APPROVAL SHEET

Title of Dissertation: Characterization of the Monomeric Conformation of the 5' Leader of HIV-1

Name of Candidate: Sarah Ann Monti

Doctor of Philosophy, 2015

Dissertation and Abstract Approved: _____

Dr. Michael Summers

Professor

Department of Chemistry and Biochemistry

Date Approved: 8/25/15

ABSTRACT

Title of Document: CHARACTERIZATION OF THE MONOMERIC CONFORMATION OF THE HIV-1 5' LEADER

Sarah Ann Monti, Doctor of Philosophy, 2015

Directed By: Professor Michael Summers, Chemistry and Biochemistry

The structured 5' leader (5'-L) of the HIV genome regulates several important steps in the HIV life cycle and is the most highly conserved region of the genome. Previous work identified that the 5'-L exists in two mutually exclusive conformations: a monomeric and a dimeric conformation. This work focuses on enhancing our understanding of the monomeric conformation of the 5'-L, and its role in the HIV life cycle. Structural studies of the 5'-L are hindered by its large size: at 356 nucleotides it is approximately thirteen times larger than the average nuclear magnetic resonance (NMR) derived RNA structure of 27 nucleotides. To overcome these difficulties, a novel NMR strategy, long-range probing by Adenosine Interaction Detection (lr-AID), was utilized to identify the interactions in the 5'-L that stabilize the monomeric conformation. In contrast to previous predictions, we discovered that the monomeric conformation of the

5'-L is stabilized by sequestration of the dimerization initiation site (DIS) in base pairing with residues 105 through 109 of the unique 5' region. Identification of the base pairing interactions that stabilize the monomeric conformation allowed the structure of the monomeric 5'-L to be probed by NMR. Two-dimensional proton-proton NOESY spectra of predicted secondary structure elements were compared to the spectrum of the monomeric 5'-L. The presence of a number of these elements including the top of the TAR stemloop, the intact DIS stem, the splice-donor stemloop, and an extended ψ stemloop were confirmed by NMR. These findings map out large portions of the monomeric 5'-L secondary structure, allowing development of a smaller construct for structural studies of a monomeric core. Additionally, inconsistencies in the reported 5' transcription start site (TSS) of the HIV genomic RNA were identified. As some of the differences were reported to result from the presence of the 5'-7-methylguanosine cap, the capped 5'-L corresponding to each TSS was synthesized and compared. Studies of the monomer-dimer equilibrium of the capped 5'-L corresponding to the different TSSs revealed that the TSS affects the dimerization propensity of the 5'-L. Our increased understanding of the structure of the monomeric 5'-L provides insights into how this region regulates important events in the HIV life cycle.

CHARACTERIZATION OF THE MONOMERIC CONFORMATION OF THE HIV-1 5' LEADER

By

Sarah Ann Monti

Dissertation submitted to the Faculty of the Graduate School of the

University of Maryland, Baltimore County, in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

2015

Acknowledgements

The work presented in this dissertation represents the hard work and dedication of an enormous number of people. First of all, I am immensely grateful to my parents, Craig and Linda Monti, for their support, love, and faith in me from birth. Many thanks are also due to my amazing boyfriend, Dr. Steve Storck, whose exceptional amount of support made this possible. Steve, your dedication to your own education and love of research helped inspire me to continue mine. I would not have been able to do this without you. All of my family and friends have been an invaluable resource providing me with love, support, encouragement, and a connection to the real world. My sister, Hannah Monti, managed to live with me throughout this process and was a continual source of love, encouragement, and humor. Steve's family, his parents, Bob and Sandy Storck, and his brother, Rob Storck, have been a second family to me and supported me in every way possible. My friends, those I had before this program and those that I was lucky enough to gain during it, are amazing. Thank you Maggie Mitchell, Mari Markkula, Pauliina Markkula, Dr. Carolyn Black, Dr. Clare Lau, Curtis Adamo, Criselle Anderson, Justin Meserve, and so many others for managing to maintain a friendship with a mostly absentee friend during this process. Thank you Dr. Lola Brown, Deborah Girma, Dr. Sarah Keane, Dr. Xiao Heng, Dr. Yuanyua Liu, Dr. Thao Tran, Dr. Jan Marchant, Dr. Peter Mercredi, Josh Brown, Christy Gaines, Janae Baptiste, and Julie Nyman for being such fabulous labmates and friends. The most amazing part of the Summers lab is the support that we get from each other. I have relied on every single one of you, and you are family now.

An enormous debt of gratitude is due to Dr. Michael Summers who has mentored and supported me throughout this process. You have walked the difficult and fine line of providing me with the freedom to pursue a number of different avenues in this project, while also being there to guide me every time I show up in your office needing a pep talk and direction. I look forward to being able to contact you in the future and ask, "Anything new?" You create collaborative and supportive lab environment that has made these last several years both educational and fun.

I'm deeply grateful to Dr. Bruce Johnson for offering me endless help with my project, NMR processing, programming, career choices, and life in general. I have always felt like I can go to you with any problem and receive help and advice. I am inspired by your ingenuity and independence. Once again I am relying on your guidance as you read this dissertation and recommend edits for me.

I am similarly grateful to Dr. Alice Telesnitsky, who has also agreed to read and advise me on this dissertation. Your presence on my committee has guided this project to really consider the biological applications of our structural findings - I think to our immense success. Your excitement has served to buoy me through.

Many thanks are due to Dr. Zeev Rosenzweig and Dr. A-Lien Lu-Chang both of whom agreed to serve on my defense committee to make this dissertation defense possible. Additional thanks are due to Dr. Elsa Garcin and Dr. Alex Drohat who served on my proposal committee and offered valuable advice for my project.

I would not have confidence in the validity of my research without the help of my amazing collaborators, Dr. Ioana Boeras in Dr. Kathy Boris-Lawrie's lab, and Dr. Sergei Kharytonchyk in Dr. Alice Telesnitsky's lab. Your studies confirming the biological relevance of my constructs were invaluable.

Much of the work in this dissertation depended on the efforts of a number of talented undergraduate and high school students. Amar Kaneria and Verna Van were especially wonderful not only as students and labmates, but as exceptional people and researchers that I am lucky to know. Many others have contributed to this project including Megan Milstein, Canessa Swanson, Alex Shuey, Erran Briggs, Nicholas Bolden, Emily Russo, Chang-Wu Mungai, Emily Diaz, Charles Oliver, Jessica Smith, Zachary Osunsade, and Tyler Parenzan. I am grateful for the chance to interact with such a wonderful group of students.

The support staff in the Summers Lab and the UMBC Chemistry department have been invaluable. Cindy Finch, I would be completely lost without you. Yu Chen, I couldn't possibly count the number of times I've required your computer help. Rob Edwards, you keep everything that I need for the research running all the time - even when you're supposed to be on vacation. Dr. Holly Summers, you are the jack of all trades, helping with organization, communication, acting as a lab technician, and a sympathetic ear all at the same time. Without the guidance of Patty Gagne I would never have been registered for classes or managed the forms and paperwork involved in graduate school. Sandy Wilkens, I could always depend on you to point me in the right direction.

I have many people to thank for my training. Mrs. DeWitt at Tome School made chemistry sound like the most interesting and exciting field in the world. Dr. James Fishbein accepted me into his organic chemistry lab as an undergraduate where, under the tutelage of Dr. Nick Zink, I learned that research is what I love. Dr. Daniele Fabris inspired my love of RNA structure, and Dr. Kevin Turner taught me virtually everything I know about mass spectrometry and analytical chemistry. Foyeke Daramola and Dr. David Weber made it possible for me to come back to the Ph.D. program after my leave of absence, and Dr. Summers was kind enough to take me in to support me and advise me for the rest of my time here. Dr. Lianko Garyu and her students Daniel Tumillo and Azra Hosic began my training in the Summers Lab, and this training was continued throughout my career by virtually everyone in the lab, but especially Drs. Xiao Heng, Sarah Keane, and Jan Marchant. Sarah and Jan, I can't thank you enough for all of your help preparing this manuscript. Dr. Lola Brown, the friendship that I developed with you in this lab may be the most enduring and valuable thing to come from my time here. Thank you. I am grateful to be leaving this project in the capable hands of Joshua Brown. You've got this, Josh!

Table of Contents

List of Figures

List of Abbreviations

5'-L	5' Leader
7mG	7-methyl Guanosine
AD	Active Domain

AIDS	Acquired Immune Deficiency Syndrome
ARM	Arginine Rich Motif
BMH	Branch Multiple Hairpin
CA	Capsid
CD4	*Cluster of Differentiation 4*
CDK9	Cyclin Dependent Kinase
CycT	CyclinT1
DIS	Dimerization Initiation Site
DNA	Deoxyribonucleic Acid
Env	Envelope
Gag	Group-specific Antigen
gp120	Glycoprotein 120
HAART	Highly Active Anti-retroviral therapy
HIV	Human Immunodeficiency Virus
HMQC	Heteronuclear Multiple-Quantum Correlation
IN	Integrase
ITC	Isothermal Titration Calorimetry
LDI	Long Distance Interaction
lr-AID	Long Range Probing by Adenosine Interaction Detection
Lys	Lysine
MA	Matrix
mRNA	Messenger RNA
NC	Nucleocapsid
Nef	Negative Regulator Factor
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy

PAS	Primer Activation Signal
PBS	Primer Binding Site
Pol	Polymerase
polyA	Polyadenylation
P-TEFb	Positive transcription elongation factor
Rev	Regulator of expression of virion proteins
RNA	Ribonucleic Acid
RNAPII	RNA polymerase II
RRE	Rev Response Element
RT	Reverse Transcriptase
SD	Splice Donor
SHAPE	Selective 2'-Hydroxyl Acylation Analyzed by Primer Extension
TAR	Trans-activating Response
Tat	Trans-Activator of Transcription
TFIID	Transcription factor IID
TOCSY	Total Correlation Spectroscopy
tRNA	Transfer RNA
TSS	Transcription start site
Vif	Viral Invectivity Factor
Vpr	Viral protein R
Vpu	Viral Protein Unique
ψ	Packaging Signal

Chapter 1: Introduction

HIV Health Impact

The human immunodeficiency virus (HIV) is a pandemic posing a serious health threat to the human population. HIV is the causative agent of acquired immune deficiency syndrome (AIDS), a disease in which the immune system is systematically suppressed until the patient is unable to mount a sufficient immune response and succumbs to a secondary infection. In the 2013 global summary of the AIDS epidemic, the World Health Organization estimated that there were approximately 35 million people living with HIV, with 6000 new infections occurring every day.

HIV Life Cycle

Extensive research on HIV has revealed a great deal of information about the life cycle of this virus, which has been used to develop therapeutics to target the disease. Figure 1.1 shows a cartoon depiction of the HIV life cycle. HIV typically attaches to human T-cells via interactions between the glycoprotein 120 receptors on the membrane of the virion and CD4 receptors on the host cell. Additional binding to host chemokine receptors results in a conformational rearrangement of the glycoprotein 120, resulting in membrane fusion[1; 2]. Upon membrane fusion, the contents of the viral particle are released into the cytoplasm of the host cell[3]. As a retrovirus, HIV encodes its genetic information as a single-stranded RNA genome which is packaged as a pseudo-diploid dimer[4]. A viral enzyme, reverse transcriptase, is used to reverse transcribe the RNA genome into a double-stranded DNA copy of the genome, which contains a promoter region upstream of the DNA encoding the genome[5]. This proviral DNA is transported

into the nucleus of the cell and incorporated into the host genome through the action of another viral enzyme, integrase. Upon integration of the proviral DNA into the host genome, the host cell is permanently infected with the virus, making HIV a life-long, persistent infection[3].

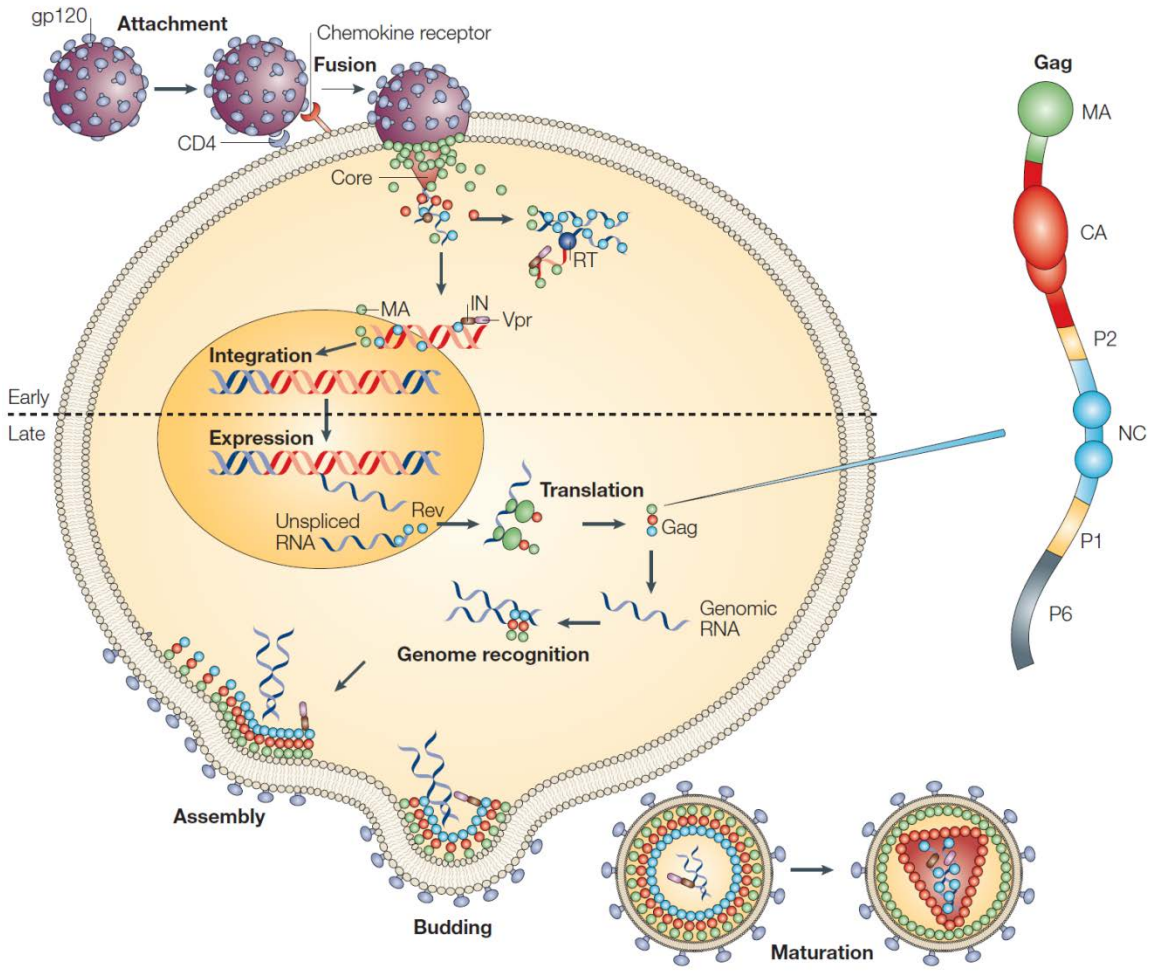

Figure 1.1. Cartoon depiction of the HIV life cycle[6].

The virus then co-opts the host cellular transcription machinery to generate its mRNA. The mRNA is transcribed by RNA polymerase II (RNAPII). Co-transcriptional splicing can occur, leading to three mRNA classes from the original HIV mRNA transcript: fully spliced RNAs which encode proteins Rev, Tat, and Nef; singly spliced

RNAs which encode the Env, Vpu, Vif, and Vpr proteins; and unspliced transcripts which include the entire 9 kb genome and encode the Gag and Gag-Pol proteins[3]. Initially only the fully spliced transcripts can be exported from the nucleus using host cell trafficking machinery, but, upon translation, Rev returns to the nucleus and mediates the export of the singly spliced and unspliced HIV mRNAs through interactions with a region of the genome labeled as the Rev Response Element (RRE)[7]. After Rev exports the unspliced RNA from the nucleus, the unspliced RNA has two roles in the viral life cycle: it is used as a template for the translation of the Gag polyprotein, and it dimerizes and is packaged into new viral particles where it serves as genetic material for the new virion[4]. It is unknown whether the same molecule of RNA can serve both roles, or if there are two pools of RNA. This is an area that requires further study, but I will address our preliminary findings in Chapter 4. It has been shown, however, that dimerization is directed by a palindromic sequence known as the Dimerization Initiation Site (DIS) in the 5' leader (5'-L) of the genome[8; 9]. The Gag polyprotein consists of three major domains: matrix (MA), capsid (CA), and nucleocapsid (NC)[10]. MA targets Gag to assembly sites on the plasma membrane. Capsid, upon cleavage, assembles into the conical core of the virion. NC is the major RNA binding domain of Gag. It contains two zinc knuckle motifs that bind exposed guanosines with high specificity[11]. Figure 1.2 shows how the exposed guanosine binds in the hydrophobic cleft of a zinc knuckle[11]. The 5'-L contains a number of high affinity NC binding sites[12; 13]. The monomeric 5'-L contains approximately six high affinity NC sites, while the dimeric 5'-L contains approximately 16 high affinity sites per strand, or 32 total[13]. It is this high specificity for the dimeric 5'-L that allows selective packaging of the HIV genome[6]. NC also acts as an RNA folding

chaperone[14]. It is not been determined *in vivo* whether the 5'-L dimerizes and then interacts with the NC domain of Gag, or if NC interactions are responsible for promoting dimerization. If the 5'-L is not present, or if the dimeric conformation cannot form, the virus will indiscriminately package cellular RNAs and form a non-infectious particles[15].

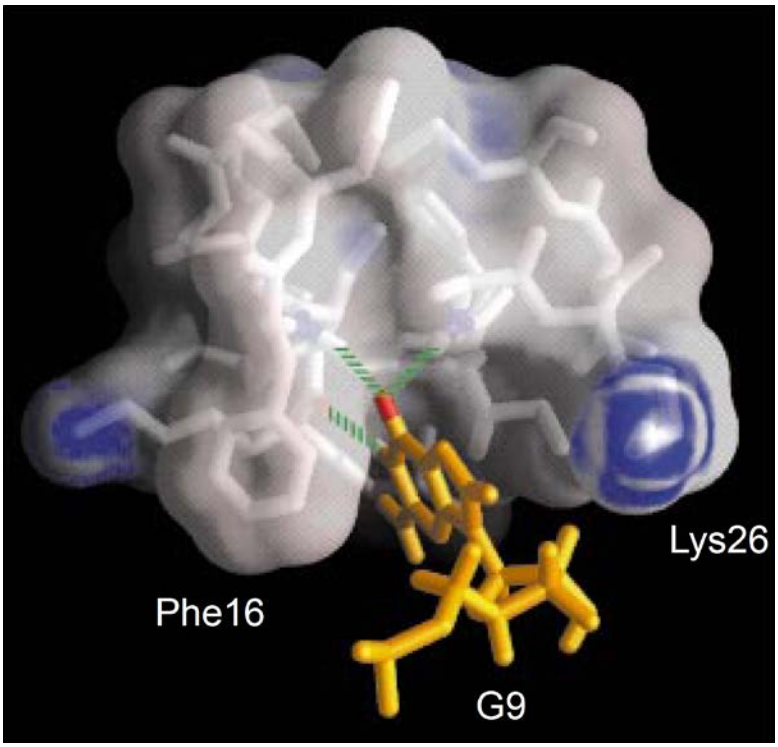

Figure 1.2. Exposed guanosine bound in the hydrophobic cleft of an NC knuckle[11].

Upon formation of the Gag-RNA complex, the MA domain of Gag directs the complex to assembly sites on the inner leaflet of the plasma membrane[16]. The nascent virion buds off from the host cell as an immature viral particle. A viral enzyme, protease, cleaves the Gag polyprotein into the individual MA, CA, and NC proteins which form the structure that characterizes a mature HIV virion with MA associated with the membrane, CA forming a conical core, and NC associated with the RNA inside of the CA core[3]. At this point the cycle can begin again.

Antiretroviral Therapies

Currently there are four main targets in the HIV life cycle for therapeutics: fusion, reverse transcription, integration, and maturation. Anti-retroviral therapy development represents an excellent example of the power of structural biology. The structure of protease was solved in 1989 by Wlodawer *et al.*[17], allowing rational drug design to be used to develop the protease-inhibitor class of drugs. Anti-retroviral drugs are used in combination, as highly active anti-retroviral therapy (HAART), because the mutation rate of the HIV genome is sufficiently high, due to the lack of proofreading activity of HIV reverse transcriptase, such that drug resistant strains develop upon treatment with a single inhibitor. While HAART therapy can be very effective, drug resistant strains still emerge due to patient non-compliance, resulting in a continuous need for new anti-retroviral therapies. All of the current treatments target proteins important in the HIV life cycle. The structured 5'-L of the HIV genome is an intriguing alternative drug target for two reasons: it controls several different important steps in the HIV life cycle (discussed below), and it is the most highly conserved region of the genome. However, the lack of structural information on this large, functional region of the genome has thus far hindered any structure-based therapeutic development.

Elements of the HIV-1 5'-Leader

The 5'-L is a 356 nucleotide region at the 5' end of the genome which does not encode protein. Instead this region of the RNA is highly structured, and these RNA structures interact with proteins to control a number of important events in the HIV life cycle. Figure 1.3 shows a dimer-competent secondary structure predicted by Abbink *et*

al. based on chemical probing and free energy calculations. This structure identifies a number of commonly predicted stemloops which are reported to have regulatory roles.

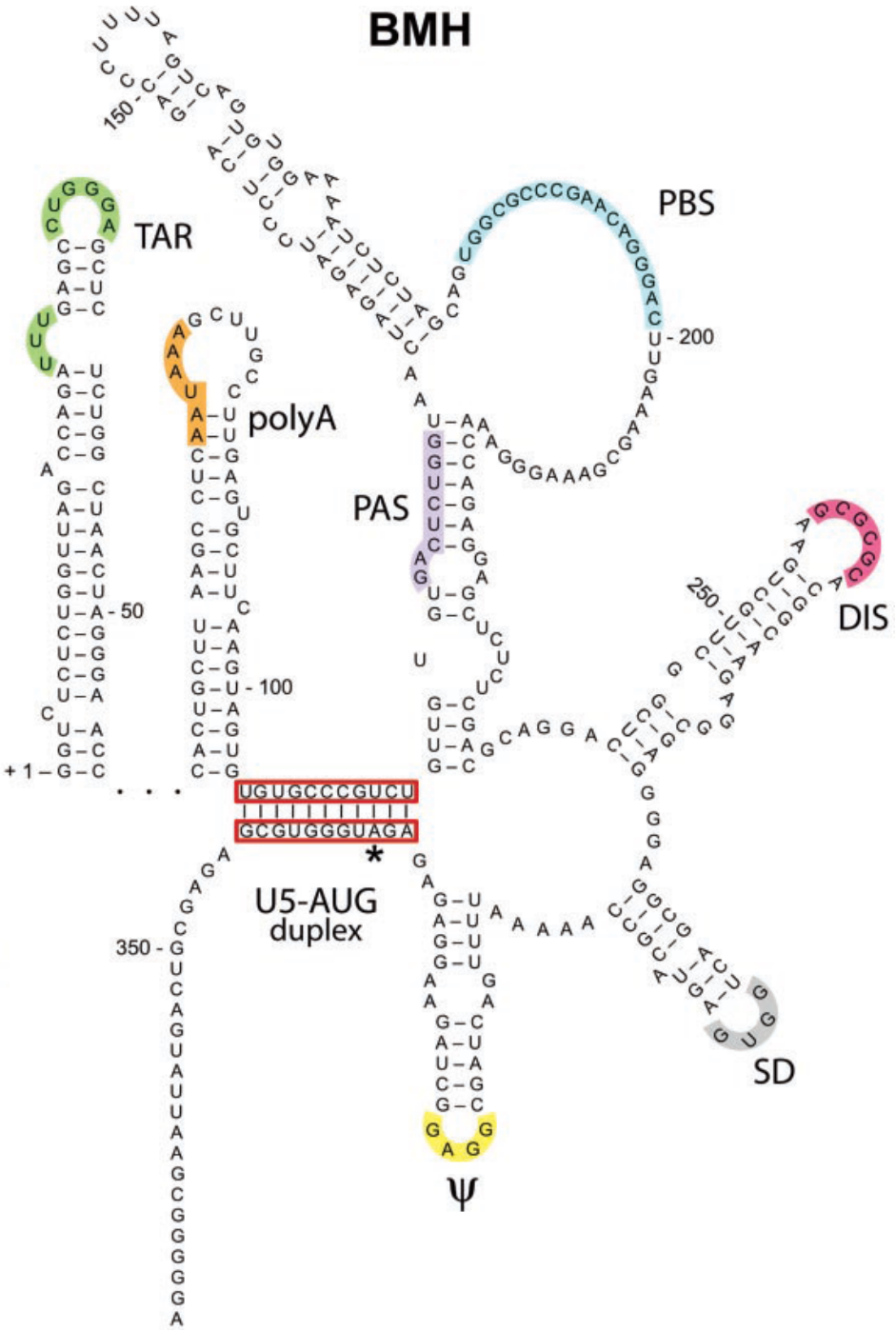

Figure 1.3. The Branch Multiple Hairpin (BMH) model of the HIV-1 5'-L predicted by Abbink *et al.* with the stemloops labeled[18].

The *trans* activating response (TAR) hairpin is important for transcriptional activation of the genome[19]. TAR is part of the repeat region of the 5'-L, and as such it is repeated at the 3' end of the genome, as is the polyA region. The polyA hairpin contains the polyadenlyation signal which is utilized at the 3' end of the genome to recruit cellular machinery to attach the polyA tail to the mRNA; however its role in the 5'-L is not well understood. Downstream of polyA is the beginning of the unique 5' region (U5). This region has been implicated in base pairing with distal parts of the 5'-L to stabilize the dimeric conformation of the 5'-L[18]. The primer binding site (PBS) loop is the binding site for the tRNA Lys, which is used as a primer for reverse transcription[20]. The DIS stemloop contains a palindromic sequence in its loop, and it is the interaction between the DIS of one strand of RNA and the DIS of another strand of RNA that initiates dimerization of the 5'-L[8]. The splice-donor (SD) stemloop contains the major splice donor sequence. All spliced transcripts utilize this splice site, meaning that all spliced and unspliced transcripts contain the same nucleotide sequence from nucleotides 1-288[21]. The ψ hairpin contains a high affinity nucleocapsid binding site[11] and was originally thought to be the main structural determinant for genome packaging[22]. A structure of the ψ hairpin bound to NC has been solved and is shown in Figure 1.4[11]. The AUG stemloop contains the start codon for the Gag polyprotein. It has also been suggested to interact with the U5 region to stabilize the dimeric structure (as shown in Figure 1.3)[18].

The 5'-L exists in a monomeric or a dimeric conformation, and these two different conformations are expected to play different roles in the HIV life cycle[18]. Because dimerization is not required until packaging is initiated, it is reasonable to predict that the monomeric conformation is the transcribed conformation, which may undergo splicing,

and is translated. The dimeric conformation promotes RNA dimerization, has a large number of NC binding sites, and is selectively packaged[6]. The structure of the dimeric conformation has been extensively studied[13; 23; 24], and recently the three dimensional structure of the dimeric core packaging signal was published[25]. This dissertation will focus on furthering our knowledge of the structure of the monomeric conformation.

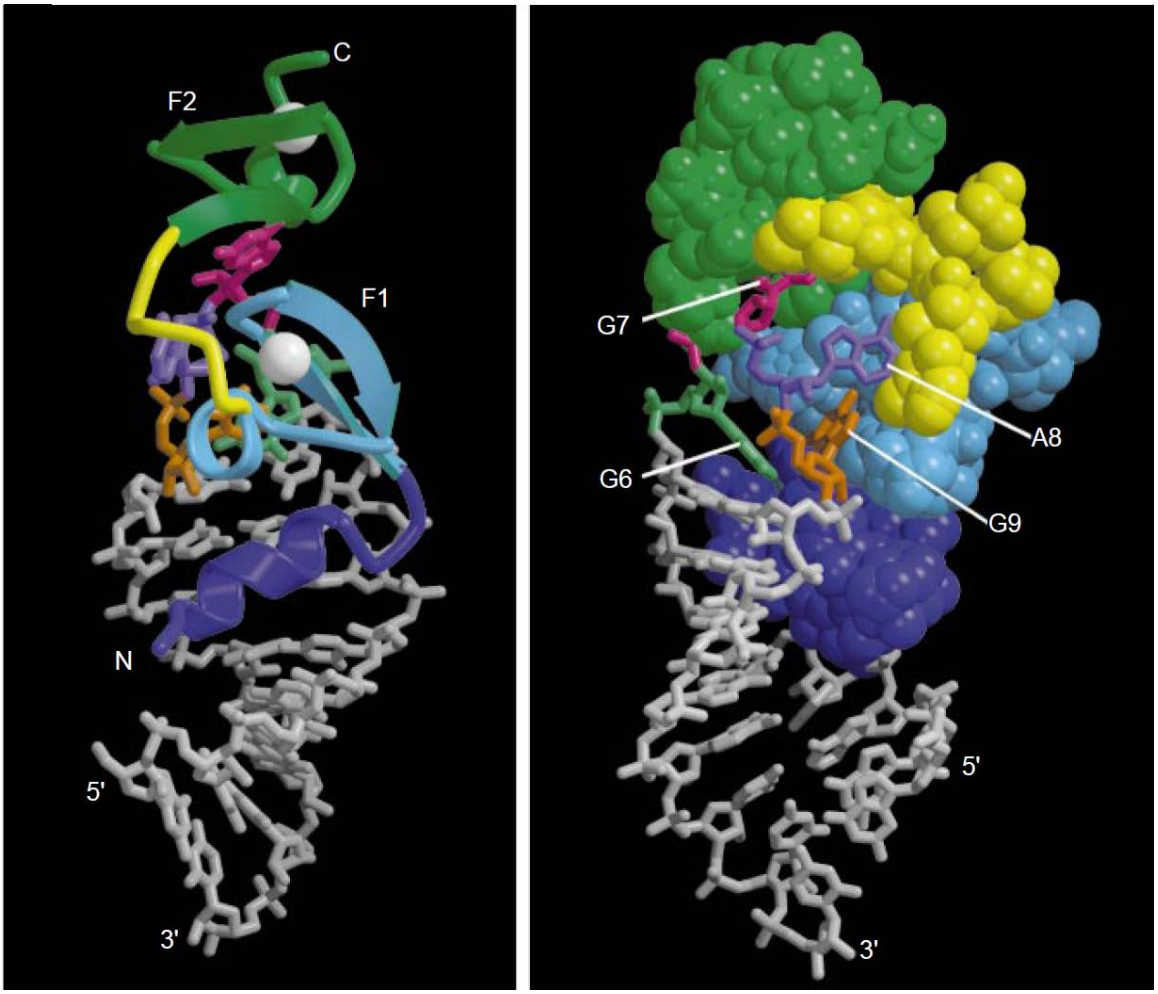

Figure 1.4. Three-dimensional structure of the NC protein bound to its high affinity site on the ψ hairpin of the 5'-L[11]. Left: HIV-1 ψ RNA (sticks) with NC protein (ribbon) bound. The 3_{10} helix is is purple, the F1 knuckle is in blue, the linker segment is in yellow, and the F2 knuckle is in green. Zinc atoms are showns as white spheres, and the RNA is gray sticks except for G^6, which is light green, G^7, which is pink, A^8, which is violet, and G^9, which is orange. Right: HIV-1 ψ RNA (sticks) with NC protein (space-filling)

bound, rotated 90˚ counter-clockwise from the left image. The interactions between G^9, A^8 and F1 are shown, as are the G^7 and F2 interactions.

It is possible to glean some structural information from the biological role of the monomeric RNA. For example, for transcriptional activation it is essential that the TAR hairpin forms, to allow binding of Tat to promote transcriptional elongation[19]. Before Tat is translated and returns to the nucleus, HIV has a very low basal level of transcription because the majority of transcripts prematurely terminate at the end of the TAR hairpin. Tat recruits the human positive transcription elongation factor (P-TEFb) to TAR through interactions with its cyclinT1 (CycT) domain, forming a Tat/TAR/CycT complex in which CycT interacts with the loop of the TAR (Figure 1.5)[26; 27].

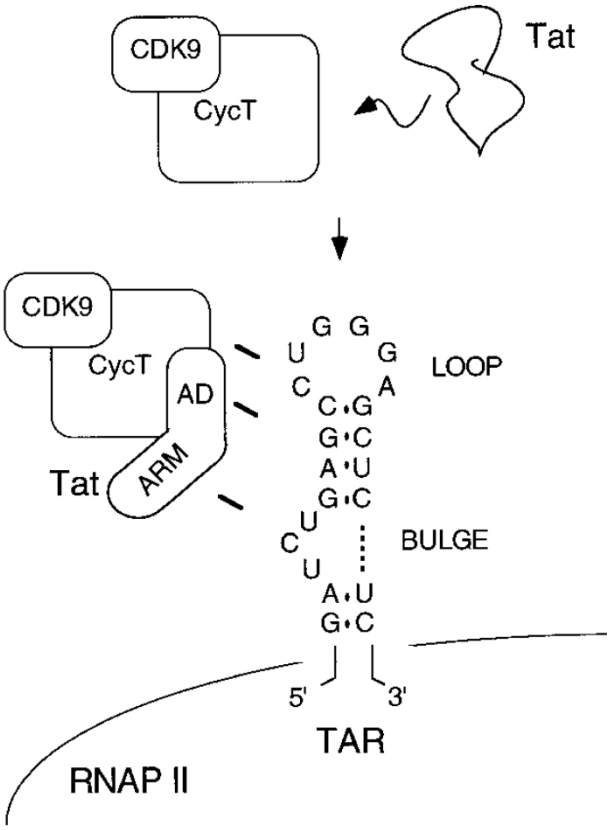

Figure 1.5. Model showing that the Activation Domain (AD) of Tat interacts with the CyclinT1 (CycT) domain of P-TEFb, allowing the Arginine Rich Motif (ARM) domain of Tat to bind to the bulge of TAR while CycT binds to the loop[27]. These interactions allow the C-terminal domain of RNAPII to be phosphorylated, releasing RNAPII from pausing.

