 34, no. 11, pp. 867–882, 2018, doi: 10.1016/j.tig.2018.08.001.

[3] S. T. Cross, D. Michalski, M. R. Miller, and J. Wilusz, “RNA regulatory processes in RNA virus biology,” Wiley Interdiscip. Rev. RNA, vol. 10, no. 5, pp. 1–24, 2019, doi: 10.1002/wrna.1536.

[4] L. Ye et al., “RNA Structures and Their Role in Selective Genome Packaging,” Viruses, vol. 13, no. 9, 2021, doi: 10.3390/v13091788.

[5] L. R. Newburn and K. Andrew White, “Trans-acting RNA–RNA interactions in segmented RNA viruses,” Viruses, vol. 11, no. 8, 2019, doi: 10.3390/v11080751.

[6] D. E. Alvarez, M. F. Lodeiro, S. J. Ludueña, L. I. Pietrasanta, and A. V. Gamarnik, “Long-Range RNA-RNA Interactions Circularize the Dengue Virus Genome,” J. Virol., vol. 79, no. 11, pp. 6631–6643, 2005, doi: 10.1128/jvi.79.11.6631-6643.2005.

[7] R. G. Huber et al., “Structure mapping of dengue and Zika viruses reveals functional long-range interactions,” Nat. Commun., vol. 10, no. 1, 2019, doi: 10.1038/s41467-019-09391-8.

[8] W. N. Moss, L. I. Dela-Moss, E. Kierzek, R. Kierzek, S. F. Priore, and D. H. Turner, “The 3′ splice site of influenza A segment 7 mRNA can exist in two conformations: A pseudoknot and a hairpin,” PLoS One, vol. 7, no. 6, pp. 1–11, 2012, doi: 10.1371/journal.pone.0038323.

[9] N. Mueller, B. Berkhout, and A. T. Das, “HIV-1 splicing is controlled by local RNA structure and binding of splicing regulatory proteins at the major 5′ splice site,” J. Gen. Virol., vol. 96, no. 7, pp. 1906–1917, 2015, doi: 10.1099/vir.0.000122.

[10] B. Dadonaite et al., “The structure of the influenza A virus genome,” Nat. Microbiol., vol. 4, no. 11, pp. 1781–1789, 2019, doi: 10.1038/s41564-019-0513-7.

[11] O. Ziv, “The short- and long-range RNA-RNA Interactome of SARS-CoV-2,” 2020.

[12] S. Kharytonchyk et al., “Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome,” Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 47, pp. 13378–13383, 2016, doi: 10.1073/pnas.1616627113.

[13] P. J. Lukavsky, “Structure and function of HCV IRES domains,” Virus Res., vol. 139, no. 2, pp. 166–171, 2009, doi: 10.1016/j.virusres.2008.06.004.

[14] S. M. Villordo, D. E. Alvarez, and A. V. Gamarnik, “A balance between circular and linear forms of the dengue virus genome is crucial for viral replication,” Rna, vol. 16, no. 12, pp. 2325–2335, 2010, doi: 10.1261/rna.2120410.

[15] V. Le Sage, J. P. Kanarek, D. J. Snyder, V. S. Cooper, S. S. Lakdawala, and N. Lee, “Mapping of Influenza Virus RNA-RNA Interactions Reveals a Flexible Network,” Cell Rep., vol. 31, no. 13, p. 107823, 2020, doi: 10.1016/j.celrep.2020.107823.

[16] R. S. Russell, C. Liang, and M. A. Wainberg, “Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably?,” Retrovirology, vol. 1, no. Mlv, pp. 1–14, 2004, doi: 10.1186/1742-4690-1-23.

[17] E. W. Abd El-Wahab et al., “Specific recognition of the HIV-1 genomic RNA by the Gag precursor.,” Nat. Commun., vol. 5, no. 1, p. 4304, Jul. 2014, doi: 10.1038/ncomms5304.

[18] E. Mailler et al., “The Life-Cycle of the HIV-1 Gag-RNA Complex.,” Viruses, vol. 8, no. 9, p. 248, Sep. 2016, doi: 10.3390/v8090248.

[19] P. Ding et al., “Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal,” Proc. Natl. Acad. Sci., vol. 117, no. 30, pp. 17737–17746, Jul. 2020, doi: 10.1073/PNAS.2008519117.

[20] M. Kuzembayeva, K. Dilley, L. Sardo, and W. S. Hu, “Life of psi: How full-length HIV-1 RNAs become packaged genomes in the viral particles,” Virology, vol. 454–455, no. 1, pp. 362–370, 2014, doi: 10.1016/j.virol.2014.01.019.