This places P-TEFb in close proximity to RNAPII, allowing the cyclin-dependent kinase 9 (CDK9) domain of P-TEFb to phosphorylate the C-terminal domain of RNAPII. Furthermore P-TEFb phosphorylates factors that inhibit elongation, releasing RNAPII from pausing. In the absence of the TAR bulge and loop, HIV RNA transcription is largely blocked at the elongation step[28]. This suggests that in order for the genome to be transcribed (monomeric conformation), the top of the TAR stemloop is must be present.

Additionally, it has been suggested that splicing levels are affected by the secondary structure in the vicinity of the major splice-donor[29]. Splicing efficiency is regulated by the stability of the splice-donor (SD) stemloop. Stabilizing the stemloop reduces splicing, while destabilizing the stemloop increases splicing, however both SD stabilization and SD destabilization are detrimental to overall viral fitness[29; 30]. Intriguingly, the recently published structure of the HIV-1 packaging signal reveals that in the dimeric conformation, the splice-donor hairpin does not form, and instead the residues that have been predicted to constitute the splice-donor hairpin are involved in alternative base pairing to form a double three-way junction[25]. This suggests that the splice-donor region is one of the areas of the 5'-L that undergoes a major secondary structure rearrangement during the conformational change from monomer to dimer.

Overall 5'-Leader Structure

The structure of the complete 5'-L has been extensively studied by chemical probing and mutagenesis, and numerous different secondary structures have been posited[18; 31; 32; 33; 34; 35; 36; 37; 38; 39]. The techniques utilized by these groups to research the structure suffer from the fact that chemical probing, mutagenesis, and biochemical assays provide no direct structural information. Chemical probing, which was extensively utilized in previous studies, can provide data suggesting what nucleotides are accessible (assumed to be single-stranded) and what nucleotides are occluded (assumed to be base paired). From this information, secondary structure calculations can be performed. However, as chemical probing data does not provide information specifying the base pairing partners for each nucleotide, there are a number of secondary structure options that have the possibility of satisfying the data. Additionally, because RNA structure is more complex than a simple A helix, it is possible to have residues that are involved in secondary and tertiary structures that may be chemically accessible while participating in base pairing, or maybe be inaccessible to chemical probing, but not base paired. Biochemical and mutagenesis experiments suffer from similar limitations. Simply because a mutation has altered a phenotype of the RNA as predicted by a hypothesis does not guarantee that the change in phenotype occurred due to the reasoning predicted by the hypothesis. Furthermore, studies of the 5'-L are complicated by the fact that in solution it exists in a monomer-dimer equilibrium. Studies that fail to separate the monomer and dimer would give data from a mixture of the two species and would therefore likely not be informative about the structure of either conformation individually. A recent study showed that separation of the monomer and dimer of the 5'-L by gel electrophoresis prior

to applying SHAPE chemical probing provided different results for the monomeric and dimeric conformations[40]. This suggests that previous results from SHAPE probing of the mixture probably provides an average of the reactivities of the monomeric and dimeric reactivities, which complicates interpretation of the data[38; 39].

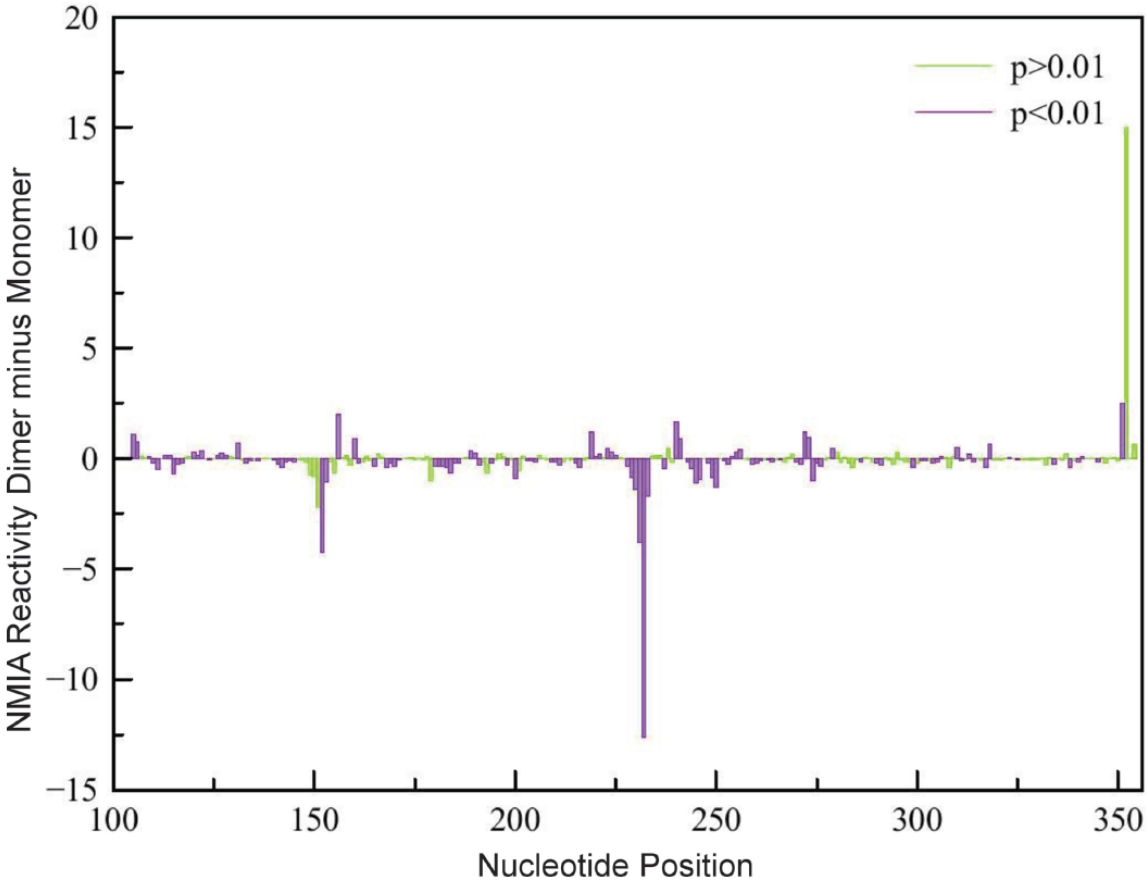

Figure 1.6. Difference in the SHAPE reactivity between the dimer and monomer of the 5'-L. Many nucleotides show significant differences in reactivity between the monomeric and dimeric conformations. Purple bars show statistically significant differences in nucleotide reactivity, while green bars are not significant, with $P < 0.1$[40].

Direct structural information can be obtained by techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. X-ray crystallography has not yet been successfully used to study the complete 5'-L because

conformational heterogeneity and flexibility causes difficulties for crystallization of large RNAs. For analysis of large RNAs, NMR spectroscopy suffers from severe spectral overlap and disadvantageous relaxation properties[41]. As a result, while isolated hairpin structures have been solved by NMR spectroscopy and X-ray crystallography studies[11; 42; 43; 44; 45], the overall dimeric and monomeric structures remain elusive. Using a combination of techniques developed to aid in the assignment of large RNAs by NMR spectroscopy, the Summers lab was able to solve the structure of the core packaging signal of the 5'-L[25]. The core packaging signal represents the all of the elements required for efficient packaging in the dimeric conformation[13]. While this 155 nucleotide RNA is less than half the size of the complete 5'-L, it provides a plethora of direct structural data on the dimeric conformation of the 5'-L, and the techniques developed for this project will allow the structure of other large RNAs to be solved[25].

NMR Strategies for Large RNAs

The NMR experiment that provides the most useful information for structural studies of RNA is nuclear overhauser effect spectroscopy (NOESY). NOESYs provide information about through space interactions. If two ^{1}H are within 6 Å in space, they will generate a cross peak, representing an NOE, on a 2D ^{1}H-^{1}H NOESY spectrum. This provides information about sequential nuclei that are close in space which allows one to walk through the sequence of the RNA: for example, in an A helix, the H8 of an adenosine or guanosine has an NOE to its own H1' as well as the H1' of the preceding residue (Figure 1.7).

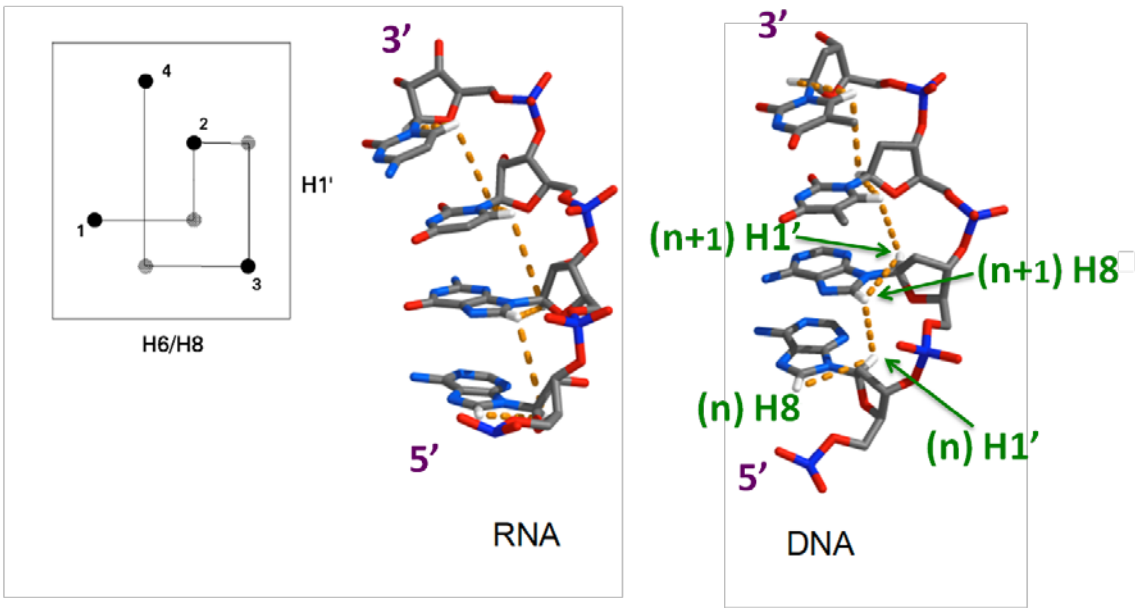

Figure 1.7. Demonstration of how NOE connectivities between H8s and H1's of sequential residues can be used to assign a nucleotide NMR spectrum by a NOESY walk. Dashed yellow lines indicate NOEs. Figure courtesy of Dr. Xiao Heng.

Because NOEs result from through space interactions, it is also possible to have NOEs between two nuclei that are very far apart in sequence, but close in space. For example, in an A helix, the H2 of an adenosine generates an NOE to its own H1', the following H1', to the cross-strand H1', and to the H1' of the cross-strand plus one H1' (Figure 1.8). This provides information about what bases are close enough in space to be participating in base pairing interactions. With these and other rules it is possible to perform a NOESY walk for very small RNAs in an A helix[46], but spectral overlap causes it to be very difficult as RNAs increase in size.

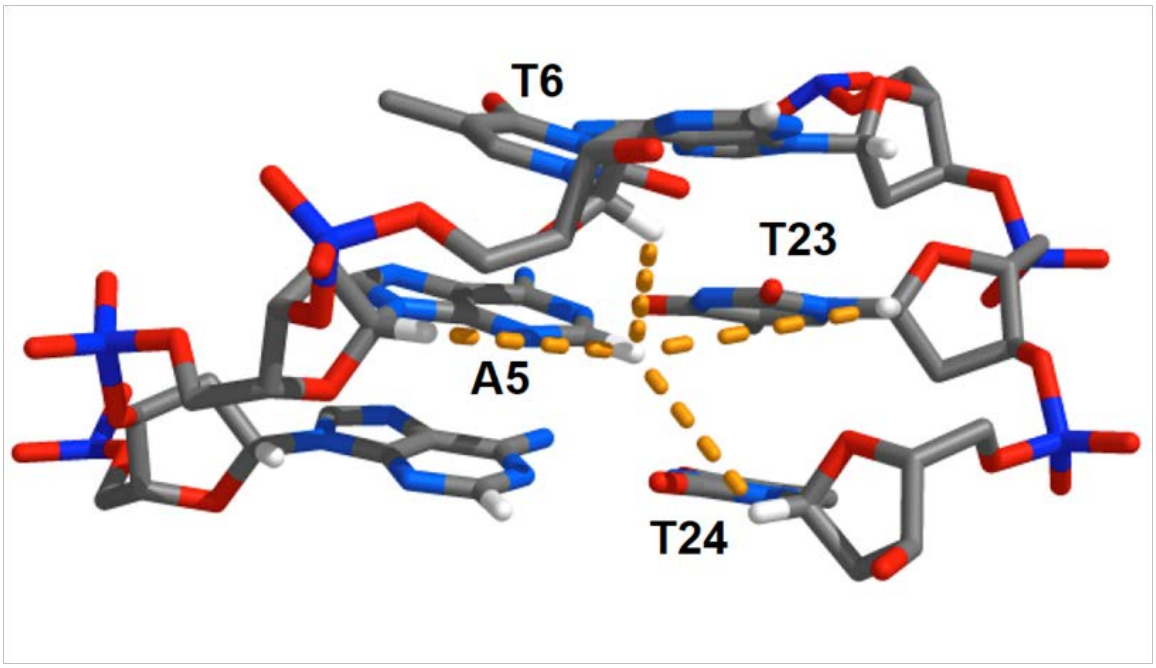

Figure 1.8. NOEs from an adenosine H2 in a helix to residues not only close in sequence (T6) but also close in space (T23 and T24) which provide information about cross-strand residues. Figure courtesy of Dr. Xiao Heng.

One of the techniques utilized to reduce spectral overlap of large RNAs is selective deuteration. This improves the spectral quality of large RNAs in two ways. First, the signal line widths are narrowed due to reduced ^{1}H-^{1}H spin diffusion. Additionally, deuterium (^{2}H) is essentially "invisible" under ^{1}H NMR conditions[41]. Therefore, incorporating ^{2}H into certain positions of the RNA greatly reduces spectral complexity[41], as seen in Figure 1.9.

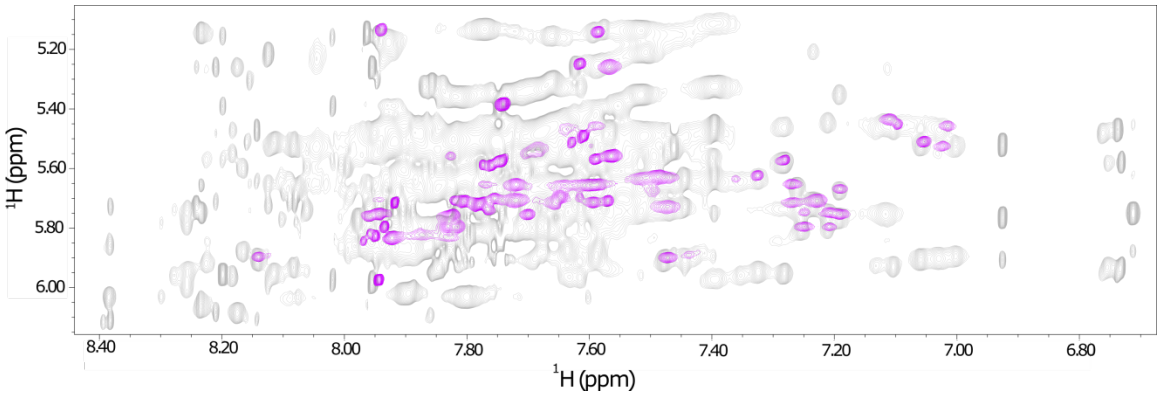

Figure 1.9. Overlay of a fully protiated NMR sample of the HIV-1 core packaging signal (gray) and a sample with only guanosines protiated while all other nucleotides are deuterated (purple). Figure courtesy of Dr. Sarah Keane.

This is typically accomplished by transcribing the RNA with some combination of partially or fully deuterated nucleotide triphosphates[41]. Figure 1.10 shows the numbering for the protons on an adenosine, guanosine, cytosine, and uridine, bases, as well as the ribose ring.

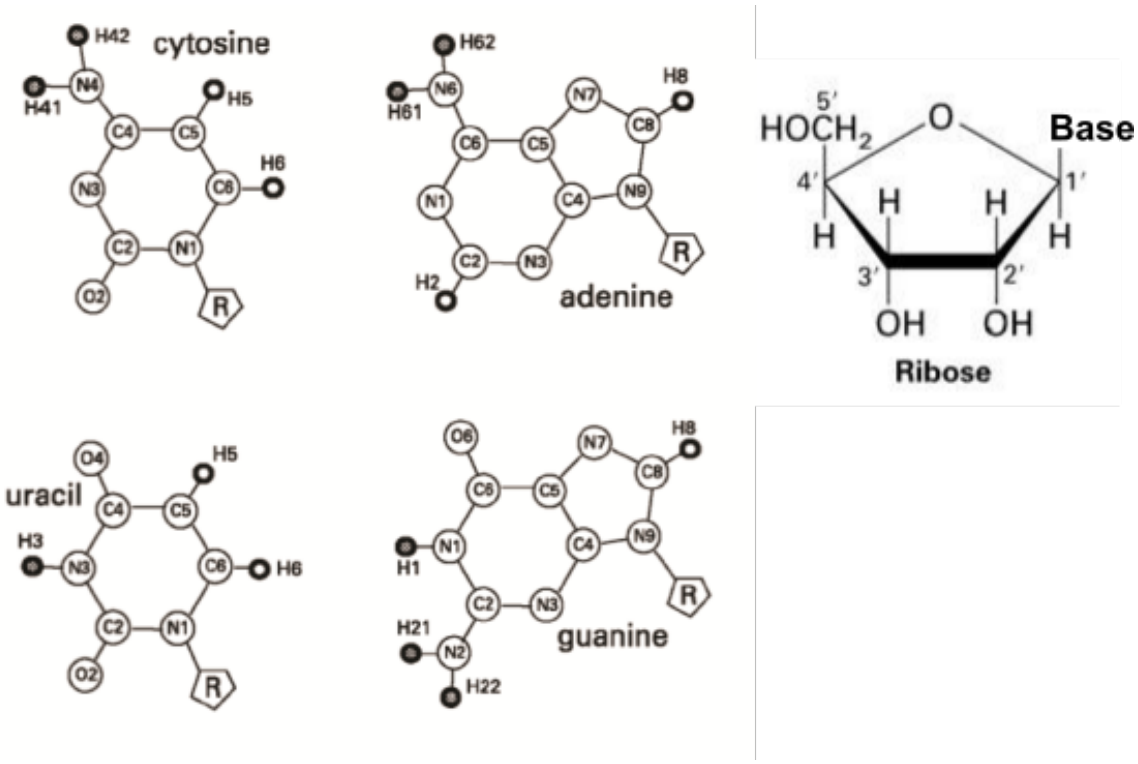

Figure 1.10. Illustration of the nucleotide bases with the carbons, nitrogens, and hydrogens numbered. The ribose ring is also shown with the carbon positions numbered.

Partially deuterated samples are named by the protiated positions. For example, an A2rGr sample is an RNA that only has protons on the ribose ring and C2 position of the adenosine, and the ribose ring of the guanosine. All other positions have been replaced with deuterium. While employing a deuterium-edited approach is very valuable for reducing spectral overlap, it does limit the information gained from a single experiment. To obtain full sequence coverage, NOESYs can be performed on multiple samples with different ^{2}H incorporation schemes.

Another method that is extremely helpful in assigning large RNAs is to make small oligonucleotides which contain a structure expected to appear in the large RNA. These small RNAs can be readily assigned using a traditional suite of NOESY, TOCSY

and HMQC experiments. These controls can then be used for as comparison against the larger RNA, if the chemical shift and NOES pattern identified in the small RNA matches the large RNA, it suggests that the isolated structure occurs in the large RNA NOESY structure. This methodology will be utilized in Chapters 2 and 3 of this dissertation.

Additional techniques for assigning large RNA include segmental labeling. This allows two or more sections of the same RNA to be differentially labeled. This can be accomplished by enzymatic cleavage and ligation of the RNA[41], or by creating two RNA fragments that have complimentary base pairs to allow annealing[25]. Differential labeling has many advantages. It can reduce spectral complexity by allowing small sections of the RNA to be protiated and assigned, while the rest remains deuterated. Additionally, it can be used to provide information about the section of the RNA sequence from which NOEs originate. For example, when solving the structure of the core packaging signal, a fragment based segmental labeling scheme was used to distinguish NOEs that were from 5' or 3' fragments of the core packaging signal (Figure 1.11). The disadvantages to segmental labeling are that it increases the number of samples that must be made to solve a structure, enzymatic cleavage and ligation are often low efficiency, resulting in reduced RNA yield, and the fragmentation approach requires structural knowledge of a hairpin that can serve as the site of fragmentation and annealing.

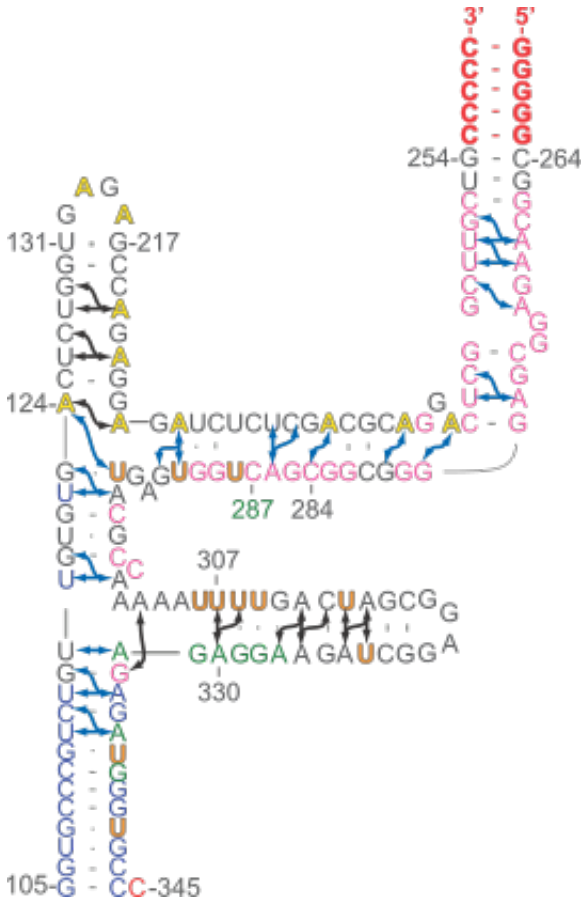

Figure 1.11. Fragment based segmental labeling strategy used for NOE assignments of the HIV-1 core packaging signal. The 5' fragment extends from nucleotides 105 through 254, with a run of five cytosines at the 3' end. The 3' fragment begins with a run of five guanosines, and proceeds from nucleotides 264 through 345. Adenosine H2 NOEs are shown by arrows. Blue arrows represent NOEs assigned in fragmentally labeled samples. Black arrows represent NOEs assigned from uniformly labeled samples. NOE Assignment made possible by the fragmentation scheme guide the structure of the upper three-way junction[25].

Composition of This Dissertation

The remainder of this dissertation comprises four chapters. Chapter 2 describes the studies which identified the interactions that stabilize the monomeric conformation of the HIV-1 5'-L. Chapter 3 discusses what elements of secondary structure can be

identified in the monomeric 5'-L by NMR. Chapter 4 examines the effects of heterogeneity at the 5' start site and the presence of the 5'-methylguanosine cap on the monomer-dimer equilibrium of the 5'-L. Chapter 5 offers conclusions and possible directions for future work.

References

1. Clapham, P. R. & Weiss, R. A. (1997). Immunodeficiency viruses. Spoilt for choice of co-receptors. *Nature* **388**, 230-1.

2. Feng, Y., Broder, C. C., Kennedy, P. E. & Berger, E. A. (1996). HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* **272**, 872-7.

3. Turner, B. G. & Summers, M. F. (1999). Structural biology of HIV. *J Mol Biol* **285**, 1-32.

4. Johnson, S. F. & Telesnitsky, A. (2010). Retroviral RNA dimerization and packaging: the what, how, when, where, and why. *PLoS Pathog* **6**, e1001007.

5. Hu, W. S. & Hughes, S. H. (2012). HIV-1 reverse transcription. *Cold Spring Harb Perspect Med* **2**.

6. D'Souza, V. & Summers, M. F. (2005). How retroviruses select their genomes. *Nat Rev Microbiol* **3**, 643-55.

7. Malim, M. H., Hauber, J., Le, S. Y., Maizel, J. V. & Cullen, B. R. (1989). The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338**, 254-7.

8. Paillart, J. C., Marquet, R., Skripkin, E., Ehresmann, B. & Ehresmann, C. (1994). Mutational analysis of the bipartite dimer linkage structure of human immunodeficiency virus type 1 genomic RNA. *J Biol Chem* **269**, 27486-93.

9. Skripkin, E., Paillart, J. C., Marquet, R., Blumenfeld, M., Ehresmann, B. & Ehresmann, C. (1996). Mechanisms of inhibition of in vitro dimerization of HIV type I RNA by sense and antisense oligonucleotides. *J Biol Chem* **271**, 28812-7.

10. Coffin, J. M., Hughes, S. H. & Varmus, H. (1997). *Retroviruses*, Cold Spring Harbor Laboratory Press, Plainview, N.Y.

11. De Guzman, R. N., Wu, Z. R., Stalling, C. C., Pappalardo, L., Borer, P. N. & Summers, M. F. (1998). Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* **279**, 384-8.

12. Rein, A. (1994). Retroviral RNA packaging: a review. *Arch Virol Suppl* **9**, 513-22.

13. Heng, X., Kharytonchyk, S., Garcia, E. L., Lu, K., Divakaruni, S. S., Lacotti, C., Edme, K., Telesnitsky, A. & Summers, M. F. (2012). Identification of a Minimal

Region of the HIV-1 5'-Leader Required for RNA Dimerization, NC Binding, and Packaging. *J Mol Biol*.

14. Egelé, C., Schaub, E., Ramalanjaona, N., Piémont, E., Ficheux, D., Roques, B., Darlix, J. L. & Mély, Y. (2004). HIV-1 nucleocapsid protein binds to the viral DNA initiation sequences and chaperones their kissing interactions. *J Mol Biol* **342**, 453-66.

15. Rulli, S. J., Hibbert, C. S., Mirro, J., Pederson, T., Biswal, S. & Rein, A. (2007). Selective and nonselective packaging of cellular RNAs in retrovirus particles. *J Virol* **81**, 6623-31.

16. Sundquist, W. I. & Kräusslich, H. G. (2012). HIV-1 assembly, budding, and maturation. *Cold Spring Harb Perspect Med* **2**, a006924.

17. Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J. & Kent, S. B. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* **245**, 616-21.

18. Abbink, T. E. & Berkhout, B. (2003). A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J Biol Chem* **278**, 11601-11.

19. Jakobovits, A., Smith, D. H., Jakobovits, E. B. & Capon, D. J. (1988). A discrete element 3' of human immunodeficiency virus 1 (HIV-1) and HIV-2 mRNA initiation sites mediates transcriptional activation by an HIV trans activator. *Mol Cell Biol* **8**, 2555-61.

20. Rhim, H., Park, J. & Morrow, C. D. (1991). Deletions in the tRNA(Lys) primer-binding site of human immunodeficiency virus type 1 identify essential regions for reverse transcription. *J Virol* **65**, 4555-64.

21. Bohne, J., Wodrich, H. & Kräusslich, H. G. (2005). Splicing of human immunodeficiency virus RNA is position-dependent suggesting sequential removal of introns from the 5' end. *Nucleic Acids Res* **33**, 825-37.

22. Hayashi, T., Shioda, T., Iwakura, Y. & Shibuta, H. (1992). RNA packaging signal of human immunodeficiency virus type 1. *Virology* **188**, 590-9.

23. Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A. & Summers, M. F. (2011). NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-5.

24. Lu, K., Heng, X. & Summers, M. F. (2011). Structural determinants and mechanism of HIV-1 genome packaging. *J Mol Biol* **410**, 609-33.

25. Keane, S. C., Heng, X., Lu, K., Kharytonchyk, S., Ramakrishnan, V., Carter, G., Barton, S., Hosic, A., Florwick, A., Santos, J., Bolden, N. C., McCowin, S., Case, D. A., Johnson, B. A., Salemi, M., Telesnitsky, A. & Summers, M. F. (2015). RNA structure. Structure of the HIV-1 RNA packaging signal. *Science* **348**, 917-21.

26. Mancebo, H. S., Lee, G., Flygare, J., Tomassini, J., Luu, P., Zhu, Y., Peng, J., Blau, C., Hazuda, D., Price, D. & Flores, O. (1997). P-TEFb kinase is required for HIV Tat transcriptional activation in vivo and in vitro. *Genes Dev* **11**, 2633-44.

27. Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H. & Jones, K. A. (1998). A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451-62.

28. Ott, M., Geyer, M. & Zhou, Q. (2011). The control of HIV transcription: keeping RNA polymerase II on track. *Cell Host Microbe* **10**, 426-35.

29. Abbink, T. E. & Berkhout, B. (2008). RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *J Virol* **82**, 3090-8.

30. Mueller, N., van Bel, N., Berkhout, B. & Das, A. T. (2014). HIV-1 splicing at the major splice donor site is restricted by RNA structure. *Virology* **468-470**, 609-20.

31. Harrison, G. P. & Lever, A. M. (1992). The human immunodeficiency virus type 1 packaging signal and major splice donor region have a conserved stable secondary structure. *J Virol* **66**, 4144-53.

32. Baudin, F., Marquet, R., Isel, C., Darlix, J. L., Ehresmann, B. & Ehresmann, C. (1993). Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *J Mol Biol* **229**, 382-97.

33. Clever, J., Sassetti, C. & Parslow, T. G. (1995). RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1. *J Virol* **69**, 2101-9.

34. Clever, J. L., Miranda, D. & Parslow, T. G. (2002). RNA structure and packaging signals in the 5' leader region of the human immunodeficiency virus type 1 genome. *J Virol* **76**, 12381-7.

35. McBride, M. S. & Panganiban, A. T. (1996). The human immunodeficiency virus type 1 encapsidation site is a multipartite RNA element composed of functional hairpin structures. *J Virol* **70**, 2963-73.

36. Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. (2004). RNA interactions in the 5' region of the HIV-1 genome. *J Mol Biol* **336**, 369-79.
37. Paillart, J. C., Dettenhofer, M., Yu, X. F., Ehresmann, C., Ehresmann, B. & Marquet, R. (2004). First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J Biol Chem* **279**, 48397-403.

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

39. Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Swanstrom, R., Burch, C. L. & Weeks, K. M. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-6.

40. Kenyon, J. C., Prestwood, L. J., Le Grice, S. F. & Lever, A. M. (2013). In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader. *Nucleic Acids Res* **41**, e174.

41. Lu, K., Miyazaki, Y. & Summers, M. F. (2010). Isotope labeling strategies for NMR studies of RNA. *J Biomol NMR* **46**, 113-25.

42. Ennifar, E. & Dumas, P. (2006). Polymorphism of bulged-out residues in HIV-1 RNA DIS kissing complex and structure comparison with solution studies. *J Mol Biol* **356**, 771-82.

43. Ennifar, E., Yusupov, M., Walter, P., Marquet, R., Ehresmann, B., Ehresmann, C. & Dumas, P. (1999). The crystal structure of the dimerization initiation site of genomic HIV-1 RNA reveals an extended duplex with two adenine bulges. *Structure* **7**, 1439-49.

44. Amarasinghe, G. K., De Guzman, R. N., Turner, R. B. & Summers, M. F. (2000). NMR structure of stem-loop SL2 of the HIV-1 psi RNA packaging signal reveals a novel A-U-A base-triple platform. *J Mol Biol* **299**, 145-56.

45. Kerwood, D. J., Cavaluzzi, M. J. & Borer, P. N. (2001). Structure of SL4 RNA from the HIV-1 packaging signal. *Biochemistry* **40**, 14518-29.

46. Wüthrich, K. (1986). *NMR of proteins and nucleic acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University, Wiley, New York.

Chapter 2: The U5:DIS Interaction

Introduction

As discussed in Chapter 1, the structure of the HIV-1 5' leader (5'-L) is known to direct a number of important events in the viral life cycle. In 2001 Huthoff and Berkhout proposed that the two alternating structures of the HIV-1 leader RNA may act as a structural switch[1]. These studies were extended in 2003 to propose secondary structures for the full 5'-L with a long-distance interaction (LDI) monomeric structure, stabilized by long distance interactions between residues previously predicted to be part of the polyA and dimerization initiation site (DIS) stemloops, and a branch multiple hairpin (BMH) dimeric structure, in which base pairing between the unique 5' (U5) and *gag* start codon (AUG) regions stabilize the dimeric structure[2]. The proposed LDI and BMH structures are shown in Figure 2.1.

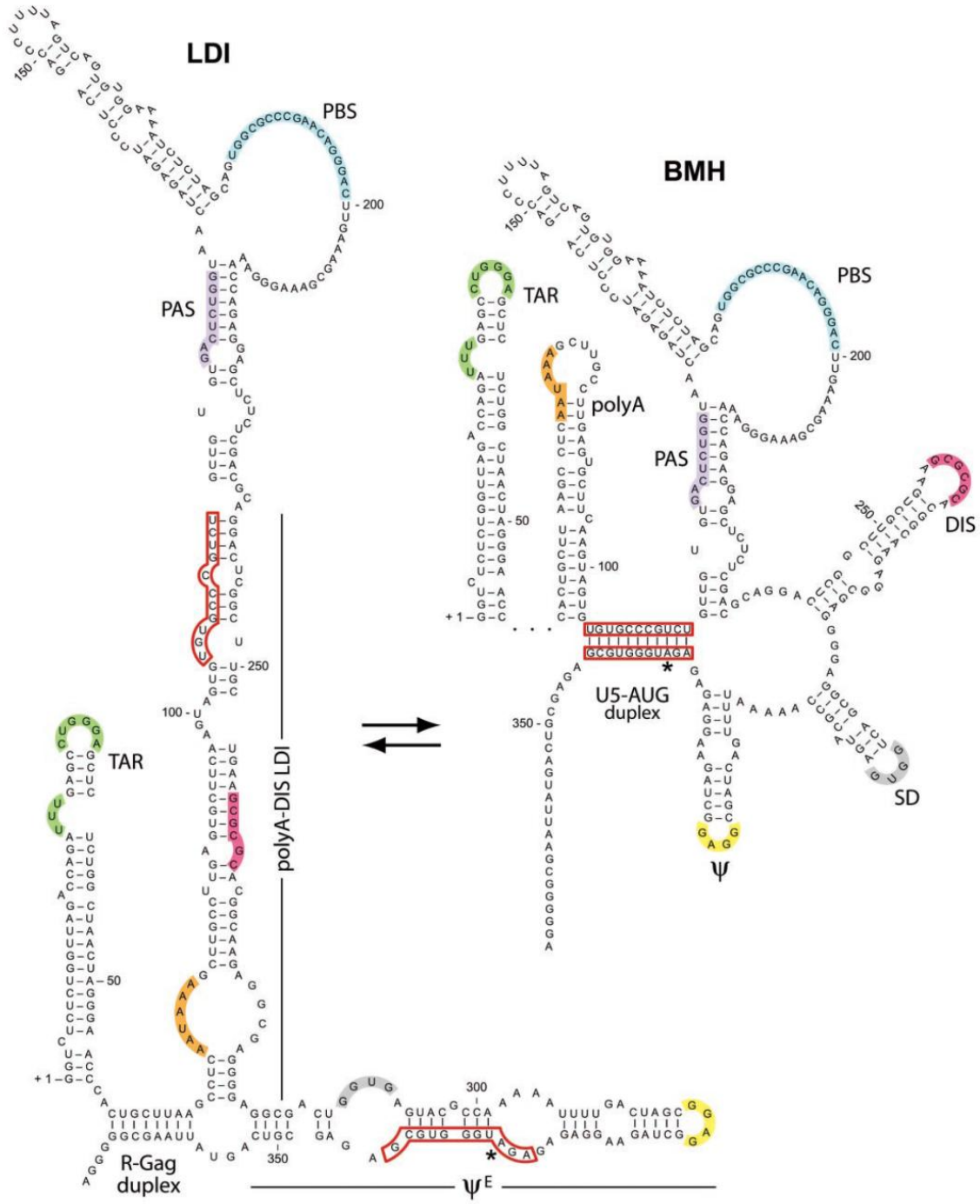

Figure 2.1. The monomeric (LDI) and dimeric (BMH) secondary structures proposed by Abbink and Berkhout[2]. The LDI model lacks the commonly predicted polyA, DIS, and SD hairpins.

While *ex vivo* chemical probing provided evidence for the BMH structure, no evidence was found for the LDI model[3]. Segmental ^{13}C isotopic labeling studies conducted by Dr. Kun Lu revealed that in the monomer the AUG region formed a hairpin, while in the dimer these residues changed conformation[4]. To test whether the AUG base paired with U5 in the dimer, as predicted in the BMH model, a new strategy was developed in the Summers laboratory to detect interactions consistent with formation of an A helix in large RNA. This technique, named long-range probing by Adenosine Interaction Detection (lr-AID)[4], takes advantage of the naturally occurring outlier in which an adenosine (red, underlined) H2 has an unusually upfield chemical shift when located in the sequence 5'-UAA-3' base paired with 5'-UUA-3'. This sequence occurs in the HIV-1 5'-L natively in the TAR hairpin and results in a "signature peak" in the NMR spectrum (Figure 2.2). The adenosine H2 in this base paired sequence is in a unique chemical environment such that it is shifted sufficiently far upfield to be in a rather sparse region of the proton (^{1}H) spectrum between the majority of the protons associated with the nucleotide base (downfield) and those associated with the ribose (upfield). The principle of the lr-AID technique is to incorporate a 5'-UUA-3' sequence and a 5'-UAA-3' sequence in two regions of an RNA that are predicted to base pair. If these two regions interact to form an A helix, the upfield shifted H2 signal will be present.

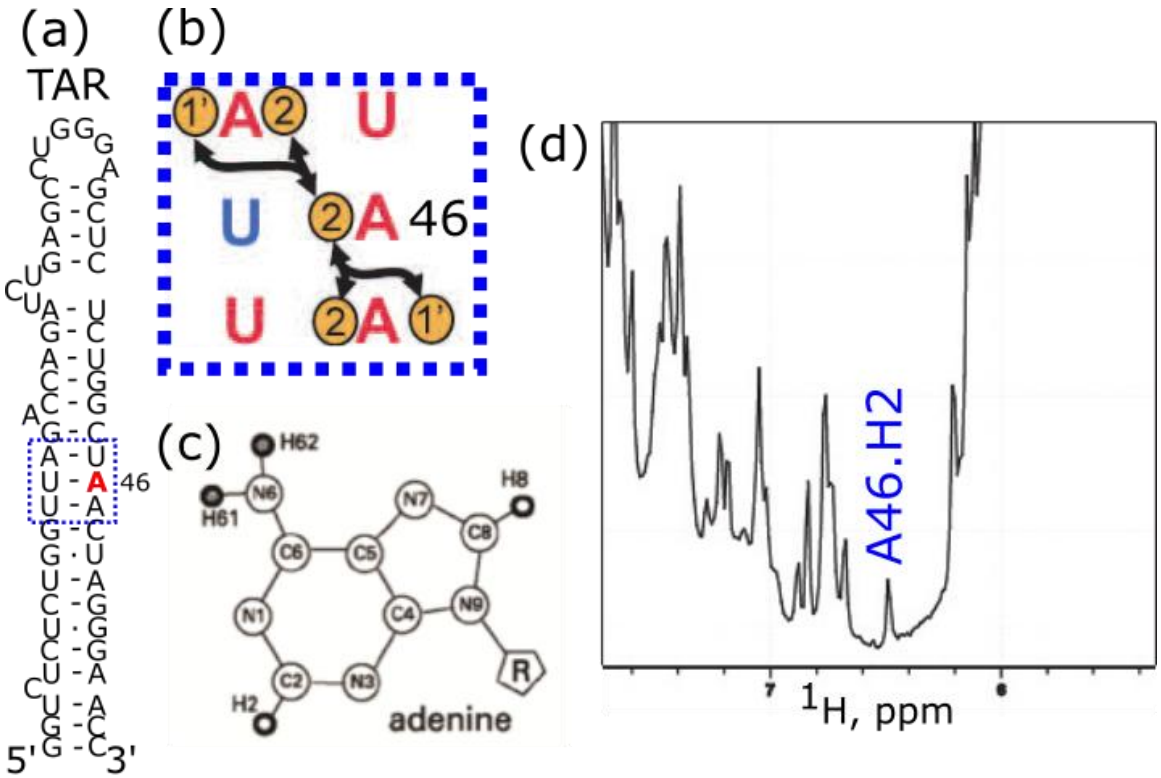

Figure 2.2. The basis of the lr-AID strategy. (a) The native TAR hairpin, the blue square highlights the base-paired 5'-UUA-3'-5'-UAA-3'. A46 (red) is numbered for clarity. (b) The observed A46.H2 NOEs seen for this base pairing interaction. The H2 of A46 sees the H1' and H2 protons of both the following and the cross-strand plus one adenosine. (c) The C2 to which the H2 is bonded and 1' carbon to which the H1' is bonded on the ribose ring are illustrated. (d) The A46.H2 signal is observable in a 1D proton spectrum of the leader RNA. The A46.H2 chemical shift is isolated and easily identified.

Using the lr-AID strategy, Lu *et al.* were able to provide structural evidence for the U5:AUG interaction in the dimeric 5'-L[4]. Experiments to probe the effect of mutations to the AUG region on packaging showed that mutations to the AUG region that interfered with the U5:AUG interaction severely reduced packaging of the viral genomic RNA[5]. The extent of the packaging defect suggested that perhaps AUG competed with the DIS for base pairing to these residues. As such, when the AUG region was mutated to prevent binding to this region, the DIS remained sequestered, preventing dimerization.

This would suggest that the sequestration of the DIS was directly related to the availability of the U5 residues for base pairing. This chapter describes the experiments that identify the base pairing interactions between the U5 and DIS regions that stabilize the monomeric conformation of the HIV-1 5'-L.

Results

Location of DIS Sequestration by U5

As shown in Figure 2.1, the U5:AUG interaction predicted in the BMH model extends from residue 105 through residue 115 in the U5 region. To determine which, if any, of these residues were important for sequestration of the DIS, we introduced point mutations in this region, one at a time and in combination. To eliminate complications resulting from alterations of the U5:AUG interaction, these mutations were performed in a T-SL3 construct in which only extends through residue 327 (Figure 2.3).

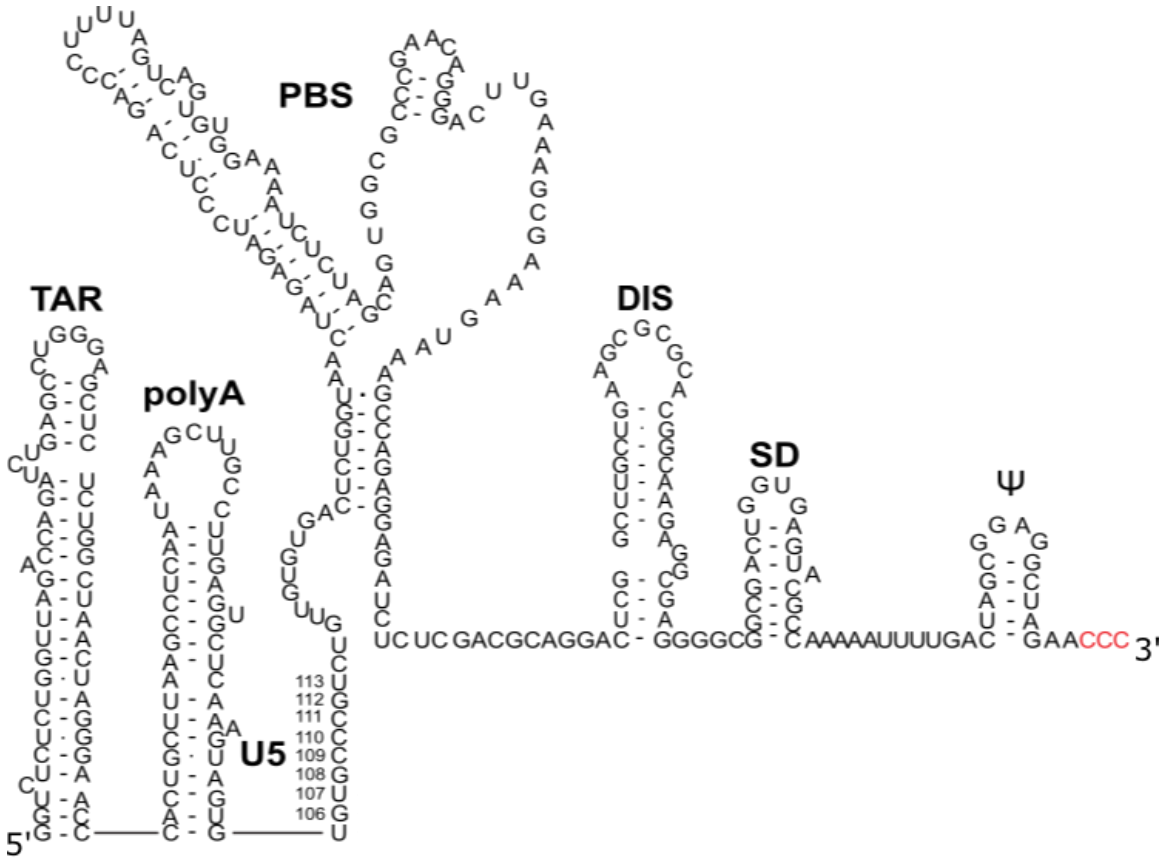

Figure 2.3. T-SL3 construct used to assess the monomeric stability of the HIV-1 5'-L. The leader is truncated at residue 327, but three non-native cytosines were added to the 3' end (red) to allow for a SmaI restriction enzyme recognition sequence for template linearization.

The T-SL3 (ΔAUG) construct was used so that stabilization of the monomeric conformation would be a direct result of stabilizing the monomeric base pairing rather than an indirect effect of destabilizing the U5:AUG interaction. The following mutations were introduced: U113G, U113C, G112A, C111G, C111A, G108U/C109A, G108U, G108A, U107C, and U107C/U105C (Figure 2.4).

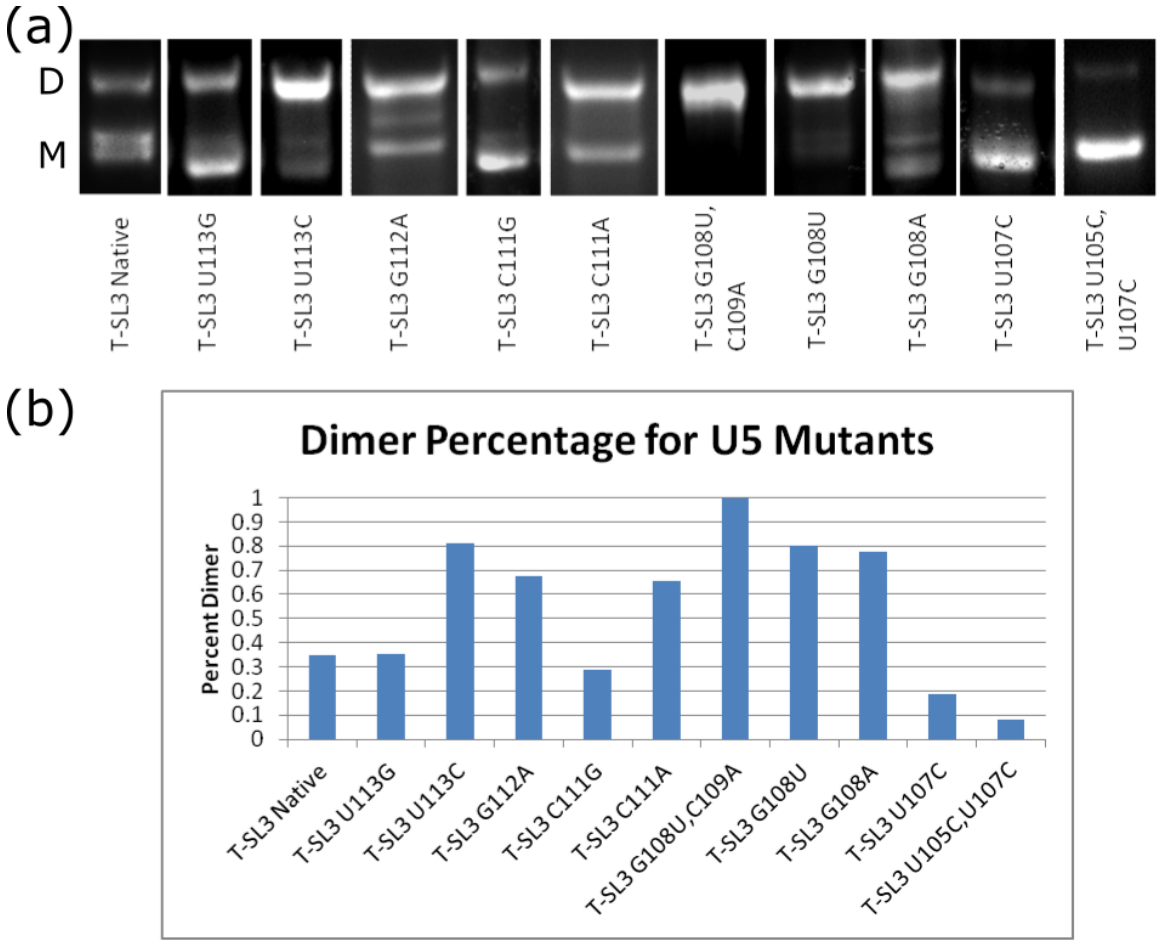

Figure 2.4. (a) Native agarose Tris-Borate gel electrophoresis allows comparison of the monomer-dimer equilibria of T-SL3 native to mutants. Numbering for the mutations corresponds to the numbering shown on the T-SL3 construct in Figure 2.3. The upper band (D) is the dimer band, and the lower band (M) is the monomer band. Many of the point mutations greatly affected the monomer-dimer equilibrium, but U107C and U105C/U107C showed the most significant stabilization of the monomeric conformation. (b) Quantification of gel from (a) showing the percentage of dimer for each construct.

Two mutations, U107C and U105C/U107C, were identified as stabilizing the monomeric conformation (Figure 2.4). Inspection of the sequence of U5 in this area revealed complementarity between residues 105-109 (5'-UGUGC-3') in U5 and residues 257-261 (5'-GCGCG-3') in the DIS loop (Figure 2.2.5), suggesting that this could be the

region of U5 responsible for sequestration of the DIS. The U5:DIS complementarity is enhanced by the U107C and U105C mutations (Figure 2.5), which would explain why these mutations enhanced the stability of the monomer.

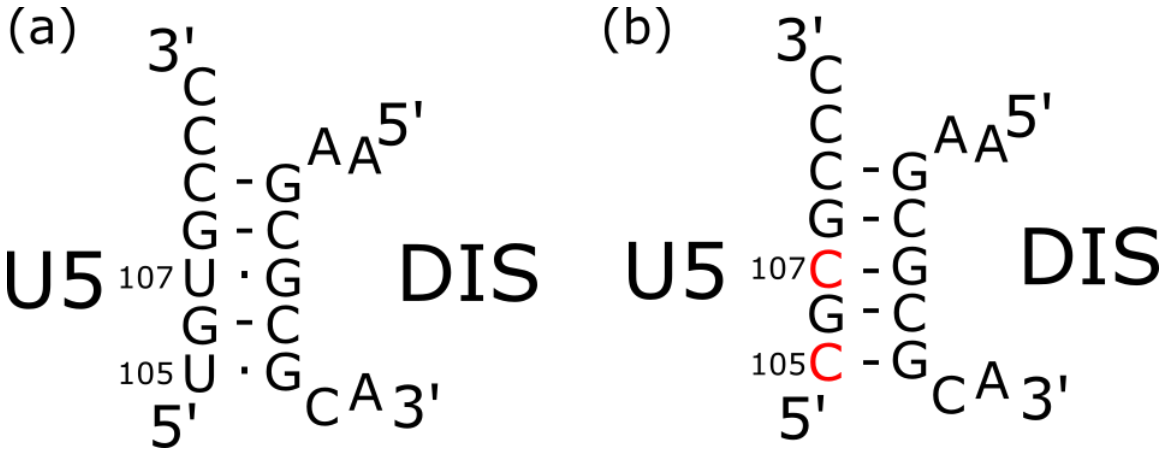

Figure 2.5. Inspection of the sequence of U5 the DIS loop revealed sequence complementarity. Furthermore, the U105C and U107C mutations enhanced this complementarity.

These findings led to the hypothesis that in the monomeric conformation the DIS loop is sequestered in base pairing with residues 105 - 109 in the U5 region and therefore unavailable for dimerization. The increased monomeric stability caused by the U105C and U107C mutations was maintained in a T-SL4 construct which contains the AUG hairpin (Figure 2.6).

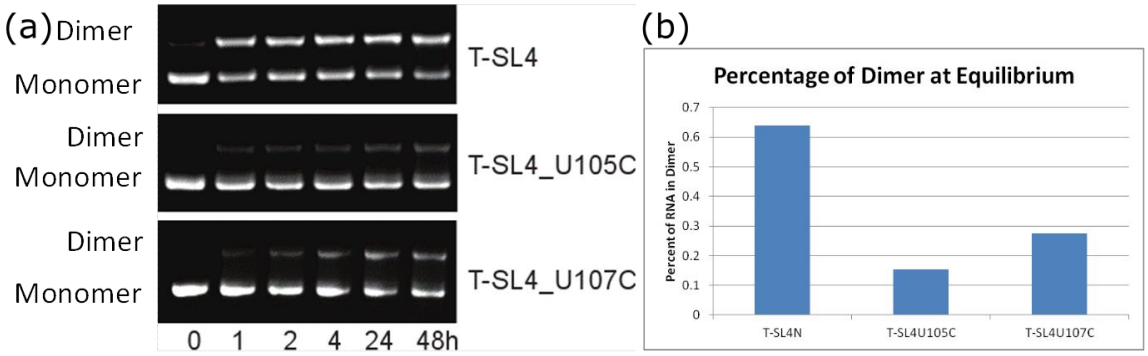

Figure 2.6. (a) The U105C and U107C single point mutations maintain their effect of increasing monomeric stability in the context of the full leader (T-SL4). T-SL4 exists at a monomer-dimer equilibrium favoring the dimer. With either the U105C or the U107C mutation, the equilibrium is shifted to favor the monomeric conformation. (b) Quantification of the 48 hour band for each construct from (a), showing the relative percentage of RNA in the dimeric conformation at equilibrium. Figure 2.6(a) courtesy of Dr. Xiao Heng.

Evidence for the U5:DIS Interaction by lr-AID

To confirm the U5:DIS interaction, the new lr-AID strategy was again employed. This technique made it possible to probe for the U5:DIS interaction in the context of the monomeric leader. It was unclear, however, whether the DIS loop base pairing with the single-stranded U5 region would create a normal A helix to provide the upfield shifted lr-AID peak. To confirm that base pairing between a single-stranded region and the DIS loop could occur and would give rise to the lr-AID signal with the sequences incorporated, small oligo controls were designed. Various sequences representing U5 and the DIS stemloop were ordered from Dharmacon and tested for binding (Figure 2.7).

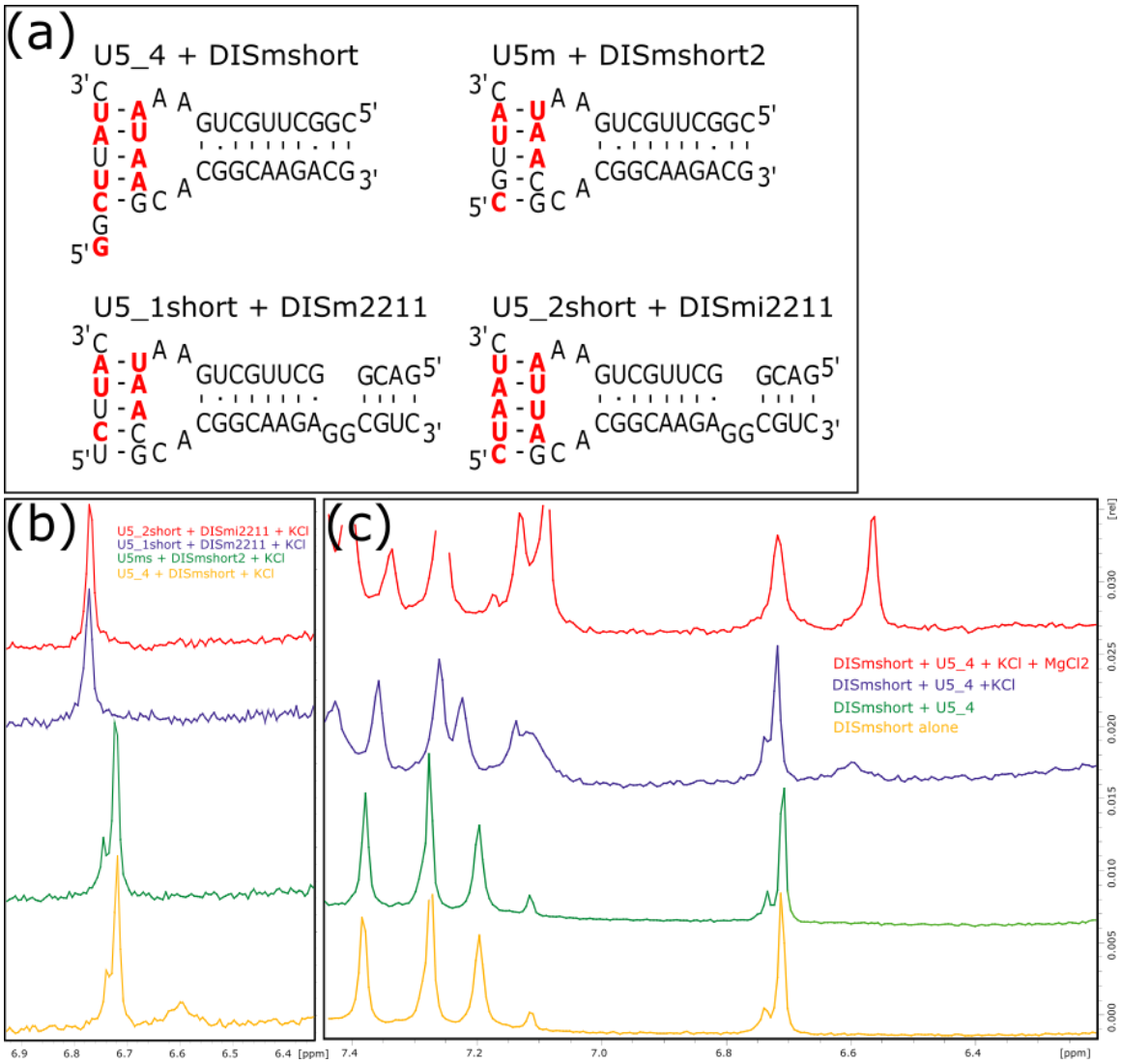

Figure 2.7. Control oligos suggest that lr-AID can be used to test the U5:DIS interaction. (a) The short oligo controls that were used to represent that single-stranded region of U5 and the DIS stemloop. (b) The short oligo control 1D proton spectra in 10 mM Tris, pH 7 and 140 mM KCl. Only U5_4 + DISmshort gave evidence of a peak - the broad hump seen at ~6.6 ppm. (c) lr-AID peak indicative of A helical formation by U5_4 and DISmshort is sharpest in the presence of potassium and magnesium.

This pairing gave rise to the lr-AID chemical shift (Figure 2.7c, red spectrum). Because similar sequences did not successfully give the lr-AID chemical shift, the U5_4 and DISmshort sequences were used to probe for the U5:DIS interaction in the context of the 5'-L. A T-SL3 construct was used to prevent competition for U5 between the DIS

and AUG regions. A naturally occurring mutant in 7% of reported viruses eliminates the lr-AID signal from the native TAR sequence with an A46G mutation. The T-SL3 construct contained an A46G mutant to eliminate the TAR A46.H2 signal and also incorporated the following mutations, as dictated by the control experiment (Fig. 2.7); U103G, U105C, G106U, G108A, and C109U. The construct also included the following mutations in the DIS loop; G257A, C258U, G259A, and C260A. Using this construct, shown in Figure 2.8, it was possible to directly probe for the U5:DIS interaction.

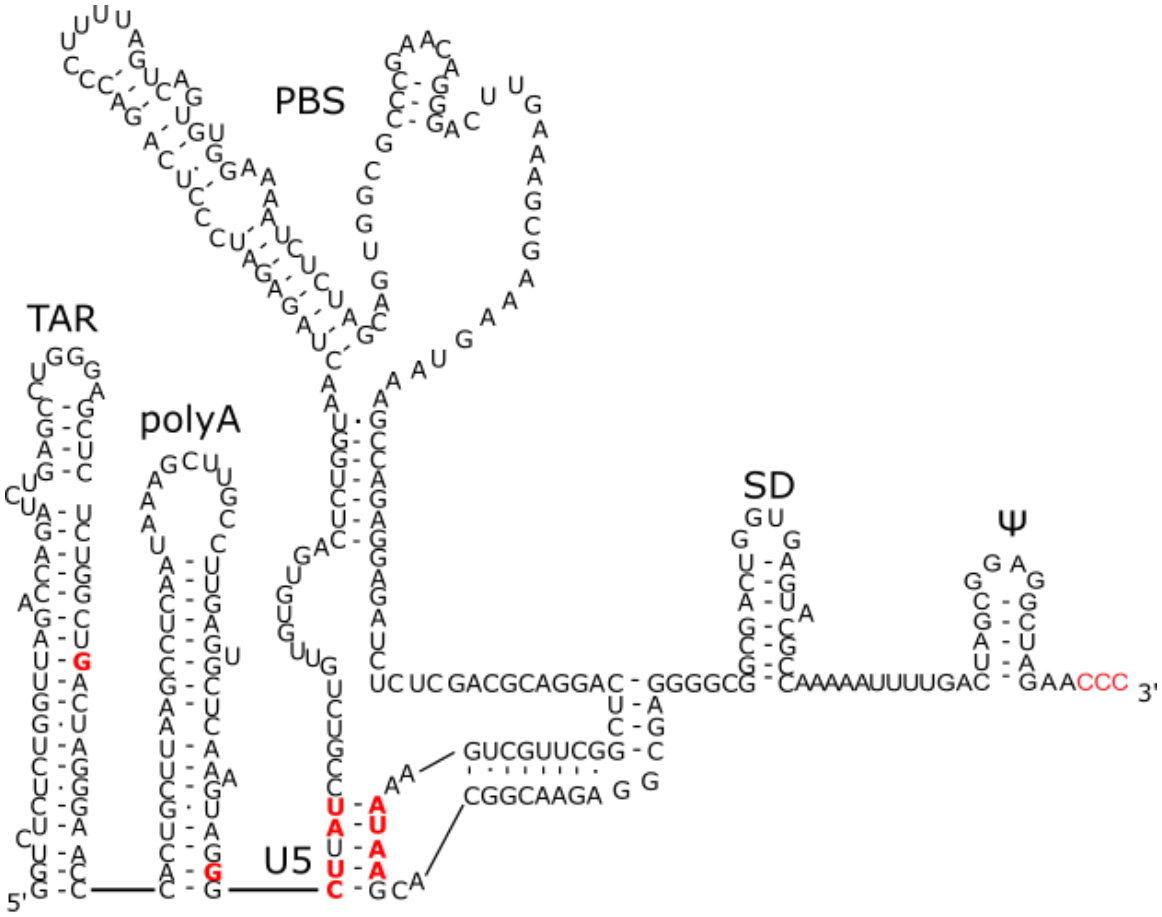

Figure 2.8. The construct used to directly probe for the U5:DIS interaction in the monomer. Mutations are shown in red. The A46G mutation in TAR eliminates the TAR A46.H2 signal to prevent overlap. The mutations to the U5 and DIS were made such that if these two regions form an A helix as predicted an upfield shifted lr-AID signal will be present in the NMR spectrum.

Proton-proton NOESY spectrum of this construct revealed the signature upfield shifted lr-AID peak as well as NOEs from cross-strand protons consistent with formation of an A helix (Figure 2.9).

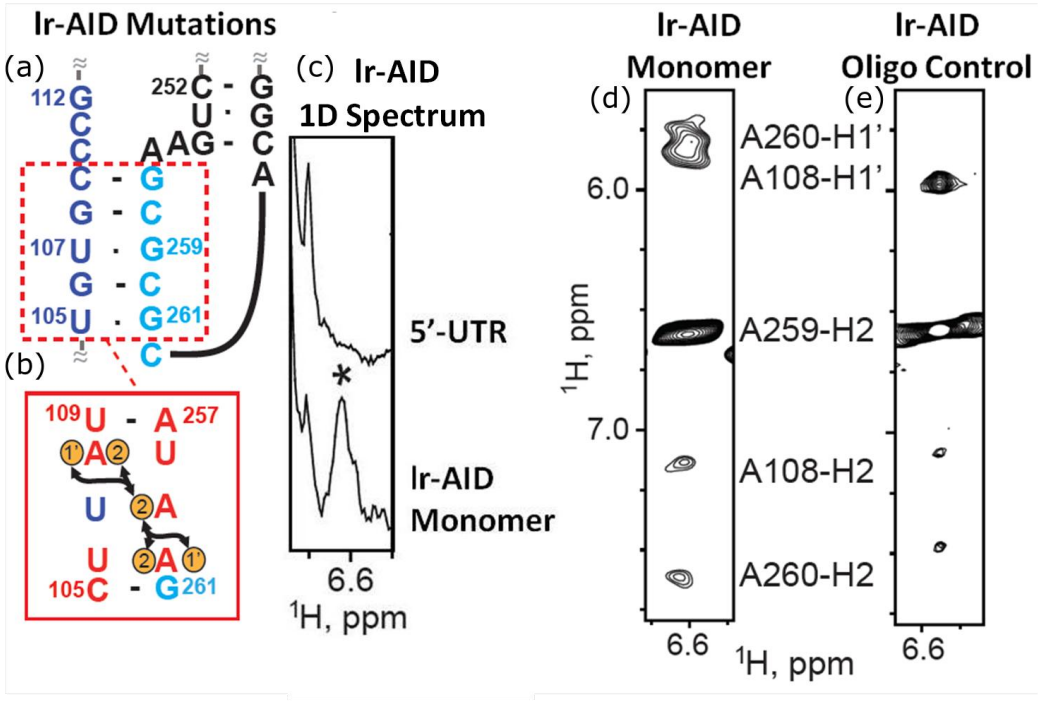

Figure 2.9. The monomeric conformation is stabilized by a U5:DIS interaction validated by lr-AID[4]. (a) The native sequence of U5 and DIS predicted to base pair. (b) The lr-AID mutations incorporated into the U5 and DIS regions as well as the NOEs expected if base pairing occurs. (c) The presence of the upfield shifted lr-AID peak in the mutated construct. (d) A strip of the 2D NOESY spectrum from the mutated leader construct. (e) Comparison to a strip of the 2D NOESY spectrum of the short oligo control. The chemical shift for the lr-AID H2 is similar to the control and a similar pattern of NOEs is detected.