[21] J. Chen et al., “HIV-1 RNA genome dimerizes on the plasma membrane in the presence of Gag protein,” Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 2, pp. E201–E208, 2016, doi: 10.1073/pnas.1518572113.

[22] X. Heng et al., “Identification of a minimal region of the HIV-1 5′-leader required for RNA dimerization, NC binding, and packaging,” J. Mol. Biol., vol. 417, no. 3, pp. 224–239, 2012, doi: 10.1016/j.jmb.2012.01.033.

[23] J. Sakuragi, S. Sakuragi, and T. Shioda, “Minimal Region Sufficient for Genome Dimerization in the Human Immunodeficiency Virus Type 1 Virion and Its Potential Roles in the Early Stages of Viral Replication,” J. Virol., vol. 81, no. 15, pp. 7985–7992, 2007, doi: 10.1128/jvi.00429-07.

[24] M. D. Moore and W. S. Hu, “HIV-1 RNA dimerization: It takes two to tango,” AIDS Rev., vol. 11, no. 2, pp. 91–102, 2009.

[25] N. Dubois, R. Marquet, J. C. Paillart, and S. Bernacchi, “Retroviral RNA dimerization: From structure to functions,” Front. Microbiol., vol. 9, no. MAR, pp. 1–19, 2018, doi: 10.3389/fmicb.2018.00527.

[26] J. C. Paillart, R. Marquet, E. Skripkin, C. Ehresmann, and B. Ehresmann, “Dimerization of retroviral genomic RNAs: Structural and functional implications,” Biochimie, vol. 78, no. 7, pp. 639–653, 1996, doi: 10.1016/S0300-9084(96)80010-1.

[27] R. Song, J. Kafaie, and M. Laughrea, “Role of the 5′ TAR stem--loop and the U5-AUG duplex in dimerization of HIV-1 genomic RNA.,” Biochemistry, vol. 47, no. 10, pp. 3283–93, Mar. 2008, doi: 10.1021/bi7023173.

[28] S. C. Keane et al., “RNA structure. Structure of the HIV-1 RNA packaging signal.,” Science, vol. 348, no. 6237, pp. 917–21, May 2015, doi: 10.1126/science.aaa9266.

[29] T. E. M. Abbink and B. Berkhout, “A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon.,” J. Biol. Chem., vol. 278, no. 13, pp. 11601–11, Mar. 2003, doi: 10.1074/jbc.M210291200.

[30] J. D. Brown et al., “Structural basis for transcriptional start site control of HIV-1 RNA fate.,” Science, vol. 368, no. 6489, pp. 413–417, 2020, doi: 10.1126/science.aaz7959.

[31] J. C. Kenyon, L. J. Prestwood, S. F. J. Le Grice, and A. M. L. Lever, “In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader,” Nucleic Acids Res., vol. 41, no. 18, pp. 1–11, 2013, doi: 10.1093/nar/gkt690.

[32] T. Noda et al., “Importance of the 1+7 configuration of ribonucleoprotein complexes for influenza A virus genome packaging,” Nat. Commun., vol. 9, no. 1, pp. 1–10, 2018, doi: 10.1038/s41467-017-02517-w.

[33] E. Fournier et al., “A supramolecular assembly formed by influenza A virus genomic RNA segments,” Nucleic Acids Res., vol. 40, no. 5, pp. 2197–2209, 2012, doi: 10.1093/nar/gkr985.

[34] T. Noda et al., “Three-dimensional analysis of ribonucleoprotein complexes in influenza A virus,” Nat. Commun., vol. 3, 2012, doi: 10.1038/ncomms1647.

[35] T. Noda et al., “Architecture of ribonucleoprotein complexes in influenza A virus particles,” Nature, vol. 439, no. 7075, pp. 490–492, 2006, doi: 10.1038/nature04378.

Publications during candidature

1) **Ye L**, Ambi UB, Olguin-Nava M, Gribling-Burrer AS, Ahmad S, Bohn P, Weber MM, Smyth RP. RNA Structures and Their Role in Selective Genome Packaging. Viruses. 2021 Sep 8;13(9):1788. doi: 10.3390/v13091788. PMID: 34578369; PMCID: PMC8472981.

2) Zimmer MM, Kibe A, Rand U, Pekarek L, **Ye L**, Buck S, Smyth RP, Cicin-Sain L, Caliskan N. The short isoform of the host antiviral protein ZAP acts as an inhibitor of SARS-CoV-2 programmed ribosomal frameshifting. Nat Commun. 2021 Dec 10;12(1):7193. doi: 10.1038/s41467-021-27431-0. PMID: 34893599; PMCID: PMC8664833.