This provided the first direct evidence for the U5:DIS interaction which stabilizes the monomeric conformation of the 5'-L by sequestering the DIS by base pairing[4]. One concern is that the lr-AID mutations could cause misfolding of the RNA. Isothermal titration calorimetry (ITC) studies were performed on the native T-SL3 construct, on the

T-SL3 U105C, T-SL3 U107C, and the T-SL3 construct incorporating the lr-AID mutations (Figure 2.10). The ITC studies revealed that the mutated constructs bound nucleocapsid (NC) protein in a manner indistinguishable from the native T-SL3 construct. This suggests that the overall fold of the RNA is the same as it exposes the same profile of NC binding sites. Additionally, *in vitro* translation assays conducted by our collaborator Dr. Ioana Boeras in Dr. Boris-Lawrie's lab indicate that the lr-AID mutations incorporated into the 5'-L of HIV-1 enhance translation, consistent with stabilization of the monomeric conformation (manuscript in preparation).

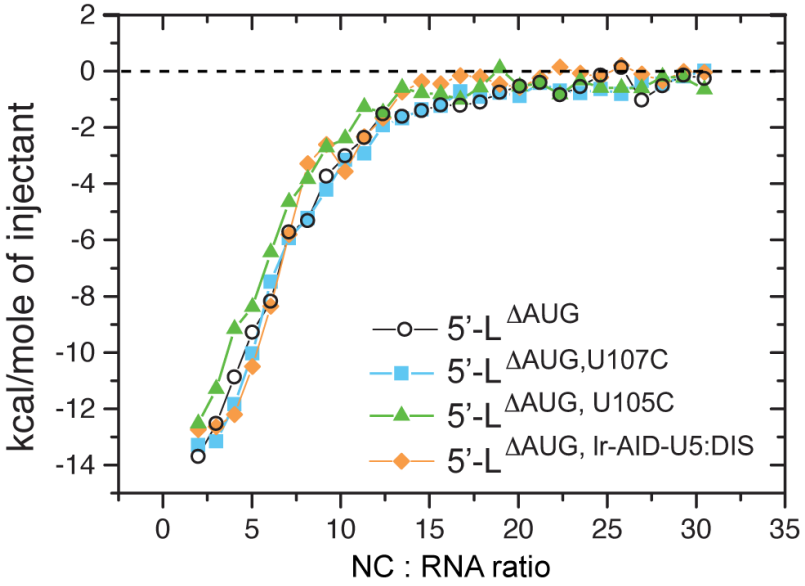

Figure 2.10. ITC studies of the native T-SL3 construct (5'-$L^{\Delta AUG}$) and mutations reveal an indistinguishable nucleocapsid binding profile which suggests that the RNA is adopting its native conformation even in the presence of the mutations. These experiments were conducted by Dr. Xiao Heng.

Placement of the U5:DIS Interaction

An area of interest was the exact placement of the U5:DIS interaction. While the interaction could exist as shown in Figure 2.11(a), it is also possible to shift the interaction down two base pairs such that the U5:DIS interaction occurs from G102-U107

("Monomer Down") rather than from U105-C109 ("Monomer Up") as previously tested (Figure 2.11). To test whether lr-AID could be used to distinguish between these possibilities, "Monomer Down" lr-AID mutations were made in a full-length leader construct, T-SL4Blunt. The native 5'-L in the "Monomer Up" conformation and Monomer Down construct are shown in Figure 2.11 (a) and (b) respectively.

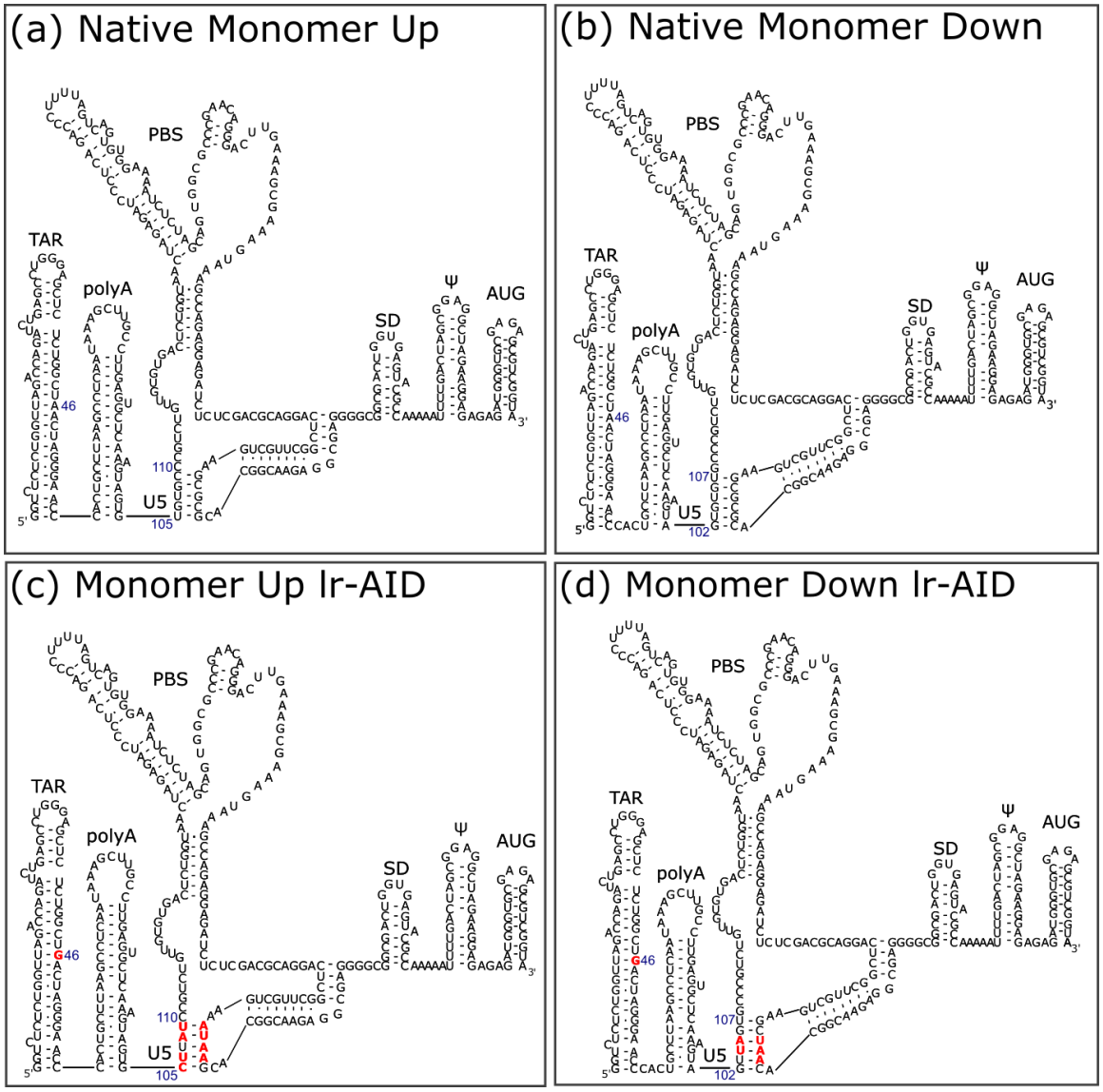

Figure 2.11. Monomer Up and Monomer Down constructs. (a) The native 5'-L shown in the Monomer Up conformation. (b) The native 5'-L shown in the Monomer Down conformation. In this conformation

nucleotides predicted to be at the base of PolyA are opened up to allow the U5:DIS interaction to occur two base pairs lower. (c) The lr-AID mutations made to the U5 and DIS regions for the Monomer Up lr-AID construct. (d) The lr-AID mutations made to the U5 and DIS regions used to test for the Monomer Down conformation.

The lr-AID region of the Monomer Down construct spectrum is compared to the native TAR oligo spectrum in Figure 2.12.

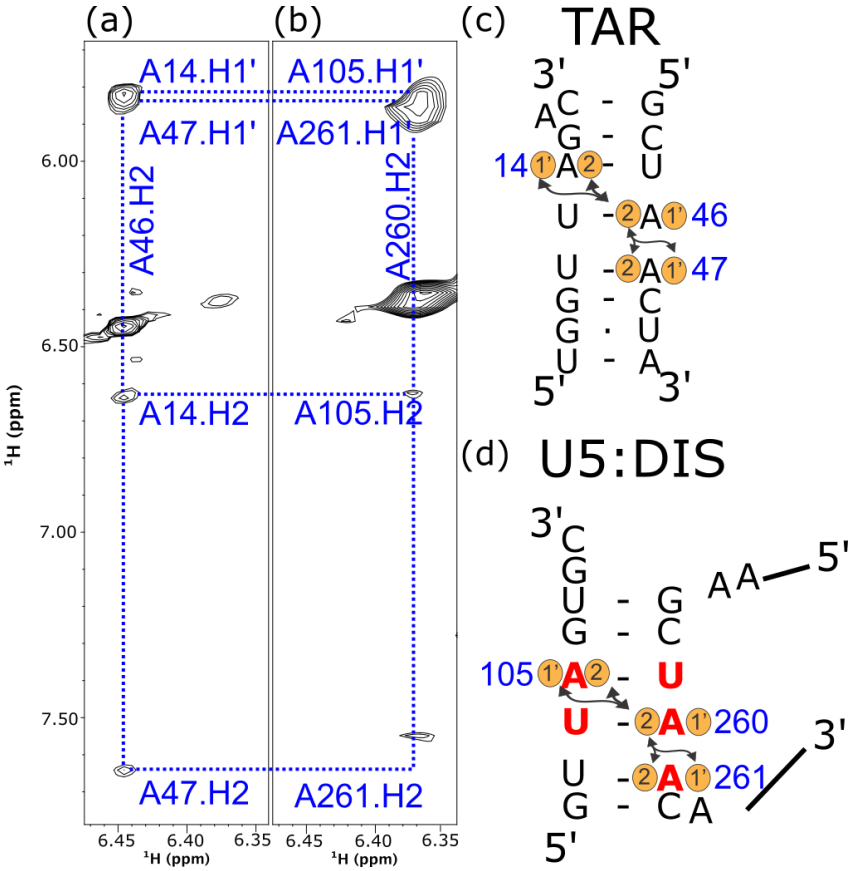

Figure 2.12. Regions of 2D ^{1}H-^{1}H NOESY spectra corresponding to Monomer Down lr-AID signal. (a) The strip of the native TAR oligo NOESY spectrum showing the A14.H2 connectivities. This serves as the control for the Monomer Down lr-AID signal which has the same local base pairing. (b) The lr-AID signal from the Monomer Down spectrum, which matches the NOE pattern from the native TAR hairpin. (c) The NOE connectivities observed in the TAR spectrum (a) are shown on the TAR secondary structure. (d) The

NOE connectivities observed for the Monomer Down lr-AID signal (b) are shown on the relevant portion of the Monomer Down secondary structure.

Both the Monomer Up and Monomer Down conformations show the lr-AID NOEs consistent with the U5:DIS interaction. This indicates the lr-AID cannot distinguish between these two conformations. It is possible that both are formed.

Conclusions

The evidence presented suggests an alternative to the LDI model of the monomer in which the monomeric conformation of the HIV-1 5'-L is stabilized by interactions between the U5 and DIS regions. NMR data are consistent with a model in which the loop of the DIS stemloop forms a short A helix with residues from the U5 region. These findings, in combination with other studies, allowed our lab to propose a structural switch mechanism in which the monomeric conformation is stabilized by sequestration of the DIS by base pairing with residues in U5, while the Gag start codon is available for translation at the base of a semi-stable AUG stemloop. Dimerization occurs as the result of a conformational rearrangement in which the AUG stemloop opens up and base pairs with the U5 region to displace the DIS, freeing the DIS to participate in dimerization, sequestering the Gag start codon to inhibit translation, and exposing more high affinity nucleocapsid binding sites to allow selective packaging of the dimeric genome. The biological relevance of the mutations used for these studies has been supported by *in vitro* translation assays. In 2013 in-gel SHAPE studies specifically testing the reactivity of the monomeric species were found to be more consistent with the U5:DIS model of the monomer than the LDI model of the monomer[6]. It is likely that previous studies which relied heavily on Mfold free energy predictions were unable to predict this model because

Mfold lacks the ability to predict pseudoknot structures[7]. While RNA folding prediction algorithms are very valuable, currently the scope of such programs is limited by the restricted structural information regarding large RNA structural motifs. This emphasizes the importance of developing new strategies, such as the lr-AID strategy described in this work, to allow structural characterization of large RNAs.

Methods

RNA preparation: A pUC19 plasmid containing the a T7 promoter followed by the T-SL3 construct (ΔAUG) was mutated to contain the mutations visualized in Figure 2.3 using the QuikChange site-directed mutagenesis kit (Agilent Technologies). The same original T-SL3 plasmid was mutated using the QuikChange multi site-directed mutagenesis kit (Agilent Technologies) for the lr-AID studies. The plasmids were sequenced (Genewiz) to confirm the mutations. The plasmids were transformed into DH5α cells (invitrogen) according to manufacturer's directions. To create DNA templates for RNA transcription 1 L cultures of the DH5α cells containing the plasmid were grown overnight in Luria broth at 37 ˚C at 250 rpm. Plasmids were purified using the QIAGEN Megaprep kit according to manufacturer's directions. Purified plasmids were digested by SmaI (T-SL3) or BstZ17I (Monomer Down) (enzymes from NEB) overnight, according to manufacturer's protocol. Digestion was checked on a 2% TAE gel. After digestion, DNA template was purified by PCI extraction followed by ethanol precipitation. After purification the DNA template was washed by serial dilution and concentration three times with four milliliters of ultrapure water on a 30k Amicon Ultra

centrifugal filter device. *In vitro* transcription using in-house purified T7 RNA polymerase was used to synthesize the RNA in a reaction mixture containing ~50 ng/µL purified DNA template, 2mM spermidine, 80mM Tris-HCl (pH 8), 2mM DTT, 10-20 mM $MgCl_2$, and 3-6 mM NTPs. 30 µL trial reactions were used to optimize $MgCl_2$ and NTP ratios for each construct. Transcription reactions were incubated for 2.5 hrs at 37 ˚C. Transcription reactions were halted with the addition of 1 mmole of EDTA followed by boiling for three minutes and snap cooling on ice for three minutes. RNA was purified by gel electrphoresis on a 6% denaturing acrylamide gel (SequaGel, National Diagnostics) using the FisherBiotech DNA sequencing system at 20W overnight to achieve the best resolution. The gel bands were visualized by UV-shadowing, excised, and eluted using the Elutrap ® electroelution system (Whatman) at 100 V overnight. The eluted RNA was collected, then washed two times with 4 mL of 2M high purity NaCl (99.999%, Acros), followed by 8 times with 4 mL of ultrapure water on a 30k Amicon Ultra centrifugal filter device.

Native gel electrophoresis studies: RNAs were prepared for native gel electrophoresis from purified stocks in water by diluting the RNA to 1.21 times the desired concentration (1 µM) in RNA water followed by addition of 10% of the final volume of a 100 mM Tris buffer, pH 7.0. 90% of this volume was transferred to a fresh RNase-free 1.5 mL centrifuge tube (Eppendorf), and the remaining volume was used to verify the concentration. Sufficient 10x phsyiological ionic (PI) salts were added to this volume to give 1 µM RNA and 1x PI salts (140 mM KCl, 10 mM NaCl, and 1 mM $MgCl_2$). The RNA in the physiological buffer was incubated at 37 ˚C overnight to allow equilibration. To a 10 µL aliquot of sample 1 µL of all-purpose, native agarose gel loading solution

(life technologies) was added. 10 µL of sample was then loaded into the well of a 1% native agarose TB gel. TB gels were prepared by boiling 0.5 g of high gelling temperature agarose (Fisher) dissolved in 50 mL of 44 mM Tris Base - Boric Acid (Fisher) buffer, pre-stained with 0.5 µg ethidium bromide per milliliter of gel. The gels and 44mM Tris Base- Boric Acid buffer were chilled for one hour before use. The gels were run at 115V for approximately 1 hour and 15 minutes. Gels were visualized and photographed using a Kodak Gel Logic 200 Imaging System.

NMR studies: NMR data were collected with a Bruker Avance 800 MHz spectrometer, processed with NMRPIPE/NMRDraw[8] and analyzed with NMRView[9]. NMR signals were assigned by standard methods[10]. RNA concentrations for 2D NOESY spectra were 125 µM. The 2D NOESY data forT-SL3 lr-AID NMR sample used to originally identify the U5:DIS interaction was collected in a buffer of 10 mM Tris, pH 7.5, 140 mM KCl, and 1 mM $MgCl_2$. The later NMR studies of the Monomer Down conformation were collected in a 10 mM Tris buffer, pH 7.5, with no added salts.

References

1. Huthoff, H. & Berkhout, B. (2001). Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143-57.

2. Abbink, T. E. & Berkhout, B. (2003). A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J Biol Chem* **278**, 11601-11.

3. Paillart, J. C., Dettenhofer, M., Yu, X. F., Ehresmann, C., Ehresmann, B. & Marquet, R. (2004). First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J Biol Chem* **279**, 48397-403.

4. Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A. & Summers, M. F. (2011). NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-5.

5. Nikolaitchik, O., Rhodes, T. D., Ott, D. & Hu, W. S. (2006). Effects of mutations in the human immunodeficiency virus type 1 Gag gene on RNA packaging and recombination. *J Virol* **80**, 4691-7.

6. Kenyon, J. C., Prestwood, L. J., Le Grice, S. F. & Lever, A. M. (2013). In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader. *Nucleic Acids Res* **41**, e174.

7. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-15.

8. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-93.

9. Johnson, B. A. & Blevins, R. A. (1994). NMR View: A computer program for the visualization and analysis of NMR data. *J Biomol NMR* **4**, 603-14.

10. Wüthrich, K. (1986). *NMR of proteins and nucleic acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University, Wiley, New York.

Chapter 3: Probing the Structure of the Monomeric HIV-1 5'- Leader RNA

Introduction

The findings discussed in Chapter 2 and published by Lu *et al.* confirmed the existence of an interaction between residues of the unique 5' region (U5) and residues of the dimer initiation site (DIS) which stabilizes the monomeric conformation of the HIV-1 5' leader (5'-L)[1]. The mutations incorporated to allow for the long-range probing by Adenosine Interaction Detection (lr-AID) of the U5:DIS interaction include mutations to the DIS which disable dimerization. These mutations provided a monomeric construct of the 5'-L which maintains its monomeric conformation at the concentrations of RNA necessary for structural studies by nuclear magnetic resonance (NMR) spectroscopy. A continuous challenge in structural studies of the monomeric HIV-1 5'-L has been the fact that even in the absence of added salts, the high concentrations required for structural NMR studies (125 µM RNA) cause the 5'-L to dimerize (Figure 3.1). Incorporation of the lr-AID mutations discussed in Chapter 2 disables dimerization at the DIS and provides a stable monomeric construct even at 125 µM RNA (Figure 3.1). These constructs provided the foundation for the first NMR studies mapping the secondary structure of the monomeric 5’-L.

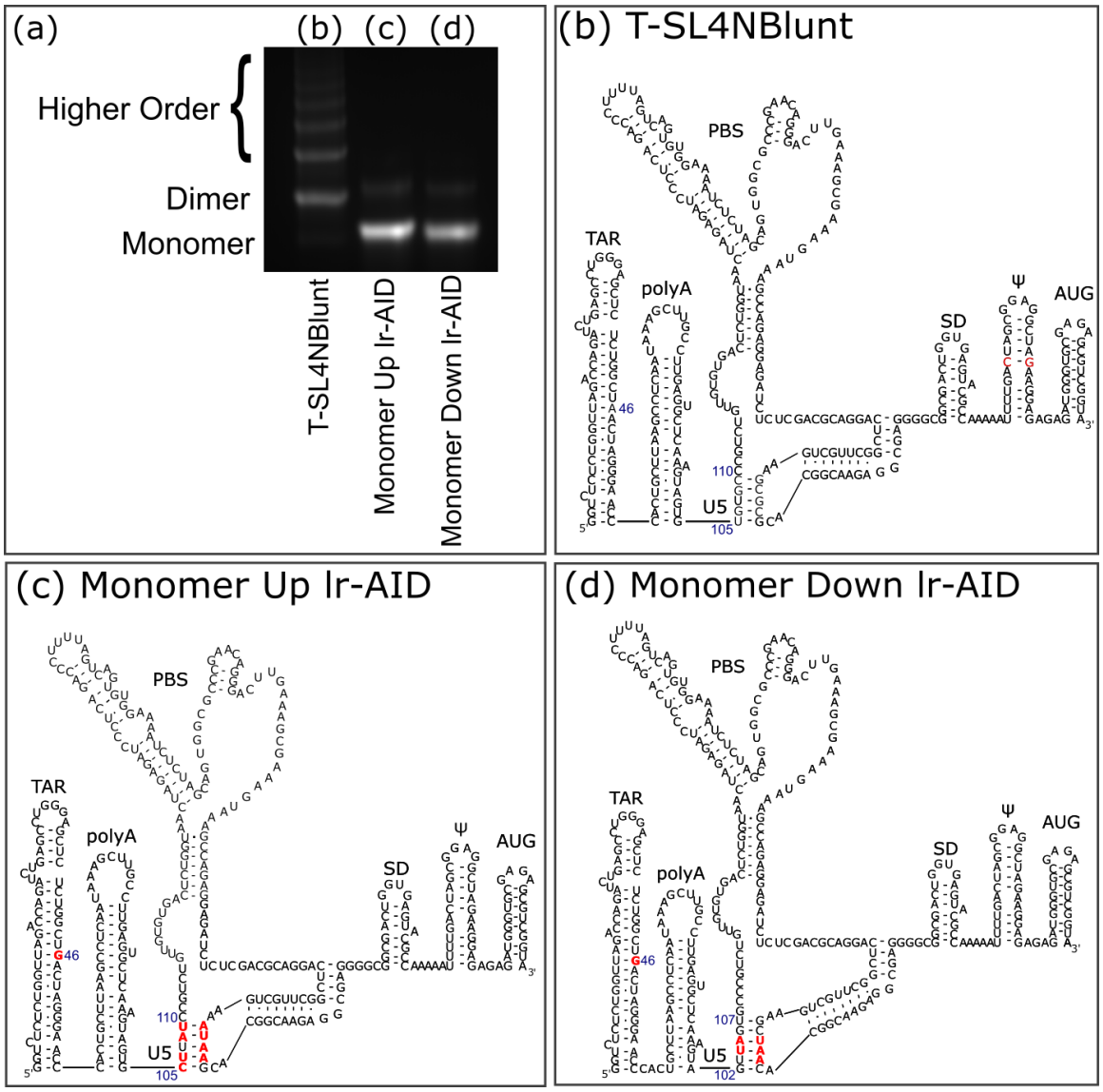

Figure 3.1. lr-AID mutations allow NMR studies of the 5'-L in a monomeric conformation. (a) The native 5'-L (T-SL4NBlunt), monomeric secondary structure (shown in (b)) exists primarily as a dimer or in a higher order complex in NMR conditions (125µM RNA, 10mM Tris, pH 7.5). However, the 5'-L incorporating the lr-AID mutations, either Monomer Up lr-AID (shown in (c), 91% monomeric) or Monomer Down lr-AID (shown in (d), 91% monomeric), is almost entirely monomeric under NMR conditions, as monitored by native agarose TB gel electrophoresis.

A number of secondary structural elements have been predicted in the 5'-L[1; 2; 3; 4; 5; 6; 7; 8; 9], many have known biological function. Figure 3.2, shows a consensus

secondary structure of the 5'-L with many of the most commonly predicted secondary structural elements identified. The TAR hairpin is required for transcriptional activation of the genome through interactions with Tat and P-TEFb, described in the introduction[2; 10]. The polyA region contains the "AAUAAA" polyadenylation signal which is repeated at the 3' end of the genome to signal for the addition of the polyA tail. The PBS region contains the primer binding site for the lysine tRNA which is required to initiate reverse transcription[5]. The DIS hairpin contains the dimerization initiation sequence, a palindromic sequence which initiates dimerization of the genome through base pairing interactions with the DIS of another 5'-L[6]. Dimerization is a requisite step in genome packaging[11]. The SD hairpin contains the major splice-donor site which is required for splicing of the genome[7]. The ψ hairpin contains a high affinity nucleocapsid (NC) binding site and was originally thought to be sufficient to direct packaging[9; 12]. The AUG hairpin contains the start codon for the Gag polyprotein.

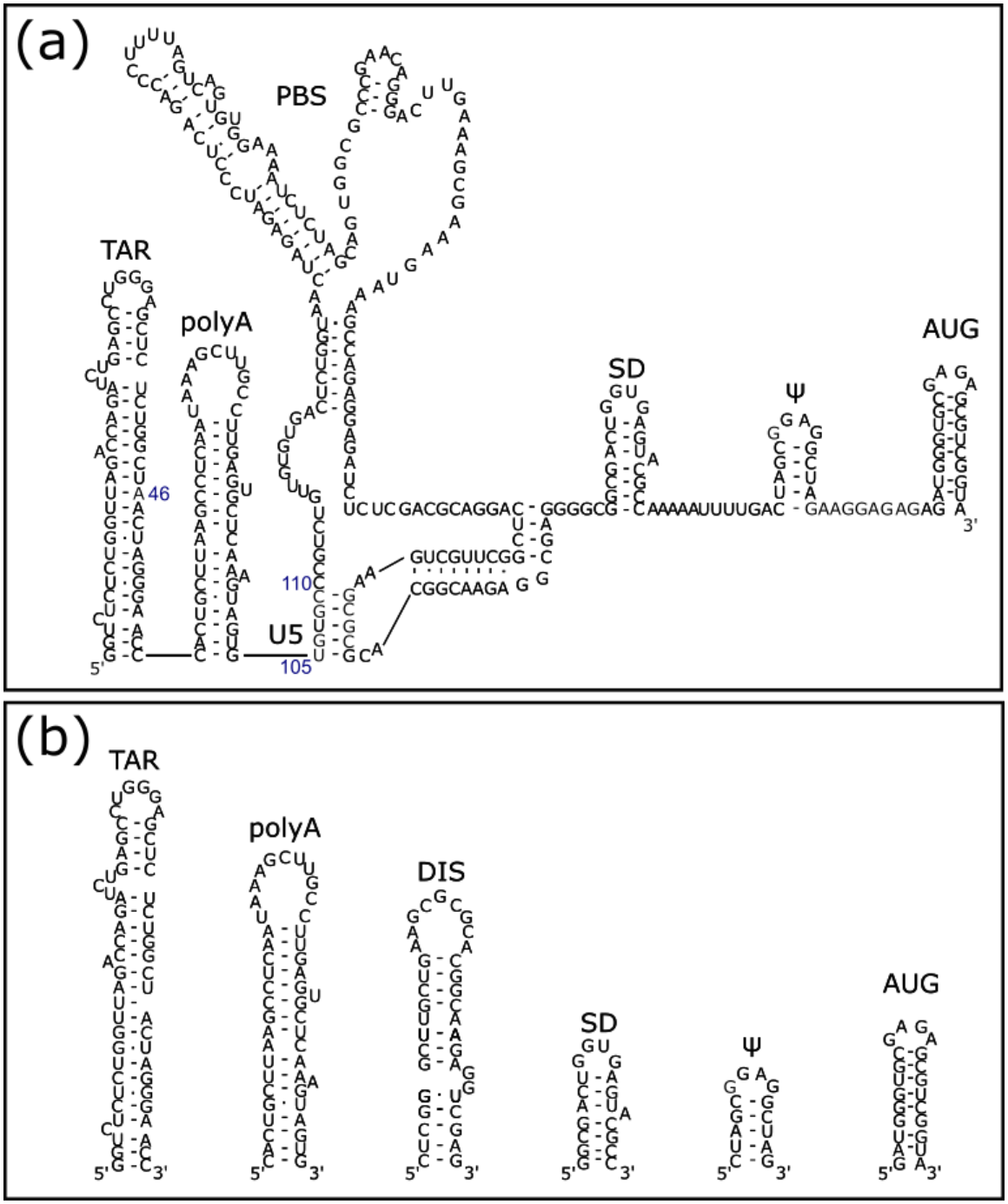

Figure 3.2. (a) Predicted secondary structure elements of the 5'-L. (b) The Summers group has transcribed, purified, and made assignments for the following isolated hairpins: TAR, polyA, DIS, SD, and ψ. These assignments make it possible to look for signals consistent with formation of these secondary structure elements in the 5'-L.

Because the Summers' group has assigned spectra for a number of the hairpins predicted to form in the 5'-L (TAR, polyA, DIS, SD, ψ, and AUG, shown in Figure 3.2)[8; 9; 13], it is possible to compare easily assigned outlying chemical shifts in the 2D ^{1}H-^{1}H NOESY spectrum of the monomeric 5'-L to the 2D NOESY spectra of the isolated stemloops to identify the existence of certain secondary structural elements in the monomeric leader. If there are matching patterns of NOEs in the spectrum of an isolated hairpin and that of the monomeric leader, it suggests that the RNA in the leader is adopting the same fold as the isolated RNA. Incorporating the U5:DIS mutations identified in Lu *et al.*[1] in the context of the full-length 5'-L (T-SL4Blunt) allowed inspection of the chemical shifts of the monomeric 5'-L in an A2rGr labeled sample. In A2rGr samples the only protons are on the C2 position of the adenine base and on the ribose ring of the adenosines and the ribose ring of the guanosines. All other protons have been substituted with deuterium to reduce spectral complexity, as discussed in Chapter 1. In such a large RNA there is a great deal of chemical shift overlap even in highly deuterated samples, however, many chemical shifts could be assigned based on comparisons to previously assigned hairpin spectra.

Results

TAR Hairpin

Figure 3.3 shows a comparison between strips of the isolated TAR hairpin NOESY spectrum (Figure 3 a,e), the Monomer Up lr-AID NOESY spectrum (Figure 3.3 b,f), and the Monomer Down lr-AID NOESY spectrum (Figure 3.3 c,g) corresponding to the chemical shifts of A26.H2 and A34.H2 in the isolated TAR spectrum.

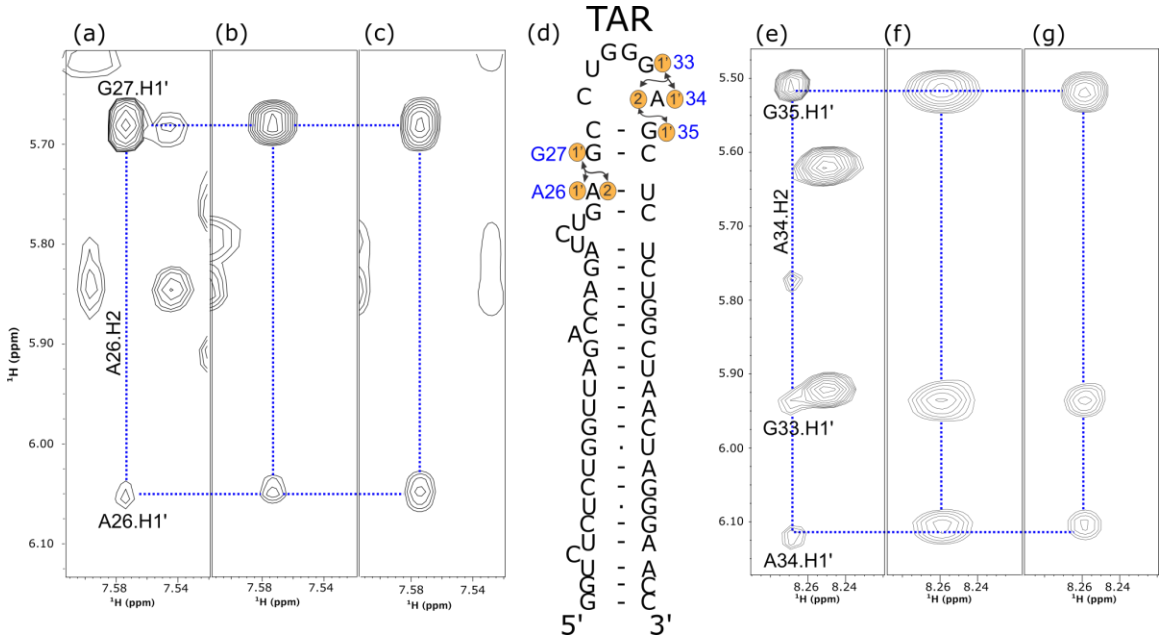

Figure 3.3. NOEs consistent with the top of the TAR stemloop are seen in the monomeric 5'-L. (a) NOEs assigned in the isolated TAR hairpin from A26.H2 to G27.H1' and A26.H1'. (b) The same pattern of NOEs is present in the Monomer Up lr-AID spectrum. (c) The same pattern of NOEs is present in the Monomer Down lr-AID spectrum. (d) The observed NOEs from A26.H2 and A34.H2 are shown on the TAR hairpin secondary structure. (e) NOEs assigned in the isolated TAR hairpin from A34.H2 to A34.H1', G33.H1', and G35.H1'. (f) This pattern is reproduced in both the Monomer Up lr-AID spectrum (f) and the Monomer Down lr-AID spectrum (g).

The pattern of NOEs at the A26.H2 chemical shift in the TAR hairpin is consistent with the pattern of NOEs found at the same chemical shift in the monomeric 5'-L. The TAR spectra (Figure 3.3 a,e) show additional NOEs because the isolated TAR hairpin is fully protiated. Figure 3.3e-g shows a comparison between the region of the NOESY spectrum at the chemical shift of the TAR hairpin's A34.H2 and the same regions in monomeric 5'-L spectra. Again, the pattern of NOEs assigned in the TAR hairpin (Figure 3.3e) matches the pattern of NOEs in the monomeric spectra. These data suggest that the structure that produces these chemical environments in the TAR hairpin

is conserved in the monomer, or more simply put, that these bases in the monomeric 5'-L adopt the same structure as they do in the isolated TAR hairpin. Taken together these data suggest that the top of the TAR stemloop is maintained in the monomeric 5'-L. This finding is consistent with the biological role of the TAR hairpin. Transcriptional activation of the HIV genome requires the bulge and loop to be exposed in TAR as it is transcribed to allow binding of Tat to the bulge and a P-TEFb binding to the loop[10; 14].

PolyA Hairpin

The 2D NOESY of the isolated polyA hairpin shows an upfield shifted signal, assigned as A73.H2 (Figure 3.4a), which we expect to be resolved in the monomeric 5'-L spectrum. However, NOEs comparable to these signals are seen only very weakly in the Monomer Up lr-AID sample (Figure 3.4b), and not at all in the Monomer Down lr-AID sample (Figure 3.4c). It is possible that the dynamics of the polyA hairpin are sufficiently slow in the context of the monomeric 5'-L to significantly broaden these signals. Alternatively, these residues from polyA in the context of the monomeric 5'-L may adopt a different conformation than that of the isolated hairpin. It is unclear if the weak signals seen in the Monomer Up lr-AID sample could result from a small percentage of 5'-L in a dimeric conformation in the NMR sample.

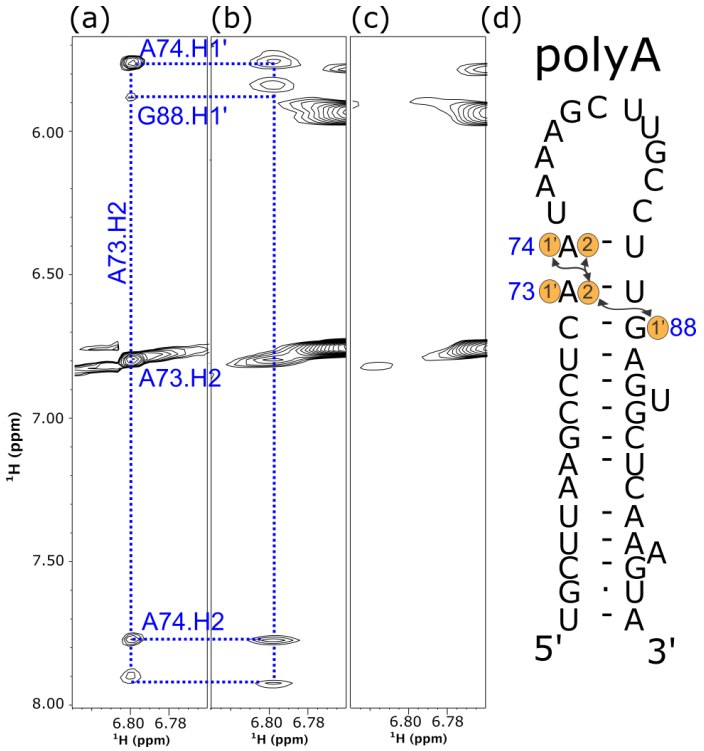

Figure 3.4. Presence of the polyA hairpin is not confirmed in the monomer. (a) The A73.H2 chemical shift in the isolated polyA is upfield shifted, and should be resolvable in the monomer. (b) A similar pattern of peaks occurs very weakly in the Monomer Up lr-AID sample, whereas these peaks are not found even at very low levels in the Monomer Down lr-AID sample (c). It is not clear whether the peaks are weak in the Monomer Up lr-AID because they are broad, and simply so broad in the Monomer Down lr-AID to be beyond detection, or if the peaks are weak in the Monomer Up lr-AID because they come from a small proportion of 5'-L folded in the dimeric conformation not present in the Monomer Down lr-AID sample. (d) The NOE connectivities from A73.H2 identified in the isolated polyA hairpin.

DIS Hairpin

One area of interest in the monomeric 5'-L spectrum is the DIS hairpin. Berkhout's LDI model[4] predicts that the DIS hairpin opens up to base pair with residues from the polyA hairpin in the monomer, whereas the U5:DIS interaction proposed by the Summers lab[1] suggests that the DIS stem remains base paired. The isolated DIS stemloop has an upfield shifted signal from A268.H2. This region from the DIS NOESY

spectrum can be seen in Figure 3.5a. A268.H2 sees the cross-strand plus one G251.H1' and sees forward to A269.H2. The crosspeak with the G251.H1' occurs in a fairly crowded region of the 5'-L spectrum, with overlap with peaks from the ψ hairpin and peaks thought to be from the PBS region, however the A268.H2-A269.H2 cross peak is well resolved. Figure 3.5b shows the pattern of NOEs is present in the Monomer Up lr-AID sample, although the G251.H1' crosspeak appears weak and is not well resolved. Figure 3.5c shows the A268.H2 pattern of NOEs is clearly present in the Monomer Down lr-AID sample, although there appears to be overlap with another peak. This suggests that the DIS hairpin is formed in the monomeric 5'-L as it is in the isolated DIS hairpin.

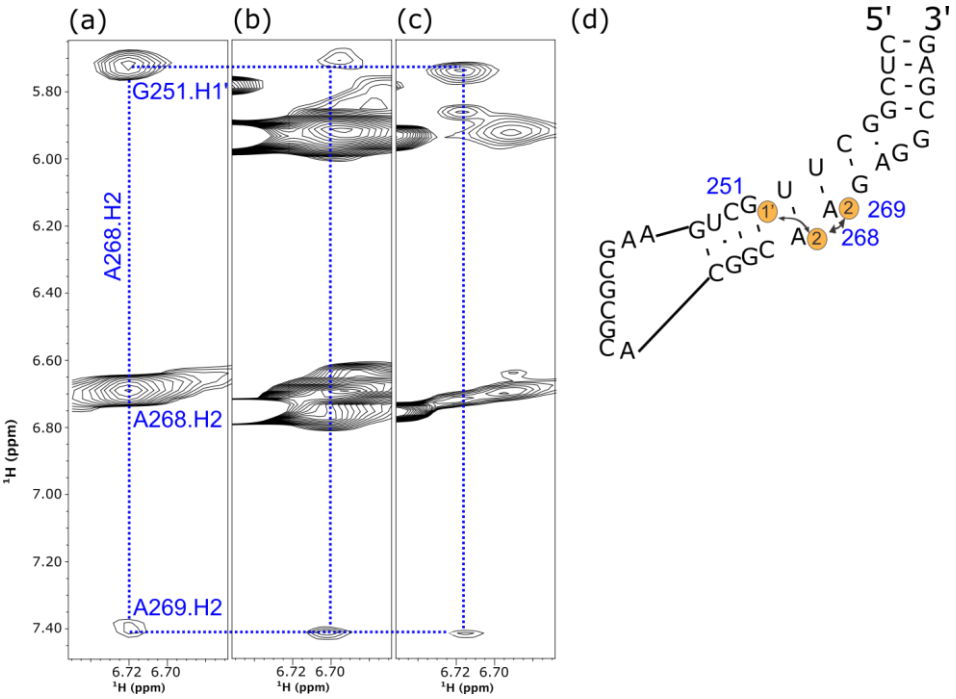

Figure 3.5. NOEs consistent with formation of the DIS stem are found in the monomeric leader. (a) The A268.H2 region of the isolated DIS hairpin NOESY spectrum. (b) The same pattern of NOEs is found in the Monomer Up lr-AID spectrum. (c) The same pattern of NOEs is found in the Monomer Down lr-AID spectrum. The G251.H1' peak in the Monomer Up lr-AID spectrum appears weak, but it is broad peak in a

crowded region of the spectrum. (d) The NOE connectivities from A268.H2 assigned in the isolated DIS hairpin.

ψ Hairpin

The ψ hairpin has two upfield shifted A.H2 signals, A314.H2 and A324.H2 shown in Figure 3.6a. Both of these A.H2s see NOEs to the following G.H1', as well as to the cross-strand plus one H1' (in this case, each other's H1'). Figure 3.6 shows the same pattern of peaks is found in both monomeric 5'-L samples, suggesting that the ψ hairpin is formed in the monomeric 5'-L.

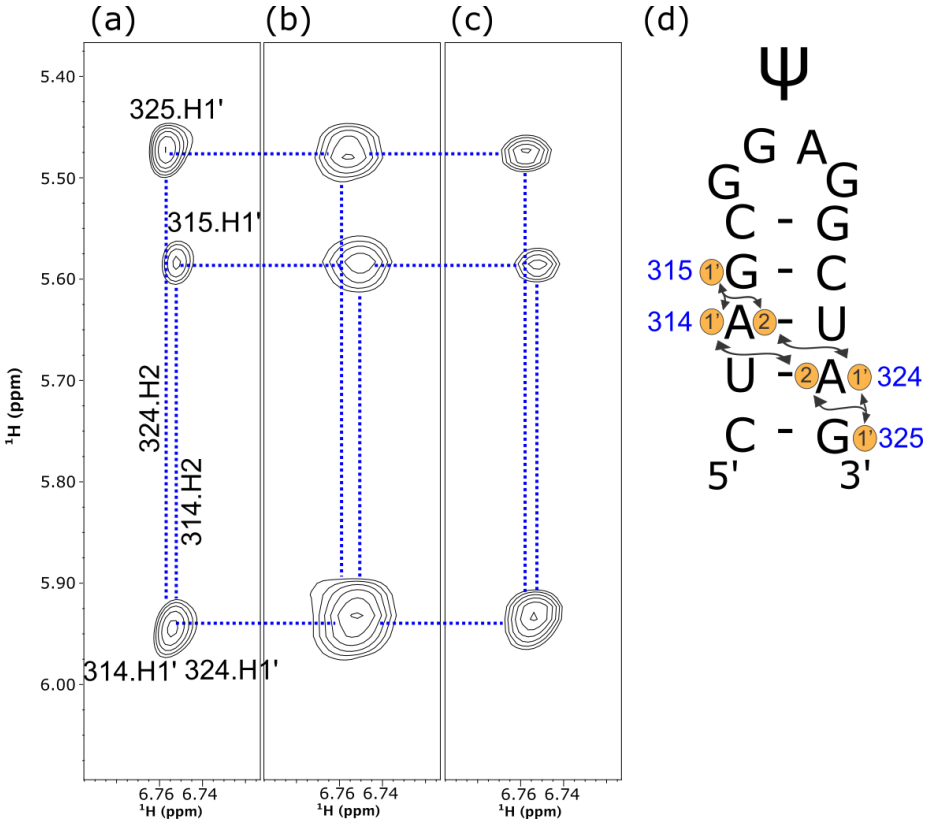

Figure 3.6. NOEs consistent with formation of the ψ hairpin are seen in both monomeric 5'-L spectra. (a) The region of the isolated ψ hairpin NOESY spectrum containing the A324.H2 and A314.H2 signals. The NOE pattern matches in both the Monomer Up lr-AID spectrum (b) and the Monomer Down lr-AID

spectrum (c), suggesting that this hairpin is formed in the monomer. (d) The NOE connectivities from A314.H2 and A324.H2 assigned in the isolated ψ hairpin.

Studies of the dimeric 5'-L revealed that it contained an extended ψ hairpin. Figure 3.7a shows a strip of the isolated extended ψ hairpin spectrum that appears to be matched in both monomeric 5'-L samples (Figure 3.7b and 3.7c). While this is suggestive of the formation of the extended ψ hairpin, the peaks occur in a relatively crowded region of the 5'-L spectra. It is possible to perform a NOESY walk including these peaks (shown for Monomer Up lr-AID in Figure 3.8 and for Monomer Down lr-AID in Figure 3.9), to confirm a network of NOEs consistent with the formation of the extended ψ hairpin. It is unusual to be able to assign a NOESY walk from a single, highly deuterated sample of a large RNA, but there is a high percentage of adenosines and guanosines in the ψ hairpin, and the peaks are very sharp and strong compared to the majority of the spectrum.

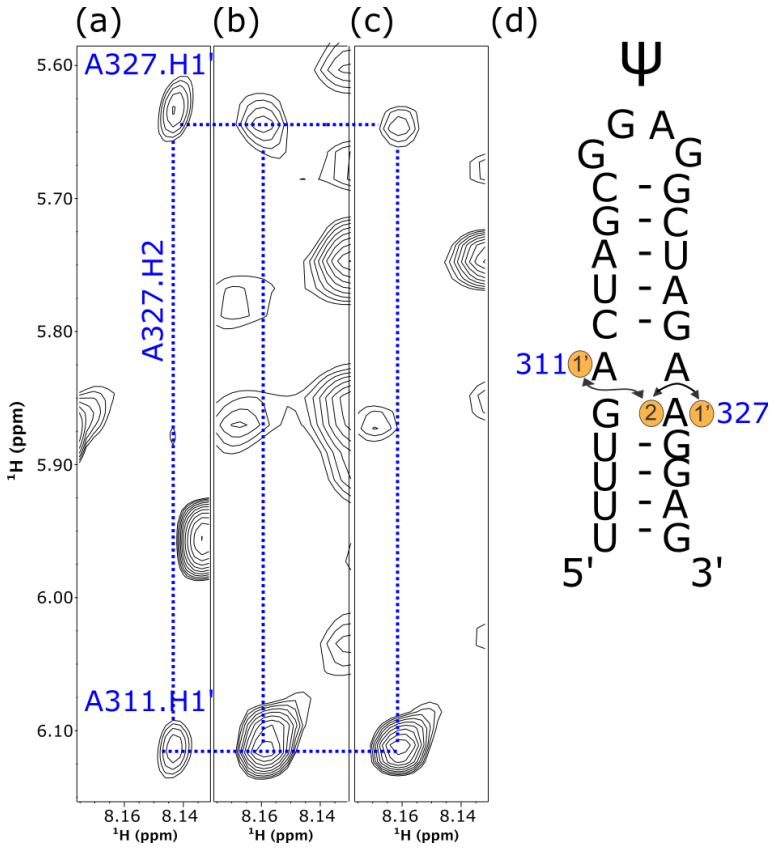

Figure 3.7. NOEs consistent with the formation of an extended ψ hairpin are seen in the monomeric 5'-L. (a) The A327.H2 slice of the isolated ψ hairpin NOESY spectrum. A matching pattern of peaks is found in both the Monomer Up lr-AID spectrum (b) and the Monomer Down lr-AID spectrum (c). These peaks are resolved, but exist in a densely populated area of the spectrum. Finding two NOEs matching the same pattern is suggestive but not conclusive of the existence of the extended ψ hairpin. (d) The NOE connectivities from A327.H2 assigned in the isolated extended ψ hairpin.

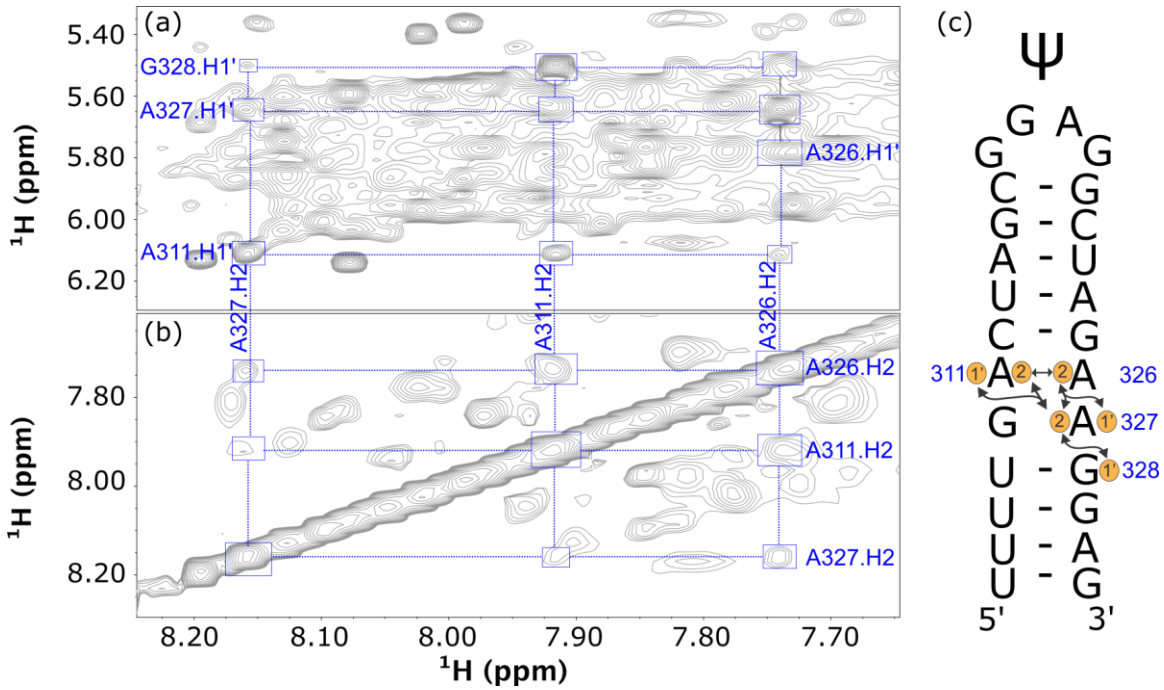

Figure 3.8. The NOESY walk consistent with the formation of the extended ψ hairpin in the Monomer Up lr-AID spectrum. (a) The H2-H1' NOE connectivities. (b) The H2-H2 NOE connectivities. (c) The NOE connectivities on the extended ψ hairpin secondary structure.

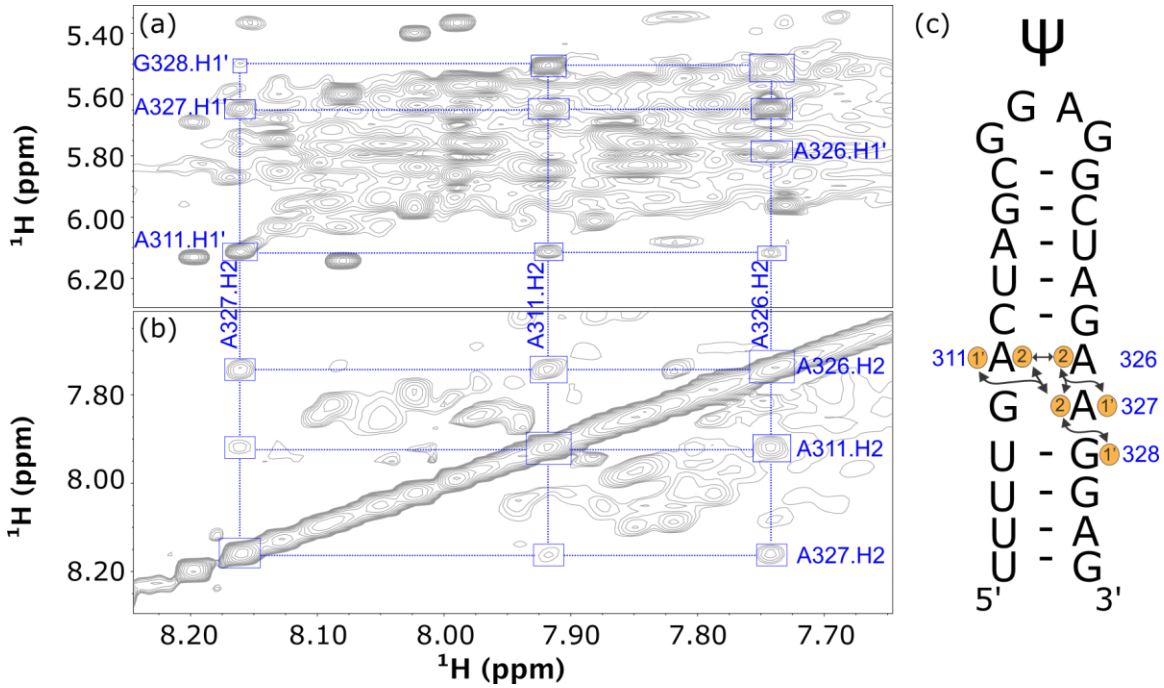

Figure 3.9. The NOESY walk consistent with the formation of the extended ψ hairpin in the Monomer Down lr-AID spectrum. (a) The H2-H1' NOE connectivities. (b) The H2-H2 NOE connectivities. (c) The NOE connectivities on the extended ψ hairpin secondary structure.

Splice-Donor Hairpin

Another area of interest in the monomeric 5'-L structure is the splice-donor (SD) region. The SD residues are predicted to form a hairpin in nearly all models of the 5'-L, excepting Berkhout's LDI model[4]. Recent studies from the Summers' lab have shown that the SD residues do not form a hairpin in the dimeric conformation of the 5'-L, instead these residues contribute to part of a novel double three-way junction[15]. Berkhout showed that the formation of a semi-stable hairpin with the splicing consensus sequence in the upper part of the hairpin is important for splicing efficiency of the HIV-1 5'-L[16]. As splicing occurs co-transcriptionally in cells, it is expected that the monomeric conformation of the 5'-L should be splicing-competent, but, as the dimeric conformation of the genome will provide the genetic information for a new virion, it would not be advantageous to have the dimeric conformation of the 5'-L capable of undergoing splicing. Comparison of the isolated SD stemloop spectrum to the monomeric 5'-L spectrum revealed that all of the chemical shifts from the native SD hairpin occur in crowded regions of the monomeric 5'-L spectrum, preventing determination of chemical shifts consistent with the existence or absence of the SD hairpin in the monomeric 5'-L from being identified. To determine if the SD hairpin formed as predicted in the monomeric 5'-L, lr-AID was again utilized. lr-AID mutations were incorporated into the SD hairpin (Figure 3.10c) in order to preserve the splicing consensus sequence and hairpin stability to allow *in vivo* splicing studies of this construct in the future. These

mutations were incorporated into the Monomer Down lr-AID construct. Figure 3.10 shows that the pattern of NOEs seen by A295.H2 in the control SD lr-AID hairpin spectrum is also found in the NOESY spectrum of the Monomer Down lr-AID construct incorporating the SD lr-AID mutations. The SD lr-AID signals are very sharp and intense, consistent with the SD hairpin having dynamics expected for an isolated hairpin.

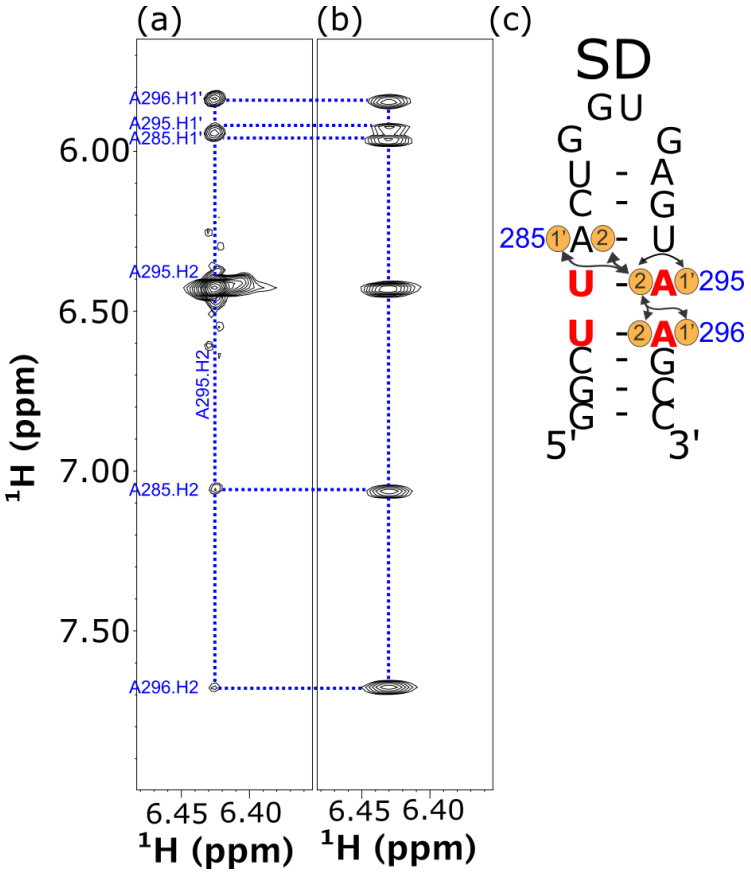

Figure 3.10. lr-AID mutations confirm the existence of the SD hairpin in the monomeric 5'-L. (a) The A295.H2 region of NOESY spectrum of the small SD lr-AID oligo shown in (c). The NOE connectivites are shown on the spectrum (a) as well as on the oligo (c). (b) The pattern of NOEs is repeated in the spectrum of the Monomer Down lr-AID construct incorporating the SD lr-AID mutations. This suggests that the SD hairpin is formed in the monomeric 5'-L.

Figure 3.11 shows that the addition of the SD lr-AID mutations does not alter the rest of the spectrum, which suggests that the addition of the SD lr-AID mutations does not change the fold of the RNA.

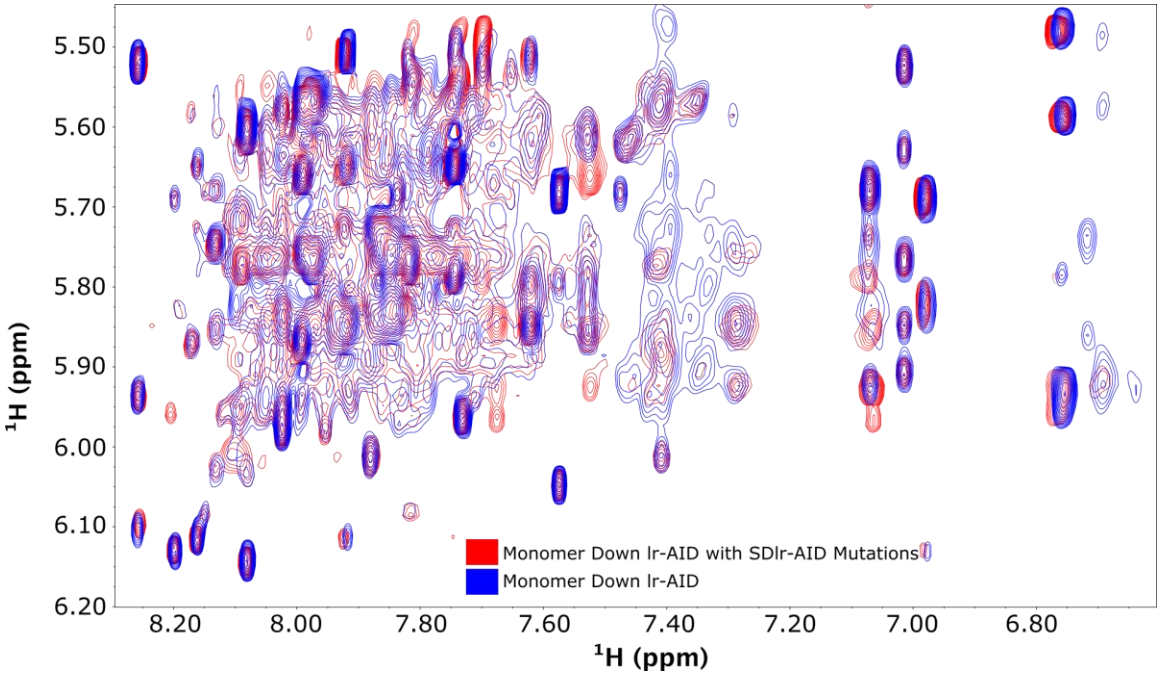

Figure 3.11. Mutation of the SD hairpin in the Monomer Down lr-AID construct does not change the overall fold of the RNA. The 2D NOESY spectrum of the Monomer Down lr-AID (blue) is overlaid with the spectrum of the Monomer Down lr-AID construct with additional SD-lrAID mutation (red). Overall the spectra are very similar with outlying peaks matching well.

Conclusions

In previous work, Dr. Lu successfully showed that the AUG hairpin is formed in the monomer using segmental C^{13} labeling[1]. Figure 3.12 shows an updated prediction of the monomeric 5'-L secondary structure taking into consideration what can be confirmed by NMR. The highlighted green areas were confirmed by outlying chemical shifts from the Monomer Up and Monomer Down lr-AID spectra. The highlighted blue areas were confirmed directly by lr-AID mutations. The highlighted purple area was confirmed by

Dr. Lu's segmental labeling studies. The top of the polyA hairpin is highlighted in yellow to indicate the uncertainty concerning formation of this secondary structure element.

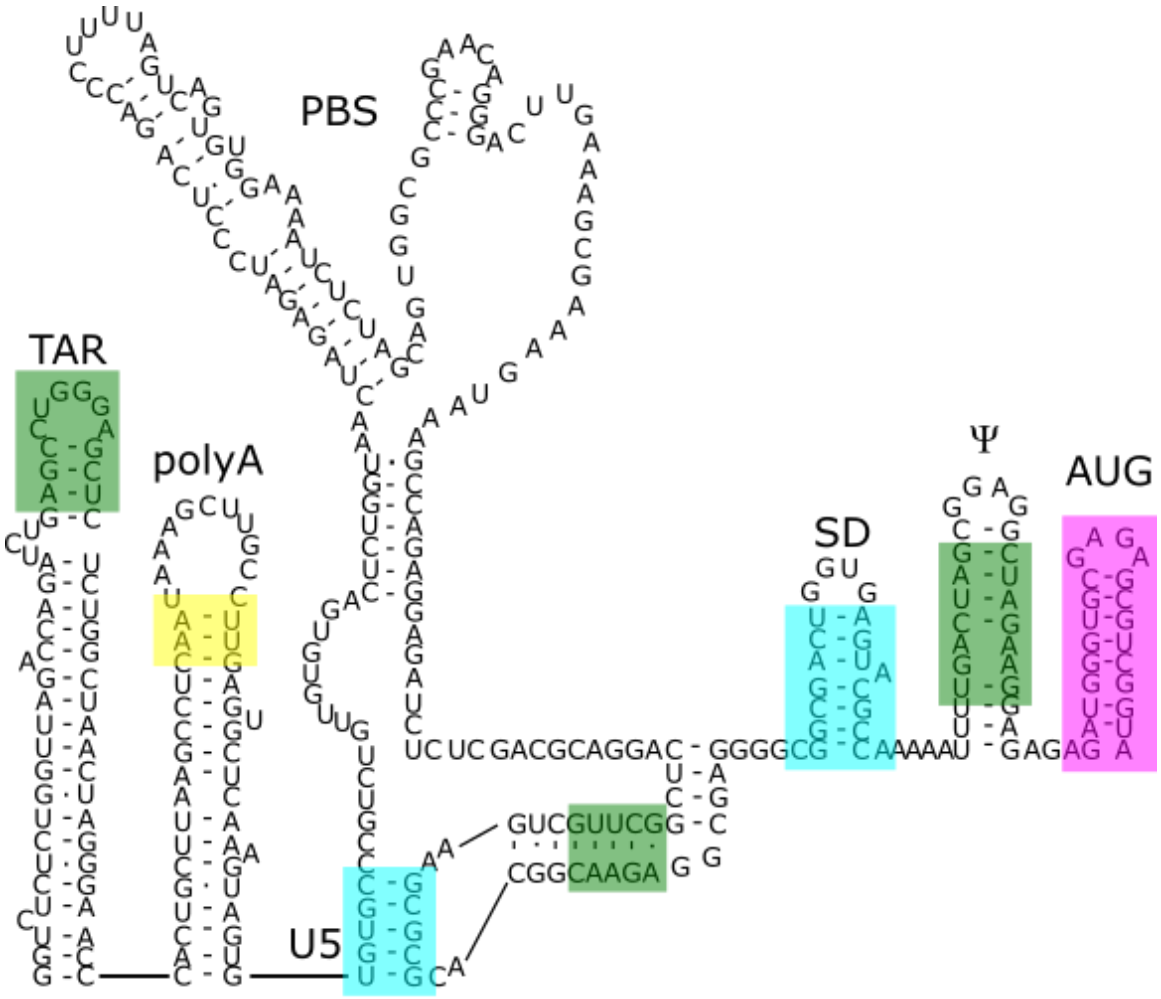

Figure 3.12. Summary of the secondary structure elements confirmed to exist in the 5'-L. Secondary structure elements assigned using outlying chemical shifts in the Monomer Up and Monomer Down lr-AID spectra are highlighted in green. Secondary structures identified by lr-AID mutations are highlighted in teal. Secondary structure elements identified by segmental C^{13} labeling[1] are highlighted in purple. The top of the polyA hairpin is highlighted in yellow to indicate the NOEs consistent with formation may have been identified, but were not consistently found.

In summary, a combination of strategies including lr-AID, comparison of control isolated hairpins to highly deuterated 5'-L constructs, and segmental isotopic labeling

have now provided evidence to suggest the secondary structure for large portions of the monomeric 5'-L.