3) **Ye L**, Gribling-Burrer AS, Bohn P, Kibe A, Börtlein C, Ambi UB, Ahmad S, Olguin-Nava M, Smith M, Caliskan N, von Kleist M, Smyth RP. Short- and long-range interactions in the HIV-1 5' UTR regulate genome dimerization and packaging. Nat Struct Mol Biol. 2022 Apr;29(4):306-319. doi: 10.1038/s41594-022-00746-2. Epub 2022 Mar 28. PMID: 35347312; PMCID: PMC9010304.

Abbreviations

2IL	two-internal loop model
3IL	three-internal loop model
3WJ	three-way junction
AFM	atomic force microscopy
AMT	4’-aminomethyltrioxsalen
BMH	branched multiple hairpin
BMV	brome mosaic virus
BSA	Bovine serum albumin
CDI	1,1′-carbonyldiimidazole
cDNA	complementary DNA
CLASH	cross-linking, ligation and sequencing of hybrid
COMRADES	Cross-linking of matched RNAs and deep sequencing
CoV	Coronaviruses
CP	coat protein
CRE	cis-acting replicating element

cRNA	complementary RNA
cryo-EM	cryogenic electron microscopy
CS	cyclization sequence
DAR	downstream of AUG region
DASH	Depletion of Abundant Sequences by Hybridization
DENV	dengue virus
DEPC	Diethylpyrocarbonate
DIS	dimerization initiation site
DMEM	Dulbecco's Modified Eagle Medium
DMS	Dimethyl sulfate
DMS-MaPseq	dimethyl sulfate mutational profiling with sequencing
DMSO	Dimethyl sulfoxide
DMS-seq	dimethyl sulfate sequencing
DNA	deoxyribonucleic acid
dsRNA	double-stranded RNA
DTT	Dithiothreitol

eIF	eukaryotic initiation factor
EM	electron microscopy
ER	endoplasmic reticulum
FA	Formaldehyde
FARS-seq	Functional Analysis of RNA Structure
Frag-seq	fragmentation sequencing
Gag	Group-specific antigen
HA	hemagglutinin
HBV	hepatitis B virus
HCV	hepatitis C virus
hiCLIP	hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation
HIV-1	Human immunodeficiency virus 1
hpi	hour post infection
icSHAPE	in vivo click selective 2-hydroxyl acylation and profiling experiment
IDR	intrinsically disordered regions
IRES	internal ribosome entry sites

IVT	*in vitro* transcription
LDI	long-distance interaction
LIGR-seq	Ligation of Interacting RNA followed by high-throughput sequencing
LLPS	liquid-liquid phase separation
MARIO	Mapping RNA interactome in vivo
MDCK cells	Madin-Darby Canine Kidney cells
MEM	Minimal Essential Medium
MOI	multiplicity of infection
MoMLV	Moloney murine leukemia virus
mRNA	messenger RNA
MST	microscale thermophoresis
NA	neuraminidase
NC	nucleocapsid
NHS beads	N-hydroxy-succinimide-Activated Magnetic Beads
NMR	nuclear magnetic resonance
PAGE	Polyacrylamide gel electrophoresis

PARIS	Psoralen Analysis of RNA Interactions and Structures
PARS	Parallel analysis of RNA structure
PARTE	parallel analysis of RNA structures with temperature elevation
PAS	primer activation sequence
PBS	primer binding site
PBS buffer	Phosphate-buffered saline buffer
PCA	principal component analysis
pCp-biotin	biotinylated cytidine (bis) phosphate
PCR	polymerase chain reaction
PFU	plaque-forming unit
PIP-seq	protein interaction profile sequencing
PR8	A/Puerto Rico/8/1934(H1N1)
qPCR	quantitative polymerase chain reaction
RAP/RIA-seq	RNA antisense purification and RNA interactome analysis followed by deep sequencing
RBP	RNA binding protein
RdRp	RNA-dependent RNA polymerase