NMR data supports that the top of the TAR stemloop is present in the monomeric 5'-L. This is consistent with the biological role of the TAR bulge and loop which bind Tat and P-TEFb to phosphorylate the C-terminal domain of RNA polymerase II (RNAPII)[2; 10; 14]. This releases RNAPII from transcriptional pausing and allows it to recruit factors that form the elongation complex[17].

The U5:DIS interaction which stabilizes the monomeric conformation by sequestering the DIS through base pairing was demonstrated in Chapter 2 and Lu *et al*[1]. This model is consistent with the findings in Nikolaitchik *et al.*[18] in which mutations to the AUG hairpin had a highly deleterious effect on packaging. Nikolaitchik *et al.* proposed that the region of U5 that participates in the U5:AUG interaction which stabilizes the dimer[4] may also participate in sequestration of the DIS. Furthermore, chemical probing applied to the isolated monomer, rather than a mixture of monomer and dimer, gave reactivities more consistent with the U5:DIS interaction than the previously LDI model of the monomer previously proposed by Abbink *et al*[4; 19].

NOEs consistent with formation of the DIS stem are found in the monomeric 5'-L. This, combined with the lr-AID data for the U5:DIS interaction suggests that the DIS hairpin does not open up to base pair with polyA, as proposed in the LDI model[4].

The presence of the splice-donor hairpin was confirmed by lr-AID studies. This is consistent with the structural role of a semi-stable hairpin for splicing regulation[16; 20]. The splice-donor hairpin is not present in the core packaging signal[15], indicating that

these residues are part of the large conformational rearrangement that occur during dimerization.

NMR data consistent with the formation of an extended ψ hairpin is shown. The loop of the ψ hairpin contains a high affinity NC binding site[9]. Previous studies indicate that there are approximately six high affinity NC sites in the monomer, but 16 per strand in the dimer[13]. This suggests that the ψ hairpin loop is one of the NC sites available in the monomer. As NC is known to have RNA chaperone activity, it is possible that NC must bind to the monomer to high affinity sites such as the one present on the ψ hairpin to promote unwinding and refolding into the dimeric conformation[21].

The findings presented here should guide future studies of the monomeric 5'-L structure. It appears that the splice-donor hairpin, extended ψ hairpin, and AUG hairpin do not interact with the remained of the monomeric 5'-L structure. This suggests that it should be possible to create a monomeric construct from the TAR hairpin through the DIS stemloop, which would greatly decrease the size of the RNA, making structural studies more tractable. Further evidence that the core monomeric structure is contained by the sequence from the TAR hairpin through the DIS stemloop is provided by the fact that spliced RNAs contain the 5'-L sequence from the TAR hairpin through the major splice-donor site in the loop of the SD hairpin. Preliminary work to characterize the structure of a spliced RNA is described in Appendix A-2.

Methods

RNA preparation: A pUC19 plasmid containing the a T7 promoter followed by the T-SL4 construct was mutated to contain the Monomer Up or Monomer Down mutations

using the QuikChange multi site-directed mutagenesis kit (Agilent Technologies). The following primers (IDT) were used for the Monomer Up mutagenesis: 5'-GCT CTC TGG CTG ACT AGG GAA CCC-3', 5'-GTG CTC AAA GTA GTG CTT ATC CGT CTG TTG TGT GAC TC-3', and 5'-GGA CTC GGC TTG CTG AAA TAA GCA CGG CAA GAG GCG AG-3'. The following primers (IDT) were used for the Monomer Down mutagenesis: 5'-GCT CTC TGG CTG ACT AGG GAA CCC-3', 5'-GAG TGC TCA AAG TAG TTA GTG CCC GTC TGT TGT GTG-3' , and 5'-TCG GCT TGC TGA AGC TAA CAC GGC AAG AGG CGA-3'. The lr-AID mutations to confirm the SD hairpin were made to the Monomer Down plasmid using the QuikChange site-directed mutagenesis kit (Agilent Technologies) and the following primers (IDT): 5'-AGA GGC GAG GGG CGG **TTA** CTG GTG AG**T AA**C CAA AAA TTT TGA CT-3' and 5'-AGT CAA AAT TTT TGG TTA CTC ACC AGT AAC CGC CCC TCG CCT CT-3'. The plasmids were sequenced (Genewiz) to confirm the mutations. The plasmids were transformed into DH5α cells (invitrogen) according to manufacturer's directions. To create DNA templates for RNA transcription 1 L cultures of the DH5α cells containing the plasmid were grown overnight in Luria broth at 37 ˚C at 250 rpm. Plasmids were purified using the QIAGEN Megaprep kit according to manufacturer's directions. Purified plasmids were digested by BstZ17I (NEB) overnight, according to manufacturer's protocol. Digestion was checked on a 2% TAE gel. After digestion, DNA template was purified by PCI extraction followed by ethanol precipitation. After purification the DNA template was washed by serial dilution and concentration three times with four milliliters of ultrapure water on a 30k Amicon Ultra centrifugal filter device. The DNA template for the isolated splice-donor hairpin containing lr-AID

mutations was made using a DNA from IDT (5'-mGmGT TAC TCA CCA GTA ACC TAT AGT GAG TCG TAT TA-3') annealed to Top17 from IDT (5'-TAA TAC GAC TCA CTA TA-3'). *In vitro* transcription using in-house purified T7 RNA polymerase was used to synthesize the RNA in a reaction mixture containing ~50 ng/µL purified DNA template, 2mM spermidine, 80mM Tris-HCl (pH 8), 2mM DTT, 10-20 mM $MgCl_2$, and 3-6 mM NTPs. 30 µL trial reactions were used to optimize $MgCl_2$ and NTP ratios for each construct. Transcription reactions were incubated for 2.5 hrs at 37 ˚C. Transcription reactions were halted with the addition of 1 mmole of EDTA followed by boiling for three minutes and snap cooling on ice for three minutes. RNA was purified by gel electrphoresis on a 6% denaturing acrylamide gel (SequaGel, National Diagnostics) for the large RNAs and a 20% denaturing acrylamide gel (SequaGel, National Diagnostics) for the isolated splice donor hairpin containing the lr-AID mutations using the FisherBiotech DNA sequencing system at 20W overnight to achieve the best resolution. The gel bands were visualized by UV-shadowing, excised, and eluted using the Elutrap ® electroelution system (Whatman) at 100 V overnight. The eluted RNA was collected, then washed two times with 4 mL of 2M high purity NaCl (99.999%, Acros), followed by 8 times with 4 mL of ultrapure water on a 30k Amicon Ultra centrifugal filter device.

Native gel electrophoresis studies: RNAs were prepared for native gel electrophoresis from purified stocks in water lyophilizing the sufficient RNA to give 125 µM in 10 µL. The lyophilized RNA was resuspended in 9 µL of water. Once the RNA was fully resuspended, 1 µL of 100 mM Tris, pH 7.5 was added to give a final concentration of 125 µM RNA and 10 mM Tris. The RNA in the 10mM Tris buffer was incubated at 37 ˚C

overnight to allow equilibration. To prevent overloading the gel, samples were diluted prior to running. To prevent dissociation during dilution, 9 µL aliquots of 10 mM Tris were chilled on ice. 1 µL of sample was added to the 9 µL of buffer, immediately mixed by pipetting up and down, and 1 µL of of all-purpose, native agarose gel loading solution (life technologies) was added. The samples were then immediately loaded into the wells of a 1% native agarose TB gel. TB gels were prepared by boiling 0.5 g of high gelling temperature agarose (Fisher) dissolved in 50 mL of 44 mM Tris Base - Boric Acid (Fisher) buffer, pre-stained with 0.5 µg ethidium bromide per milliliter of gel. The gels and 44mM Tris Base- Boric Acid buffer were chilled for one hour before use. The gels were run at 115V for approximately 1 hour and 15 minutes. Gels were visualized and photographed using a Kodak Gel Logic 200 Imaging System.

NMR Studies: NMR data were collected with a Bruker Avance 800 MHz spectrometer, processed with NMRPIPE/NMRDraw[22] and analyzed with NMRView[23]. NMR signals were assigned by standard methods[24]. RNA concentrations for 2D NOESY spectra were 125 µM. NMR samples were buffered with10 mM Tris buffer, pH 7.5, with no added salts.

References

1. Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A. & Summers, M. F. (2011). NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-5.

2. Jakobovits, A., Smith, D. H., Jakobovits, E. B. & Capon, D. J. (1988). A discrete element 3' of human immunodeficiency virus 1 (HIV-1) and HIV-2 mRNA initiation sites mediates transcriptional activation by an HIV trans activator. *Mol Cell Biol* **8**, 2555-61.

3. Huthoff, H. & Berkhout, B. (2001). Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143-57.

4. Abbink, T. E. & Berkhout, B. (2003). A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J Biol Chem* **278**, 11601-11.

5. Rhim, H., Park, J. & Morrow, C. D. (1991). Deletions in the tRNA(Lys) primer-binding site of human immunodeficiency virus type 1 identify essential regions for reverse transcription. *J Virol* **65**, 4555-64.

6. Paillart, J. C., Marquet, R., Skripkin, E., Ehresmann, B. & Ehresmann, C. (1994). Mutational analysis of the bipartite dimer linkage structure of human immunodeficiency virus type 1 genomic RNA. *J Biol Chem* **269**, 27486-93.

7. Bohne, J., Wodrich, H. & Kräusslich, H. G. (2005). Splicing of human immunodeficiency virus RNA is position-dependent suggesting sequential removal of introns from the 5' end. *Nucleic Acids Res* **33**, 825-37.

8. Amarasinghe, G. K., De Guzman, R. N., Turner, R. B. & Summers, M. F. (2000). NMR structure of stem-loop SL2 of the HIV-1 psi RNA packaging signal reveals a novel A-U-A base-triple platform. *J Mol Biol* **299**, 145-56.

9. De Guzman, R. N., Wu, Z. R., Stalling, C. C., Pappalardo, L., Borer, P. N. & Summers, M. F. (1998). Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* **279**, 384-8.

10. Mancebo, H. S., Lee, G., Flygare, J., Tomassini, J., Luu, P., Zhu, Y., Peng, J., Blau, C., Hazuda, D., Price, D. & Flores, O. (1997). P-TEFb kinase is required for HIV Tat transcriptional activation in vivo and in vitro. *Genes Dev* **11**, 2633-44.

11. Johnson, S. F. & Telesnitsky, A. (2010). Retroviral RNA dimerization and packaging: the what, how, when, where, and why. *PLoS Pathog* **6**, e1001007.

12. Hayashi, T., Shioda, T., Iwakura, Y. & Shibuta, H. (1992). RNA packaging signal of human immunodeficiency virus type 1. *Virology* **188**, 590-9.

13. Heng, X., Kharytonchyk, S., Garcia, E. L., Lu, K., Divakaruni, S. S., Lacotti, C., Edme, K., Telesnitsky, A. & Summers, M. F. (2012). Identification of a Minimal Region of the HIV-1 5'-Leader Required for RNA Dimerization, NC Binding, and Packaging. *J Mol Biol*.

14. Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H. & Jones, K. A. (1998). A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451-62.

15. Keane, S. C., Heng, X., Lu, K., Kharytonchyk, S., Ramakrishnan, V., Carter, G., Barton, S., Hosic, A., Florwick, A., Santos, J., Bolden, N. C., McCowin, S., Case, D. A., Johnson, B. A., Salemi, M., Telesnitsky, A. & Summers, M. F. (2015). RNA structure. Structure of the HIV-1 RNA packaging signal. *Science* **348**, 917-21.

16. Abbink, T. E. & Berkhout, B. (2008). RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *J Virol* **82**, 3090-8.

17. Ott, M., Geyer, M. & Zhou, Q. (2011). The control of HIV transcription: keeping RNA polymerase II on track. *Cell Host Microbe* **10**, 426-35.

18. Nikolaitchik, O., Rhodes, T. D., Ott, D. & Hu, W. S. (2006). Effects of mutations in the human immunodeficiency virus type 1 Gag gene on RNA packaging and recombination. *J Virol* **80**, 4691-7.

19. Kenyon, J. C., Prestwood, L. J., Le Grice, S. F. & Lever, A. M. (2013). In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader. *Nucleic Acids Res* **41**, e174.

20. Mueller, N., van Bel, N., Berkhout, B. & Das, A. T. (2014). HIV-1 splicing at the major splice donor site is restricted by RNA structure. *Virology* **468-470**, 609-20.

21. Egelé, C., Schaub, E., Ramalanjaona, N., Piémont, E., Ficheux, D., Roques, B., Darlix, J. L. & Mély, Y. (2004). HIV-1 nucleocapsid protein binds to the viral DNA initiation sequences and chaperones their kissing interactions. *J Mol Biol* **342**, 453-66.

22. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-93.

23. Johnson, B. A. & Blevins, R. A. (1994). NMR View: A computer program for the visualization and analysis of NMR data. *J Biomol NMR* **4**, 603-14.

24. Wüthrich, K. (1986). *NMR of proteins and nucleic acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University, Wiley, New York.

Chapter 4:Effects of Capping and Heterogeneity at the 5' mRNA Start Site of HIV-1 on the Monomer-Dimer Equilibrium of the 5' leader

---

Introduction

In eukaryotic transcription, the TATA box is a common core promoter element[1]. The human TATA box is bound by the transcription factor IID complex (TFIID)[2], and factors that make up TFIIB and RNA polymerase II place the transcription start site (~25-30 nucleotides downstream)[3; 4]. Heterogeneity in transcription start sites (TSS) is commonly reported for human mRNAs[5]. It has been suggested that TSS variation contributes more to mRNA diversity than alternative splicing[6]. In prokaryotes variations in TSS have been linked to changes in mRNA secondary structure which regulate translation of the transcript[7]. The HIV proviral DNA provides a promoter for viral RNA transcription. This promoter includes a TATA box positioned approximately 30 nucleotides upstream of three guanosine residues, all of which have been suggested as HIV-1 viral RNA transcription start sites by different groups (Figure 4.1A)[8; 9; 10; 11; 12; 13]. The beginning of the HIV RNA genome, the 5' leader (5'-L) is an important region of the genome which controls a number of steps in the HIV life cycle. One key step controlled by the 5'-L is dimerization to allow for selective packaging of the dimeric HIV-1 genome into newly formed viral particles[14]. We investigated whether the different 5' start sites reported in the literature had an effect on the monomer-dimer equilibrium of the 5'-L. 5'-L constructs beginning with the guanosine numbered 454 in the proviral DNA (G^{454}), the guanosine numbered 455 (G^{455}), and the guanosine numbered 456 (G^{456}) were transcribed *in vitro*.

Results

The monomer-dimer equilibria of the constructs are presented in Figure 4.1B. It is apparent that there are differences in the equilibria of these RNAs. Unexpectedly, the construct beginning with G^{454} showed considerable stabilization of the monomeric species. While a 2007 paper by Menees *et al.* suggested that most viral RNA has a start site that corresponds with G^{456}, the paper also revealed evidence for a 5' 7-methylguanosine cap (m7G cap) on the mRNA[9]. Capping of the viral mRNA is expected because the host cellular machinery which transcribes the HIV genome caps the mRNAs that it produces. However, *in vitro* transcribed RNAs lack a 5' m7G cap. The additional monomeric stability provided by an additional 5' guanosine residue prompted us to determine whether or not the addition of a 5' m7G cap would similarly aid in monomeric stability. A Vaccinia virus capping system was utilized to cap the *in vitro* transcribed RNA. This allowed the monomer-dimer equilibria of all 3 possible start sites, with and without the presence of a 5' m7G cap to be interrogated (Figure 4.1).

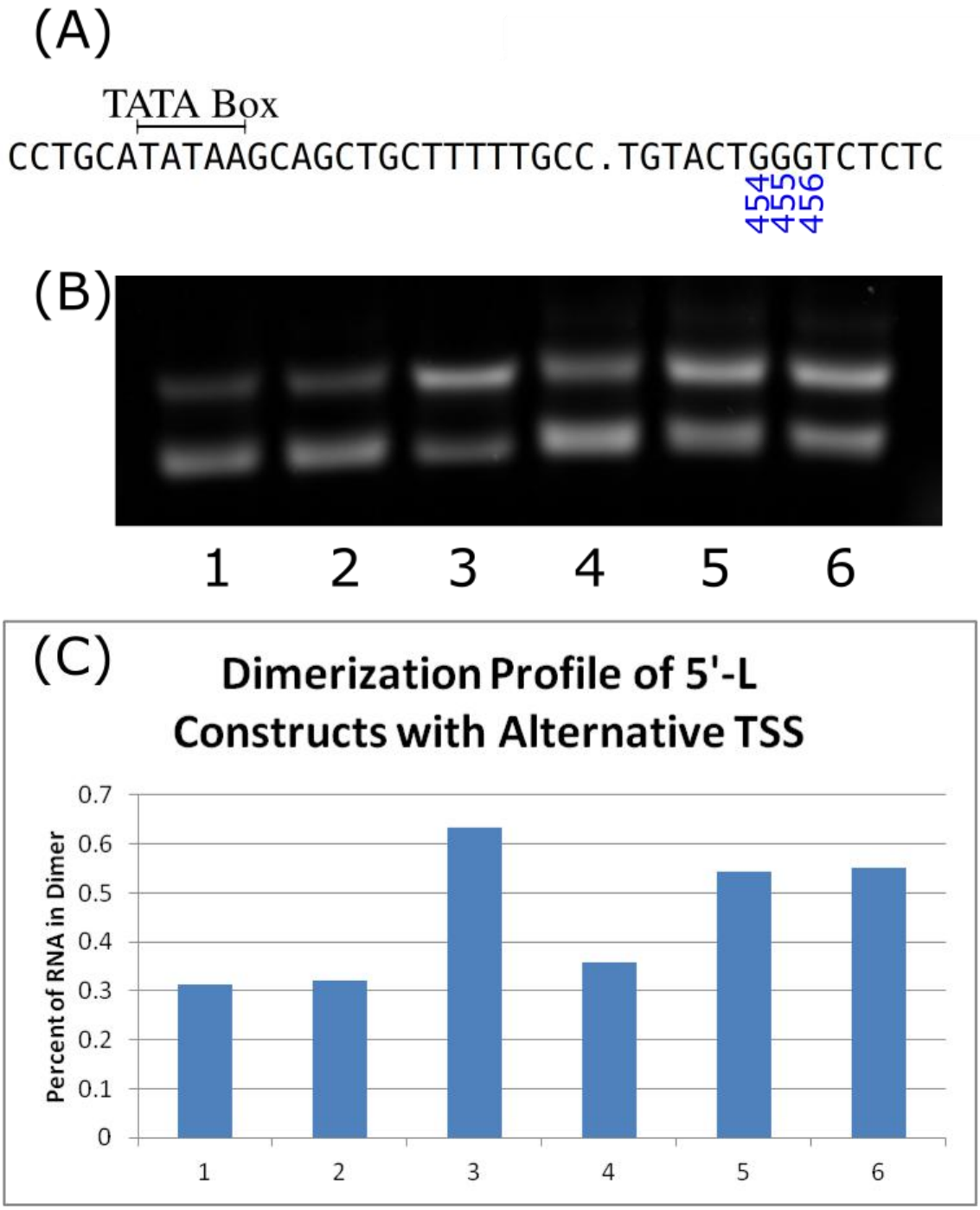

Figure 4.1. Inconsistencies in the start site of the HIV-1 +1 mRNA transcript. **(A)** The proviral DNA sequence[15] of HIV-1 B.FR.83.HXB2 showing the TATA box and three guanosine residues (454, 455, 456) in appropriate positions to serve as the mRNA +1 start site. (B) The monomer-dimer equilibrium of 5'-L constructs with varying +1 start sites with and without the addition of the 5'-7-methylguanosine cap (m^7G cap) as determined by native gel electrophoresis. 5'-L constructs starting with (1) the 454 start site, (2) the

454 site with the m^7G cap, (3) the 455 start site, (4) the 455 start site with the m^7G cap, (5) the 456 start site, and (6) the 456 start site with the m^7G cap. The RNAs in lanes 1, 2, and 4 favor the monomeric conformation while the RNAs in lanes 3, 5, and 6 favor the dimeric conformation. (C) Quantification of the gel from (B) showing the percent of RNA in the dimeric conformation for each of the alternative 5' TSS with and without the presence of the m^7G cap.

If only the capped RNAs (lanes 2,4, and 6), which represent the possible RNAs found *in vivo*, are compared, it is clear that the G^{456} start site favors the dimeric conformation, while the G^{455} and G^{456} start sites favor the monomeric conformation. It is worth noting that the presence of the 5' m7G cap seems to be mimicked by incorporation of a 5' guanosine. Thus the G^{456} start site with the 5' m7G cap favors the dimer as does the G^{455} start site, and the G^{455} start site with the 5' m7G cap favors the monomer as does the G^{454} start site. This suggests that the m7G cap is base pairing in some manner that changes the structure of the 5'-L to either favor the dimeric conformation (with the G^{456} start site) or the monomeric conformation (with the G^{455} start site). Figure 4.2 proposes a possible model for this phenomenon.

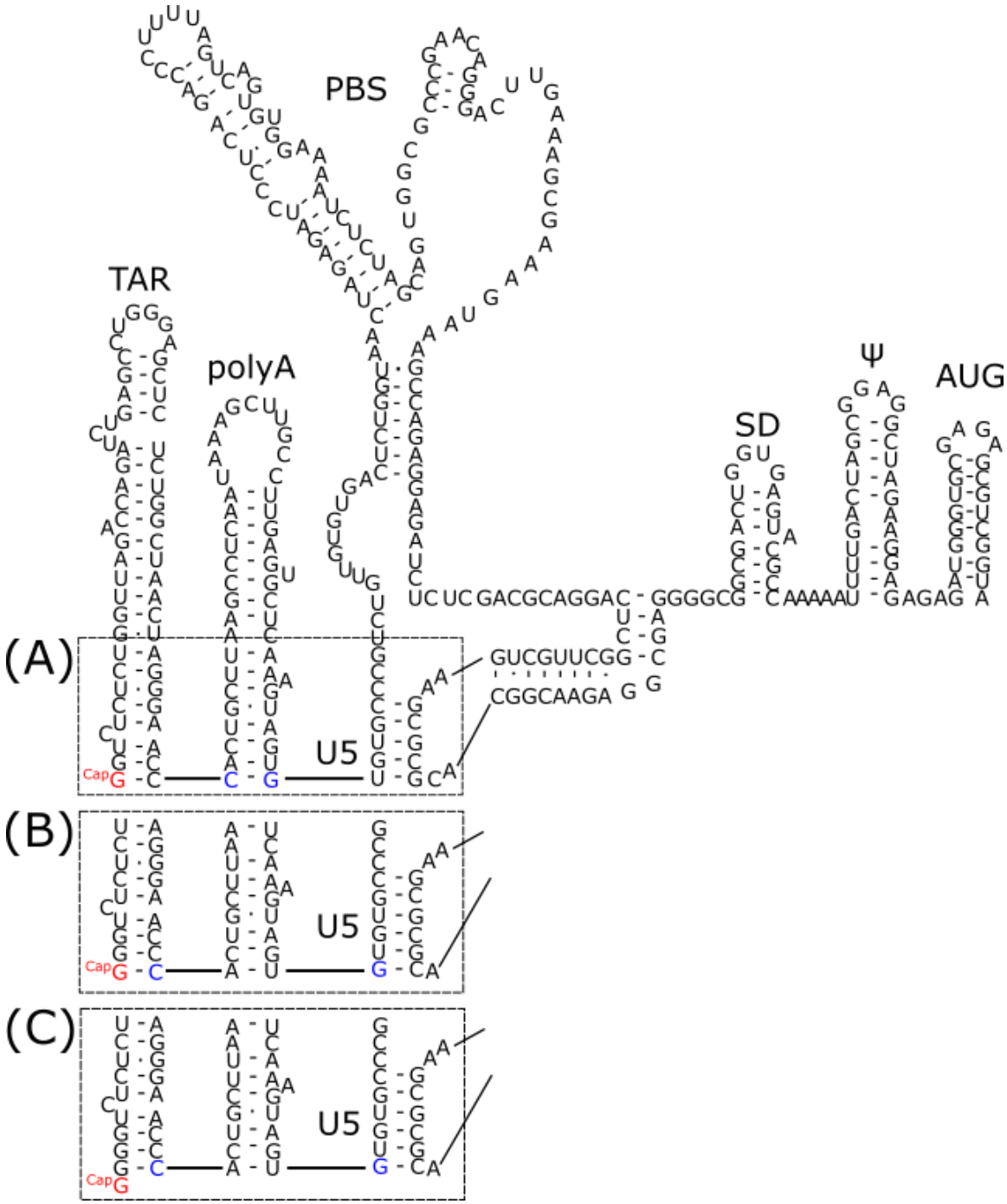

Figure 4.2. Shift in 5'-L secondary structure proposed to account for change in monomer-dimer equilibrium with differing 5' start sites. (A) Proposed secondary structure for the 5'-L beginning at G456 with the m^7G cap. The m^7G cap (red) is expected to base pair with cytosine 57 (black). (B) Proposed secondary structure for 5'-L beginning at G455 with the m^7G cap. The m^7G cap (red) is expected to base pair with the cytosine 58 (blue) previously predicted to stabilize the base of the PolyA hairpin (as shown in

A). This opens up the base of the PolyA such that guanosine 104 (blue) is able to base pair with the DIS to enhance the U5:DIS interaction that stabilizes the monomer. (C) Proposed secondary structure for the 5'-L beginning at G454 with the m^7G cap. The m^7G cap (red) is not expected to base pair, and as such the secondary structure is otherwise identical to (B).

The proposed structures in Figure 4.2 illustrate the possibility that the addition of the 5' 7mG cap to the G^{455} start site could allow the base of the polyA hairpin to open up such that C58 base pairs with the 7mG cap, allowing G104 to base pair to C262, which would enhance the U5:DIS interaction by adding an additional G-C base pair. The increased stability of the U5:DIS interaction would explain the increased monomeric stability for this construct. If it is the destabilization of the base of the polyA hairpin which enhances the monomeric stability, it would be expected that mutations that destabilize the polyA hairpin would enhance monomeric stability. To test this hypothesis, mutations were made to A59 and U103 to promote destabilization of the base of the polyA hairpin. Figure 4.3 shows that these mutations stabilize the monomeric conformation of the 5'-L.

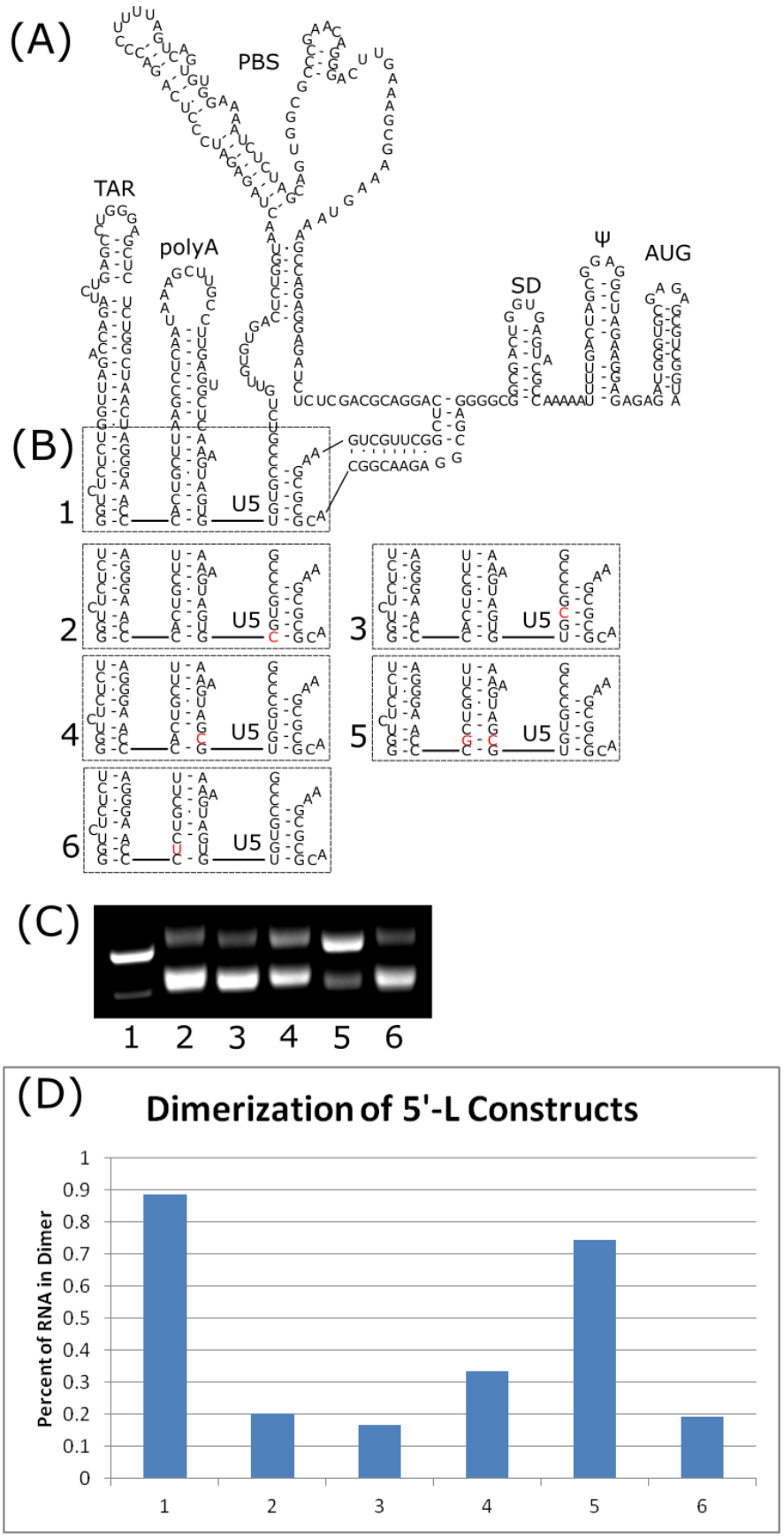
(A)
PBS
TAR
polyA
SD
Ψ
AUG
(B)
U5
GUCGUUCGG
CGGCAAGA
1
2
3
4
5
6
(C)
1 2 3 4 5 6
(D)
Dimerization of 5'-L Constructs
Percent of RNA in Dimer
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6

Figure 4.3. Mutations that destabilize the base of PolyA stabilize the monomeric conformation of the 5'-L. (A) A proposed secondary structure of the native 5'-L. Dashed box (1) outlines the region that will contain mutations and is duplicated in (B). (B) Point mutations made to the 5'-L. (2) The U105C mutation and (3) the U107C mutation stabilize the monomer by improving the U5:DIS base pairing. (4) The U103C mutation destabilizes the base of the polyA stemloop. (5) Mutations A59G and U103C restabilize the base of the polyA stemloop. (6) The A59U mutation destabilizes the base of the polyA stemloop. (C) Native gel electrophoresis studies of constructs 1-6 reveals that the mutations that stabilize the U5:DIS interaction stabilize the monomeric conformation of the 5'-L, as expected, but mutations that destabilize the base of the polyA stemloop also stabilize the monomeric conformation of the 5'-L. Mutations that restabilize the base of the polyA stemloop revert to a wildtype phenotype favoring the dimeric conformation of the 5'-L. (D) Quantification of the gel from (C) showing the percent of RNA in the dimeric conformation for each construct.

Previous work indicated that U105C and U107C mutants stabilized the monomeric conformation of the 5'-L, but these mutations were assessed in constructs that contained non-native 3' cytosine residues[16]. Therefore the U105C and U107C mutants were tested in the context of the native 5'-L. Both mutations enhanced the stability of the monomer relative to wild-type, which is consistent with the U5:DIS interaction being enhanced by swapping a U-G base pair with a C-G base pair. To test whether destabilization of the polyA hairpin showed a similar effect, a U103C mutant was assessed. The U103C mutation was found to stabilize the monomeric conformation. When the base of the polyA hairpin was restabilized by making an A59G mutation in combination with the U103C mutation, the wild-type monomer-dimer equilibrium was restored. Additionally, an A59U mutation which destabilizes the base of the polyA hairpin also stabilizes the monomeric conformation. These results support the model proposed in Figure 4.2.

While these results present a clear *in vitro* difference behavior of the 5'-L depending upon the transcription start site, the *in vivo* significance remained to be determined. The work by Menees *et al*. to determine the 5' transcription start site for HIV-1 suggested that the RNA primarily began in position G^{456}, but results consistent with a transcription start site of G^{455} were shown, but the authors proposed that this resulted from the non-templated addition of an extra cytosine by reverse transcriptase[9]. To allow more precise identification of the 5' transcription start site, our collaborator, Dr. Siarhei Kharytonchyk developed riboprobes to differentiate between the G^{454}, G^{455}, and G^{456} transcription start sites. The specificity of these riboprobes was tested on *in vitro* synthesized and enzymatically capped 5'-L RNAs corresponding with each of the start sites, which I provided. Figure 4.4 shows that the riboprobes are able to distinguish between the different 5'-L start sites. When these riboprobes were used to identify the 5' start site of RNA extracted from HIV-1 infected cells, two populations were seen (Figure 4.4, lane 2), suggesting that there is heterogeneity in the 5' start site of HIV-1 mRNA present in cells. However, when these riboprobes were used to identify the 5' start site of RNA extracted from HIV-1 virions, a single population of mRNAs consistent with the G^{456} start site was identified, suggesting that HIV-1 preferentially packages mRNA with the G^{456} start site.