RIC-seq	RNA In situ Conformation sequencing
RNA	ribonucleic acid
RNP	ribonucleoprotein
rRNA	ribosomal RNA
rSAP	Shrimp Alkaline Phosphatase
RVFV	rift valley fever virus
SA	sialic acid
SAXS	small angle X-ray scattering
SBV	Schmallenberg virus
SHAPE	selective 2′-hydroxyl acylation analyzed by primer extension
SHAPE-Map	selective 2’-hydroxyl acylation analyzed by primer extension and mutational profiling
SHARC	Spatial 2'-Hydroxyl Acylation Reversible Crosslinking
SL	stem-loop
smFISH	single molecule fluorescent in situ hybridisation
SPF	specific pathogen-free
SPLASH	Sequencing of Psoralen crosslinked, Ligated, And Selected Hybrids

SSC	saline-sodium citrate buffer
ssRNA	single-stranded RNA
T4 PNK	T4 Polynucleotide Kinase
TAE	Tris-Acetate-EDTA
TAR	transactivation response
TBE	Tris-Borate-EDTA
TBM	Tris-borate magnesium
TNV	tobacco necrosis virus
tRNA	transfer RNA
U5	unique-5’ element
UAR	upstream of AUG region
UTR	untranslated region
UV	Ultraviolet
vRNA	viral RNA
WT	wild-type

List of Figures

List of Tables

Statement of individual author contributions to figures/tables of manuscripts included in the dissertation

Manuscript 1 (complete reference): Ye, Liqing et al. “RNA Structures and Their Role in Selective Genome Packaging.” Viruses vol. 13,9 1788. 8 Sep. 2021, doi:10.3390/v13091788					
Figure	**Author Initials,** Responsibility decreasing from left to right				
1	Patrick Bohn	Redmond P Smyth			
2	Redmond P Smyth				
3	Marco Olguin-Nava	Redmond P Smyth			
4	Liqing Ye,	Melanie M.Weber	Redmond P Smyth		
5	Uddhav B Ambi	Redmond P Smyth	Shazeb Ahmad	Liqing Ye	
6	Anne-Sophie Gribling-Burrer	Redmond P Smyth	Liqing Ye		
Table	**Author Initials,** Responsibility decreasing from left to right				
1					
2					

Explanations (if applicable):

Manuscript 2 (complete reference): Ye, Liqing et al. “Short- and long-range interactions in the HIV-1 5' UTR regulate genome dimerization and packaging.” Nature structural & molecular biology vol. 29,4 (2022): 306-319. doi:10.1038/s41594-022-00746-2					
Figure	**Author Initials,** Responsibility decreasing from left to right				
1	Liqing Ye	Redmond P Smyth			
2	Liqing Ye	Redmond P Smyth	Patrick Bohn		
3	Redmond P Smyth	Liqing Ye			
4	Redmond P Smyth	Liqing Ye			
5	Redmond P Smyth	Liqing Ye			
6	Liqing Ye	Redmond P Smyth			
7	Liqing Ye	Anne-Sophie Gribling-Burrer	Anuja Kibe	Redmond P Smyth	
Table	**Author Initials,** Responsibility decreasing from left to right				
1					
2					

Explanations (if applicable):

Manuscript 3 (complete reference): Ye, Liqing et al. Chapter 4 in the thesis: Defining the architecture of the influenza RNA genome by RNA-RNA-seq					
Figure	**Author Initials,** Responsibility decreasing from left to right				
1	Liqing Ye				
2	Liqing Ye				
3	Liqing Ye				
4	Liqing Ye				
5	Liqing Ye				
6	Liqing Ye				
7	Liqing Ye				
8	Liqing Ye				
9	Liqing Ye				
10	Liqing Ye				
11	Liqing Ye				
12	Liqing Ye				
13	Liqing Ye	Redmond P Smyth			
14	Liqing Ye				
15	Liqing Ye				
16	Liqing Ye	Redmond P Smyth			
17	Liqing Ye	Redmond P Smyth			
18	Liqing Ye				
19	Liqing Ye				
Table	**Author Initials,** Responsibility decreasing from left to right				
1	Liqing Ye				
2	Liqing Ye				

Explanations (if applicable): Not published yet

I also confirm my primary supervisor's acceptance.

Liqing Ye

_____

Doctoral Researcher's Name Date Place Signature

Statement of individual author contributions and of legal second publication rights to manuscripts included in the dissertation

Manuscript 1 (complete reference): Ye, Liqing et al. “RNA Structures and Their Role in Selective Genome Packaging.” Viruses vol. 13,9 1788. 8 Sep. 2021, doi:10.3390/v13091788					
Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	Redmond P Smyth Liqing Ye	Uddhav B Ambi Marco Olguin-Nava Anne-Sophie Gribling-Burrer Shazeb Ahmad Patrick Bohn Melanie M Weber			
Data Collection	Redmond P Smyth Liqing Ye	Uddhav B Ambi Marco Olguin-Nava Anne-Sophie Gribling-Burrer Shazeb Ahmad Patrick Bohn Melanie M Weber			
Data Analysis and Interpretation					
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	Redmond P Smyth Liqing Ye	Uddhav B Ambi Marco Olguin-Nava Anne-Sophie Gribling-Burrer Shazeb Ahmad Patrick Bohn Melanie M Weber			

Explanations (if applicable): This is a review manuscript, so there is no material and methods as well as data analysis

Manuscript 2 (complete reference): Ye, Liqing et al. “Short- and long-range interactions in the HIV-1 5' UTR regulate genome dimerization and packaging.” Nature structural & molecular biology vol. 29,4 (2022): 306-319. doi:10.1038/s41594-022-00746-2					
Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	Redmond P Smyth Anne-Sophie Gribling-Burrer Liqing Ye				
Data Collection	Liqing Ye	Anne-Sophie	Anuja Kibe	Uddhav B Ambi	

		Gribling-Burrer	Charlene Börtlein	Shazeb Ahmad Marco Olguin-Nava	
Data Analysis and Interpretation	Liqing Ye Redmond P Smyth Patrick Bohn	Maureen Smith	Anne-Sophie Gribling-Burrer		
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	Redmond P Smyth Liqing Ye	Anne-Sophie Gribling-Burrer Patrick Bohn	Neva Caliska Max von Kleist		

Explanations (if applicable):

Manuscript 3 (complete reference): Defining the architecture of the influenza RNA genome by RNA-RNA-seq					
Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	Redmond P Smyth Liqing Ye				
Data Collection	Liqing Ye				
Data Analysis and Interpretation	Liqing Ye Redmond P Smyth				
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	Liqing Ye	Redmond P Smyth			

Explanations (if applicable): This manuscript is not published yet.

If applicable, the doctoral researcher confirms that she/he has obtained permission from both the publishers (copyright) and the co-authors for legal second publication.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Liqing Ye

_____

Doctoral Researcher's Name Date Place Signature

Jun. Prof. Redmond Smyth

_____

Primary Supervisor's Name Date Place Signature

Acknowledgements

First of all, I would like to express my special appreciation to my supervisor Jun Prof. Dr. Redmond Smyth for his support throughout my PhD. Redmond is a super cool scientist and always encourages us to try new techniques and be confident. I learned a lot from him.

I would like to thank my colleagues in GARV: Dr. Anne-Sophie Gribling, Patrick Bohn, Uddhav Ambi, Shazeb Ahmad, Marco Antonio Olguin Nava, Melanie Weber for their valuable scientific advises and emotional support during the daily lab work and life.

My sincere thanks also goes to Dr. Victoria McParland, Till Balla, Sofiya Rachkevych, Charlene Börtlein for their daily help and support in the lab, especially preparing the high-quality enzymes and making mutants for my projects, which really accelerate the experimental progress.

I am thankful to Prof. Dr. Sibylle Schneider-Schaulies, Prof. Dr. Cynthia Sharma for being my thesis committee members and spend the time with me to discuss the projects and give pertinent suggestions and comments.

Thanks to Prof. Dr. Max von Kleist, Dr. Maureen Smith, Jun Prof. Dr. Neva Caliskan, Anuja Kibe, Jun Prof. Dr. Mathias Munschauer, Sabina Ganskih, I am grateful for the productive collaboration, technical support and scientific input.

Thanks to Alice Hohn, Julia Miriam Mendorff, Michael Kütt, Dr Tim Schnyder from HIRI and Dr. Irina Pleines-Meinhold, Dr. Gabriele Blum-Oehler, Katharina Bötsch from GSLS for their help on administration and registration affairs.

Thanks to my kind and cute friends and colleagues in HIRI and IMIB: Melanie Weber, Stefan Buck, Nora Schmidt, Xiangyi Wang, Yan Zhu, Mingjing Kang, Matthias Zimmer, Sara Santos, Anuja Kibe, Ricada Riegger, Lukáš Pekárek, Chunyu Liao, Chunlei Jiao, Jiaqi Yu, Franziska Wimmer, Mastura Neyazi, Sebastian Zielinski, Jens Aydin, Sabina Ganskih, Yuanjie Wei, Shuba Varshini Alampalli, Yanying Yu, Elisa Venturini and many others, I am lucky to meet you in Würzburg and thanks for the kind help in the lab and outside of the lab.

Last but not least, thanks to my family and friends in China, especially to Fan, for their love, support and patience throughout the whole PhD and my life in general.