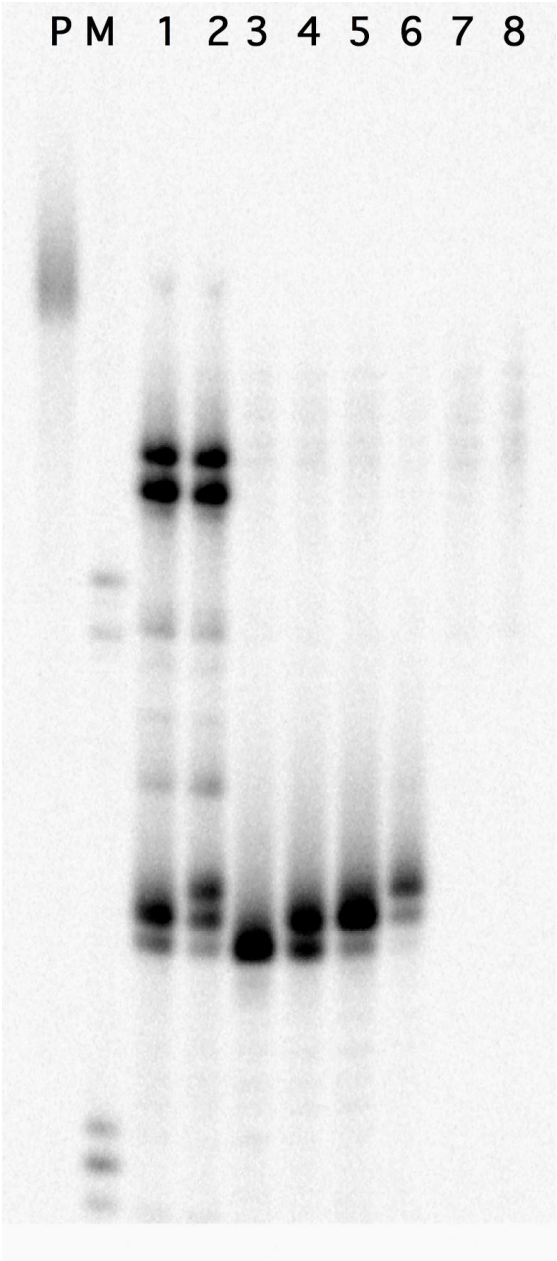

Figure 4.4. 5' start site heterogeneity is seen in the HIV-1 genome mRNA of cells consistent with start site 456 plus a cap and start site 455 plus a cap, but virions are enriched for mRNA with start site 456 plus a cap. (1) RNA extracted from HIV-1 virions. (2) RNA extracted from HIV-1 infected cells. Riboprobe assays can distinguish between *in vitro* transcribed RNAs corresponding to (3) 456 start site, (4) 456 start site with the m^7G cap, (5) 455 start site, and (6) 455 start site with the m^7G cap. (7) RNA extracted from media from mock infected cells. (8) RNA extracted from mock infected cells. The RNA extracted from HIV-1 virions (1) contains bands consistent with lane 4, suggesting that only the RNA corresponding to the

456 start site with the m^7G cap is packaged into virions. However, the RNA extracted from HIV-1 infected cells (2) contains bands consistent with lanes 4 and 6 suggesting that there is a heterogenous population of RNA present in HIV-1 infected cells corresponding to both the 456 and 455 start sites. This work was conducted by Dr. Siarhei Kharytonchyk in Dr. Alice Telesnitsky's laboratory.

Conclusions

The work presented suggests that HIV-1 produces mRNAs with heterogeneous 5' transcription start sites. Additionally, it has been shown that mRNAs with a transcription start site that has been determined to confer increased stability of the dimeric conformation *in vitro* are preferentially selected for packaging into virions. This suggests that variations in the 5' start site can affect the structure of the 5'-L. A proposed model for this effect has been presented, although further work will be required to verify the hypothesis. These findings suggest that small differences in mRNA sequences can have a large influence on RNA structure and trafficking of mRNAs. Additionally, the effect of the 5'-L start site and 5' 7mG cap on the 5'-L monomer-dimer equilibrium highlights the necessity for *in vitro* studies of the 5'-L to utilize the biologically relevant start sites and incorporation of the 5' 7mG cap. Furthermore, if the 5' 7mG cap can participate in base pairing interactions, it may be unavailable for binding to the nuclear cap binding complex or the cytosolic cap binding protein. If this is true, alternative RNA trafficking pathways must be utilized to traffic the HIV-1 unspliced RNA, which may offer a selective drug target.

Methods

RNA preparation: A pUC19 plasmid containing the a T7 promoter followed by the T-SL4 construct that started with two guanosines (G^{455} start site) was mutated to contain

either a single G (G^{456} start site) or three guanosines (G^{454} start site) using the QuikChange site-directed mutagenesis kit (Agilent Technologies) and the following primers (IDT): 5'-ATA CGA CTC ACT ATA **G**GG TCT CTC TGG TTA G-3' and 5'-CTA ACC AGA GAG ACC CTA TAG TGA GTC GTA T-3' for the G^{454} start site and 5'-ATA CGA CTC ACT ATA GTC TCT CTG GTT AG-3' and 5'-CTA ACC AGA GAG ACT ATA GTG AGT CGT AT-3' for the G^{456} start site. The plasmids were sequenced (Genewiz) to confirm the mutations. The plasmids were transformed into DH5α cells (invitrogen) according to manufacturer's directions. To create DNA templates for RNA transcription 1 L cultures of the DH5α cells containing the plasmid were grown overnight in Luria broth at 37 ˚C at 250 rpm. Plasmids were purified using the QIAGEN Megaprep kit according to manufacturer's directions. Purified plasmids were digested by BstZ17I (NEB) overnight, according to manufacturer's protocol. Digestion was checked on a 2% TAE gel. After digestion, DNA template was purified by PCI extraction followed by ethanol precipitation. After purification the DNA template was washed by serial dilution and concentration three times with four milliliters of ultrapure water on a 30k Amicon Ultra centrifugal filter device. *In vitro* transcription using in-house purified T7 RNA polymerase was used to synthesize the RNA in a reaction mixture containing ~50 ng/µL purified DNA template, 2mM spermidine, 80mM Tris-HCl (pH 8), 2mM DTT, 10-20 mM $MgCl_2$, and 3-6 mM NTPs. 30 µL trial reactions were used to optimize $MgCl_2$ and NTP ratios for each construct. Transcription reactions were incubated for 2.5 hrs at 37 ˚C. Transcription reactions were halted with the addition of 1 mmole of EDTA followed by boiling for three minutes and snap cooling on ice for three minutes. RNA was purified by gel electrophoresis on a 6% denaturing acrylamide gel (SequaGel, National

Diagnostics) using the FisherBiotech DNA sequencing system at 20W overnight to achieve the best resolution. The gel bands were visualized by UV-shadowing, excised, and eluted using the Elutrap ® electroelution system (Whatman) at 100 V overnight. The eluted RNA was collected, then washed two times with 4 mL of 2M high purity NaCl (99.999%, Acros), followed by 8 times with 4 mL of ultrapure water on a 30k Amicon Ultra centrifugal filter device.

Enzymatic Capping: An in-house Vaccinia virus capping system was developed based on the NEB system. Plasmid containing the His-tagged Vaccinia virus capping enzyme was a kind gift from Stephen Cusack's lab at the European Molecular Biology Laboratory (EMBL)[17]. The plasmid was transformed into BL21(DE3)pLysS cells (life technologies) by thawing competent cells on ice and aliquoting 50 μL into a prechilled 1.5 mL eppendorf. 3 μL of 150 ng/μL plasmid DNA was added and gently mixed by stirring, this mixture sat on ice for 30 minutes, then was heat shocked at 42C for 25 seconds. 90 μL of room temperature SOC media was added into the mixture, and this was incubated for 60 min at 37 ˚C and 250 rpm. 75 uL was plated onto a luria agar plate with ampicillin and chlorampicillin, and incubated overnight at 37 ˚C.

The plasmid was also transformed into DH5α cells (life technologies) by thawing competent cells on ice and aliquoting 50 μL into a prechilled 1.5mL eppendorf. 1 μL of 17.6 ng/μL plasmid DNA was added and gently mixed by stirring. This mixture sat on ice for 30 minutes, then was heat shocked at 42 ˚C for 45 seconds. 100 μL of prewarmed SOC media was added, and this was incubated for 60 min at 37 ˚C and 250 rpm. 50 μL was plated onto an luria agar and ampicillin plate and incubated overnight at 37C.

A 200 mL Terrific Broth starter culture with ampicillin and chlorampicillin was innoculated overnight from a glycerol stock of the transformed BL21(DE3)pLysS cells and incubated at 37 ˚C, 250 rpm overnight. The culture was spun down in 50 mL falcon tubes at 5500 rpm for 10 min at 4 ˚C. Each cell pellets was resuspended in 25 mL from a 1 L Terrific broth culture with ampicillin and chlorampicillin (25 mL were used for each 1 L culture). Four 4 L flasks containing 1 L of Terrific broth culture with ampicillin and chlorampicillin were innoculated from the starter culture and grown at 37 ˚C at 150 rpm to an OD of ~0.88. The flasks were then placed on ice for 15-45 minutes. Next the flasks were induced with 0.5 mM IPTG (500 µL of 1 M) at 20 ˚C overnight at 150 rpm. The cells were spun down at 8000 rpm for 10 min at 4 ˚C. The supernatant was poured off, and the cell pellets were frozen at -20 ˚C for at least 30 min. The cell pellets were thawed and resuspended in 100 mL lysis buffer (40 mM Tris-HCl, 200 mM NaCl, 10 mM Imidazole, 5mM TCEP, pH 8.0) and 300 µL of protease inhibitor was added. The cells were lysed by microfluidizing three times. The lysate was centrifuged for 30 minutes at 17000 rpm at 4 ˚C. 10 mL of cobalt resin slurry was equilibrated in lysis buffer by washing twice with a 10 mL volume. The lysate was bound to the resin for 1 hr at 4 ˚C. The lysate was allowed to flow through and then the resin was washed two times with a 50 mL volume of lysis buffer. Next the resin was washed four times with 10 mL aliquots of lysis buffer, and the capping enzyme was eluted with five 10 mL elutions of elution buffer (40 mM Tris-HCl, 200 mM NaCl, 250 mM imidazole, 5mM TCEP, pH 8.0). The concentrations of all elutions were measured, and all elutions with concentrations of over 0.05 mg/mL were combined and dialyzed overnight at 4 ˚C in dialysis buffer (20 mM Tris-HCl, 100 mM NaCl, 0.100 mM EDTA, 10% glycerol, 1 mM

DTT, pH 8.0). Elutions were tested for activity by comparison of capping efficiency to the NEB Vaccinia capping system enzyme on a 19 nucleotide hairpin. Capping was determined by gel shift on a 20% denaturing acrylamide gel (SequaGel, National Diagnostics) run at 220 V for two hours. Capping was performed based on the instructions provided for the NEB capping system, using SAM purchased at NEB. Capping reactions were stopped with the addition of 1 mmole of EDTA followed by boiling for three minutes and snap cooling on ice for three minutes. RNA was purified by gel electrphoresis on a 6% denaturing acrylamide gel (SequaGel, National Diagnostics) using the FisherBiotech DNA sequencing system at 20W overnight to achieve the best resolution. The gel bands were visualized by UV-shadowing, excised, and eluted using the Elutrap ® electroelution system (Whatman) at 100 V overnight. The eluted RNA was collected, then washed two times with 4 mL of 2M high purity NaCl (99.999%, Acros), followed by 8 times with 4 mL of ultrapure water on a 30k Amicon Ultra centrifugal filter device.

Native gel electrophoresis studies: RNAs were prepared for native gel electrophoresis from purified stocks in water by diluting the RNA to 1.21 times the desired concentration (0.1 µM final concentration) in RNA water followed by addition of 10% of the final volume of a 100 mM HEPES buffer, pH 7.0. 90% of this volume was transferred to a fresh RNase-free 1.5 mL centrifuge tube (Eppendorf), and the remaining volume was used to verify the concentration. Sufficient 10x phsyiological ionic (PI) salts were added to this volume to give 0.15 µM RNA and 1x PI salts (140 mM KCl, 10 mM NaCl, and 1 mM $MgCl_2$). The RNA in the physiological buffer was incubated at 37 ˚C overnight to allow equilibration. To a 20 µL aliquot of sample 2 µL of all-purpose, native agarose gel

loading solution (life technologies) was added. 20 µL of sample was then loaded into the well of a 1% native agarose TB gel. TB gels were prepared by boiling 0.5 g of high gelling temperature agarose (Fisher) dissolved in 50 mL of 44 mM Tris Base - Boric Acid (Fisher) buffer, pre-stained with 0.5 µg ethidium bromide per milliliter of gel. The gels and 44mM Tris Base- Boric Acid buffer were chilled for one hour before use. The gels were run at 115V for approximately 1 hour and 15 minutes. Gels were visualized and photographed using a Kodak Gel Logic 200 Imaging System.

References

1. Breathnach, R. & Chambon, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem* **50**, 349-83.

2. Nakajima, N., Horikoshi, M. & Roeder, R. G. (1988). Factors involved in specific transcription by mammalian RNA polymerase II: purification, genetic specificity, and TATA box-promoter interactions of TFIID. *Mol Cell Biol* **8**, 4028-40.

3. Li, Y., Flanagan, P. M., Tschochner, H. & Kornberg, R. D. (1994). RNA polymerase II initiation factor interactions and transcription start site selection. *Science* **263**, 805-7.

4. Smale, S. T. & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449-79.

5. Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., Okubo, K., Sakaki, Y., Nakamura, Y., Suyama, A. & Sugano, S. (2001). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep* **2**, 388-93.

6. Rojas-Duran, M. F. & Gilbert, W. V. (2012). Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**, 2299-305.

7. Liu, J. & Turnbough, C. L. (1994). Effects of transcriptional start site sequence and position on nucleotide-sensitive selection of alternative start sites at the pyrC promoter in Escherichia coli. *J Bacteriol* **176**, 2938-45.

8. Soto-Rifo, R., Rubilar, P. S., Limousin, T., de Breyne, S., Décimo, D. & Ohlmann, T. (2012). DEAD-box protein DDX3 associates with eIF4F to promote translation of selected mRNAs. *EMBO J* **31**, 3745-56.

9. Menees, T. M., Müller, B. & Kräusslich, H. G. (2007). The major 5' end of HIV type 1 RNA corresponds to G456. *AIDS Res Hum Retroviruses* **23**, 1042-8.

10. Huthoff, H. & Berkhout, B. (2001). Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143-57.

11. Parkin, N. T., Cohen, E. A., Darveau, A., Rosen, C., Haseltine, W. & Sonenberg, N. (1988). Mutational analysis of the 5' non-coding region of human immunodeficiency virus type 1: effects of secondary structure on translation. *EMBO J* **7**, 2831-7.

12. Feng, S. & Holland, E. C. (1988). HIV-1 tat trans-activation requires the loop sequence within tar. *Nature* **334**, 165-7.

13. Dingwall, C., Ernberg, I., Gait, M. J., Green, S. M., Heaphy, S., Karn, J., Lowe, A. D., Singh, M. & Skinner, M. A. (1990). HIV-1 tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure. *EMBO J* **9**, 4145-53.

14. Heng, X., Kharytonchyk, S., Garcia, E. L., Lu, K., Divakaruni, S. S., Lacotti, C., Edme, K., Telesnitsky, A. & Summers, M. F. (2012). Identification of a Minimal Region of the HIV-1 5'-Leader Required for RNA Dimerization, NC Binding, and Packaging. *J Mol Biol*.

15. Foley B, Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, Rambaut A, Wolinsky S & Eds., K. B. **HIV Sequence Compendium 2014**. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 13-26007.

16. Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A. & Summers, M. F. (2011). NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-5.

17. De la Peña, M., Kyrieleis, O. J. & Cusack, S. (2007). Structural insights into the mechanism and evolution of the vaccinia virus mRNA cap N7 methyl-transferase. *EMBO J* **26**, 4913-25.

Chapter 5: Conclusions and Future Work

Conclusions

The studies presented in this dissertation have enhanced our understanding of the mechanisms by which the 5' leader (5'-L) acts in the life cycle of HIV-1. While there was evidence for two mutually exclusive conformations of the HIV-1 5'-L *in vitro*, the structure of the monomeric conformation proposed at the onset of this work lacked evidence for a number of secondary structure elements with known biological function and was not supported by *ex vivo* chemical probing studies[1; 2].

In this work we sought to understand how the monomeric conformation of the 5'-L is involved in the life cycle of HIV-1 through structural characterization of this RNA. In order to perform structural studies on the monomeric 5'-L, it was important to first determine how the monomeric conformation was stabilized. Prior studies suggested that the U5 region of the 5'-L may be involved in sequestration of the dimer initiation site (DIS)[3]. We performed mutagenesis on this region of the 5'-L and used native gel electrophoresis to characterize the monomer-dimer equilibria of the resultant constructs. These studies suggested that nucleotides U105-C110 may be involved in sequestration of the DIS in the monomeric conformation. A new NMR strategy to probe for base pairing interactions in the context of large RNAs, long-range probing by Adenosine Interaction Detection, was developed by the lab, allowing us to probe for the proposed U5:DIS interaction in the context of the 5'-L. These experiments allowed us to confirm the existence of a U5:DIS interaction in the monomeric 5'-L. Further studies by collaborators in the Boris-Lawrie lab demonstrated the biological relevance of this

construct and the role of the monomeric 5'-L as the translated conformation of the 5'-L by showing that the lr-AID mutations made to test U5:DIS interaction also enhance translation of the genome compared to the wild type and dimeric constructs of the 5'-L. These results indicate that the structure of the overall 5'-L determines its function in the HIV life cycle[4].

Confirmation of the U5:DIS interaction and its role promoting translation of the genome allowed structural studies of the monomeric 5'-L by NMR to proceed. A challenge in assessing the structure of the monomeric 5'-L by NMR has been that at the concentrations of RNA required for structural studies by NMR, the monomer-dimer equilibrium of the native 5'-L strongly favors the dimeric conformation. The lr-AID mutations made to probe for the U5:DIS interaction allow us to overcome this difficulty by disabling dimerization at the DIS due to mutation of the self-complementary palindrome, while enabling the formation of the U5:DIS interaction to maintain the native monomeric fold. This allowed us to probe the structure of the monomeric 5'-L by NMR for the presence of proposed secondary structure elements. These studies indicate that the top of the TAR stemloop is formed, which is consistent with the structural requirements to promote transcription of the genome. There is evidence that the stem of the DIS stemloop remains intact, rather than opening in the monomeric conformation such that residues from the DIS stemloop and polyA region base pair to sequester the DIS as predicted previously[1; 5]. The presence of a splice-donor hairpin, demonstrated to be important for regulated splicing efficiency of HIV-1, was verified by lr-AID studies[6; 7]. As the recently published structure of the core packaging signal shows that the dimeric conformation of the 5'-L does not contain the splice-donor hairpin, this represents one of

the areas of the HIV-1 5'-L that undergoes a major changes in secondary structure upon dimerization[8]. Additionally, NMR evidence was found for the formation of an extended ψ stemloop in the monomeric 5'-L. As the ψ stemloop is known to contain a high affinity nucleocapsid binding site, it is likely that this is one of the initial binding sites for nucleocapsid, allowing it to operate as a chaperone promoting dimerization of the 5'-L.

These structural studies of the monomeric 5'-L will guide future studies. The line widths of the splice-donor and ψ NOEs suggest that these stemloops are isolated and unimpeded by interactions with the remainder of the 5'-L. The ^{13}C studies conducted by Dr. Lu on the AUG stemloop suggest that it is also an isolated stemloop. This suggests that future studies of the monomeric 5'-L core structure can focus on constructs containing TAR through the DIS stemloop. This is consistent with the expectation that spliced RNAs contain the structural elements to promote translation, because spliced RNAs contain the elements of the 5'-L through the splice-donor, located in the loop of the splice-donor stemloop. The reduced size of a construct for NMR structural studies which eliminates the splice-donor, ψ, and AUG stemloops would help reduce spectral overlap and rotational correlation times.

Additional studies in this dissertation suggest a structural role for the 5' 7-methylguanosine (7mG) cap of the 5'-L. While *in vivo* the 5'-L of HIV-1 is capped by the eukaryotic transcriptional machinery, RNAs transcribed *in vitro* using a T7 RNA polymerase are not 5' capped. *In vitro* transcribed RNAs can have a native 7mG cap applied enzymatically using a Vaccinia virus capping enzyme in the presence of GTP and *S*-Adenosyl methionine. Using this system we measured the monomer-dimer equilibria of the capped and uncapped 5'-L and found that the presence of the 5' 7mG cap stabilizes

the monomeric conformation of the 5'-L. Mutagenesis studies support that this monomeric stabilization could be caused by destabilization of the base of the polyA stemloop.

Furthermore, a literature search revealed inconsistencies in the reported 5' transcription start site for the HIV-1 5'-L. When we investigated 5'-L RNAs with the alternative start sites, we found that the transcription start site determined the monomer-dimer ratio of the 5'-L. In combination with the 5' capping studies, we showed that one transcription start site favored the dimeric conformation, while the other 5' start sites favored the monomeric conformation. Futher studies by our collaborators in the Telesnitsky lab confirmed that two populations of RNA corresponding to two different transcription start sites exist in HIV-1 infected cells, however, the population of RNAs with the transcription start site that favors dimerization *in vitro* (G^{456}) is enriched in virions. This suggests that the fate of the HIV-1 unspliced RNA is highly influenced by its 5' transcription start site, most likely due to a structural effect.

Future Work

The studies presented here suggest a number of exciting avenues of investigation for the future. The structure of the monomeric 5'-L core remains to be solved, and a construct truncated after the DIS stemloop is recommended for future work. Additionally, the stabilizing effect of the 5' m7G cap should be investigated at a number of levels. It would be very exciting to study the interactions between the cap and the 5'-L RNA using NMR. For initial studies a small hairpin representing the TAR stemloop could be transcribed and its structure analyzed by NMR in the presence and absence of

the 5' 7mG cap. This would allow us to determine if the 5' 7mG cap can base pair as predicted by the model presented in Chapter 4. Furthermore, the 5' 7mG cap may be able to stabilize the monomeric 5'-L sufficiently for structural studies without requiring mutations to the DIS. Additional studies could be conducted on the nucleocapsid binding properties of the capped and uncapped 5'-Ls of differing transcription start sites to help elucidate the mechanism by which HIV-1 preferentially packages RNAs containing the G^{456} transcription start site. These studies could help answer long-standing questions about when the fate of the unspliced HIV RNA is determined. If it is confirmed that the 5' 7mG cap affects structure by base pairing with regions of the 5'-L, it will be interesting to see if the cap interacts with the nuclear or cytoplasmic cap binding protein in this conformation. If it is not able to, that would suggest the use of cap-independent translation for the unspliced HIV-1 genome, providing another area to target therapeutic development.

References

1. Abbink, T. E. & Berkhout, B. (2003). A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J Biol Chem* **278**, 11601-11.

2. Paillart, J. C., Dettenhofer, M., Yu, X. F., Ehresmann, C., Ehresmann, B. & Marquet, R. (2004). First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J Biol Chem* **279**, 48397-403.

3. Nikolaitchik, O., Rhodes, T. D., Ott, D. & Hu, W. S. (2006). Effects of mutations in the human immunodeficiency virus type 1 Gag gene on RNA packaging and recombination. *J Virol* **80**, 4691-7.

4. Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A. & Summers, M. F. (2011). NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-5.

5. Huthoff, H. & Berkhout, B. (2001). Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143-57.

6. Abbink, T. E. & Berkhout, B. (2008). RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *J Virol* **82**, 3090-8.

7. Mueller, N., van Bel, N., Berkhout, B. & Das, A. T. (2014). HIV-1 splicing at the major splice donor site is restricted by RNA structure. *Virology* **468-470**, 609-20.

8. Keane, S. C., Heng, X., Lu, K., Kharytonchyk, S., Ramakrishnan, V., Carter, G., Barton, S., Hosic, A., Florwick, A., Santos, J., Bolden, N. C., McCowin, S., Case, D. A., Johnson, B. A., Salemi, M., Telesnitsky, A. & Summers, M. F. (2015). RNA structure. Structure of the HIV-1 RNA packaging signal. *Science* **348**, 917-21.

Appendix A-1: DISMAL Construct

The strain of HIV that is used for our 5'-L studies is NL4-3. The 5' leader (5'-L) of HIV-1 is highly conserved, but there are variations in the DIS sequence for different strains. NL4-3 has a 5'-GCGCGC-3' DIS, but other strains have different palindromic sequences. For example, a different strain, MAL, has a 5'-GUGCAC-3' DIS sequence (Figure A-1.1). I hypothesized that the difference in DIS sequence could affect the dimerization profile of the 5'-L.

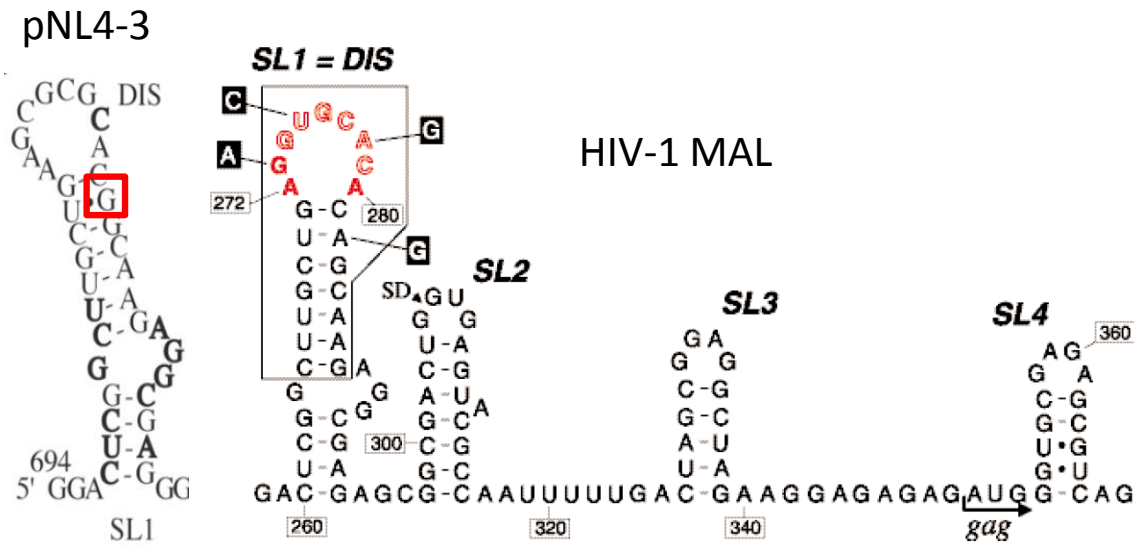

Figure A-1.1. Comparison of the DIS stemloop from NL4-3 (left)[1] to the DIS stemloop of MAL (right)[2]. There are differences in the base pairings that close the DIS loop, the sequences in the loop flanking the DIS, and the DIS itself.

The MAL 5'-L would theoretically have a weaker DIS:DIS interaction than the NL4-3 5'-L. Additionally, with the "Monomer Up" orientation, the MAL 5'-L would theoretically have a stronger U5:DIS interaction that the NL4-3 5'-L. This would suggest

that the use of the MAL DIS palindrome sequence would favor the monomeric conformation of the 5'-L (Figure A-1.2).

MAL/NL4-3 U5: **MAL DIS:**	**5'-GUGUGUGCCCG-3'** **3'-ACACGUGGA-5'**	**8 base pairs for U5:DIS interaction and slightly weaker DIS:DIS interaction (two A-U base pairs)**
MAL/NL4-3 U5: **NL4-3 DIS:**	**5'-GUGUGUGCCCG-3'** **3'-ACGCGCGAA-5'**	**7 base pairs for U5:DIS interaction and stronger DIS:DIS interaction (all 6 G-C base pairs)**

Free Energy Calculations

MAL DIS-DIS: -7.9	**NL4-3 DIS-DIS: -10.9**
MAL U5-DIS: -11.0	**NL4-3 U5-DIS: -9.9**
ΔΔG: -3.1	**ΔΔG: -1**

Figure A-1.2. Comparison of the U5:DIS base pairing that would occur in the MAL strain versus the NL4-3 strain of HIV-1. The free energy of the DIS:DIS and U5:DIS interaction for each strain as calculated by RNAstructure[3] are given, as are the difference in free energy for each strain. By these calculations, and assuming the U5:DIS interaction stabilizes the monomer, the MAL strain is expected to have a more stable monomer than the NL4-3 strain.

There are differences in sequence between the MAL and NL4-3 strains outside of the DIS loop. Because the effects of these sequence differences on structure would be hard to predict, I made mutations to NL4-3 to contain the DIS loop of MAL instead of simply using the MAL 5'-L (Figure A-1.3), to create a DISMAL construct. If there were differences in the dimerization profile of this construct, it would be reasonable to suggest that they occurred because of the changes to the DIS loop.

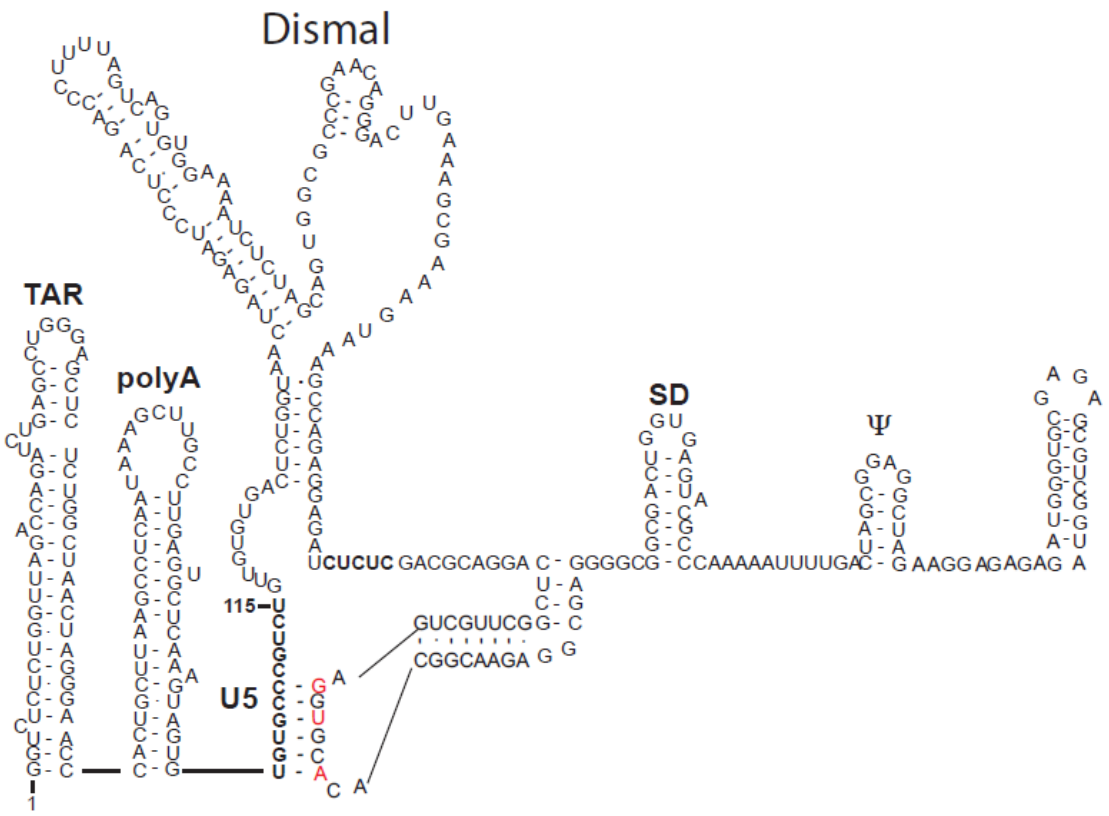

Figure A-1.3. The DISMAL construct which is the NL4-3 5'-L mutated to contain the DIS loop of the MAL strain.

To characterize the DISMAL construct, a time dependence dimerization gel was run to determine how long the construct must incubate to come to equilibrium. Figure A-1.4 shows that the construct appears to have come to equilibrium by 24 hours. Because dimerization is concentration dependent, a concentration dependence study was conducted on RNA that had incubated for 24 hours. As shown in Figure A-1.4, the DISMAL construct maintains a predominately monomeric conformation at the concentrations tested. This is in contrast to the NL4-3 5'-L which has a dimerization K_d of approximately 0.4 µM.

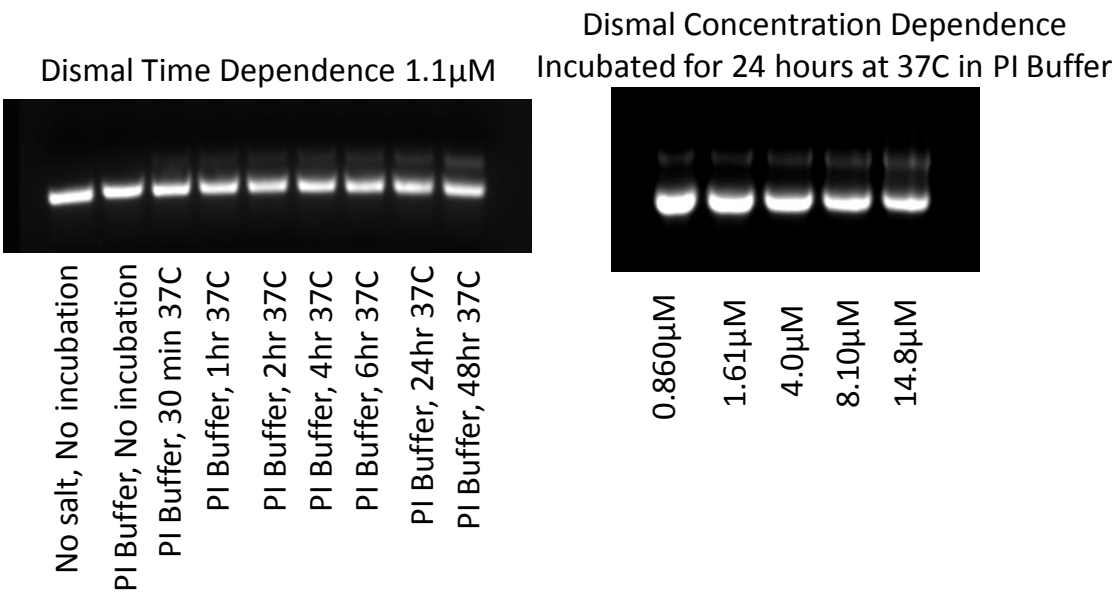

Figure A-1.4. Native agarose gel electrophoresis of the DISMAL construct reveals that it favors the monomeric conformation compared to NL4-3. 1% agarose gels run with a 44mM Tris-Borate buffer indicate that the RNA has come to equilibrium by 24 hours (left), and that the RNA is primarily monomeric at concentrations from 0.86 - 14.8 µM (right), when incubated at 37 ˚C in a physiological ionic strength buffer (10mM Tris pH 7, 10mM NaCl, 140 mM KCl, 1 mM $MgCl_2$).

The initial findings were very promising, so characterizations of nucleocapsid (NC) binding to the DISMAL construct using gel shifts (Figure A-1.5) and isothermal titration calorimetry (ITC) (Figure A-1.6) were performed.

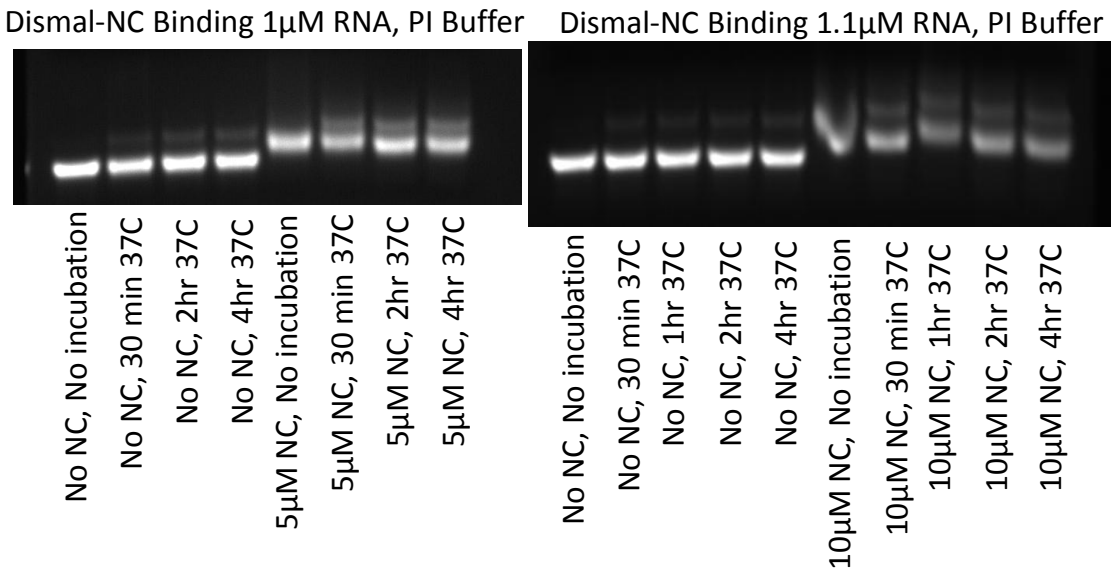

Figure A-1.5. Native 1% Tris-borate agarose gel electrophoresis of DISMAL in the absence and presence of NC. NC binds to both the monomer and dimer to shift both bands, and it increases the amount of dimer present in the sample. The increase in the amount of dimer is proportional to the amount of NC in the sample.

The gel binding studies show that NC binds to the DISMAL 5'-L and that it acts as a riboswitch to promote dimerization of the 5'-L. This is behavior consistent with a properly folded 5'-L and demonstrates that the DISMAL construct is not dimerization incompetent, but appears simply to be a stabilized monomer.

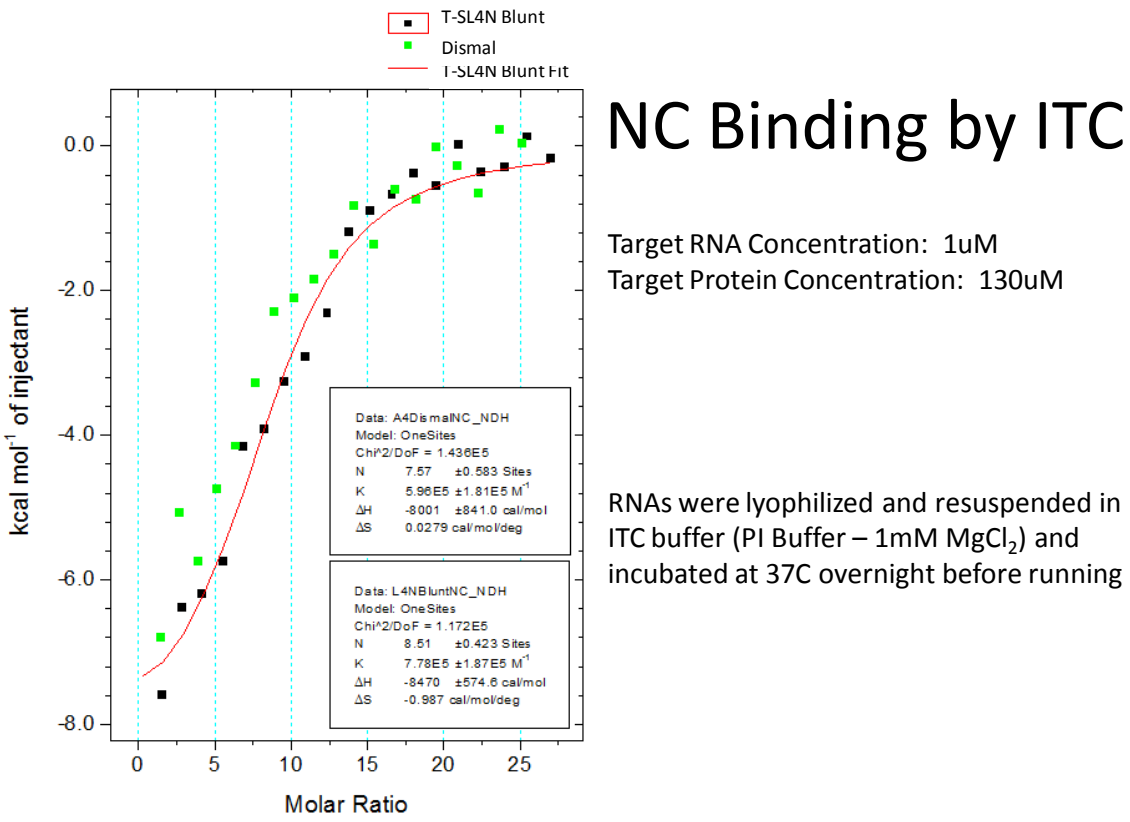

Figure A-1.6. ITC experiments comparing NC binding to the NL4-3 5'-L (T-SL4NBlunt - black squares) and the DISMAL 5'-L (Dismal - green squares) were conducted. The binding profiles appear very similar, and the number of binding sites and affinity for NC are very similar as well. This suggests that the mutations have not changed the overall fold of the 5'-L such that the native binding sites are exposed.

NC binding to the DISMAL construct is very similar to NC binding to the NL4-3 construct. The number of binding sites, the affinity of the binding sites, and the overall ITC profile were very comparable. This was somewhat unexpected, because if the DISMAL construct is a stabilized monomer, one would expect it to bind less NC at first than the NL4-3 5'-L, since a larger number of the RNA molecules at a given concentration would be in a monomeric conformation. As the monomeric conformation of the 5'-L has been shown to have less NC binding sites, I would expect a difference in the NC binding curve. This led me to suspect that perhaps the DISMAL construct forms

a labile dimer which requires magnesium in the gel and running buffer to be visualized using native agarose gel electrophoresis, which has been characterized by Dr. Thao Tran and Dr. Yuanyuan Liu in their dissertations regarding SIV and HIV-2 5'-Ls. To determine whether or not this was the case I ran a TBM gel of the DISMAL construct (Figure A-1.7) and saw dimerization behavior very similar to the NL4-3 5'-L.

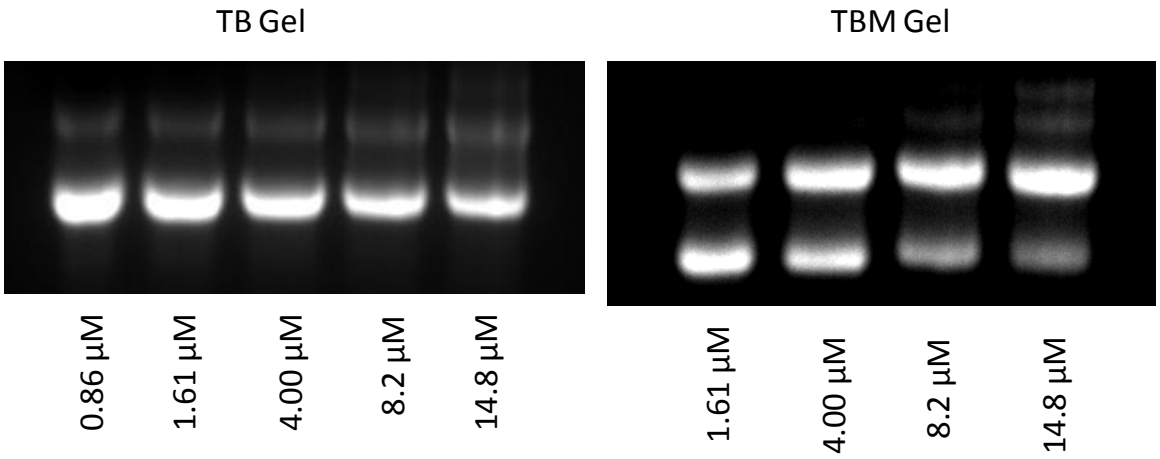

Figure A-1.7. 1% agarose TB (44 mM) and TBM (44 mM Tris-Borate, 0.15 mM $MgCl_2$) gels of the DISMAL construct. While the TB gels shows primarily monomer at all concentrations (as shown previously), the TBM gel shows an increasing percentage of dimer.

These results suggest that the DISMAL construct forms primarily a labile dimer species which can only be visualized with magnesium in the agarose gel and running buffer. This is consistent with the findings for HIV-2 which has a similar DIS sequence. It would be best to find an alternative method to assess dimerization that detects labile dimer as dimeric, such as size exclusion chromatography or capillary electrophoresis. These will be tested in the future.

1. Lawrence, D. C., Stover, C. C., Noznitsky, J., Wu, Z. & Summers, M. F. (2003). Structure of the intact stem and bulge of HIV-1 Psi-RNA stem-loop SL1. *J Mol Biol* **326**, 529-42.

2. Ennifar, E., Walter, P., Ehresmann, B., Ehresmann, C. & Dumas, P. (2001). Crystal structures of coaxially stacked kissing complexes of the HIV-1 RNA dimerization initiation site. *Nat Struct Biol* **8**, 1064-8.

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

Appendix A-2: Spliced RNA Construct

As described in Chapter 1, the HIV genome produces three classes of RNA transcripts: fully spliced, singly spliced, and unspliced. Although the spliced transcripts contain the 5' leader (5'-L) through the major splice-donor, and as such contain the DIS sequence, spliced RNAs are packaged at less than 1% of the efficiency of the unspliced RNA[1]. The spliced RNAs are utilized for translation[2], which suggests that the 5' ends of these RNAs may form a structure similar to the monomeric 5'-L. To study this possibility, an RNA construct was designed to represent the 5' end of the singly spliced *Vpu/Env* transcript. The final construct was designed based off of a 401 nucleotide construct suggested by Dr. Siarhei Kharytonchyk, however, RNA folding predictions for a number of different HIV-1 strains suggested the formation of a stable 3' hairpin that could be incorporated into the structure by extending the construct to a total length of 413 nucleotides. This construct, Spliced 413, is shown in Figure A-2.1 below.

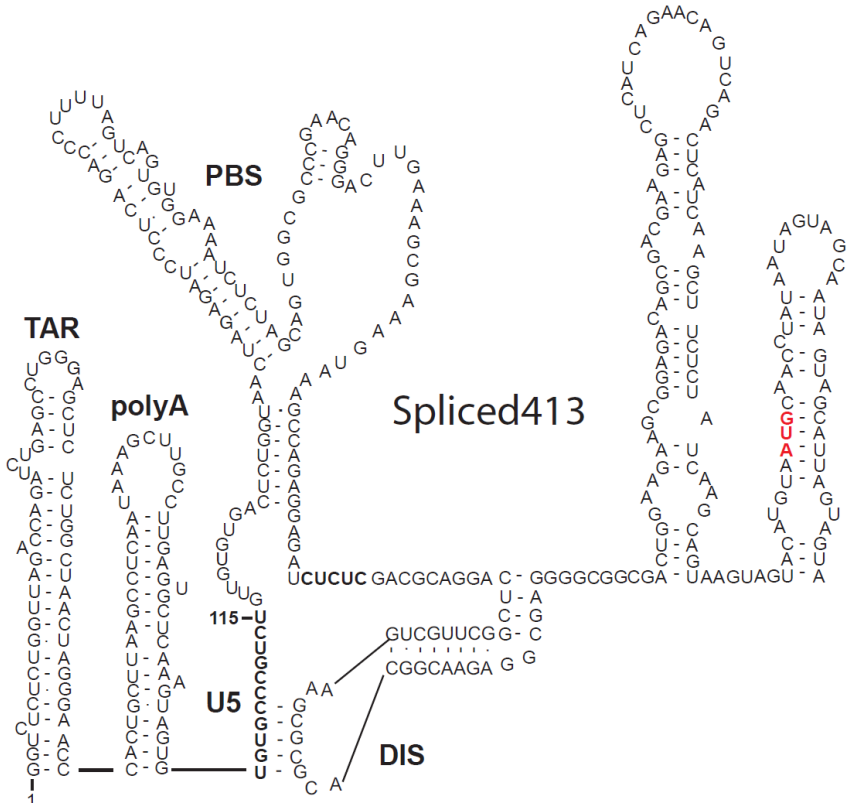

Figure A-2.1. Proposed secondary structure of the Spliced 413 construct. It is proposed to be identical to the monomeric 5'-L through the DIS stemloop and to be followed by two stemloops.

We hypothesized that the Spliced 413 construct should be a stable monomer because it lacks the AUG region to promote the U5:AUG interaction which stabilizes the dimer. Figure A-2.2 shows a dimerization study of the Spliced 413 RNA at approximately 1 µM. While it does show more monomer at 1 µM than is seen for the 5'-L, there is still dimer present.

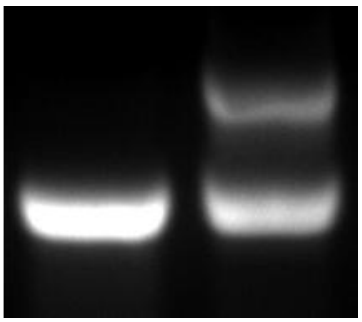

Figure A-2.2. 1% Tris-borate native agarose gel of Spliced 413 construct. Lane one is a monomeric control with no salt present. Lane two is the Spliced 413 construct at 1 μM incubated overnight at 37 ˚C in physiological ionic strength buffer.

While further studies of this construct are warranted, it does not give a sufficiently stable monomer to use directly for NMR studies. An experiment that would be useful would be to segmentally label a 5'-L NMR sample with some degree of protiation (for example an A2rGr labeling scheme) through the end of the DIS stemloop and deuterate the remainder of the RNA. Similarly a Spliced 413 NMR sample could be segmentally labeled to have the same protiation through the end of the DIS stemloop and deuterated for the remainder of the RNA. If 2D NOESY experiments of each of these RNAs were collected and overlaid it should be straightforward to determine if the 5' portion of each of these RNAs are folding into the same structure. If the NOEs match up, then it is likely that the RNAs adopt the same structure through the regions of the RNA with the same sequence. However, if there are significant differences in the NOESY spectra, then it is likely that the RNAs adopt different structures.

1. Houzet, L., Paillart, J. C., Smagulova, F., Maurel, S., Morichaud, Z., Marquet, R. & Mougel, M. (2007). HIV controls the selective packaging of genomic, spliced viral and cellular RNAs into virions through different mechanisms. *Nucleic Acids Res* **35**, 2695-704.

2. Bohne, J., Wodrich, H. & Kräusslich, H. G. (2005). Splicing of human immunodeficiency virus RNA is position-dependent suggesting sequential removal of introns from the 5' end. *Nucleic Acids Res* **33**, 825-37.

Appendix A-3: T-SL4NBlunt Dimerization Dissertation Constants

RNA Structure predicts that the presence of three non-native cytosine residues at the 3' end of the T-SL4N RNA transcripts could cause misfolding of the AUG hairpin, as shown in Figure A-3.1.[1] The misfolded AUG hairpin is a more stable hairpin (ΔG= -9.1) than the native AUG hairpin (ΔG= -5.1). This non-native stabilization of the AUG hairpin could impact the overall stability of the dimer by reducing the availability of the AUG residues to participate in the U5:AUG base pairing. To eliminate this effect a T-SL4NBlunt construct was designed, which instead of a using SmaI restriction enzyme site (CCC|GGG), utilizes a BstZ17I restriction enzyme site (GTA|TAC), allowing a native 3' end to be formed.

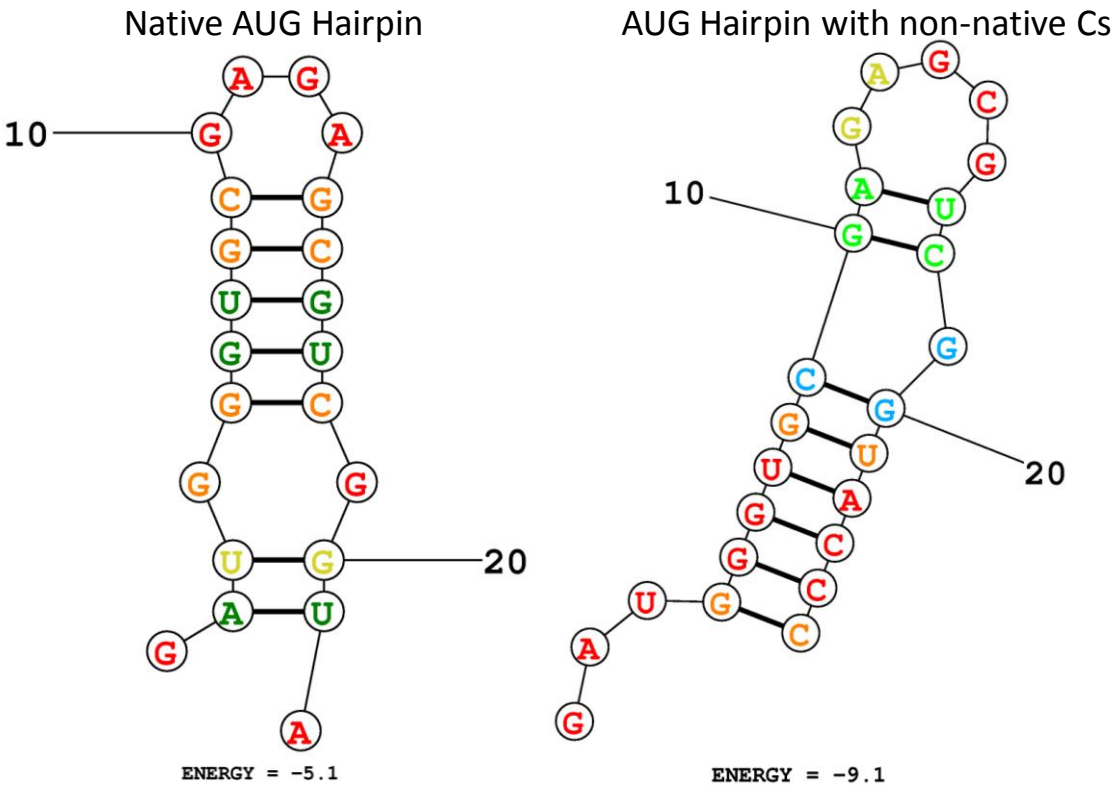

Figure A-3.1. The native AUG hairpin is shown on the left and has a predicted free energy of -5.1. The AUG hairpin incorporating three 3' non-native cytosines (shown on the right is predicted to misfold to create a hairpin of greater stability (predicted free energy of -9.1) than the native AUG hairpin.

Although Dr. Xiao Heng had previously characterized the dimerization K_d of T-SL4N with the 3' cytosines[2], the K_d of T-SL4NBlunt could be different due to the differential stability of the AUG hairpin. To determine the K_d of T-SL4NBlunt, the amount of dimer and monomer present at different concentrations was determined by native agarose gel electrophoresis. T-SL4NBlunt was incubated in a physiological ionic strength buffer (10 mM HEPES, 10 mM NaCl, 140 mM KCl, 1 mM $MgCl_2$, pH 7) overnight at concentrations ranging from 0.05 µM to 5 µM. The samples were loaded into a 1% agarose Tris-borate (44 mM) gel, which had been pre-chilled and was run with chilled buffer to reduce the amount of heat that the sample was exposed to while running. The intensity of the upper (dimer) and lower (monomer) bands was measured using ImageJ software[3]. The each molecule of dimer is expected to give twice the intensity of each molecule of monomer, because the dimer has twice as much RNA. The percentage of monomer and dimer can be calculated based on the intensities. If the total intensity is I_T, the intensity of the dimer band is I_D and the intensity of the monomer band is I_M, then the fraction of the RNA that is in the dimeric conformation ($Frac_D$) given by:

$$Frac_D = (I_D)/(I_T)$$

and the fraction of the RNA that is in monomeric conformation ($Frac_M$) is given by:

$$Frac_M = (I_M)/(I_T)$$

using these equations it is possible to calculate the concentration of monomeric RNA and the concentration of dimeric RNA because we know the total concentration of RNA. The concentration of monomer [M] is given by:

$$[M] = Frac_M * [RNA]$$

where [RNA] is the concentration of RNA in the sample as calculated by the A_{260} and the extinction coefficient. The concentration of dimer [D] is given by:

$$[D] = [(Frac_D)/2] * [RNA]$$

because the number of molecules of dimer is 1/2 the number of molecules of RNA strands in the dimeric conformation. If we assume that the equilibrium is given by the equation:

$$M + M \leftrightarrow D$$

then the equilibrium constant is given by the equation:

$$K_d = [D]/[M]^2$$

which can be rearranged to:

$$[D] = K_d[M]^2$$

Therefore, if we plot [D] versus $[M]^2$ we should get a straight line with a slope of the K_d. However, if we plot our data in this manner we get the plot shown in Figure A-3.2 below. Clearly our plot is not linear. This non-linear response is reproducible, and is also seen

for other constructs. It was also seen in Dr. Xiao Heng's studies of T-SL4N with the additional 3' cytosines.

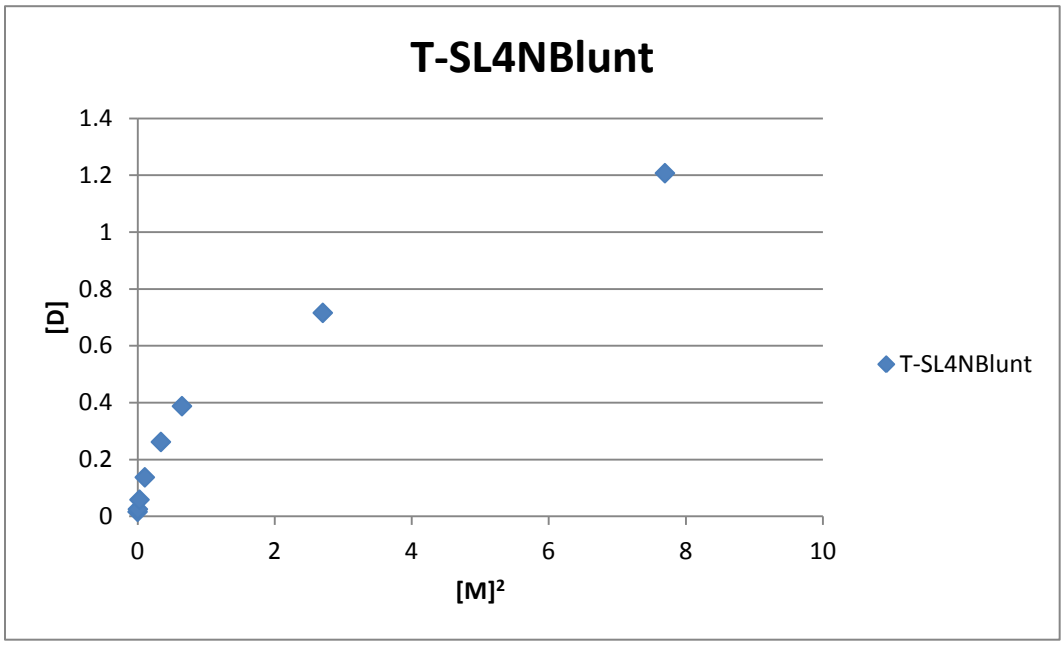

Figure A-3.2. Plot of the [D] versus the $[M]^2$ for T-SL4NBlunt. The plot is not linear, although the model that we were using predicts that it should be.

If we consider a more detailed equilibrium model:

$$2M_M \underset{K_1}{\leftrightarrow} 2M_D \underset{K_2}{\leftrightarrow} D_K \underset{K_3}{\leftrightarrow} D_E$$

where M_M is monomer in the monomeric conformation, M_D is monomer in the dimeric conformation (dimer-ready), D_K is kissing (labile) dimer, and D_E is extended dimer, we would get the following equations for each of the association constants:

$$K_1 = [M_M]/[M_D]$$

$$K_2 = [M_D]^2/[D_K]$$

$$K_3 = [D_K]/[D_E]$$

What we find is that if we plot [D] versus the [M] we get a linear plot, as shown in Figure A-3.3, below.

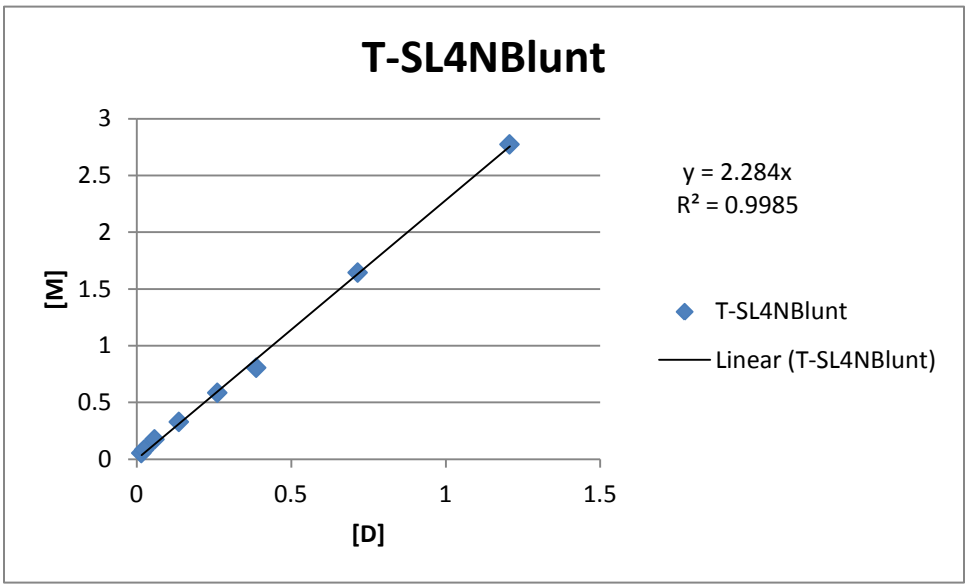

Figure A-3.3. A plot of [M] versus [D] gives a linear plot with an R^2 value of 0.9985.

So if we say that

$$[M] = m[D]$$

$$[D] = [(Frac_D)/2] * [RNA]$$

$$[M] = Frac_M * [RNA]$$

$$[RNA] = [M_M] + [M_D] + 2[D_K] + 2[D_E]$$

If we assume that the kissing (labile) dimer comes apart on the gel and runs as a monomer, which we suspect from other systems, then the concentration of apparent monomer is given by:

$$[M] = Frac_M * [RNA] = [M_M] + [M_D] + 2[K_D]$$

and the apparent concentration of dimer (which really only represents the extended dimer) is given by:

$$[D] = Frac_D * [RNA] = 2[D_E]$$

Then

$$[M] = m[D]$$

can be rewritten as

$$[M_M] + [M_D] + 2[K_D] = m * 2[D_E]$$

$$m = \frac{[M_M] + [M_D] + 2[D_K]}{2[D_E]}$$

$$m = \frac{[M_M]}{2[D_E]} + \frac{[M_D]}{2[D_E]} + \frac{[D_K]}{[D_E]}$$

$$K_1 = [M_M]/[M_D]$$

$$K_2 = [M_D]^2/[D_K]$$

$$K_3 = [D_K]/[D_E]$$

$$m = \frac{[M_M]}{2[D_E]} + \frac{[M_D]}{2[D_E]} + K_3$$

$$[M_M] = [M_D] * K_1$$

$$[M_D] = \sqrt{[D_K] * K_2}$$

$$[M_M] = K_1\sqrt{[D_K] * K_2}$$

$$[D_K] = K_3[D_E]$$

$$[M_M] = K_1\sqrt{K_3[D_E] * K_2}$$

$$m = \frac{K_1\sqrt{K_2K_3[D_E]}}{2[D_E]} + \frac{\sqrt{K_2K_3[D_E]}}{2[D_E]} + K_3$$

$$m = \frac{(K_1 + 1)\sqrt{K_2K_3[D_E]}}{2[D_E]} + K_3$$

$$m = \frac{\sqrt{(K_1 + 1)^2K_2K_3[D_E]}}{2[D_E]} + K_3$$

$$m = \sqrt{\frac{(K_1 + 1)^2K_2K_3}{4[D_E]}} + K_3$$

If we substitute *m* back into

$$[M] = m[D]$$

we get

$$[M] = (\sqrt{\frac{(K_1+1)^2 K_2 K_3}{4[D_E]}} + K_3) * [D]$$

$$[M] = (\sqrt{\frac{(K_1+1)^2 K_2 K_3}{2[D]}} + K_3) * [D]$$

$$[M] = (\sqrt{\frac{(K_1+1)^2 K_2 K_3}{2}})\sqrt{[D]} * K_3[D]$$

When we plot this equation versus our data using gnuplot we see that it fits the data quite well (Figure A-3.4).

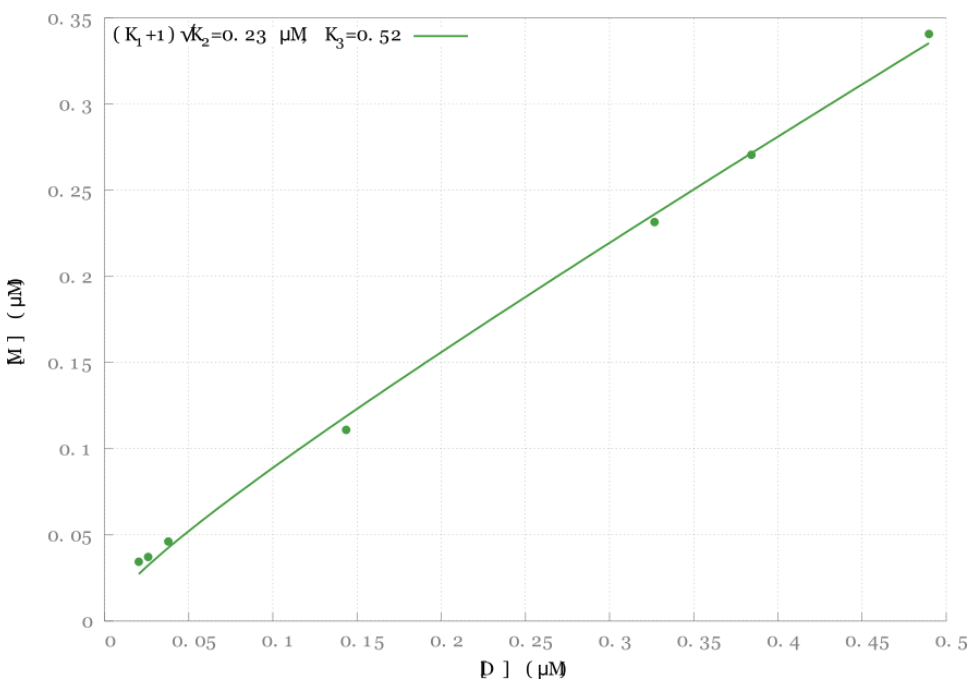

Figure A-3.4. T-SL4NBlunt dimerization data fit using the equations described in the text.

The quality of the fit based on the assumption that the kissing (labile) dimer appears to run as a monomer suggests that this model could be correct. This work was done with the assistance of Dr. Jan Marchant who contributed to the derivation of the equations as well as the use of gnuplot.

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

2. Lu, K., Heng, X., Garyu, L., Monti, S., Garcia, E. L., Kharytonchyk, S., Dorjsuren, B., Kulandaivel, G., Jones, S., Hiremath, A., Divakaruni, S. S., LaCotti, C., Barton, S., Tummillo, D., Hosic, A., Edme, K., Albrecht, S., Telesnitsky, A. & Summers, M. F. (2011). NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-5.

3. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671-5.